UNIVERSITY OF THE WITWATERSRAND

DOCTORAL THESIS

Computational approaches to characterizing morbidity and mortality patterns in rural South Africa

---

Candidate:
Michael T. MAPUNDU

Supervisors:
Professor Turgay CELIK
Dr Chodziwadziwa W. KABUDULA
Professor Eustasius MUSENGE

A thesis submitted to the Faculty of Health Sciences, University of the Witwatersrand in fulfilment of the requirements for the Degree of Doctor of Philosophy

by publication majoring in Public Health Informatics *in the*

School of Public Health



October 22, 2024

Declaration of Authorship

I, Michael T. MAPUNDU, declare that this research report titled, "Computational approaches to characterizing morbidity and mortality patterns in rural South Africa" and the work presented in it are my own, except where specific references cited to the work of others. This work has not been submitted elsewhere for any other degree or award except for this current fulfilment of the Doctor of Philosophy degree at The University of Witwatersrand. This research work reflects my independent work except in instances where explicit references have acknowledged other sources.

Signed: 

Date: 22 October 2024

"The secret of getting ahead is getting started." "

Mark Twain

University of the Witwatersrand

Abstract

Faculty of Health Sciences

School of Public Health

Doctor of Philosophy

Computational approaches to characterizing morbidity and mortality patterns in rural South Africa

by Michael T. MAPUNDU

Background: Verbal autopsies (VAs) are commonly used in Low to Middle Income Countries, as a way of determining the cause of death in cases where deaths occur outside health facilities and there is no medically certified cause of death. The VA process is usually done by conducting interviews with relatives of the deceased to elicit information about circumstances and events surrounding the death. The compiled VA narratives are then given to two doctors, and supplemented by the full set of responses, both from structured questions for assessment, in order to reach a consensus on the cause of death. In instances where they disagree, a third physician is consulted, a process known as Physician Coded Verbal Autopsy (PCVA). PCVA is the most used process for determining cause of death. However, it is widely criticized because of its lack of robustness, cost, time, inconsistencies, and inaccuracies as it is subjective and prone to errors among many drawbacks. Therefore, these challenges affect the accuracy of verbal autopsy results. Consequently, this results in PCVAs mostly employed for the training and validation of computational approaches. Despite these challenges, VAs have been employed successfully to estimate mortality rates and causes of death in settings where vital registration systems are weak or non-existent. Therefore, efforts are ongoing to improve the validity and reliability of verbal autopsies, including the use of computational approaches for analysing the data. There has been a growing interest from the VA community to apply automated algorithms that are artificially intelligent in order to improve cause of death determination using VA data, thus closing the civil registration gap. It has been proven that the use of machine learning (ML) and natural language processing (NLP) has helped identify patterns and trends in the data that might be missed by manual analysis, but are key in transforming the data into actionable insights that can help improve health outcomes.

Study Objectives: The overall aim of this study was to utilize advanced computational methodologies to gain a comprehensive understanding of the complex health dynamics within rural South African communities. The study aims to bridge the gap between traditional statistical and epidemiological approaches and the unique challenges faced by these communities, thereby contributing to more informed public health strategies and interventions. As such, we sought to understand the determinants and circumstances of events leading to cause of death in rural north-east South Africa, using predictive and descriptive analysis.

The main focus of this study was to answer the following questions crucial in VA monitoring and decision making: 1) what are the common prevalent topical diseases that led to death at Agincourt Health and Demographic Surveillance System (HDSS) in rural north-east South Africa between 1993 and 2015?; 2) how is mortality clustered by cause of death within households, and what characteristics of households are associated with high mortality at Agincourt HDSS? and 3) to what extent can machine learning and deep learning techniques accurately classify cause of death as compared to physician classification?

In the process of our investigation we will also address the following sub questions: 1) how can machine learning and statistical modelling be effectively applied to analyse and predict mortality patterns within rural South African communities?; 2) what are the key determinants and socio-economic factors that contribute to variations in mortality rates?; 3) can computational techniques uncover hidden correlations between specific diseases and socio-economic indicators, providing insights into potential causal relationships?; 4) how can spatial analysis techniques be used to identify clusters of high mortality rates in rural South Africa, and what underlying factors might be driving these patterns?; 5) what role do access to healthcare resources, healthcare infrastructure, and healthcare-seeking behaviour play in shaping mortality outcomes in rural South African communities?; 6) to what extent do traditional health beliefs, cultural practices, and community dynamics influence morbidity and mortality patterns, and how can computational methods account for these factors; 7) how effective are data-driven models in predicting future morbidity and mortality trends in rural South

Africa, and how might these predictions inform healthcare planning and resource allocation? and 8) what are the potential barriers and opportunities for scaling up successful computational approaches to other rural regions within South Africa or similar global contexts?

The specific objectives of this study were;

1. To determine the most prevalent diseases that led to deaths using VA narrative datasets and text mining techniques at Agincourt HDSS in rural north-east South Africa between 1993 and 2015.
2. To identify mortality clusters by cause of death, establish determinants of mortality clusters and investigate the mortality characteristics associated with households at Agincourt HDSS between 1993 and 2015.
3. To establish ML accuracy in automating VA classification and achieve at least the same level of accuracy with that of physician classification on cause of death in rural north-east South Africa between 1993 and 2015.

One of the aims of this study was to apply text mining techniques to derive implicit knowledge that is hidden in the unstructured VA narratives and present it in an explicit form. This allowed us to discover, mortality causes and most prevalent diseases which caused the population to succumb to death. Secondly, we sought to establish ML accuracy in automating VA classification and achieve at least the same level of accuracy with that of physician classification on cause of death. This was done through a comparative performance evaluation of ML methods and Computer Coded Verbal Autopsy (CCVA) algorithms on South African VA narratives data. Additionally, we also explored with novel deep learning architectures in order to generate cause of death prediction in a timely, cost effective and error free way. These computational techniques will make us achieve our aim of determining events and circumstances leading to cause of death by identifying morbidity and mortality occurrences in rural north-east South Africa from 1993 to 2015. As such, the study will ease the design, development, implementation and sustainment of tailored health intervention programmes. Consequently, this will improve life expectancy, turnaround time for diagnosis, and enforce a standardised VA reporting approach. This will therefore close the civil registration gap.

Method: This study was a secondary data analysis of routinely collected VA data at Agincourt HDSS, for the period of 1993 to 2015. Agincourt HDSS is a surveillance site that specifically provides evidence based health monitoring that seeks to strengthen health priorities, practice and inform policy. In this study, we used three types of datasets. The first dataset is the structured responses from the standard questionnaire, second dataset is the VA narratives, and the third dataset is a combination of the responses and the narratives. The three datasets had 287 columns/features and 16338 records/observations. For the responses only, we took all features that had responses from the standard questionnaire as our predictors and the cause of death assigned by physicians using International Classification of Diseases-10 (ICD-10) code for each record in the dataset as our target variable. Ultimately, we had 231 predictors (all symptoms, age at death and gender) and 1 target variable, and all our features were in English. The predictions using the narratives were done using age at death, gender and the narrative feature and 1 target variable. For the combined VA dataset we used 232 predictors and 1 target variable. We only added the VA narrative feature to the responses dataset in order to have our combined dataset. We further created twelve cause of death categories with corresponding labels, class distribution with number of samples for each class before and after data balancing for our training dataset. The cause of death categories were derived based on InterVA user guide. The text mining and deep learning studies used the narratives only dataset, and the ML study employed all datasets.

Results: ML models could accurately determine the cause of death from VA narratives, producing results comparable to expert diagnosis, with our optimal models attaining accuracies around 96%, with significant statistical differences in algorithmic performance ($p < 0.0001$). In the same way our robust novel stacked ensemble deep learning methods (SEDL) performed optimally than conventional DL approaches attaining an accuracy of 82% and employed Local Interpretable Model-agnostic Explanations (LIME) to enhance the interpretability of DL models, thus fostering trust in their use in healthcare. Our empirical results suggest that our automated approaches can be integrated in the CoD pipeline for identifying mortality causes, alongside human annotation, and interpretation. Additionally, through mortality trend and pattern analysis, we discovered that in the first decade of the civil registration system in South Africa, the average life expectancy was approximately 50 years. However, in the second decade, the life expectancy significantly dropped, and the population was dying at a much younger average age of 40 years. This suggests that the HDSS population succumbed to death due to mortality causes such as; *vomiting/diarrhoea, chest/stomach pain, fever, coughing and high blood pressure*. Interestingly, we found out that the most prevalent diseases entailed *human immunodeficiency virus (HIV), tuberculosis (TB), neurological disorders, malaria, diabetes, high blood pressure, chronic ailments (kidney, heart, lung, liver), maternal and accident related deaths*. Noteworthy, in the third decade, we see a gradual improvement in life expectancy, possibly attributed to effective health intervention programmes. Our sequential modelling on patient care seeking patterns, suggests that most people in the HDSS seek traditional ways than western ways for their healthcare, when faced with terminal illnesses. Additionally, we noticed that the narratives entail additional variables which can ease the cause of death diagnosis using sequential modelling and semantic and structure analysis, with a retrieval rate of approximately $n > 2$, per every case where n is number of terms. Through a structure and semantic analysis of narratives where experts disagree, we also demonstrate the most frequent terms of traditional healer consultations and visits. Therefore, this can possibly assist in determining cause of death specifically in the unknown category.

Conclusion: This research study explores the utilization of computational techniques to analyse and comprehend morbidity and mortality patterns in rural South Africa, by leveraging large-scale VA data and complex computationa approaches to discern prevalent diseases, health disparities, and mortality trends within this specific context. These computational approaches, avail nuanced insights into disease prevalence, risk factors, and potential correlations between socio-economic factors and health outcomes. The findings of this study flags the potential of computational approaches in uncovering intricate health dynamics in rural settings, shedding light on areas for targeted interventions and policy enhancements. The study enforces the significance of harnessing data-driven strategies to inform public health strategies tailored to the unique challenges faced by communities in rural South Africa. The study represents a significant advancement in the field of CoD determination from VA narratives by introducing innovative ML and DL techniques that offer accurate and interpretable results. The findings suggest the potential for these models to streamline the VA reporting process, ultimately benefiting healthcare systems in LMICs by reducing diagnosis turnaround time, costs and improving the accuracy of CoD determination. Therefore, this research bridges the gap between the amount of data available and conducting research that can lead to practical actions, thus supporting multi- disciplinary research in civil registration systems using VA data. Consequently, it provides a baseline for future studies, generalising these findings to other domains of interest, thus highlighting the importance of improving health intervention programs in LMICs to increase life expectancy, and contribute to the understanding of mortality patterns and prevalent diseases in LMICs by harnessing the power of computational techniques.

Keywords: Cause of Death, CNN, deep learning, LIME, LSTM, Machine Learning, RNN, Verbal Autopsy

Acknowledgements

The writing of this Doctoral thesis has been one of the most significant academic challenges I have ever had to face. Without the support, patience and guidance of the following people, this study would not have been completed. It is to them that I owe my deepest gratitude. First and foremost I offer my sincerest gratitude to my supervisors, Professor Turgay Celik, Professor Eustasius Musenge and Doc Chodziwadziwa Kabudula and Mr Victor Olago as a collaborator, for their continuous unwavering support, for their patience, motivation, enthusiasm and immense knowledge. I attribute the level of my thesis to their tremendous encouragement and effort, without them this thesis would not have been a success. Furthermore, I would like to thank my colleague, Victor Olago for his assistance with novel ideas. Lastly I would love to thank my wife Pamela for helping me and supporting me throughout this study. Your love, support and constant patience have taught me so much about sacrifice, discipline and compromise.. . .

Contents

Declaration of Authorship ii

Abstract v

Acknowledgements viii

1 Chapter 1 1
- 1.1 Background on VA . . . 1
- 1.2 VA process . . . 1
- 1.3 ML in VA . . . 2
- 1.4 Problem Statement . . . 4
- 1.5 Justification . . . 5
- 1.6 Research aims, questions and objectives . . . 6
  - 1.6.1 General aim . . . 6
  - 1.6.2 Research questions . . . 6
  - 1.6.3 Objectives . . . 7
  - 1.6.4 Data source . . . 7
- 1.7 Conceptual Framework and overview . . . 7
- 1.8 Proposed Analytical ML Framework and Architecture . . . 8
- 1.9 List of publications . . . 9
- 1.10 Conference presentations . . . 9
- 1.11 Structure of the Thesis . . . 11

2 Statistical and Machine Learning Techniques in Verbal Autopsy: A Systematic Literature Review 12
- 2.1 Abstract . . . 12
- 2.2 Background . . . 13
  - 2.2.1 **Feature selection and feature engineering** . . . 15
    - **Feature extraction** . . . 15
    - **Feature value Representation Techniques** . . . 16
    - **Feature selection** . . . 17
    - **Automated feature selection** . . . 18
  - 2.2.2 **Semi Supervised learning** . . . 18
    - **Semi supervised learning problem** . . . 19
    - **Statistical models** . . . 19
    - **Machine learning techniques** . . . 20
    - Ensemble classifiers . . . 22
    - VA algorithms . . . 23
  - 2.2.3 Deep learning . . . 24

Distributed vector representations . . . 25
Character embeddings . . . 26
2.2.4 ML Performance Metrics in VA . . . 27
2.2.5 ML Optimisation techniques in VA . . . 29
k-fold cross validation . . . 29
2.2.6 Imbalanced and fragmented datasets . . . 30
2.2.7 Model limitations and imposed assumptions . . . 30
2.2.8 Emerging Trends: Deep Learning in VA . . . 32
Other Recurrent Neural Networks approaches . . . 32
2.2.9 Limitations in VA literature . . . 33
2.3 Methods . . . 35
2.3.1 Existing literature survey on VA . . . 36
2.4 Results . . . 36
2.4.1 Preprocessing . . . 36
2.4.2 Transparency . . . 38
2.4.3 Data balancing . . . 39
2.4.4 Feature extraction . . . 39
2.4.5 Feature value representation schemes . . . 41
2.4.6 Feature selection . . . 41
2.4.7 Comparative analysis of studies reviewed in this paper . . . 42
2.4.8 Evaluation metrics . . . 44
2.4.9 Tag cloud of common and frequent terms of reviewed studies . . . 45
2.4.10 Models . . . 45
2.4.11 Machine learning framework guidelines for VA . . . 45
2.5 Discussion . . . 50
2.6 Conclusion . . . 53
2.7 Declarations . . . 53
2.7.1 Ethics approval and consent to participate . . . 53
2.7.2 Consent for publication . . . 53
2.7.3 Availability of data and materials . . . 53
2.7.4 Competing interests . . . 54
2.7.5 Declaration of interest . . . 54
2.7.6 Funding . . . 54
2.7.7 Authors' contributions . . . 54
2.7.8 Corresponding authors . . . 54
2.7.9 Acknowledgements . . . 55

3 Performance evaluation of Machine Learning and Computer Coded Verbal Autopsy (CCVA) algorithms for cause of death determination: A comparative analysis of data from rural South Africa 56
3.1 Abstract . . . 56
3.2 Introduction . . . 57
3.2.1 Computer Coded Verbal Autopsy Algorithms . . . 58
3.2.2 Machine Learning in VA . . . 58
3.3 Methods . . . 59
3.3.1 Study design . . . 59

3.3.2 Population . . . 60
3.3.3 Data Source . . . 60
3.3.4 Data Pre-processing and Encoding . . . 61
3.3.5 Feature Engineering . . . 61
3.3.6 Data Balancing and feature scaling . . . 62
3.3.7 Machine Learning Models for CoD Prediction . . . 62
3.3.8 Model Training, Validation and Testing . . . 62
3.4 CCVA algorithms . . . 63
3.5 Identification of contradicting cases and best model predictors . . . 64
3.5.1 ML Techniques Model Evaluation . . . 64
3.5.2 CCVA Techniques Model Evaluation . . . 64
3.6 Statistical Analysis . . . 65
3.7 Results . . . 65
3.7.1 Performance Evaluation of ML Classifiers . . . 65
Results from only the VA Narrative Predictors . . . 65
Results from using questionnaire responses only as predictors . . . 65
Results from using combined narratives and questionnaire responses . . . 66
3.8 CCVA algorithm evaluation using CSMFs . . . 67
3.9 Trend and pattern analysis using ML and CCVA approaches . . . 68
3.9.1 CCVA algorithms . . . 68
3.9.2 ML techniques . . . 69
3.9.3 Analysis of contradicting cases . . . 75
3.9.4 Model best predictors . . . 75
3.10 Discussion . . . 76
3.11 Conclusion . . . 78
3.12 Conflict of Interest Statement . . . 78
3.13 Author Contributions . . . 78
3.14 Funding . . . 78
3.15 Acknowledgments . . . 78
3.16 Data Availability Statement . . . 79

4 Text mining of verbal autopsy narratives to extract mortality causes and most prevalent diseases using Natural Language Processing 80
4.1 Abstract . . . 80
4.2 Introduction . . . 81
4.3 Literature review . . . 82
4.3.1 Related work . . . 82
4.3.2 N-gram language processing . . . 83
4.3.3 Topic modelling . . . 83
Topic modelling using LDA . . . 83
Topic modelling using BERTopic . . . 83
4.4 Materials and methods . . . 84
4.4.1 Data pre-processing . . . 86
4.4.2 Information extraction . . . 86
4.4.3 Topic modelling using LDA . . . 86

LDA algorithm . . . 87
4.4.4 Topic modelling using BERTopic . . . 87
BERTopic algorithm . . . 87
Optimisation of LDA and BERTopic algorithms . . . 88
4.4.5 Knowledge discovery . . . 89
4.5 Data analysis . . . 89
4.6 Statistical packages . . . 89
4.7 Results . . . 90
4.7.1 Term frequencies using word cloud . . . 90
4.7.2 Term frequencies using TF-IDF . . . 90
4.7.3 Term frequencies using n-grams . . . 90
4.7.4 Semantic and structure analysis of additional and similar variables . . . 92
4.7.5 Topic modelling results using LDA approach . . . 95
4.7.6 Topic modelling results using BERTopic approach . . . 95
4.7.7 Results from the knowledge discovery process: Identified relevant topics . . . 102
4.8 Discussion . . . 103
4.9 Future work . . . 105
4.10 Conclusion . . . 105
4.11 Acknowledgements . . . 105

5 Explainable Stacked Ensemble Deep Learning (SEDL) Framework to Determine Cause of Death from Verbal Autopsies 106
5.1 Abstract . . . 106
5.2 Introduction . . . 107
5.3 Deep Learning in VA . . . 108
5.3.1 Ensemble DL . . . 109
5.3.2 Explainable AI . . . 110
5.4 Materials and Methods . . . 111
5.4.1 Study Design . . . 111
5.4.2 Study Population . . . 111
5.4.3 Data source and description . . . 111
5.4.4 Pre-processing . . . 112
5.4.5 Word Embedding and representation . . . 112
5.4.6 Text Classification with Deep Learning architectures . . . 112
Long Term Short Memory . . . 112
Convolutional Neural Network . . . 113
5.4.7 Stacked ensemble DL models . . . 113
5.4.8 Explainable AI using LIME . . . 113
5.5 Experiments . . . 113
5.5.1 Deep Learning models . . . 113
5.6 DL Model evaluation . . . 116
5.7 Results . . . 117
5.7.1 Model performance . . . 117
5.7.2 Model explainability . . . 118
5.8 Discussion . . . 120

5.9 Conclusions . . . 122
5.10 Author contributions . . . 123
5.11 Funding . . . 123
5.12 Ethics . . . 123
5.13 Data availability . . . 123
5.14 Acknowledgments . . . 123
5.15 Conflicts of interest . . . 123
5.16 Abbreviations . . . 123

6 Summary and conclusion **125**
6.1 Introduction . . . 125
6.2 Strengths of this study . . . 126
6.3 Limitations of this study . . . 127
6.4 Future Research Directions . . . 127
6.5 Conclusion . . . 127

A Appendices **129**
A.1 Equations . . . 129
A.1.1 Feature reduction techniques . . . 129
Pearson correlation . . . 129
Chi-squared . . . 129
Information gain . . . 129
Principal component analysis . . . 129
A.1.2 Statistical equations . . . 130
Logistic regression . . . 130
Naive Bayes . . . 130
A.1.3 Machine Learning equations . . . 130
k-nearest neighbour . . . 130
Random Forests . . . 130
Support Vector Machine . . . 130
A.1.4 Artificial Neural Networks . . . 131
A.1.5 k-means clustering . . . 132
Boosting . . . 132
Bagging . . . 132
A.1.6 Deep Learning equations . . . 133
A.1.7 Distributional vectors . . . 133
A.1.8 Neural Language Model . . . 133
A.2 Word embeddings . . . 133
Word2vec . . . 133
CBOW model . . . 134
Sentence2vec . . . 135
Global Vectors . . . 135
A.2.1 Character embeddings . . . 136
Convolutional Neural Networks (CNN) . . . 136
Simple Recurrent Neural Networks . . . 136
Recurrent Neural Networks . . . 137

A.2.2 Long Short-Term Memory RNN . . . 137
Gated RNN . . . 138
Matrix Vector Recursive Neural Network . . . 138
Recursive Neural Tensor Network . . . 139
A.2.3 Transformer architecture . . . 139
A.2.4 Performance Metrics . . . 140
Macro precision . . . 140
Macro recall . . . 141
Macro F-measure (F-measure . . . 141
Overall accuracy . . . 141
Area under curve (AUC) . . . 141
Overall chance-corrected concordance (CCC) . . . 141
Cause Specific Mortality Fractions (CSMF) accuracy . . . 142
Top cause accuracy . . . 142
Top 3 cause accuracy . . . 142
A.2.5 Validation . . . 142

B Appendices **144**

C Appendices **165**

D Appendices **192**

E Appendices **212**

F Appendices **245**

G Appendices **247**

References **304**

List of Figures

1.1 Completeness of CoD coverage (2007-2016), Source: https://www.who.int/images/default-source/maps/global_civilregistrationdeaths_2007_2016.png?sfvrsn=5f760b4f_0 . . . 1
1.2 Autopsy rates, Source [6] . . . 2
1.3 VA Process and tools, Source: https://www.publichealthnotes.com/verbal-autopsy-and-mpdsr/
1.4 ML types . . . 4
1.5 Conceptual framework illustrating the structure of the thesis.TM: Text mining, ML:Machine learning, DL: Deep learning . . . 9
1.6 ML Framework structure and Architecture . . . 10

2.1 ANN structure of single neuron. . . . 22
2.2 Comparison of ANN and deep learning architecture. . . . 25

2.3 Convolutional neural network framework. . . . 27
2.4 ML performance evaluation confusion matrix 28
2.5 Recurrent neural network framework. . . . 32
2.6 Long term short memory framework. . . . 33
2.7 Distribution of initial searched journal articles vs conference papers 36
2.8 Flowchart that details our methodological systematic analysis. . . . 38
2.9 Distribution of VA primary studies that include pre-processing. . . . 39
2.10 Pictorial distribution of VA primary studies that include transparency. . . . 39
2.11 Pictorial distribution of VA related studies that include data balancing. . . . 41
2.12 Distribution of feature extraction techniques reviewed. . . . 41
2.13 Distribution of feature value representation schemes. . . . 42
2.14 Distribution of feature selection techniques. . . . 44
2.15 Distribution of VA related studies in comparison to other fields used in this methodological systematic analysis. . . . 44
2.16 Distribution of journal articles and conference articles used in this methodological systematic analysis. . . . 45
2.17 Distribution of evaluation metrics from primary studies reviewed. . . . 48
2.18 Tag cloud of this methodological systematic analysis 49
2.19 Distribution of statistical and ML models used in primary studies. . . . 49
2.20 Proposed guidelines for the VA machine learning framework 51

3.1 Schematic diagram of ML process followed. . . . 61
3.2 AUROC of our nine classifiers using combined questionnaire responses and narratives. **(A)** AUROC for ANN. **(B)** AUROC for KNN.**(C)** AUROC for RF. **(D)** AUROC for DT. **(E)** AUROC for SVM.**(F)** AUROC for LR. **(G)** AUROC for XGBOOST. **(H)** AUROC for BG.**(I)** AUROC for BC. 69
3.3 Top twelve CoD diseases 70
3.4 CCVA mortality trends based on age, population and gender. **(A)** Cause of death by sex. **(B)** Percentage of deaths by age group.**(C)** Yearly mortality trends by gender. . . . 71
3.5 Gender and age group counts graphs. **(A)** Gender count. **(B)** Age group count per gender. **(C)** Age group count. . . . 72
3.6 Mortality trends across age groups. **(A)** Number of deaths over time. **(B)** Age at death count. **(C)** Yearly death count across age groups. **(D)** Age group CoD count. **(E)** CoD and age death. **(F)** Age at death per year. . . . 73
3.7 Yearly CoD based on gender.**(A)** CoD based on gender. **(B)** Yearly CoD based on gender. . . . 74
3.8 Tri-gram model showing frequently occurring contradicting cases. . . . 75
3.9 Bi-gram model of our best model predictors 76

4.1 Text mining steps followed in this study 85
4.2 Hyper-parameter optimization for number of topics. . . . 88
4.3 Word frequencies using word cloud 91
4.4 Most prevalent mortality causes using n-grams language processing. . . . 92
4.5 An example of a processed VA narrative 93

4.6 Inter-topic distance map using multi-dimensional scaling **Legend:** T1:Diarrhoea/Respiratory tract infection, T2:HIV, T3:TB, T4:Cervical cancer, T5:Chronic liver ailment, T6:Neurological disorder, T7:Unknown, T8:Breast cancer, T9:Chronic ailments, T10:Malaria, T11:Jaundice/malaria/skin infection, T12:Unknown, T13:Diabetes, T14:Mental confusion, T15:Accidents, T16:Heart ailment/septic wound, T17:Epilepsy, T18:Heart chronic ailments, T19:Maternal and Asthma 97
4.7 Topics generated using BERTopic model 99
4.8 BERTopic inter-topic distance map 100
4.9 Heatmap similarity matrix 101
4.10 Hierarchical clusters of similar topics 102

5.1 LSTM deep learning framework, Source: [22]. 114
5.2 CNN deep learning framework, Source: [22]. 114
5.3 Stacking of our models, Source: [174]. 115
5.4 General pipeline of XAI implementation using VA data. 115
5.5 Learning curve depicting performance evaluation of CNN 118
5.6 Learning curve depicting performance evaluation of LSTM 118
5.7 Learning curve depicting SEDL performance evaluation during model training 119
5.8 Learning curve depicting SEDL performance evaluation 119
5.9 Example of LIME correct prediction 119
5.10 Example of LIME correct prediction 120
5.11 Example of LIME incorrect prediction 120
5.12 Example of LIME incorrect prediction 120

A.1 Random Forests 131
A.2 Distributional vectors represented by a D-dimensional vector 133
A.3 Neural Language Model 133
A.4 Model for CBOW 134
A.5 CNN framework for word prediction 136
A.6 CNN modelling on text 137
A.7 Simple RNN network 137
A.8 Illustration of an LSTM and GRU gate 138
A.9 Illustration of the Transformer Architecture 139

List of Tables

2.1 Existing literature surveys on VA and their differences from this survey paper: Legend:Feature Engineering (FE), Feature Selection (FS). 37
2.2 Feature extraction methods used in primary studies. Legend: Bag of Words (BoW), n-grams (NG), Hybrid (HB), Word2vec (WV), Concept-Based (CB), Graph-Based (GB), Rule Base (RB), Doc2vec (DV), Expert driven (ED), Sentence2vec (SV), Glove (GV). 40

2.3 Feature value representation schemes from primary studies. Legend: Binary representation (BR), Term frequency (TF), Document frequency (DF), Term Frequency-Inverse Document Frequency (TF-IDF), Normalised Term Frequency-Inverse Document Frequency (NTF-IDF). 42
2.4 Feature selection techniques from primary studies. Legend: Pearson correlation (PC), Chi-squared (CS), Information gain (IG), Principal component analysis (PCA),Fisher Markov Selector (FMS), Improved Global Feature Selection (IGFS), Local Semi Supervised (LSS), Natural Language Processing (NLP) 43
2.5 Evaluation metrics techniques applied in primary studies. Legend: Specificity (Sp), Precision (Pr), Recall (Re), F-score (Fs), Accuracy (Ac), CSFM (CS), CCC, PCCC, Kappa (Ka), Case Similarity measure (CSM), Top Cause (TC), Top3CoD (T3), Area Under Curve (AUC) 46
2.6 Models applied in primary studies. 47

3.1 **Twelve disease classes and number of data samples before and after data balancing.** 60
3.2 **Model optimal hyperparameters.** 63
3.3 **Comparison of nine ML models using narratives only.** 66
3.4 **Comparison of nine ML models using questionnaire responses only.** 66
3.5 **Comparison of nine ML models using combined narratives and questionnaire responses.** . . . 67
3.6 **Statistical tests of our nine models.** 68

4.1 **Most frequently occurring terms.** 90
4.2 **Comparison of symptoms** 94
4.3 **Most prevalent diseases identified using LDA** 96

5.1 Labelling of the twelve disease classes. 115
5.2 Performance evaluation of models. 118

List of Abbreviations

AI	**A**rtificial **I**ntelligence
API	**A**pplication **P**rogramming **I**nterface
ANN	**A**rtificial **N**eural **N**etwork
CBOW	**C**ontinous **B**ag **O**f **W**ords
CNN	**C**onvolutional **N**eural **N**etwork
CBR	**C**oncept **B**ased **R**epresentation
CSMF	**C**ause **S**pecific **M**ortality **F**ractions
DFR	**D**ocument **F**requency **R**epresentation
DT	**D**ecision **T**rees
HDSS	**H**ealth **D**emographic **S**urveillance **S**ystem
HPC	**H**igh **P**eformancer **C**omputing
IDE	**I**ntergrated **D**evelopment **E**nvironment
LIME	**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations
LSTM	**L**ong **T**erm **S**hort **M**emory
ICD-10	**I**nternational **C**lassification of **D**iseases-10
ML	**M**achine **L**earning
NHLS	**N**ational **H**ealth **L**aboratory **S**ervices
NLP	**N**atural **L**anguage **P**rocessing
PHMRC	**P**opulation **H**ealth **M**etrics **R**esearch **C**onsortium
SCI	**S**ymptom **C**ause **I**nformation
SEDL	**S**tacked **E**nsemble **D**eep **L**earning
SNMCT	**S**ystematic **N**omenclature of **M**edical **C**linical **T**erms
TFR	**T**erm **F**requency **R**epresentation
TFIDFR	**T**erm **F**requency **I**nverse **D**ocument **R**epresentation
RF	**R**andom **F**orests
RNN	**R**ecurrent **N**eural **N**etwork
SVM	**S**upport **V**ector **M**achine
VA	**V**erbal **A**utopsy
UMLS	**U**nified **M**odelling **L**anguage **S**ystem

Dedicated to my late mum Agatha Chiratidzo and father Michael Pamba Mapundu…

1 Chapter 1

1.1 Background on VA

More than 65 percent of the population in the world lack high quality information on cause of death (CoD), accounting to two thirds of the of sixty million deaths worldwide, that are not assigned a medically certified cause [1]. Additionally, only 59 out of 195 countries report on autopsy rates, accounting for 30.3%. This is depicted in Figure 1.1 and 1.2. As such, most of the countries in the world fail to meet the United Nations 90 percent death registration coverage requirement, as deaths in many Low to Medium Income Countries (LMICs) are not captured in civil registration systems [2, 3]. On the contrary, the CoD information is vital for public health monitoring, informing critical health policies and priorities. Therefore, in the absence of clinically oriented sources, CoD information should be derived from alternative sources. Verbal Autopsy (VA) is the mostly used tool worldwide as an alternative source of CoD information (see Figure 1.2). VA is common in LMICs and is a process that is used to determine CoD where deaths occur outside health facilities and is not certified by a medical practitioner [4]. These sentiments are supported by [5], who also report on the importance of the VA process in determining CoD in countries where there are no active civil registration systems.

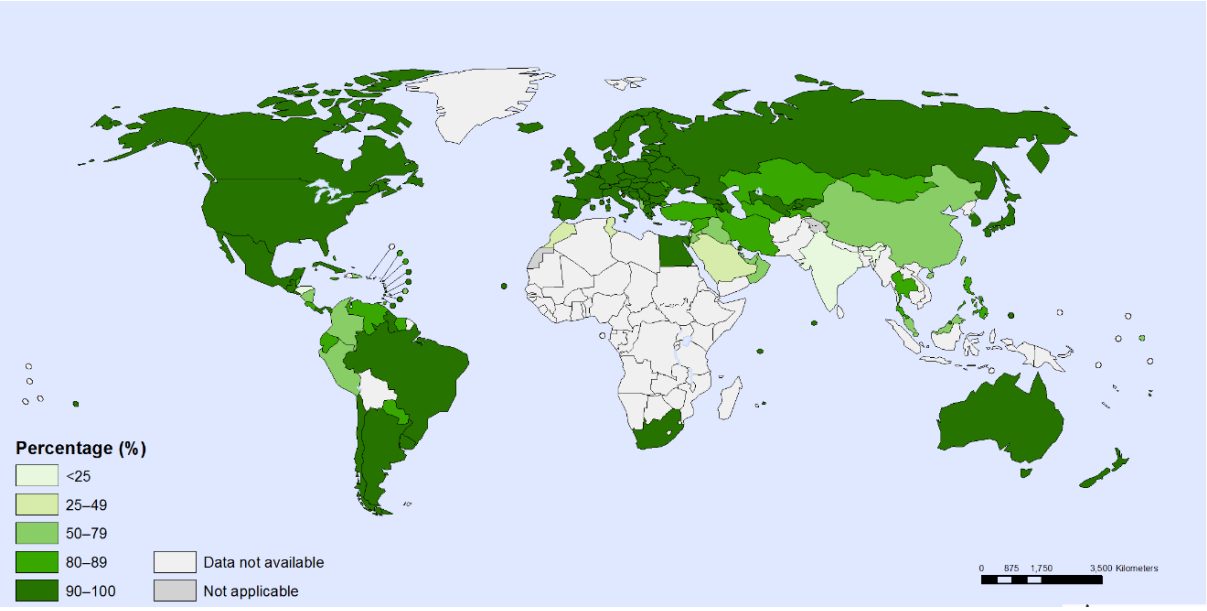


Figure 1.1: Completeness of CoD coverage (2007-2016), Source: https://www.who.int/images/default-source/maps/global_civilregistrationdeaths_2007_2016.png?sfvrsn=5f760b4f_0

1.2 VA process

The VA process is conducted by a non-medical personnel who seek to elicit valuable information using both structured questions and an open narrative section with next of kin of the deceased about circumstances and

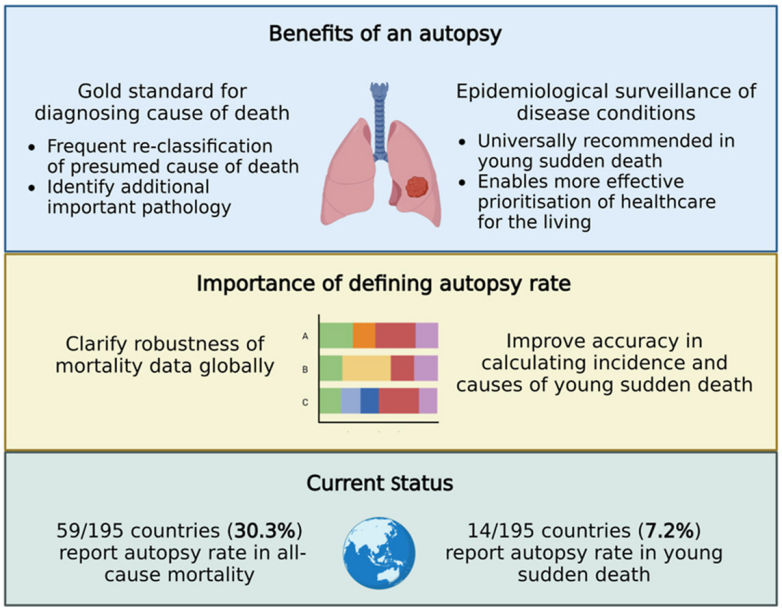


Figure 1.2: Autopsy rates, Source [6]

events that led to death [1]. Two doctors are given the full set of responses, both from structured questions and open narratives for assessment and reach a consensus on the CoD and if not a third physician is consulted, a process known as Physician Coded Verbal Autopsy (PCVA). This is shown in Figure 1.3. PCVA is the mostly used process for determining CoD. However, it is widely criticised because of lack of robustness, cost, time, inconsistencies and inaccuracies as it is subjective and prone to errors amongst many drawbacks [7]. This results in PCVAs mostly employed for training and validation of computational approaches.

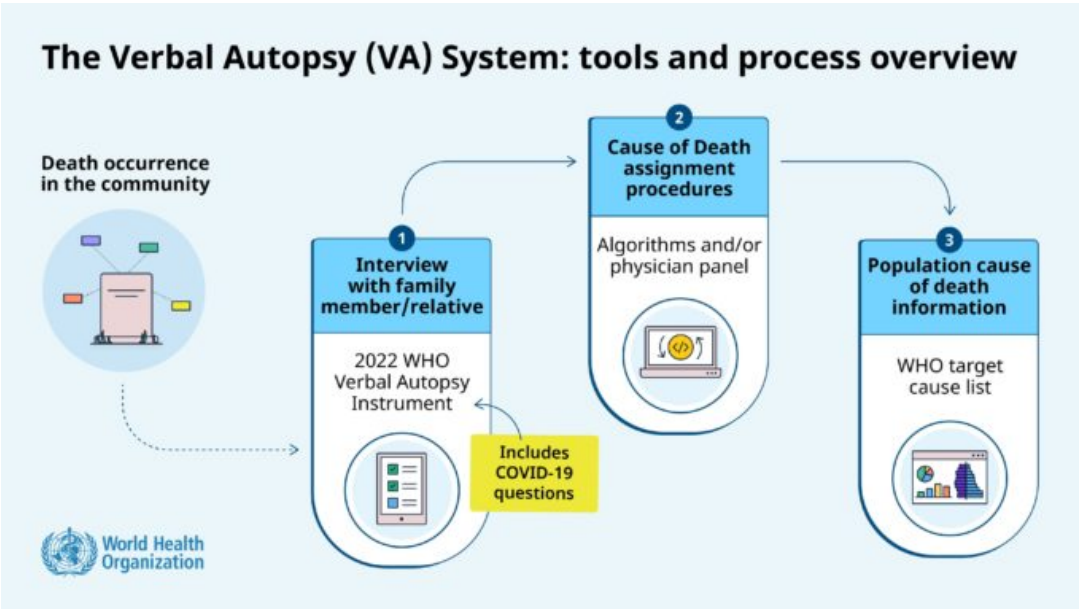


Figure 1.3: VA Process and tools, Source: https://www.publichealthnotes.com/verbal-autopsy-and-mpdsr/

1.3 ML in VA

The surge of technological advances has availed a plethora of automated methods for determining CoD which are faster, efficient and cost effective [1]. ML combined with deep learning can be powerful, effective and

efficient tools in developing better semantic structures to solve issues around automated VA classification. ML can be divided into supervised, unsupervised, semi-supervised, reinforcement and deep learning.

Supervised machine learning seeks to maximize the training data that contain the input and output variables. It takes the training data as input that is fed to the mapping function to come up with new predictions on new data, through learning. This approach has a labelled dataset done by doctors on possible CoD. The most common supervised ML problems are classification and regression. On the contrary, unsupervised ML techniques do not need a labelled dataset, they are applied to identify trends and patterns in data [8].

Semi-supervised learning is a machine learning approach that is a combination of supervised and unsupervised learning, and involves using a small amount of labelled data along with a larger amount of unlabelled data to train a model. The idea is to leverage the labelled data to help the model learn better patterns and make more accurate predictions on the unlabelled data. This approach is particularly useful when labelling data is expensive or time-consuming.

Reinforcement learning is a type of ML where an agent learns to make decisions by interacting with an environment. The agent takes actions, observes the results, and receives feedback in the form of rewards or penalties. Over time, the agent aims to maximize cumulative rewards by learning the best strategy or policy to achieve its goals. This type of learning is often used in scenarios like game playing, robotics, and autonomous systems. Deep learning entails various layers made up of neural networks that process data to learn hierarchical abstract representations of data [9]. Figure 1.3 shows the various ML types.

Most of the research that reports on ML application in the VA domain mainly use the responses from the questionnaire as the classical dataset. As such, this affects comparability and generalisability. In this study we validate the performance of various ML techniques using various VA data types for determining CoD using a comparative analysis approach. We apply enhanced data standardisation and normalisation strategies to achieve optimum transparency and accuracy through addressing most model limitations and applying recommendations that are reported in [10, 8]. We assess the robustness of several classifiers including; random forest (RF), k-nearest neighbour (KNN), decision tree (DT), support vector machine (SVM), logistic regression (LR), artificial neural network (ANN), Bayes Classifier (BC), bagging and eXtreme Gradient Boosting (XGBoost) as ensemble classifiers. We also validate our dataset using the common conventional Computer Coded Verbal Autopsy (CCVA) algorithm; InterVA.

Advances in technology specifically in ML and NLP have availed new opportunities and proficiencies in processing and classification of textual narratives. In this study we present novel robust deep learning (DL) architectures that use natural language processing, (NLP) that can effectively process VA narratives and predict probable CoD. Consequently, discovering new knowledge, presenting valuable insights and semantic relationships between the VA narratives and their corresponding probable CoD. Even though DL techniques are more accurate in terms of prediction, they have a limitation in terms of complexity and model interpretability (understanding how the model performed the way it did). Despite these challenges, DL has shown promising results because of their efficiency, superior performance, integrated feature learning and effective capabilities of attaining end to end learning from complex and multi modality data [9]. Conversely, conventional statistical and ML techniques require firstly to perform feature engineering to attain effective and robust features representative of the data, and then build predictive models. There is scanty literature, especially in applying advanced ML techniques known as DL models as approaches to solving VA problems.

In most cases VA narratives entail information that is encapsulated in the textual format, usually unstructured and is stored in large volumes. Consequently, it difficult to manage, takes time to process and suffers from

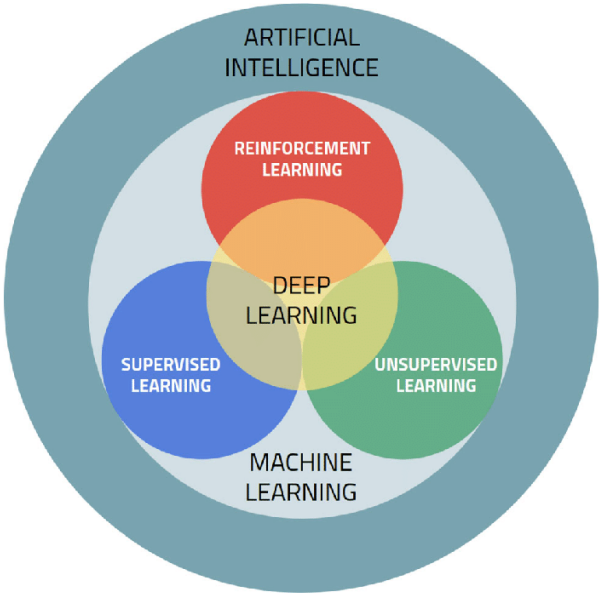


Figure 1.4: ML types

word polysemy, amongst many challenges. This therefore makes the analysis and harnessing of meaningful information difficult, even though it is vital in identifying mortality causes. As such, the application of text mining (TM) which employs natural language processing (NLP) in an automated fashion can be beneficial to extract meaningful information regarding mortality causes and most prevalent diseases from verbal autopsy narratives. There has been several TM approaches using supervised learning that have been applied to date within the verbal autopsy domain to determine cause of death, nevertheless they require large datasets which are labelled for training the machine learning models. This also makes the process tedious, time consuming, expensive and prone to human error as the experts are the ones responsible for labelling the datasets. This study seeks to address this gap by reducing the need for human intervention, by implementing supervised and unsupervised ML approaches. We detail more on the literature in chapter 2.

1.4 Problem Statement

In the context of rural South Africa, there exists a critical need to comprehensively understand the morbidity and mortality patterns that shape the health landscape of these communities. Limited access to healthcare resources, unique socio-economic dynamics, and geographical isolation contribute to distinctive health challenges that require targeted interventions. Traditional epidemiological methods often struggle to capture the complexity of these patterns, necessitating the integration of computational approaches. The problem statement of this research revolves around addressing the following questions: How can computational methodologies effectively analyze and characterize morbidity and mortality patterns in rural South Africa? How can these approaches reveal insights into prevalent diseases, health disparities, and potential correlations between socio-economic factors and health outcomes? By tackling these questions, this study aims to provide a foundation for data-driven, tailored public health strategies that can lead to improved healthcare planning and resource allocation in rural South African communities, and the world at large.

Automated understanding of domain specific narratives in order to extract useful associations for classification and decision making is a key issue. One such relationship is that of cause of death and associated descriptions as VA narratives. The process of performing VA using narratives has become a key issue in terms of lack of standardisation, accuracy, quality of data, incompleteness among many drawbacks. Furthermore, there is

no good means of comparing and distinguishing best performing algorithms as narratives of VA are usually captured and stored in text format. This information is key for HDSS sites and civil registration systems for policy making and health priorities. Although the VA narratives provide useful information, there is need to apply alternative methods for VA narrative classification that are more efficient, consistent, and accurate. The proposed ML computational techniques will enable efficient and effective standardisation of VA classification, and this will give us an insight on the determinants of events and circumstances leading to cause of death to identify morbidity and mortality occurrences in rural north-east South Africa from 1993 to 2015.

1.5 Justification

The justification for conducting a study on computational approaches to characterizing morbidity and mortality patterns in rural South Africa stems from the critical need to address the unique healthcare challenges faced by these communities. Traditional epidemiological methods often fall short in capturing the multifaceted nature of health dynamics in rural settings due to their complex interplay with socio-economic, cultural, and geographical factors. By harnessing computational techniques, this study aims to provide a more nuanced understanding of disease prevalence, health disparities, and mortality trends within these communities. Moreover, the study's justification lies in the potential for data-driven insights to inform targeted interventions and policy enhancements. By uncovering hidden correlations between health outcomes and socio-economic variables, decision-makers can tailor interventions to address the specific needs of rural South African populations. This study's approach holds the promise of providing a comprehensive and accurate portrayal of health challenges, enabling evidence-based strategies for improved healthcare planning and resource allocation.

Automated processing and reasoning through algorithms have been applied in VA [11, 12, 13, 14, 15, 16, 17, 18], in order to come up with data driven predictions to classify causes of death by using statistical and mathematical models based on captured narratives. Algorithms take input of narratives in text which is processed by a natural language systems to transform textual data to numeric format which can be used along with machine learning algorithms to perform cause of death classification.

It is key to apply ML computational techniques as it uses large datasets to perform data pattern analysis. New rules can be learnt from data and applied to diagnose and predict outcomes. [19] points out that machine learning models perform better if large datasets are employed, and the model can then understand better the input-output relationship. Moreover they ML approaches are; effective, use independent parameters which are non parametric, fast in testing, produce results in complex domains, reduce problem complexity, easy to implement and employ more resembling characteristics of the model [14]. Machine learning techniques are better in that they are extensive, comprehensive, highly reliable and consistent relative to changes [4]. Few studies report mainly on these conventional approaches [20, 21] as compared to deep learning. The application of shallow models which are trained on very high dimensional and sparse features has mostly been evident in VA domains. Literature reports a wide gap, especially in applying advanced machine learning techniques known as deep learning models as approaches to solving VA problems. There has not been full utilisation of deep learning and its capabilities in the health domain, even though they demonstrated great performance and potential in other areas such as computer vision, speech recognition and NLP tasks. There is great need to invest in application of Natural Language Processing (NLP) as a deep learning approach to solve VA problems. Discovering new knowledge and valuable insights in health domains is key even though the data has various challenges. Applying traditional ML techniques requires firstly to perform feature engineering to attain effective and robust features representative of the data, and then build predictive models. The new approaches of deep learning such as NLP have efficient, superior performance, integrated feature

learning and effective capabilities of attaining end to end learning from complex and multi modality data [9]. NLP's deep learning methods such as word embeddings on dense vector representations, enabling multi-level automated feature representation learning, has to date produced excellent results [22]. Recent deep learning methods such as Convolutional Neural Networks (CNN), Recurrent Neural Networks(RNN), Long Short Term Memory (LSTM) and the transformer architecture have managed to explicitly solve most NLP tasks [9].

Motivated by the recent advances in both theoretical and methodological aspects of deep learning in text vectorisation and classification, we seek a deep understanding in the determinants of events and circumstances leading to cause of death by applying ML computational techniques to identify morbidity and mortality occurrences in rural north-east South Africa from 1993 to 2015. The proposed research is not only relevant to VA domain only, it is also appealing and relevant to areas that involve text narrative conversion to numeric format for further processing, thus availing further insights and opportunities to explore with the data. Moreover, we seek to add to existing body of knowledge through developing a centralised dataset that will serve as a baseline, and can be used and shared by researchers in order to effectively compute VA models. The study will also investigate the best NLP method to convert narratives to numeric form. Furthermore, we will validate the performance of ML methods for VA classification. This implies saving time, improving on efficiency, accuracy and consistent diagnosis, thus enforcing a standard for determining cause of death. The problem will be addressed through developing and implementing an efficient ML framework for automated VA narratives.

1.6 Research aims, questions and objectives

1.6.1 General aim

The aim of this study was to employ computational approaches to delve into the complex landscape of morbidity and mortality patterns in rural South Africa. By leveraging advanced data science techniques, the research seeks to uncover hidden insights, correlations, and trends within health data. The overarching goal is to contribute to evidence-based decision-making in public health policy and practice. The study spans from 1993 to 2015, examining the determinants of events leading to causes of death, and thereby informs strategies for health system strengthening and targeted interventions. The study will seek to add novelty and contribution to the health domain in informing policy and practice. As such the study will ease the development and implementation of health priorities through identifying critical elements for strengthening of health systems and determining the key health services and programmes.

1.6.2 Research questions

The study seeks to address the following research questions:

1. What are the common prevalent topical diseases that led to death at Agincourt HDSS in rural north-east South Africa between 1993 and 2015?
2. How is mortality clustered by cause of death within households, and what characteristics of households are associated with high mortality at Agincourt HDSS?
3. To what extent can machine learning and deep learning techniques accurately classify cause of death as compared to physician classification?

1.6.3 Objectives

The proposed study will seek to achieve the following objectives:

1. To determine the most prevalent common topical diseases leading to deaths using VA narrative datasets and text mining techniques.
2. To identify mortality clusters by cause of death and investigate the mortality characteristics associated with households at Agincourt HDSS.
3. To establish ML accuracy in automating VA classification and achieve at least the same level of accuracy with that of physician classification on cause of death in rural north-east South Africa between 1993 and 2015.

1.6.4 Data source

The data collection process for health and socio-demographic data was longitudinal in nature. Various data collection approaches have been applied to date including a census that was conducted in 1992, where each household was visited and all residents were registered. Moreover, there has been efforts to collect data annually from 1993 to 1999. The other approach was to collect data on birth, death, in-and out migrations at 15-18 monthly periods. There have been special census modules introduced from the year 2000 for health and population dynamics conducted at frequent intervals, and at individual and household level. In cases where there has been a death, there has been a record of place of death, name of hospital if death occurred at a hospital and also if that death was registered. Moreover, data has been collected for all possible symptoms that might have caused death through VA interviews. If a death occurs outside clinical facilities, caregivers will interview the deceased's relatives after a month to a year interval. VA data has been used to determine the possible cause of death for each death occurrence.

The process for assigning cause of death involves two doctors independently reviewing the VA data to assign the possible cause of death. If they do not agree then a third opinion is sought. If all three assignments do not match then the cause of death is classified as "ill-defined". However there has been application of VA algorithms which are automated to do the assignment. Information collection has been done through administering a questionnaire to adult members of the households. It has been common practice to carry out the interviews at most two times and if needs be also interview neighbours in order to get the required information. The Agincourt HDSS commenced data automation using FoxPro data management system, and migrated to Microsoft Access 95 with some data model improvements being initiated in 1996. With changes being introduced to standardise according to INDEPTH population reference model, they migrated their data management system in 2002 to Microsoft SQL Server platform, and are still currently hosting their data on that platform.

1.7 Conceptual Framework and overview

This study will be within the conceptual framework of the epidemiological transition, which describes the mortality and health dynamics in populations over time and place and their relationship to the socio-economic and demographics determinants [23, 24]. The authors argue that people in society experience mortality and disease patterns that fall into three categories, which are; i) pestilence and famine phase, ii) receding pandemics and iii) degenerative and man made phase. The first phase exhibits high mortality due to wars, famines and infectious disease epidemics. The second stage has a decline in mortality where infectious disease

are now key in morbidity and death. The last phase presents chronic diseases as major causes of mortality. Furthermore after the third phase there is decline in mortality after which the levels become constant.

However, recent literature suggests otherwise, in that there is a fourth phase which applies to developed countries where mortality continued to decline. This fourth phase is known as "the age of declining cardiovascular mortality, ageing, lifestyles modification, emerging and resurgent diseases". This fourth new phase evidenced a transition in mortality from the third phase and resulted in more life expectancy for the aged population [25]. Nevertheless there has been contradictions in the earlier conceived transition theories, in that these transitions are non linear and can take various directions in evolution. This gave rise to new theories that applied in developing countries [26]. [27] report on new theories with various phases. They report on the stage of "counter transition" as changes that can occur in mortality and disease patterns, "protracted epidemiological transition" as dynamics in mortality and disease patterns with possible multiple diseases affecting same population and lastly the "epidemiological polarisation" as being the transformations in disease patterns and mortality highly evident in the poorer population. These sentiments were supported by [26] who argued that a surge in mortality in the early 1990's was witnessed and gave birth to communicable and non-communicable diseases, which have driven the need to adopt the above theories within South African populations. This study will seek to find answers to the counter transition, protractedness and polarisation in line with the epidemiological transition within the South African context. This will be achieved through applying computational approaches to identify events and circumstances that led to cause of death by characterising morbidity and mortality patterns and trends.

In this study, we used three types of datasets. The first dataset is the structured responses from the standard questionnaire, second dataset is the VA narratives, and third dataset is a combination of the responses and the narratives. The three datasets comprised of 287 columns/features and 16338 records/observations. For the responses only, we took all features that had responses from the standard questionnaire as our predictors and the cause of death assigned by physicians using International Classification of Diseases-10 (ICD-10) code for each record in the dataset as our target variable. Ultimately, we had 231 predictors (all symptoms, age at death and gender) and 1 target variable, and all our features were in English. The predictions using the narratives were done using age at death, gender and the narrative feature and 1 target variable. For the combined VA dataset we used 232 predictors and 1 target variable. We only added the VA narrative feature to the responses dataset in order to have our combined dataset. We further created twelve cause of death categories with corresponding labels, class distribution with number of samples for each class before and after data balancing for our training dataset. The cause of death categories were derived based on InterVA user guide.

Figure 1.5 below depicts the systematic structure of our conceptual framework and overview of the phases and how they relate to the study objectives. Phase I denotes unsupervised learning and extraction of mortality causes and most prevalent diseases using text mining approaches. This relates to objective 1. In this study, we explored with descriptive statistics and visualisation approaches in order to discover new information. Additionally, objective 2 was addressed through clustering to identify mortality clusters by CoD, and investigate the mortality characteristics associated with households. Phase II was based on studies related to objective 2 and entailed performance evaluation of automated algorithms in CoD determination using statistical, CCVA, ML and DL techniques. Phase II gave us insights on how to progress with Phase I work.

1.8 Proposed Analytical ML Framework and Architecture

Figure 1.6 below is a detailed proposed ML framework architecture for our study. It gives an overview of the back end and front end processes for all our studies. The back end will have our conventional ML techniques

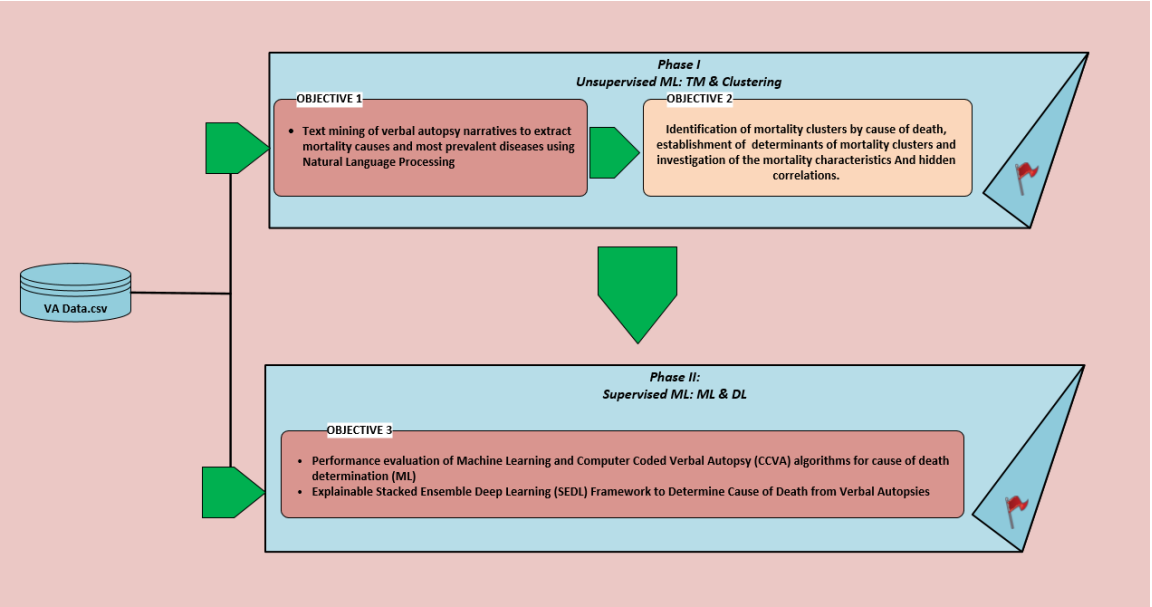


Figure 1.5: Conceptual framework illustrating the structure of the thesis.TM: Text mining, ML:Machine learning, DL: Deep learning

taking input of text narratives and pushing to the data preprocessing engine for cleaning and normalisation. Thereafter, selected features will be pushed to the data modelling and engineering engine for training, testing and validation. Using grid search and 10-cross-fold validation, we will achieve optimum classifiers for our models. Class labels for our models will be stored on a centralised numeric database and served to the web application on our front end component of the system. Therefore, the end users from our front end can push narratives to our proposed real-time VA classification framework and get instant results.

1.9 List of publications

1. **Published paper:** Mapundu MT, Kabudula CW, Musenge E, Olago V, Celik T. Performance evaluation of machine learning and Computer Coded Verbal Autopsy (CCVA) algorithms for cause of death determination: A comparative analysis of data from rural South Africa. *Front. Public Health, 27 September 2022, Digital Public Health Volume 10 - 2022*. https://doi.org/10.3389/fpubh.2022.990838.

2. **Published paper:** Mapundu MT, Kabudula CW, Musenge E, Olago V, Celik T. Explainable Stacked Ensemble Deep Learning (SEDL) Framework to Determine Cause of Death from Verbal Autopsies. *Mach. Learn. Knowl. Extr. 2023, 5(4), 1570-1588*. https://doi.org/10.3390/make5040079.

3. **Published paper:** Mapundu MT, Kabudula CW, Musenge E, Olago V, Celik T. Text mining of verbal autopsy narratives to extract mortality causes and most prevalent diseases using Natural Language Processing. *PLoS ONE 19(9): e0308452*. https://doi.org/10.1371/journal.pone.0308452.

1.10 Conference presentations

1. 1ST VUT INTERDISCIPLINARY RESEARCH AND POSTGRADUATE CONFERENCE, 09-10th November 2016 – VUT Southern Gauteng Science and Technology Park, Sebokeng Campus. Theme: *Advancing Research that Matters: Developing a Vibrant Research Culture*.

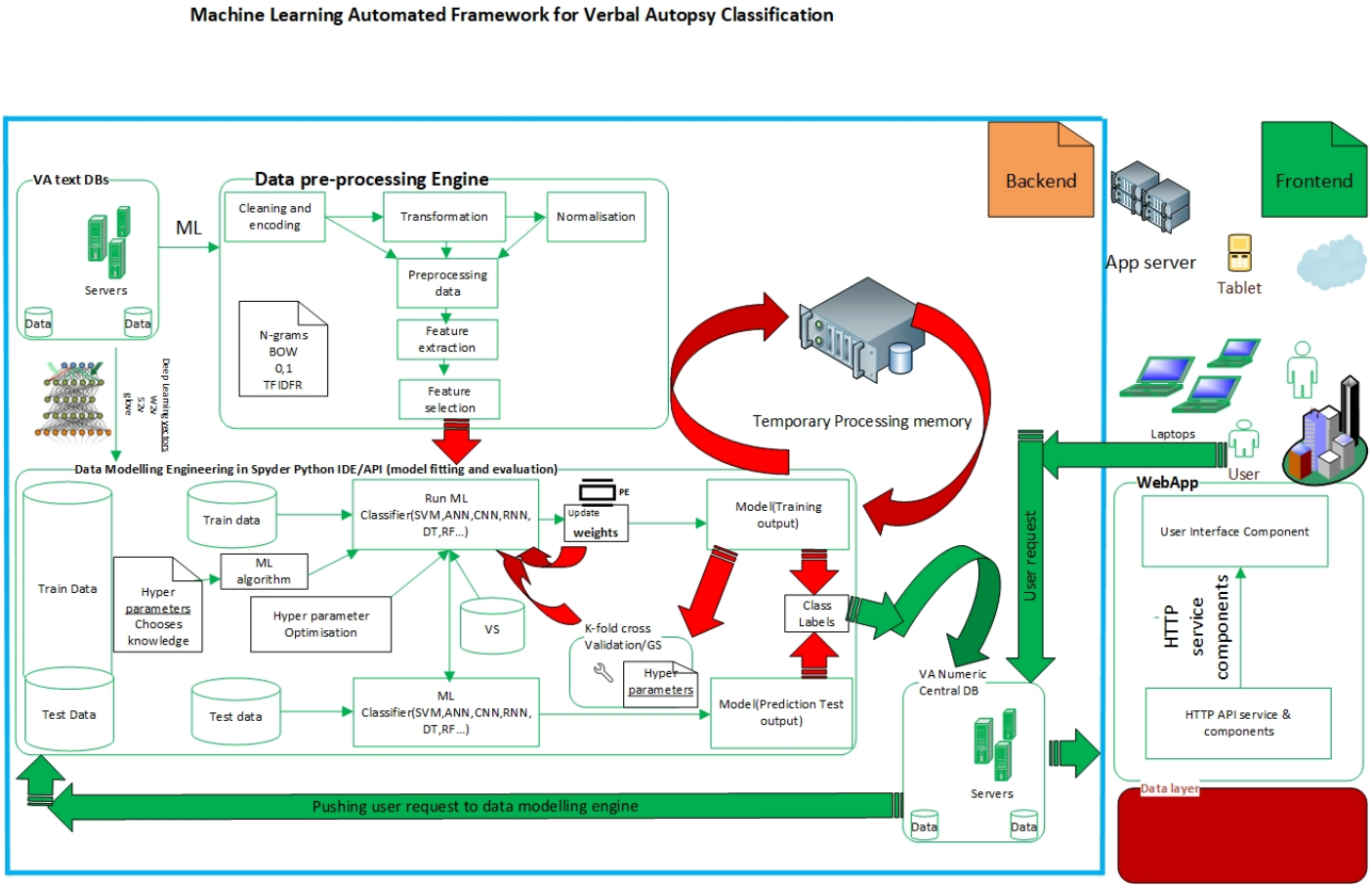


Figure 1.6: ML Framework structure and Architecture

2. Using big data techniques to improve prostate cancer reporting in the Gauteng province, South Africa. *3rd Pathology Research and Development Congress – PathReD 2019. 18-21st July 2019*. Oral presentation.
3. Using big data techniques to improve prostate cancer reporting in the Gauteng province, South Africa. *3rd Pathology Research and Development Congress – PathReD 2019. 18-21st July 2019*. Poster presentation.
4. Using supervised machine-learning algorithms to assign prostate biopsy histological findings. *Wits School of Public Health Research Day 22 August 2019, Theme: Innovative public health teaching and research methodologies. Oral presentation*. Oral presentation.
5. Explainable Ensemble Deep Learning Framework to Determine Cause of Death from Verbal Autopsies. *Wits School of Public Health Research Day 14/15 September 2023, Theme: Innovative public health teaching and research methodologies. Oral presentation*.

My involvement in the PhD research work includes being the first and corresponding author in all the thesis papers in addition to the following contributions:

- Conceptualisation and study design
- Implementation of data analysis
- Result interpretation
- Writing original manuscript
- Addressing reviewers and co-authors comments

- Proofreading all accepted manuscripts

1.11 Structure of the Thesis

The thesis is divided into eight chapters. Chapter one gives a detailed report of the research aims, and justification of the current study. Chapter two discusses a comprehensive systematic review that focusses on most ML techniques that have been applied to date. Chapter three is a comparative analysis of ML and CCVA algorithms in cause of death determination. Chapter four discusses unsupervised TM approaches that apply natural language processing approaches such as; n-gram language processing, Latent Dirichlet Allocation and BERTopic to extract mortality causes and most prevalent diseases. Chapter five explores if DL methods are feasible in determining causes of death and enforces trust in the use of ML algorithms in CoD determination using VA narratives. Chapter six presents the discussion and summary of the entire thesis and points to future directions.

2 Statistical and Machine Learning Techniques in Verbal Autopsy: A Systematic Literature Review

This study is a draft manuscript

Michael T. Mapundu[1]*, Chodziwadziwa W. Kabudula[1,2], Eustasius Musenge[1], Victor Olago [3], Turgay Celik[4,5].

[1]Division of Epidemiology and Biostatistics, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Parktown, Johannesburg, South Africa
[2]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), University of The Witwatersrand, Johannesburg, South Africa
[3]National Health Laboratory Service (NHLS), National Cancer Registry, Johannesburg, South Africa
[4]Wits Institute of Data Science, University of The Witwatersrand, Johannesburg, South Africa
[5]School of Electrical and Information Engineering, University of The Witwatersrand, Johannesburg, South Africa

2.1 Abstract

Background:

The process of determining causes of death in areas where there is limited clinical services using verbal autopsies has become a key issue in terms of accuracy on cause of death (prone to errors and subjective), quality of data among many drawbacks. This is mainly because there is no proper standard available in performing verbal autopsy, even though it is important for civil registration systems and strengthening of health priorities. Physician diagnosis is the only gold standard in reviewing verbal autopsy narratives. In practice, conventional statistical methods are used to perform verbal autopsies due to their simplicity and transparency. However, in literature complex machine learning models can be found that can replace the traditional statistical methods. There has not been much application of machine learning techniques in verbal autopsy to determine cause of death, despite the advances in technology. As such, there is a need for a thorough survey of recent literature on statistical and machine learning approaches applied in verbal autopsy to determine cause of death.

Methods:

A systematic review was conducted and included a search from six databases. Our study only included scientific articles published in last decade that reported on verbal autopsy and: (1) algorithms; (2) statistical techniques; (3) machine learning and (4) deep learning. The search yielded 110 articles, after methodological systematic analysis, we identified 85 articles as being relevant and discarded the other 25. We investigated and compared the most commonly used statistical and machine learning techniques in VAs, identified limitations of each of these techniques, proposed a guiding machine learning framework and pointed to future directions.

Results:

Eighty five studies met the inclusion criteria. Apart from physician diagnosis, statistical methods are the most currently applied tools to determine cause of death from verbal autopsies. However, there has been little application of traditional machine learning and emerging techniques, even though they have shown promising results in other domains.

Conclusions:

Technological application of machine learning to determine cause of death, should focus on effective ideal strategies of pre-processing, transparency, robust feature engineering techniques and data balancing in order to attain optimal model performance.

Keywords: Verbal Autopsy, Machine learning, Algorithms, Natural Language Processing, Deep learning

2.2 Background

Half of the countries of the world do not meet the United Nations 90 percent death registration coverage requirement because deaths in many developing countries are not captured in civil registration systems [2, 3]. In addition, 65 percent of the population in the world lack high quality information on cause of death since every year about sixty million deaths worldwide are not assigned a medically certified cause [1]. Since cause of death information is required to inform critical health policies and priorities, in the absence of the more traditional, clinically oriented sources, cause of death information should be derived from alternative sources. One widely adopted alternative source of cause of death information is Verbal autopsy (VA). As stated by [3], the VA concept emanated in Africa and Asia in the 1950's and involves having dialogue with the family member of the deceased who provides information about any possible preceding conditions that may have led to the cause of death in their opinion. Literature to date suggests that the VA approach was first adopted in India and subsequently research studies using VA surfaced in the late 1970s and early 1980s in Niakhar in Senegal and Matlab in Bangladesh respectively [3]. The use of VA as a source of cause of death information is now common in many developing countries where access to clinical services, death registration and consequently medical certification of causes of death is limited such as Bangladesh, India, Malawi, Niger, South Africa and Nigeria [4].

The process of determining causes of death from VAs is traditionally done through the Physician Certified Verbal Autopsy (PCVA) approach. This approach involves two physicians reviewing the responses to questions from VA interviews to reach an agreement on cause of death, where they do not agree a third physician decides on cause of death [28]. The PCVA approach is regarded as the gold standard for determining causes of death from VA and is used for training, testing and evaluation of automated approaches for determining causes of death from VA. However, the PCVA has a number of shortcomings as discussed in section 2.4. In

order to improve the determination of causes of death from VA and address some of the weaknesses of the PCVA approach, automated approaches also known as Computer Coded Verbal Autopsies (CCVA) are used as alternative approaches. CCVA methods are data driven and employ automated processing and reasoning. They take as input, the responses to the questions of the VA interviews and maps them to corresponding target ICD-10 cause of death codes or other recognized standard categorization [29, 30, 4]. As argued by [31], one problem with data driven methods of determining causes of death from VA is reliance on clinically confirmed causes. Handling of missing data also poses a further challenge as the norm is to treat them as unknown, thus creating a bias in the interpretations.

Driven by recent advancement in technology, there has been a surge in application of computational tools that accurately analyse data for decision making. However, the quality of data is compromised by incompleteness and inaccuracies, amongst other drawbacks [32]. Awareness of these problems is required in research communities in developing and underdeveloped countries [33, 34]. Computational tools use machine learning (ML) techniques that learn from experience relative to some tasks within a class and make use of a performance measure or weighting to improve the performance. Machine learning techniques employ some statistical, probabilistic and optimisation technique in order to discover patterns and trends in complex data through some analysis to get to a decision [35]. There are various statistical and ML techniques that have been adopted in health care to provide required services and address specific problems, such as various classification and regression models that exist to date [36, 37, 38, 39]. [1] argues that to date, ML approaches have been predominantly employed to data generated from only the structured questions of VA interviews, attaining true positive prediction scores of 0.60 for individual cause of death categorization. However, they included several numbers of cause of death classes. Available evidence indicates that ML can generate real time cause of death information that is similar to that of physicians/experts [40]. Moreover, there has not been full utilisation of deep learning and its capabilities in the health domain, even though they demonstrated great performance and potential in other artificial intelligence (AI) areas and NLP problems. Further insights and critique of existing literature on CCVA is discussed in later sections of this methodological systematic analysis.

In existing research, there is evidence of application of complex machine learning (ML) algorithms that can substitute conventional PCVA and some existing CCVA algorithms as approaches for determining cause of death from VAs [11, 15, 41, 17, 18]. Even though these ML approaches yield high accuracies in model performance in terms of classification, they lack clear explanations on model interpretability (lack of justification by the researcher on model performance) [40]. There is need to have ML models that are fair, accountable and transparent. Using only hospital based deaths [28] conducted a study where they compared PCVA and automated methods and concluded that out of the two methods, no particular method is considered best in terms of performance. There is need for further inquiries with computer coded techniques in order to improve performance, thus there is need for publicly available datasets such that comparisons, inferences and deductions can be done.

Subsequently, to the best of our knowledge as of the time of this write-up, there is little evidence that suggest existence of studies that have done a methodological systematic analysis of literature on the use of statistical and ML approaches (including deep learning) in VA. Therefore, there is a need to have a systematic and comprehensive literature review in order to have an objective understanding of these statistical and ML applications in VA. The methodological systematic analysis process highlights various studies in which statistical and ML were applied in VA, thus giving researchers and any other interested parties an idea of the problems

and issues encountered. By so doing this improves the application of statistical and ML in VA and strengthen health priorities. Although ML models often achieve high accuracy scores, they generally struggle to explain their predictions. This is due to factors such as human errors, subjectivity, and the challenges posed by fragmented and imbalanced datasets, which are used as inputs for VA narratives and mapped to corresponding cause of death target values according to ICD-10 codes. In this systematic review, our goal is to do a methodological systematic analysis of existing literature on statistical and machine learning models which were applied in VA between 2010 and 2020 and propose ML framework guidelines for VA. This paper is organized as follows. Section I is the background and has subsections of; feature engineering and feature selection techniques, statistical and machine learning models, evaluation metrics, ML optimisation techniques, imbalanced and fragmented datasets, limitations and assumptions of models, emerging trends and current technological dynamics in VA and limitations in literature. Section II briefly highlights our methodology. In Section III we present the results and proposed guiding framework for machine learning in VA. Section IV provides a critique of the discussion. Section V is the conclusion. Section VI is a list of abbreviations. All declarations are in section VII. Figure legends are presented in section VIII.

2.2.1 Feature selection and feature engineering

This extract highlights familiar and mostly applied feature selection (FS) and feature engineering (FE) approaches in VA. The aim of FS and FE is to come up with a vector space that is efficient, scalable and accurate. FS is a process of generating new relevant input features with highest score or weighting from existing ones and FE is generating a subset of features from current features [13, 14]. FE has three distinct sub-processes and commences with feature extraction, then feature value representation and lastly feature selection [8].

Feature extraction

Feature extraction is a process that seeks to pull out relevant features from a set of reports/narratives or any source of information. Thus feature extraction allocates a document from a pre-determined set of categories [15]. Various fully automated feature extraction approaches exist in literature [8]. The following extracts discuss various automated feature extraction approaches.

Bag of words (BoW) extracts words that are unique from all narratives existent in the file regardless of their categories and stores in a list (BoW). The BoW technique splits text into words and only includes useful words in the feature vector as dimensions (words) extracted. Each word represents an autonomous and particular feature [8, 42]. However, the BoW ignores the order of words and outputs a generated reduced form of occurrences of each word in text narratives using word frequency [18].

n-grams are a set of words that are sequential as they make use of continuous number of items such as characters or words from a given sequence of narratives. n-gram models can be of the form; a) n = 1 (unigram), b) when n = 2 (bigram), c) n = 3 (trigram) and d) hybrid-grams (mixture of unigram, bigram, and trigram) [8].

Word2vec, exists in two forms namely; skip-gram and continuous bag of words (CBoW) models. The skip-gram model does it's learning repeatedly using existing words presented in a sentence to predict the next word. On the contrary, the CBoW model predicts the prevailing word using nearby words. Both approaches determine a limit on number of words by applying the parameter window size technique [8].

Concept-based feature extraction is mainly used to extract medical concepts using medically based ontologies where the relationship between terms is difficult to establish. This is common when there are various terms that mean the same concept. Concept-based feature extraction can be done through a bag of concepts or bag

of phrases [8].
Graph-based or graph of word (GoW) features are a graphical representation of features using content-based features. The aim of GoW is to represent the order of words, where each given narrative is depicted in a graph structure. The vertex of each graph contains a particular selective feature of the given narrative. Edges are used to connect two co-existing features, where each edge has the weight calculated using the frequency of nearby vertices in the graph [8].
Doc2vec is an approach that applies an unsupervised learning algorithm to come up with vector representations of paragraphs or documents. The algorithm is a modification and application of the concepts of word2vec, and aims at finding similarities in paragraphs or documents [18].
Rule-based feature extraction is used in case-based reasoning which employs transfer learning (mainly applied when input has few labelled instances that can be used to train an accurate model through transferring capabilities from existing systems to untrained ones). The rule-based approach extracts features that are taken as input by the system's reasoning inference engine. Manually crafted rules are fed into the information extraction module, and searching can be done to attain features and mapped to corresponding values. Conversion can now be done to transform them into a machine readable format for processing. The whole process is automated and makes use of predefined rules [43].

Of all the above discussed feature extraction techniques, [8] suggest the BoW as the commonly used approach for feature extraction as it maintains word frequency. However, this approach does not take into account the order of words and syntax. As an alternative to the BoW and to address some of the above drawbacks, the n-gram technique for feature extraction was proposed. Nevertheless, the n-gram approach fails to acquire the meaning of opposite words and does not consider similar words (synonyms) when applied to clinical narratives. This has a negative effect of increasing the feature space with redundant features, thus the number of features increases enormously with increasing n. In trying to address these issues, the literature discusses other feature extraction approaches such as bag of phrases and bag of contents. These approaches have shown improved accuracy and attained good results in identifying words with similar meaning as matching and also capturing vagueness in words. Although these concept-based techniques are useful, nevertheless they still ignore grammar and word order. This has seen the rise of GoW feature extraction techniques which are better in terms of performance to traditional BoW and n-gram models. However, the GoW approaches are computationally expensive [8].

Feature value Representation Techniques

It is imperative for one to transform extracted features into numeric vectors after the feature extraction phase, a process known as feature value representation [8]. The aim is to represent extracted features in a master feature vector made up of rows (documents) and columns (features) [15]. Feature representation is done through indexing (allocation of indexing to documents) and weighting (assigning a weight to each term in the document that defines its importance) [13]. There are various weighting schemes and all can take any of the forms below.
Binary representation (BR) defines feature presence as "1" and not present as "0" [15, 12].
Term frequency (TF) defines the count or number of terms found in a given document. TF allocates a weight to a feature using the number of times a word repeats in a given document. This implies that mostly frequently occurring items are a true portrayal of the content and get high scores [42].
Document frequency (DF) is basically the commonness of documents that comprise a given term or the count

of documents in which that feature occurs [15, 12].

Concept-Based Representation is term based frequency that classifies conceptual features as graph-based document representation [18].
Term Frequency-Inverse Document Frequency Representation (TF-IDF) is defined as term frequency with inverse class frequency. TF-IDF identifies a term that is a distinctive feature and frequently occurs in a particular category but is less frequent in other categories, hence has little weight or magnitude [42].

Normalised Term Frequency-Inverse Document Frequency Representation (NTF-IDF) is a technique that uses a normalised factor in relation to TF-IDF as a document length. Terms of same frequency in various documents that apply the normalised factor ensure equal importance and relevance of features derived from short and long documents [12].

Feature selection

Feature selection in VA can be either expert driven or fully automated. The extracts below highlight the various feature selection schemes that exist in VA literature.
Expert driven feature selection also known as PCVA is when a team of experts is specifically responsible for identifying and discovering relevant particular selective features from the VA narratives. The most frequently counted features are extracted and ranked on the basis of their weighting (using discriminative weights) and stored in a vocabulary list for classification [8, 44]. The PCVA process has many drawbacks despite being the gold-standard in VA cause of death determination. The use of scarce various experts to reach a consensus on cause of death, makes the process expensive. Moreover, it lacks consistency as it relies on assessment and review of VA forms [45]. One problem with analysis of VA forms is the variation of performance in the population being studied relative to the true cause of death [45, 16]. Consequently, PCVA is compromised due to poor repeatability of results amongst many drawbacks [19, 46, 4]. PCVA tends to focus on one cause of death even though there might be multiple causes [4]. PCVAs follow own methods when they capture data and they do not have standard questions as features extracted are inconsistent.

[28] report on a study where they compared PCVA and CCVA, and deduced that the Random Forest approach performed better compared to PCVA methods. Moreover, they conclude that there is no optimum approach for VA coding and suggested further inquiries with large dataset and training of models especially where death is undocumented medically. [47] investigated completion rates and factors associated with undetermined cause of death and concluded that, completion rates are high in the early phases of VA, but in later phases of the VA process a significant number of deaths are still lost. The reason for this being after getting VA forms, physicians do not allocate a definitive cause of death. [48] investigated use of PCVA and InterVA and concluded that, the two approaches comparative analysis was slight at individual level and poor at population level, hence they recommended more inquiries and research on VA methods. The wide spectrum of drawbacks above, gave rise to alternative automated feature selection approaches.

Automated feature selection

Automated feature selection entails the use of computer programs applying various statistical approaches to automatically extract features from text narratives. These approaches do not need expert intervention. Previous research provides strong evidence that most investigators have applied automated feature extraction techniques to extract content-based features, concept-based features, structural features and linguistic features [8]. The next step after feature representation is to create a vector that comprises only selected relevant words useful for categorization, a process known as dimensionality reduction. Various forms of dimensionality reduction are discussed below.

Pearson Correlation (PC), which measures the correlation between two variables thus independent X and dependent variables Y [15]. The correlation coefficient r is between -1 and +1 where -1 denotes a negative correlation and +1 denotes a positive correlation between two instances [44, 8].
Chi-squared test (CS), which determines the dependency of two variables by measuring the statistical independence [15, 8]. This statistical test measures the significance level of the observed frequencies of tokens O from the expected frequencies E. Furthermore, this approach is a feature reduction technique where O denotes observed or collected data and E_i denotes expected values.
Information Gain (IG), which measures the importance of a particular feature in the feature vector space. This metric computes the reduced degree in uncertainty given the value of the uncertainty [44, 15].

Principal component analysis (PCA), generates principal components through statistical means by applying orthogonal transformation to transform correlated features [8]. The aim of PCA is to come up with a correlation matrix made up of fewer principal components that are less than or equal to initial number of observations. PCA makes it possible to reduce large datasets into smaller ones but avoiding information loss. The benefit of PCA is that it seeks to identify meaningful patterns in a dataset enabling easier exploration and visualization of data [49].
Local Semi-Supervised Feature Selection (LSFS), which is a technique that applies a boundary for every data sample in the dataset. Thus it further applies vectors using feature weights whilst incrementing margins to get the most relevant and useful features [8].
Fisher Markov Selector (FMS), which is an automated feature selection scheme can effectively and efficiently handle high dimensionality of the feature space as it can choose optimally the most useful important subset of features from various classes [44].
Improved Global Feature Selection (IGFS), which is a hybrid approach employs the ensemble method where it merges the strengths of global feature selection method and a one-sided local feature selection [44].

2.2.2 Semi Supervised learning

VA is a semi supervised learning problem. Specifically it is a multi classification problem, where the aim is to classify VA narratives to various disease categories as per ICD-10 codes. The following sections discuss the semi supervised learning problem and then highlight the frequently used statistical and ML methods in VA.

Semi supervised learning problem

Semi supervised ML approaches seek to optimise classification accuracy with few labelled data points. This approach is relevant in cases where there is a large volume of unlabelled data. However, process of labelling is costly, difficult and labour intensive [8]. On applying semi supervised machine learning approach, the aim is to maximise the training data set using input X and output y features. Learning takes place using the mapping function $[y = f(X)]$, where training data is provided as an input to the learning algorithm. As such, semi supervised learning aims to predict efficiently and optimally the mapping function, using unseen data X [8].

Classification in VA

A classification method is defined as a description of a specific analytical technique that employs a classifier with predefined rules derived from empirical data applying any classification approach [10]. Classes are derived from classification methods and have attributes that describe them. Classification approaches can fall in any of the following: i) Linear and discriminant techniques which use least squares, hyperplane, quadratic and logistic regression function to separate the dimensional feature space by fitting a line that maximises categories of data points for each class., ii) Probability density function which utilizes the k-nearest neighbour approach and applies the concept that a new data point has a high chance of being positioned close to other data points of the same class in the feature space. Therefore, new data points in the feature space are categorized to clusters using class probabilities of their positions and iii) Decision trees and rule-based methods which iteratively divides the feature space into smaller regions until branches are pure, thus there are only branches belonging to a specific single class [10].

Classification requires the creation of a classifier from training data to predict the VA classification/categories [50]. In simple terms in the VA domain, classes can be thought of as validated cause of death and attributes are symptoms/signs that are collected through interviews with the relatives of the deceased. The feature space consists of observations that have certain conditional probabilities with corresponding scores [10]. Therefore the aim of classification methods is to divide the attribute space into predominant clusters of a single class of observation [10, 51].

Statistical models

Various statistical learning models have been applied in VA domain to date. The following extracts discuss statistical approaches.

Logistic regression

The study by [37] elaborate on logistic regression models as being widely used in biomedicine due to their user friendliness, high level of interpretability and their optimum use in solving problems. In this statistical model we assume $w \in \{0, 1\}$, where 0 represents a bad class and 1 a good class. Therefore, the logistic regression model defines X as the data set and θ as a vector of coefficients derived from maximum likelihood estimation [52].

Naive Bayes

The Naive Bayes classifier (NBC) is based on the Bayes rule and makes the assumption that features in the dataset are mutually exclusive or independent from each other [14]. This implies that the probability each feature contributes, is independent to the final class probability. This approach can be trained efficiently as it

assumes value independence of a specific feature from the value of other features, when given a class variable [53]. NBC is a statistical approach that derives conditional probability between each class and corresponding attribute values by analysing the relationship between each feature of a class and the class. This computes the probability of instance x given class y assuming the features are independent [17].
[54] argue that the NBC has shown better results in terms of VA classification. [17] implemented the NBC in VA as; each VA record applies NBC to assign a possible cause of death using probabilistic means. Each possible cause of death had a label that corresponded to independent conditional probabilities given specific symptoms and the absolute probability of that same label. The label with the maximum probability is then allocated as the appropriate cause of death to the VA record. In mathematical terms, each record is given cause of death C_{j*} with the maximum probability, for any set of n recorded features in the verbal autopsy, represented by $F1, ..., Fn$.

Machine learning techniques

ML approaches specifically can be powerful, effective and efficient tools in developing better semantic structures to solve issues around automated VA classification [9]. This is evidenced through their superior capabilities such as; use independent parameters which are non parametric, fast in testing, produce results in complex domains, reduce problem complexity, easy to implement and employ more resembling characteristics of the model [14]. [19] points out that machine learning models can understand better the input-output relationship. Moreover, ML techniques are extensive, comprehensive, highly reliable and consistent relative to changes [4]. The study of [55] concurs with the above assertions and points out that, ML approaches have been employed due to their ability to handle high dimensional data and producing good results with limited training samples.

[19] applied Artificial Neural Networks (ANN) to classify cause of death from VA narratives. However, they reported issues of over fitting and suggested increasing the number of training examples could improve ANN. [1] did a study on VA classification and they reported that there is no technique of automating VA classification that can be better than the PCVA. Albeit, their study only used conventional ML approaches. [20] applied Natural Language Processing (NLP) for classification as a deep learning model (advanced ML approach). Nevertheless, they used a small sample size and their results cannot be generalised across studies. [21] did a study on "multi-task learning for interpretable cause-of-death classification using key phrase prediction" using clustering techniques. In spite of that, they reported a low score in terms of accuracy. [53] used unstructured clinical datasets applying text mining and ML for automatic identification of protected health information and got promising results. [41] did an investigation on application of character embeddings (representing each character in a narrative as a numeric in the vector space) in VA and they reported a low score in terms of accuracy. Regardless, they concluded that application of character embeddings can be useful in VA.

Machine learning approaches have been used effectively for epidemic forecasting including surveillance for health systems. Nonetheless, there is need to focus on real-time monitoring and forecasting [56]. Literature reports on application of conventional shallow ML models in VA domains which perform poorly, suggesting that as they learn they suffer from dimensionality and sparsity of the feature space [11, 12, 13, 14, 15, 16, 17, 11, 44, 18, 57, 42, 20, 21, 29]. We discuss various ML approaches in the following extracts.

k-nearest neighbour

The k- nearest neighbour model derives it's concept from the distance or similarity function for given pairs of data points [14]. The k-nearest neighbour determines similarity function by calculating the Euclidean distance or the cosine similarity measures. However, the best choice of k depends on data. Thus, a good choice can

reduce the noise effect [14]. Furthermore, it uses all available instances to categorise a new instance based on the distance function [15]. K-nearest neighbour can also learn by analogy through n dimensional numeric attributes [58]. Therefore, data points which are close to each other are called nearest neighbours. The final class of a specific data point is known as the majority class.

k-means clustering

k-means clustering is a technique that applies unsupervised learning and aims at grouping similar objects with similar characteristics together [59]. In performing this approach one has to first define required number of clusters k for n data points, allocate each data point to a cluster randomly, work-out cluster centroids, re-allocate each data point to the nearest cluster centroid, re-calculate cluster centroids and iterate on last two steps until exhaustive to get to global optima [49]. There are a few studies that have applied this approach to date. One study by [21] used clustering, however they concluded that this approach can improve cause of death classification and can be interpreted easily even though in future they have to improve on accuracy.

Decision trees

According to [15], decision trees identify feature conditions as nodes, and the outcome of a condition as a branch and class labels as leaf nodes. The start of the tree node is known as the root node and is identified as the most important feature in the feature set. The root node can be recognized as an attribute that can split the feature space in a best optimal way. Branches can also be labelled as terms with a corresponding weight. Leaf nodes represent class labels and new data point can be generated and assigned a class based on the majority vote of the leaf nodes.

Random Forests

Random Forests has been used in the VA domain because of easier implementation. This approach uses many individual learner trees to perform classification based on votes on an overall category from a given set of inputs [60]. A majority vote is applied to determine a class of a new data point through combining random forests and decision trees. Combined decision trees k are built on randomised algorithm from the bootstrap samples with n observations [16].

Support Vector Machine

The Support Vector Machine (SVM) approach has been applied in the VA field because of the ability to handle high dimensional data, non-linearity and producing good results with limited training samples. This technique employs the idea of a decision boundary known as a hyperplane that distinguishes between classes in a high dimensional space [61]. The documents that are close to the hyperplane are called the support vectors [14]. The SVM approach aims at maximising the margin. SVM computes the maximum linear distance between separated classes in feature space. In non-linear cases the SVM has the ability to use kernels which can easily handle non-linearity between classes and the feature space.

Artificial Neural Networks

Artificial Neural Networks (ANNs) which form the basis of deep learning has been applied in VA domain, and uses a neural classifier that can be thought of being a network of units where each unit can be a term and the output unit represents a category [14]. The ANN consists of the input, hidden and output layers. In document text classification, input units are allocated weights. The activation of these units is transimitted via a network to get an output value that determines the categorisation [14]. The layering of the nodes provides a map of the

decision space also known as the neural network where a program can learn rules from massive data amounts being processed [62, 36]. The ANN applies an error function at the output layer to generate updated weights using back propagation (process of fine tuning weights based on error rate computed from previous iteration). Effective and efficient back propagation makes models genaralizable and reliable. ANNs have been applied in VA and have shown promising results even though there is need to have more training examples for the model to be effective [19]. [1] applied the various ML approaches and got best results from the ANN using a feed-forward network (no back propagation) with one hidden layer and achieved sensitivity of 0.695. Figure 2.1 below depicts a typical ANN model of a single neuron in a feed-forward ANN. Input information x_i is received from other neurons, and takes the product of input information and each data point corresponding weight denoted by w_{ij}, and produces a weighted output applying an activation function $f\,(x_j)$. The process iterates taking the weighted output as input to another next layer neuron in the next layer. The same process is continues up to the output layer.

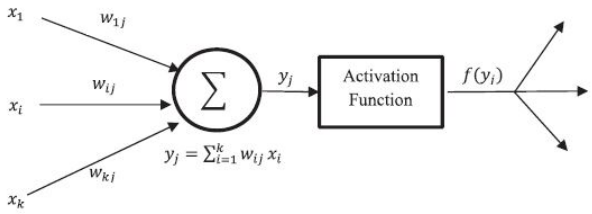


Figure 2.1: ANN structure of single neuron.

Decision rule

Decision rule is a classification approach that uses an inference rule-based engine to classify documents to their respective categories. Furthermore this approach can entail building a knowledge base with a set of terms or keywords derived from consultations with domain experts [63].

Case-based reasoning Case-based reasoning (CBR) has been applied in literature by [43] and employs reasoning through a rule-based approach to construct meaning through synthesis of information to derive a new case in an autopsy report. This new meaning constructed through extraction, is converted into features and values. The new features and values are taken as input by the system for similarity analysis. In the process these new generated feature values are then matched with the previous cases by calculating individual feature similarity scores using the k-nearest neighbour approach. It employs the Naive Bayes learner to assign weights to features based on significance and is used as a computation to the final similarity analysis. Only top scores are selected using similarity values and identified by user as the desired outcome, and thus stored in the case base [43].

Ensemble classifiers

The following extract discusses ensemble approaches, which are hybrid schemes of various ML approaches and use a voting scheme or weighting in order to do the classification. These approaches mainly use several prediction tools and aggregate their prediction, such as such as voting, associative classifiers, centroid-based classifiers discussed in the study of [14]. Implementing ensemble methods improves prediction quality [54, 29]. However, these approaches suffer from very low interpretability and high computational time [37].

Boosting

Boosting also known as an iterative ensemble classifier is a ML approach that improves a weak algorithm into a stronger one through aggregation of multiple classifiers [64]. It employs a weighted vote when labelling unlabelled instances that ultimately becomes a composite classifier [51]. The most common boosting algorithm is The ADABOOST which is short for Adaptive Boosting. The approach has been used in many studies such as [29, 64].

Bagging

Bagging or bootstrap aggregation is an approach that builds a classifier for every bootstrap based on n bootstraps from given datasets [51]. Given each trial denoted by $(t = 1, 2, 3...T)$, where the dataset is of size N. The learning environment generates a classifier C^t from the dataset and a final classifier C^* is created from aggregation of the T classifiers that are classified from instances x through voting for class k where it is recorded for every classifier $C^x = k$ and C^* which denotes the class with the most votes.

VA algorithms

VA algorithms have been the alternative traditional method of assigning cause of death apart from the PCVA technique. VA algorithms require VA data derived from real deaths and Symptom Cause Information (SCI) (which is a repository of information about symptoms that are related to each probable cause of death). The SCI may be derived from the physician's reports and logic (that entails a logical algorithm) that combines the two to identify cause-specific mortality fractions (CSMF) to assign specific causes of death. The literature to date identifies five various free online VA computational algorithms that have been widely employed and most of them apply statistical approaches [65].

The InterVA algorithm uses the Bayes rule (formula that computes conditional probabilities of X when given evidence) and Symptom Cause Information (SCI) as conditional probabilities of symptoms given a specific cause of death using Cause Specific Mortality Fractions (CSMFs). These conditional probabilities are sourced from physicians and the Population Health Metrics Research Consortium (PHMRC) gold standard dataset [66]. The PHMRC is responsible for supporting survey design, primary data collection, analysis, development, and to provide tools for quantifying population health. VA datasets cannot distinguish between the confirmed absence of a symptom and missing data. Therefore, to address this limitation of the data, InterVA-4 only utilizes the confirmed presence of a symptom. Moreover, this approach uses propensities (which are specific to a particular cause of death) and not probabilities. [17] applied the Inter VA and reported a sensitivity of 0.43 and CSMF accuracy of 0.71.

PHMRC 'gold-standard' data have both VA and medically certified cause of death. This contains information about the relationship between VA symptoms and medically certified causes. The Tariff algorithm uses fewer symptoms and it employs the SCI as a tariff score to rank causes of death, thus determining the association between specific symptoms and causes in the PHMRC gold standard dataset. CSMFs are arrived at by identifying a single cause with the highest rank for each death in the VA dataset and sum them up [65]. The standardized Tariff Score Q_{ki} for death k and cause i is the quantile value of S_{ki} in the distribution of S_{*i} formed by a set of deaths sampled with replacement from the PHMRC data set so that they have a uniform cause distribution. Values for Q_{ki} are bounded between 0 and 1 so they have the same scale and can be compared. The cause chosen for each death k is the one with the largest Q_{ki}. [67] applied the Tariff approach and reported 0.505 Chance-corrected concordance (CCC). CCC computes the degree to which the predicted

cause of death classes relate to the actual cause of death classes. They further reported a CSFM accuracy of 0.770.

The InSilicoVA algorithm uses a statistical approach employing joint probabilities to identify the most likely significant cause of death in relation to CSMFs for all deaths in a VA data set [68]. This implies that the algorithm uses SCI and derives this based on the Naive Bayes model based on conditional probabilities (symptoms given cause of death and cause of death given symptoms). The probabilities can be sourced from physicians or use a gold standard dataset that can be a de-biased physician-coded VAs. The study by [30] applied the Insilico approach and reported true positive predictions of 0.341 on all 34 cause of death classes and 0.85 CSMF accuracy.

The Naïve Bayes Classifier (NBC) algorithm applies the bayes rule to categorise cause of death. The study by [17] applied the NBC and achieved better results as compared to InterVA and the Tariff approach with a sensitivity of 0.57 and CSMF accuracy of 0.88. Nevertheless, their algorithms made use of structured questionnaire data only.
The King-Lu algorithm uses symptom conditional probabilities to approximate cause of death of a dataset over 13 classes. This approach does not avail a cause of death for individual records [1]. This algorithm relies on SCI training data which defines clusters of symptoms rather than a single symptom. Moreover, it is recommended to use gold standard deaths and if possible they should be from the same population as the VA deaths to get credible and better results [65]. [69] used the King-Lu and reported a CSMF accuracy of 0.96

[2] argue that these VA algorithms can not avail enough evidence where there is limited expert diagnosis, hence they cannot be used to guide health priorities. These VA algorithms mainly employ statistical approaches and tariff scores to rank causes of death [65, 69]. Various VA algorithms are dependent on sample size, age group, causes of death, data set size or characteristics of the sample in order to produce best results [70, 4]. There has been challenges and issues in terms of various VA techniques and there has been proposals to improve VA approaches through minimising the number of features under study (dimensionality reduction) and also combining various algorithms in order to improve on performance, accuracy and efficiency [17]. Moreover, the validity of the VA approaches performance in terms of sensitivity, specificity and predictive values vary with regard to causes of death across populations [71]. It is difficult to generalize and standardise VA classification practices, since there is no gold standard and patient's records vary in terms of socio-economic status [17, 65]. Another issue is associated with standardising VA questionnaires so that they have the same structure and content. The format and standard of VA questionnaires differs considerably, thus their administration requires appropriate training so as to elicit relevant and appropriate symptoms and causes. There are also often language barriers and the interviewer and interviewee need to speak the same language so as to derive best results. [4] recommended incorporating fully trained multiple translators. A standardised and uniformly reliable method is required to determine cause of death and also a standardised questionnaire which can be applied globally [4].

2.2.3 Deep learning

Deep learning approaches have been implemented successfully in literature for the past decades, with the aim of of improving classification. Moreover, we have witnessed the implementation of emerging deep learning frameworks that have successfully managed to extract and reveal critical information, a task that conventional statistical and machine learning models could not. Additionally, deep learning achieves learning

of data representation with varying levels of data abstraction when even computational techniques use few processing layers [8]. Therefore, these advanced approaches can be relevant in VA data with high-dimensional sparse data.

Literature reports a wide gap, especially in applying advanced machine learning techniques known as deep learning models as approaches to solving VA problems. [41, 21, 8] are some of few studies to have used deep learning approaches to date. [41] investigated "the application of character embeddings to improve cause of death classification using VA narratives." In their study they applied distributed word vectors (Sentence2vec, GLoVE and Word2vec) and combined two character embeddings techniques (CNN and GRN) for classification and they concluded an improvement in cause of death classification. Application of Natural Language Processing (NLP) as a deep learning approach to solve VA problems has shown promising results [72, 73, 1, 74, 12]. NLP is an approach that entails converting human language into formal easy to understand language for computers. It performs automated analysis and human language representation through use of a variety of automated techniques [22, 75]. NLP has efficient, superior performance, integrated feature learning and effective capabilities of attaining end to end learning from complex and multi modality data [9]. NLP word embeddings (representation of text in numeric form stored as vectors) on dense vector representations, enabling multi-level automated feature representation learning, has to date produced excellent results [22]. [76] used a method based on NLP to extract relevant clinical information and discovered that the approach was better than PCVA diagnosis. [77] investigated use of NLP on suicide note classification and discovered that NLP can assist in the classification process. [78] did a comparative study of existing ML and deep learning approaches and found out that deep learning was better in terms of performance as compared to machine learning. Recent deep learning methods such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM) and the Transformer architecture have managed to explicitly solve most NLP tasks [9]. The following sections discuss the deep learning frameworks that are used in VA. Figure 2.2 below compares the conventional ANN with one hidden layer and the deep learning architectures with multiple hidden layers.

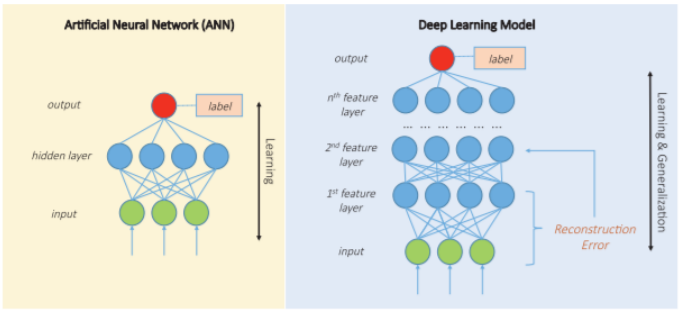


Figure 2.2: Comparison of ANN and deep learning architecture.

Distributed vector representations

Distributed representations emanated from the issue of complex and high data dimensionality for traditional statistical NLP. Their end result is the distributed representation of words in low dimensional space. The start of the data processing layer in deep learning models is represented by distributed vectors. In distributed vector representation, the logic is that analogous words usually come, about the same context. In order to compute similarity mathematically, the cosine similarity coefficient is used [22]. The study of [21] successfully applied distributed vector representations.

Word2vec

Word2vec is an approach that applies the concept that words within a vector space are of fixed length [79]. Word2vec builds vocabulary from a training text corpus and creates a vector space for each word through learning. It further groups alike words using the cosine distance function [80]. It consists two models to develop and construct distributed vector presentations which are; Continuous Bag-of -Words (CBOW) and Skip-gram models. CBOW predicts a word based on its context whilst the Skip-gram model predicts context based on the neighbouring words, thus it discovers semantic relations amongst terms in the corpus [81, 82]. The Word2vec uses the input, projection and output layers [81]. A one hot word vector of V neurons is taken by the input layer and corresponds to hidden layer N neurons. The softmax is output that contains all words in the vocabulary. Each word in the dictionary is denoted by learned vectors v_c and v_w where the connected layers use a weight matrix [22]

Sentence2vec
[83] discusses the Sentence2vec which applies the same concept of Word2vec. They argue that a sentence can be seen as text unit and the summary of the document being made up of sentences. They point out that each word in the corpus has an *id* that denotes its position in the dictionary. Sentence2vec represents each sentence in a document as a vector, unlike Word2vec which represents each word as a vector [83].

Global vectors The Glove technique is count-based and makes the supposition that the meaning of a word is implied by it's own contextualisation [81, 83]. It is mainly based on word occurrences in a textual knowledge base. This implies that words with similar contexts should also have similar embeddings.

Doc2vec
This technique uses an unsupervised learning algorithm to come up with vector representations of paragraphs or documents. The algorithm is a modification and application of the concepts of Word2vec and it aims at finding similarities in paragraphs or documents [18]. The following extracts present deep learning architectures that focus on character embeddings.

Character embeddings

Character embeddings assume each word to be composed of not more that one character. They can handle the issue of out of vocabulary words (words that appear a few times in the test set but are not in the training corpus of words), unlike distributed vector representations [22].
Convolutional Neural Networks
Convolutional Neural Networks (CNN) is an example of a character embedding technique. CNN was developed as a feature extraction function to extract high level features from n-grams, thus the features would be used for various NLP tasks [22, 84, 85, 86]. The CNN consists of input, convolutional, pooling and fully connected layers. It also has a lookup table and softmax function for classification. After taking input, a look-up table is used for transforming each word into a vector of user-defined dimensions. This implies that given an input sequence of $(s_1, s_2, ...s_n)$ of n words is changed into a series of vectors $(w_0, w_1, ...w_n)$. Learning in this network process is done using weights. Data representations and extractions are done on the convolutional and pooling layers. The convolutional layer makes use of a various numbers of filters (kernels), where one kernel extracts a particular pattern of n-gram. The pooling layer employs a maximum pooling strategy which performs sub-sampling on each input specifically by applying a maximum operation on each filter. Maximum

pooling therefore, maps the input to a fixed dimension of outputs regardless of the size of filters and thus a reduction in the output's dimensionality. Lastly, a fully connected layer employs a feed forward ANN. The fully connected layer consists of an input, hidden and an output layer(s). At the output layer, a loss/cost function is applied in this context known as Softmax classification [22]. The study of [21] successfully applied the CNN and Gated Recurrent Neural network discussed below. [8] argues that the CNN has the ability to effectively learn multifaceted features from the medical data sets as compared with traditional statistical and machine learning approaches. Figure 2.3 shows the visual representations of the CNN architecture.

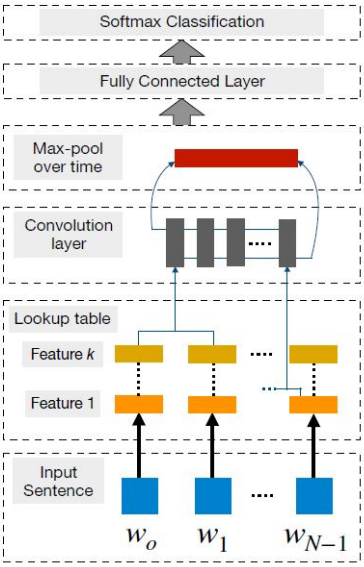


Figure 2.3: Convolutional neural network framework.

Gated Recurrent Neural network

Recurrent Neural Networks (RNNs) is a form of character embedding architecture. RNNs are employed in NLP to process sequential information. RNNs produces output that is dependent on previous tasks and results by performing same task on each instance. They make use of memory to pull previous computations for processing with current operations [22]. However, RNNs have issues of "vanishing gradient" (where weighted activation function inputs of a neural network increase or decrease whilst the derivative function approaches zero) which ultimately affects accuracy and makes it difficult for the architecture to learn and hyper tune parameters for optimum results. The Gated Recurrent Neural network (GRNN) has the rest and update gate that control flow of information. This approach further has no control of the hidden layer and has no memory unit. The GRNN has been applied in the study of [41, 21] and showed promising results.

2.2.4 ML Performance Metrics in VA

This section gives an overview of the ML evaluation metrics that have been applied in VA to date. More approaches to evaluation, accuracy and validation of algorithmic performance have already been reported on. Nonetheless, in medical data focus is given to Recall than Accuracy [32]. [66] points out that performance evaluation to determine the predictive accuracy of VA algoriths can be done at individual level or population level. At individual level we seek to determine specific cause of death categorisation, whilst at population level we seek to estimate Cause Specific Mortality Fractions (CSMFs). These processes are key to achieve comparability. Assuming a supervised learning problem with a labelled deaths dataset, there is a training set $T = (x1, y1...; x_n...; x_n)$ where each x represents a symptom and each known cause of death is indicated by

y . As such, we have a model of a VA algorithm f to derive the predicted label for a new dataset (X, Y) is denoted by function f. Given the above proposition we can therefore attain the performance of the algorithm that depends on the training data.

A confusion matrix is the benchmark of our evaluation matrices, and consists of True Positives (TP), True Negatives (TN) that denote our correct predictions, False Positives (FP) and False Negatives (FN) that signify incorrect predictions. Below we discuss various performance evaluation measures in VA.

		Predicted labels	
Actual labels		Positives	Negatives
	Positives	TP	FP
	Negatives	FN	TN

Figure 2.4: ML performance evaluation confusion matrix

Macro precision or Precision/ Positive Predictive Value (PPV) refers to the proportion of correctly predicted positive VA reports to the total positively predicted VA reports in fraction terms. This denotes how good a classifier identifies positive observations without much on the negative ones.

Macro recall is also known as Sensitivity or True Positive Rate (TPR) defines the ratio of correctly predicted positive VA reports to the all VA reports in actual positive category. This implies the rate of correctly predicted positive observations or diagnoses (cause of death) [44].

F-measure is a test score to measure accuracy of a particular model. It is a weighted combination of Precision and Recall scores and is computed as the average or mean of the Precision and Recall [44].

Overall accuracy denotes all classes with classified results that have been predicted correctly in fraction terms, thus it is the ratio of correctly predicted VA to the total VA reports [44].

The Area Under the Curve (AUC) is used to measure and validate the performance of machine learning (ML) algorithms. It is a graphical plot that depicts the True Positive Rate (TPR) against the False Positive Rate (FPR). The AUC evaluates the classifier's performance, with good performance indicated by an AUC value close to 1, meaning the curve is closer to the upper left corner. Poor performance is indicated by an AUC value of 0.5 or lower. [32, 44].

Specificity is a function that computes the percentage of actual negative medical reports that are correctly predicted as a negative [8].

Overall chance-corrected concordance (CCC) defines CCC for cause j as the overall weighted sum of cause-specific CCC [66]. [87] points out that given a proposition of measuring agreement on a cause of death, sensitivity and concordance never considers the agreement expected by chance alone. As such, given a VA algorithm that assigns randomly cause of deaths by chance, we would expect concordance of $\frac{1}{n}$.

Cause Specific Mortality Fractions (CSMF) accuracy defines accuracy as having a value between 0 and 1 and it assumes the worst possible case for predicting CSFM and assigns a weight on the least possible CSFM value that matches the total absolute error [87, 66].

Top cause accuracy is the number of correct cause of death of being first assignment divided by number of causes.

Top 3 cause accuracy is the number of correct cause of death within the first three cause assignment divided by number of causes [66].

Kappa metric applies the true test set elements under the assumption of chance prediction and is dependent on the structure of the test dataset. Thus it computes the relationship of all causes except cause specific of concordance [87].

Case similarity measure is used as performance metric in Case-Based Reasoning (CBR). CBR evaluates case similarity by using the nearest neighbour approach because of its simplicity and closeness to human judgement. Features are compared to determine the matching of all previous cases in a knowledge base of cases and all cases and feature values should exist in the system dictionary. Once the matching and comparison is done, the ecosystem performs matching cases using a ranking system based on likeliness scores applying the k-nearest neighbour. Only cases with highest weighting are considered as results [43].

2.2.5 ML Optimisation techniques in VA

ML learning entails applying a learning algorithm where a classifier is induced from training data, hence every classifier has a measure of performance (through bias and variance) known as the prediction error which is unknown and is estimated from data [50, 88].

k-fold cross validation

Validation analysis is crucial in ML as it finds optimum models to fit with new observations,thus improving accuracy and model prediction. One approach that has evidently been applied in the VA field as an optimisation technique is the k-fold cross validation. k-fold cross validation is a technique that splits data into k folds, and learning of a classifier is done through $k-1$ folds where the test error is calculated using the balance folds and gives the mean error for each fold. [89] points out that k-fold is the widely used procedure to test the effectiveness of a classification algorithm by comparing the performance between two algorithms. Moreover, performance can be evaluated through by average k-accuracies from k-fold cross validation. [1] used 10-fold cross validation and deduced the ANN as the best classifier.

2.2.6 Imbalanced and fragmented datasets

There is little literature in VA that discusses the issue of fragmented or imbalanced datasets, also known as skewed datasets, even though it is a common situation in VA domain to have these natural datasets due to the rare situation of the events under consideration. VA data has suffered from high text scarcity, imbalanced text and also in the cause of death categories which is an issue for NLP [90]. [8] points out that imbalanced class distribution (majority class out numbers minority class in terms of records) is when one class can be of more interest and is insufficiently represented compared to other class. In such instances, when categorisation is performed there can be misinterpretations and misclassification. Henceforth, imbalanced datasets are a drawback in algorithmic predictions as there is a always a bias of rules that concentrate mainly within the majority class and tend to frequent less the minority class. This leads to misinterpretation of the results in terms of accuracy in the minority class even though overall performance maybe high. As such, there is need for special techniques to address this thereof.

One approach that has been proposed is over-sampling. The effect of over-sampling in trying to create a balanced dataset without adding new data points, is that it grows the minority class size by duplicating of some positive cases in a random fashion. However, through the replication of data points over sampling can suffer from issues of model over-fitting. As a substitute to this approach and in seeking to address over-fitting is the Synthetic Minority Over-sampling Technique (SMOTE).

SMOTE uses data generation and interpolation to create synthetic minority examples of data points that are near to each other using already prevailing positive cases. On the contrary is an approach known as under-sampling which takes out negative cases from the dataset haphazardly to come up with a balanced dataset [8]. Nevertheless, under-sampling can result in information loss from the majority class. To curb information loss, one can still use random over-sampling as it can replace lost data through minority class replication of data points. However, this can result in over-fitting. However, despite a plethora of benefits such as ease of implementation and improved accuracy levels, these approaches are worse in terms of performance as compared to other techniques [37]

[37] discusses other approaches of overcoming this data imbalance, which are random sampling, applying the k-nn and selecting an appropriate way for categorization quality using techniques such as Area Under Curve or ROC can prove beneficial. Furthermore, they argue that to address issues of imbalanced data, sampling techniques are the most common solutions. Other alternative approaches of addressing imbalanced data issues such as nearest neighbours and points on the line segments use random techniques. Other sampling techniques which integrate machine learning approaches such as k-nearest neighbour is the Balance Cascade ensemble algorithm. This approach takes out past sample cases that where appropriately categorized from the main class and uses the balance of the data points to come up with a new classifier. Another approach is NearMiss algorithm that uses the k-nearest neighbour approach to choose data points that are nearest to each other for each minority case, such that every minority case is in close proximity with some majority ones. The following extract discusses limitations in statistical and ML models.

2.2.7 Model limitations and imposed assumptions

The VA community has witnessed the application of statistical approaches such as the NBC and logistic regression. The NBC assumes that features are conditionally independent and normally distributed numeric

values. However, the proposition of normal distribution in the NBC does not always hold in that there might be need to approximate other continuous distributions. The NBC model produces poor results in terms of performance if there is a high degree of feature correlation. In most cases if one does not do a rigorous feature selection process, then this can also compromise the model [14]. NBC assumes conditional independence of extracted features, however this proposition becomes more complex with the increasing number of features, which ultimately reduces the performance of NBC [11]. Moreover, if there is an imbalanced dataset in terms of training and test data, the NBC can lead to erroneous results and misinterpretations [17]. Consequently, the NBC performance is affected negatively in instances where, NBC assumes conditional independence among features which might not be the case for that specific dataset. As a result of an increase in features, it becomes difficult to manage conditional dependence of features [15].

LR is another statistical technique in VA, and is of the notion that the association between inputs and log odds is linear. On the contrary, it is not always the case in terms of linearity, for this reason we can have a non-linearity existence in the LR. To overcome the non-linearity problem in LR we recommend the SVM, which uses the kernel trick and is efficient in non-linear classification techniques [37]. As such, SVM is costly in terms of computation specifically when applying the kernel technique during learning. Moreover, achieving real space division is difficult or rather impossible, thus the SVM applies a margin that allows misclassification of some examples whilst increasing the overall performance. In such cases in biomedicine we use the soft margin SVM through introducing penalties to the objective function [37].

In addition, some studies provide evidence that one can employ successfully the k-nearest neighbour as a non-parametric technique in VA scoring. Yet, the k-nearest neighbour is costly when it comes to computation of weighting distances for each case warehoused during categorization. [15] argues that there is little of learning done by this technique during training, rather it mirrors the training set to do categorisation. Because of this, the model performs poorly and over-fitted. As a consequence, the model fails to handle noisy data and fails to identify patterns for prediction when using unseen or test data, thus it emulates the training set.

The application of the RF in VA has witnessed issues in terms of performance, specifically when the dataset contains large volume of features and very few discriminative features. This therefore, leads to trees in a forest populated by less relevant features that lead to misinterpretations and incorrect predictions [15]. [15] explored the J48 decision tree to classify cause of death from forensic autopsy reports and pointed out that, the model achieved the lowest performance. This was due to issues of having continuous data representing all features in the master feature vector, thus it was difficult to find the global minima needed to construct the J48 decision tree. They concluded that the J48 decision tree is inappropriate for categorization of VA reports. On contrary the ANNs exhibit better performance because of their ability to handle complex data, when equalled to other techniques in terms of accuracy but difficult to interpret. [19] argues that the ANN has limitation of input selection not being straight forward, time consuming in terms of designing optimal network for each cause of death, generalisation of predictions has low sensitivity and specificity and comparing and prioritising CSFM over sensitivity and specificity is a manual process.

There is little evidence that exists in terms of application of emerging deep learning frameworks in VA. Even though deep learning models suffer from interpretability, they show promising results. [8] argues that, for the most part the CNN architecture has resulted in improved accuracy, nonetheless for the model to be more effective in terms of performance it requires lots of training data to learn efficiently. As a result, with little

training cases the model is bound not to perform optimally when it comes to accuracy. In other words, the model requires a large number of features in order to iteratively learn the most useful and relevant ones to use for classification. In general within the VA domain it is difficult to find data in large volumes. In such cases, it is recommended to employ other choices of deep learning frameworks (2.2.8) to optimise VA report categorization besides CNN. The following section discusses emerging trends in deep learning.

2.2.8 Emerging Trends: Deep Learning in VA

Machine learning techniques suffer from various drawbacks, hence the need to use other emerging technologies of deep learning. [9] argues that there are various challenges with traditional ML techniques, such as supervised ML requiring a labelled dataset by experts, going through data cleaning, performing feature engineering among many. This therefore, leads to poor scaling of the feature space and consequently hides possible opportunities to discover new trends. [19] argues that conventional ML approaches lack the capability to process natural data in raw form. This is as a result of the structure, linearity and challenging transformation processes of the input feature. As an option, deep learning architectures provide representation techniques of the target value from the raw text, thus discovering new patterns and trends.

Deep learning entails various layers made up of neural networks that process data to learn hierarchical abstract representations of data. Deep learning involves NLP, which uses a variety of computational methods for automated analysis and illustration of the human language [22]. NLP tasks done by computers include among many; "parsing, tagging part-of-speech, machine translation, computer vision, pattern recognition, semantic role labelling, named-entity recognition, deep learning methods and dialogue systems [75]." The following extracts discuss other forms of character embeddings that can prove essential in the VA domain.

Other Recurrent Neural Networks approaches

Recurrent Neural Networks (RNNs) is a form of character embedding architecture that is employed in NLP to process sequential information. Therefore apart from the GRNN applied in [41], there is need to explore with other RNNs techniques in VA research such as the Long Short Term Memory (LSTM) architecture. The LSTM technique has the capability to learn and hyper tune parameters for optimum results, thus overcoming the issue of the vanishing gradient. Figure 2.5 is the RNN model.

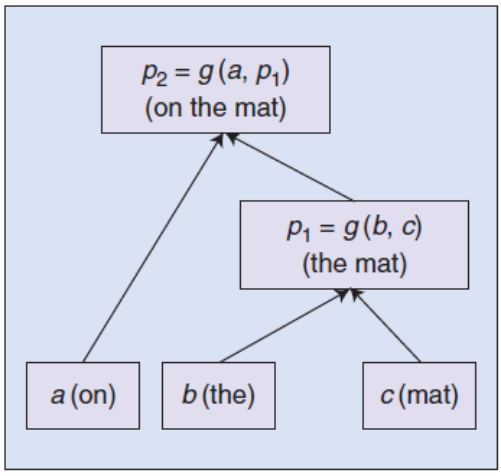


Figure 2.5: Recurrent neural network framework.

Long term short memory

Long Short Term Memory (LSTM) character embedding architecture that has the forget, input and output gate. The forget gate makes it possible for it to error back propagate a number of times and calculates the hidden layer by averaging the three layers [22]. It calculates the hidden state by combining the three gates. Figure 2.6 is the LSTM model.

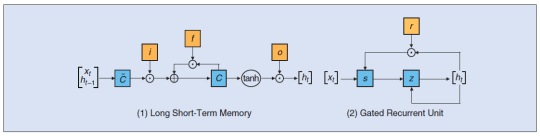


Figure 2.6: Long term short memory framework.

Matrix Vector Recursive Neural Network

The Matrix Vector Recursive Neural Network is a character embedding architecture, that seeks to represent every word and phrase as both matrix and vector [91]. It creates a combination of one matrix multiplied by the vector of the other. Recursive Neural Tensor Network (RNTN) applies less parameters and promotes more interaction between input vectors [22].

Transformer

Developments in deep learning have led to new architectures such as the transformer, which has an encoder that maps input sequences of symbols to a sequence of continuous representations, such that the decoder generates an output sequence of symbols one element at a time. The transformer takes input of previous instances to generate next feature representations [92]. [92] points out that the transformer architecture uses a decoder and encoder that are connected layers with structured attention. The encoder takes symbol representations in a sequence $(x_1, ..., x_n)$ to continuous sequenced representations $(z_1, ..., z_n)$, thus decoder uses z sequenced outputs $(y_1, ..., y_m)$ features one at a time. At each stage, the model takes previous input to generate the next symbols. The encoder has 6 stacks of similar layers which have 2 sub layers apiece for attention and feed forward functionality. The decoder consists of 6 stacked similar layers and 2 sub layers apiece. It then has a third sub layer which functions as multi head attention over output from the encoder. The attention has mappings to requests and uses keys with values that correspond to output. It creates vectors for requests/ queries, keys, values and output. The end result is a calculated weight of all values using a compatibility function of the request and the related key. It applies the scaled dot product that takes input as queries of dimension d_k and values d_v. It applies the softmax function to obtain weight by taking the dot product of the query and keys and divide each by $\sqrt{d_k}$. The multi-head attention attends to varying representations at differing positions in space by permitting the model in a combined way.

2.2.9 Limitations in VA literature

Below we discuss and analyse limitations that we identified from the VA literature;

- Missing data and varying VA input
- No model transparency performed,
- Data balancing techniques not applied,
- Absence of standardised tools and techniques that are used in VA studies,

- Scarcity and cost of experts,
- Various medical ontologies resulting in heterogeneous datasets,
- Relying on expert's diagnosis as gold standard,
- Scarce availability of publicly available biomedical datasets with standard training and testing data,
- High dimensional low sample size datasets and noisy data points,
- No exploratory data analysis is done to explore variable distribution and check for any outliers in datasets and
- Model Complexity is not taken into account.

This section synthesises the main limitations of VA literature. In general all the discussed methods take input of VA narratives, however the end result of categorisation is affected by main limitations of VA studies. We have witnessed use of smaller data cases in VA studies, and usually gives a biased outcome in terms of the precision of estimates. Apart from this, there is bias in interviewees remembering true cause of deaths, which can result in misclassification. Furthermore, because of lack of resources to determine cause of deaths, in most cases there are cases of undetermined or unknown deaths and in fact as a result these are not included in studies because of missing information [93].

Most statistical and machine models in VA do not perform model transparency and this reduces the confidence in model interpretability, hence there is need to incorporate model transparency in order to be able to justify and explain model prediction. The VA community can apply any choice of transparency models such as Local Interpretable Model-Agnostic Explanations (LIME) amongst many to cater for this limitation. Few studies report on data balancing to improve classification accuracy, even though there is scarce literature on the actual application of data balancing techniques in VA studies. Depending on the dataset type and characteristics, one can incorporate any choice of an optimal data balancing approach.

There is also a need to have standardization of tools, processes and methods used in VA studies, one example is the diverse options in data acquisition and preprocessing. Moreover, one key challenge and limitation is the scarcity of physicians that perform VAs. Thus, there is need to prioritise resources in order to address the imbalance. This synthesis of the literature revealed the one main limitation to be heterogeneous datasets. Hence, there is a great need to have standard ontologies that house VA data.

Further inquiries from the literature revealed that there is no gold standard that can be used to test the performance of VA. This lack of a true gold standard limits the explainability and interpretation of validation study findings. To add more to these limitations is the issue of scarce availability of biomedical datasets that can be used to train and test models in VA in order to improve performance. It is key going forward to have more publicly available datasets that can be used by the VA community to train and test models effectively, as this improves model performance and avail better results.

There is need to have more robust dimensionality reduction schemes in place to improve model performance. Clearly, we need to apply special preprocessing techniques and algorithms that can handle noisy data. Although previous research provides insights on how feature engineering improves model accuracy, we need to put in place other forms of feature engineering techniques in order to generate new relevant features from existing ones and improve model output. We propose to apply and explore with genetically inspired feature selection

or engineering approaches. Model complexity is a result of models that are highly parameterised, which is the norm in biomedical data. For this reason it is imperative to reduce the feature space and only retain relevant and useful features during training of the algorithms. The next extracts discuss our methodology.

2.3 Methods

In this paper we apply a methodological systematic analysis to;

1. systematically review the most commonly used statistical and ML techniques in determining causes of deaths from VAs;
2. identify limitations of each of these approaches;
3. propose machine learning framework guidelines for determining causes of death from VAs;
4. give an insight to future emerging directions;

This methodological systematic analysis takes into account statistical methods such as, LR and NB, and the ML approaches include k-nn, DT, SVM, ANN, RF, CBR, Ensemble classifiers (Boosting and Bagging), k-means clustering, distributed word vectors (Word2vec, Sentence2vec, Doc2vec, Glove), Convolutional Neural Network (CNN) and Gated Recurrent Neural Network (GRNN). The statistical and ML methods discussed in this methodological systematic analysis is not complete considering various approaches that have been applied in VAs. This literature review only highlights most of the regularly used methods. In this paper we present a flowchart that details our methodological systematic analysis in Figure 2.8. This study considers the procedure of newly published articles on VA. The study used the following search terms: "Algorithms in Verbal Autopsy", "Statistical techniques in Verbal Autopsy", "Machine Learning in Verbal Autopsy" and "Deep Learning in Verbal Autopsy" to retrieve relevant articles. The inclusion criterion included published articles on VA over the period 2010 to 2020. In this study we chose to include high quality papers from reputable journals such as; peer-reviewed journals and conference articles. As part of the filtering process we selected papers through analysis by evaluating the abstract, conclusion and reading the whole paper. Only articles that were written in English were considered. The repositories used in the search for papers were Google Scholar, PlosOne, Science Direct, IEEEXplore, BioMed Central and Springer-Link. We selected these databases as they are reputable and consists of research work that covers the whole globe. Our final list of papers did not comprise any unpublished work and dissertations. The search yielded 110 articles, of which 76 were peer-reviewed journal articles and 34 were conference papers. Furthermore, we had a database that served as a point of reference for further statistical analysis. The main intention was to quantitatively establish the research status of statistical and machine learning application in VA. The methodological systematic analysis identified 85 articles as being relevant for our study and the remaining 25 were discarded. As a result of the methodological systematic analysis of the literature we generate and present tables, pie charts and histograms. Figure 2.7 depicts a histogram of the initial peer reviewed journal articles and conference papers.

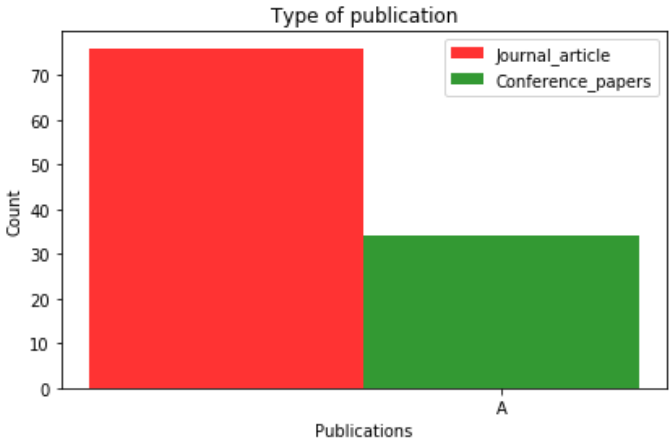


Figure 2.7: Distribution of initial searched journal articles vs conference papers

2.3.1 Existing literature survey on VA

There are few literature surveys on VA [31, 8, 12, 28]. In this study we give an insight into the key aims of each literature study and perform a comparative analysis to distinctively highlight key differences with our study. The identified studies mainly air their views in terms of model comparison in the determination of causes of deaths from VAs based on their performances. However, much of the studies do not consider critical issues in the determination of causes of deaths from VAs, such as pre-processing, nature of datasets, data exploration and model explainability (transparency). Furthermore, deep learning models are not considered in the listed table below of literature surveys. We have also included a study done by [8], which provides the most recent and the most comprehensive comparative analysis of existing methods in automated text classification in VA. The included studies cover mainly performance evaluation of traditional techniques and propose new methods for VA classification. The literature discussed in [8] covers most automated text classification techniques in VA, model performance, propose new methods for VA, feature selection, feature engineering and other issues. However, in contrast to this study, at the time of this write-up, there is little evidence to suggest any methodological systematic analysis in VA that has been published which discusses, conventional statistical and ML approaches and emerging advanced deep learning techniques. Table 2.1 provides a comparison of existing literature survey papers in VA against this study.

2.4 Results

This section gives an account of results from the primary studies, based on the meta-analysis. The frequency distributions of, feature extraction techniques, feature value representation schemes, feature selection approaches, algorithms and performance validation metrics from all principal studies are synthesized and analysed. The pie-charts for pre-processing techniques, studies that report on data transparency and balancing techniques, the distribution of selected VA articles compared to other domains are also shown. We present a tag cloud of all articles used in this methodological systematic analysis.

2.4.1 Preprocessing

Pre-processing is a key stage in cleaning data. Our methodological systematic analysis suggests that only 38% of the primary studies applied pre-processing to normalise their data. There is 62% of primary studies that did not apply pre-processing to their data. As such, this adds to noise and high dimensional data and consequently low accuracy prediction models. Figure 2.9 below denotes the distribution of studies that

Table 2.1: Existing literature surveys on VA and their differences from this survey paper: Legend:Feature Engineering (FE), Feature Selection (FS).

Comparative analysis of existing literature papers				
Survey paper	Articles searched	Years	Objectives	Differences from this study
[12]	Unspecified	1997-2009	Model performance evaluation	Deep learning models not considered
			Investigation of model types	Data balancing ignored
			Analysis of FE approaches	Not all ML approaches are discussed
				Model transparency not covered
[8]	1497	2013–2018	Model performance evaluation	Deep learning not covered
			Investigation of model types	
			Analysis of pre-processing approaches	Not all ML approaches are discussed
			Analysis of FE approaches	
			Analysis of performance metrics	
[28]	19	1997–2014	Model performance evaluation	Deep learning not covered
			Investigation of model types	
				Not all ML approaches are discussed
			Analysis of FE approaches	
			Analysis of performance metrics	
				Model transparency not covered
[31]	392	1978–2008	Model performance evaluation	Deep learning not covered
			Comparative analysis of VA methods and concepts	Nature of datasets not reported
				Not all ML approaches are discussed
				Model transparency not covered

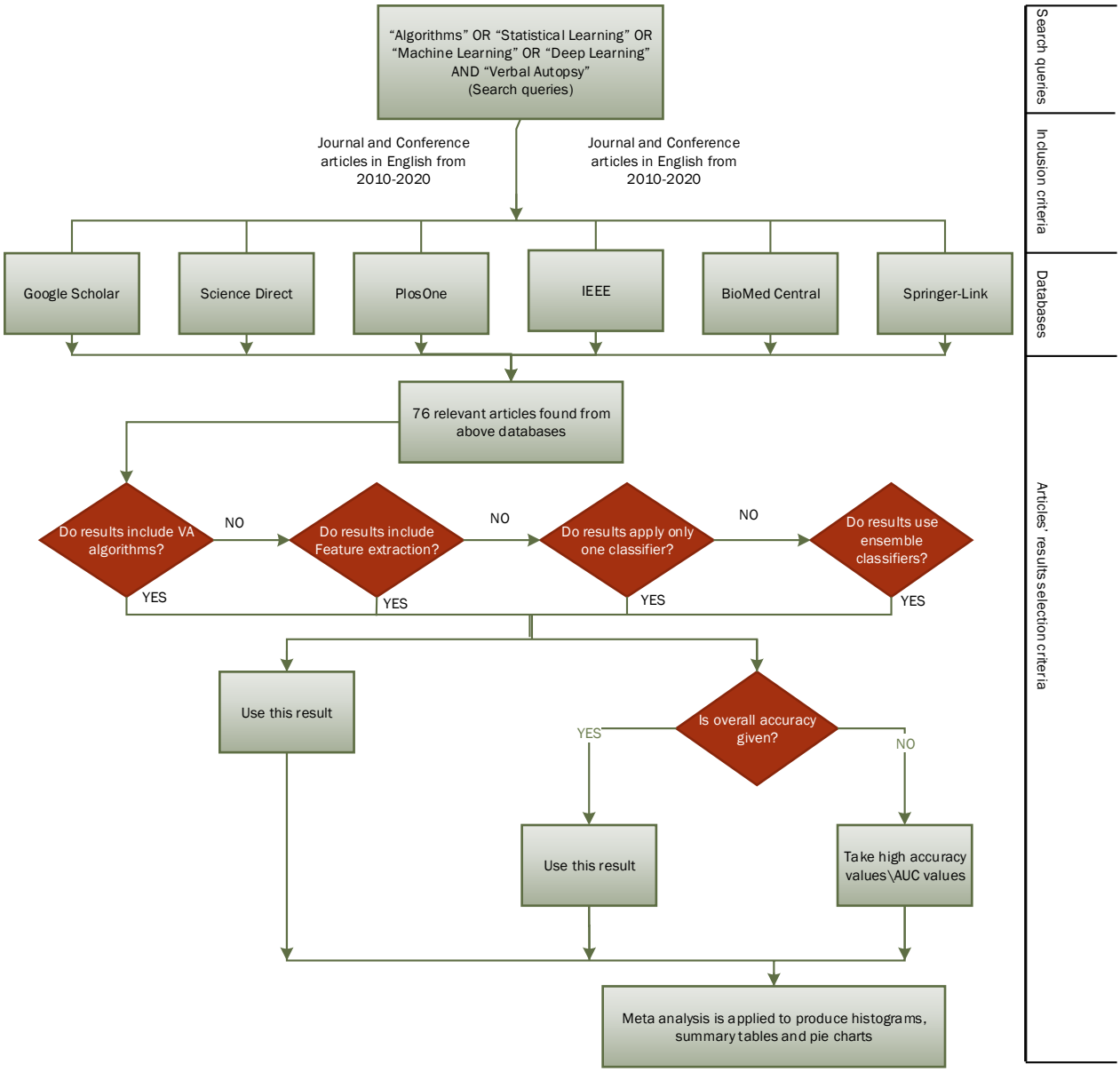


Figure 2.8: Flowchart that details our methodological systematic analysis.

applied pre-processing.

2.4.2 Transparency

Transparency is an approach that makes machine learning models more explainable and justified. As such, there is not a single study in VA that applies transparency models and we only present 5% of the studies that report on transparency. There is great need to introduce transparency in VA models. Figure 2.10 below shows the studies that discuss transparency models even though they were not specifically applied.

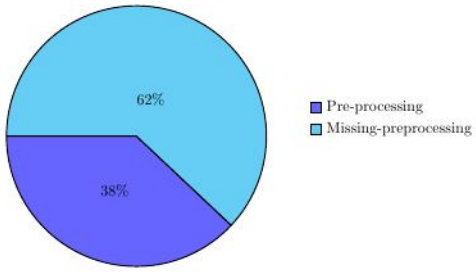


Figure 2.9: Distribution of VA primary studies that include pre-processing.

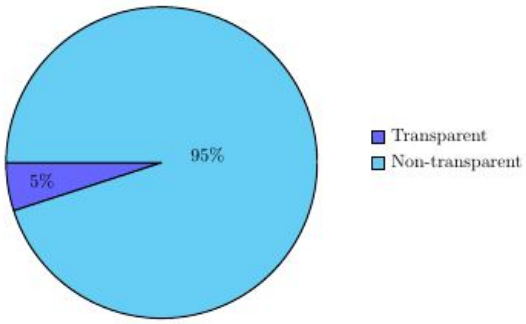


Figure 2.10: Pictorial distribution of VA primary studies that include transparency.

2.4.3 Data balancing

VA data generally entails data imbalance, where there is more bias in terms of the majority class. Approaches such as oversampling and undersampling have been applied to address this imbalance. However, few notable studies apply such approaches. Figure 2.11 shows only 6% of the primary studies applied data balancing.

2.4.4 Feature extraction

Table 2.2 and Figure 2.12 details feature extraction techniques from primary studies. Our analysis shows that out of 38 studies, most of the VA models use expert driven feature selection, followed by the bag of words and n-gram models respectively.

Table 2.2: Feature extraction methods used in primary studies. Legend: Bag of Words (BoW), n-grams (NG), Hybrid (HB), Word2vec (WV), Concept-Based (CB), Graph-Based (GB), Rule Base (RB), Doc2vec (DV), Expert driven (ED), Sentence2vec (SV), Glove (GV).

Feature extraction methods from primary studies											
Study	BoW	HB	NG	WV	CB	GB	RB	DV	SV	GV	ED
[1]	✓										
[4]											✓
[28]											✓
[65]											✓
[29]											✓
[11]		✓	✓								
[12]	✓										
[15]		✓	✓								
[16]											✓
[17]											✓
[18]	✓	✓	✓	✓	✓	✓	✓	✓			
[30]											✓
[44]											✓
[31]											✓
[66]											✓
[67]											✓
[68]											✓
[69]											✓
[70]											✓
[71]											✓
[43]							✓				
[8]	✓	✓	✓	✓	✓	✓	✓				
[11]	✓								✓		
[45]											✓
[19]											✓
[47]											✓
[48]											✓
[54]											✓
[20]	✓										
[21]				✓							
[41]				✓						✓	
[57]			✓		✓						
[63]					✓						
[72]	✓										
[73]	✓										
[77]											✓
[87]											✓
[90]											✓

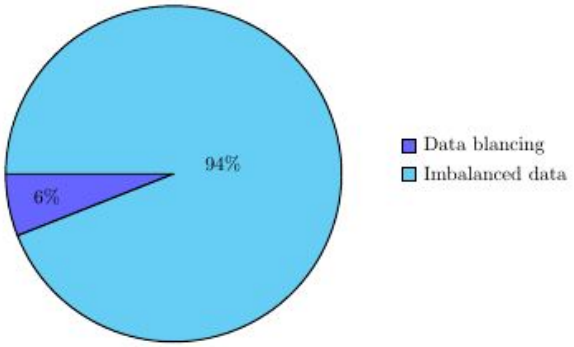


Figure 2.11: Pictorial distribution of VA related studies that include data balancing.

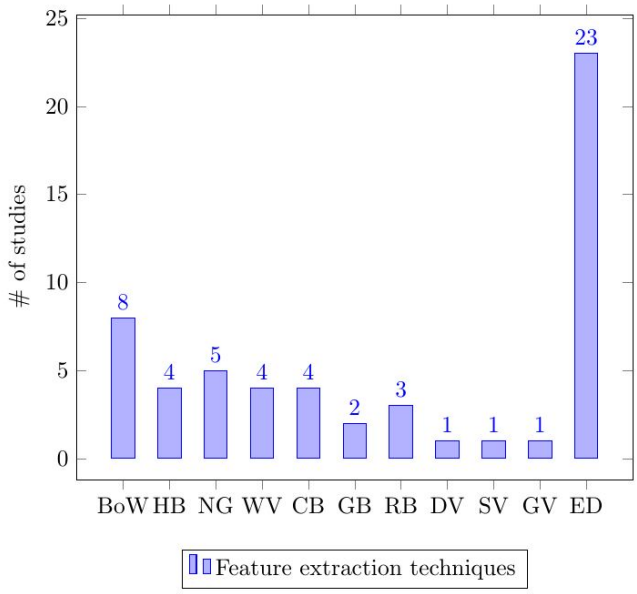


Figure 2.12: Distribution of feature extraction techniques reviewed.

2.4.5 Feature value representation schemes

Table 2.3 and Figure 2.13 shows the distribution of feature value representation schemes. This investigation suggests that out of the 11 studies analysed, term frequency value representation and binary representation are the most widely used approaches.

2.4.6 Feature selection

Table 2.4 and Figure 2.14 shows the distribution of feature selection techniques applied in primary studies. Out of 18 studies, NLP is the most common approach possibly depicting deep learning as an emerging solution, Pearson Correlation and Information Gain coming after.

Table 2.3: Feature value representation schemes from primary studies. Legend: Binary representation (BR), Term frequency (TF), Document frequency (DF), Term Frequency-Inverse Document Frequency (TF-IDF), Normalised Term Frequency-Inverse Document Frequency (NTF-IDF).

Feature value representation from primary studies				
Study	BR	TF	TF-IDF	NTF-IDF
[1]		✓		
[11]	✓	✓	✓	✓
[12]	✓	✓	✓	✓
[18]			✓	
[8]	✓	✓	✓	✓
[20]		✓		
[57]	✓	✓		
[44]		✓		
[63]	✓	✓		
[72]				✓
[90]		✓		

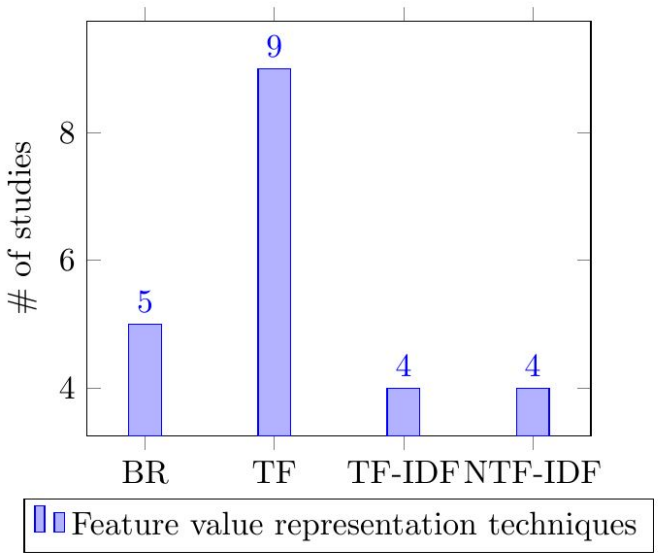


Figure 2.13: Distribution of feature value representation schemes.

2.4.7 Comparative analysis of studies reviewed in this paper

Figure 2.15 shows the VA study distribution compared to other fields. The distribution shows that we used 66% of the primary studies which are in the VA domain and the remainder were from other fields. Figure

Table 2.4: Feature selection techniques from primary studies. Legend: Pearson correlation (PC), Chi-squared (CS), Information gain (IG), Principal component analysis (PCA),Fisher Markov Selector (FMS), Improved Global Feature Selection (IGFS), Local Semi Supervised (LSS), Natural Language Processing (NLP)

Feature selection techniques from primary studies								
Study	PC	CS	IG	PCA	FMS	IGFS	LSS	NLP
[1]	✓							
[12]							✓	
[66]	✓							
[71]		✓						
[15]	✓	✓	✓					
[71]			✓					
[44]	✓	✓	✓		✓	✓		
[8]	✓	✓	✓	✓				
[11]	✓	✓	✓		✓	✓		
[20]			✓					
[21]								✓
[41]								✓
[63]								✓
[72]								✓
[73]								✓
[90]								✓
[93]								✓
[52]		✓						

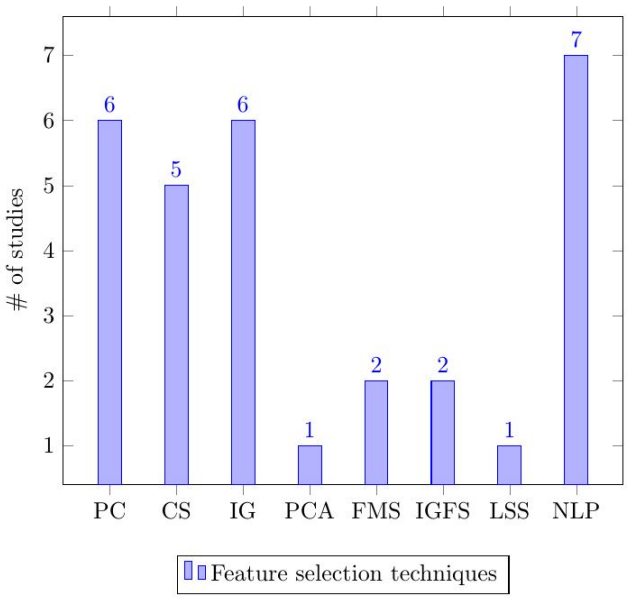


Figure 2.14: Distribution of feature selection techniques.

2.16 below depicts the journal articles and conference papers used in this methodological systematic analysis. From the primary studies synthesised, we used 66% from reputable and peer reviewed journal articles and 34% from conference proceedings.

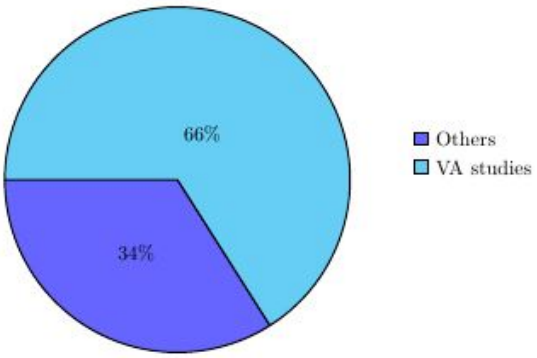


Figure 2.15: Distribution of VA related studies in comparison to other fields used in this methodological systematic analysis.

2.4.8 Evaluation metrics

The section below discusses various performance evaluation metrics used in this systematic methodological systematic analysis. Table 2.5 and Figure 2.17 show the distribution of performance evaluation approaches applied in this methodological systematic analysis. Out of the thirty four (34) studies that were reported in literature to have applied evaluation metrics, the most used evaluation metrics are Precision and Recall

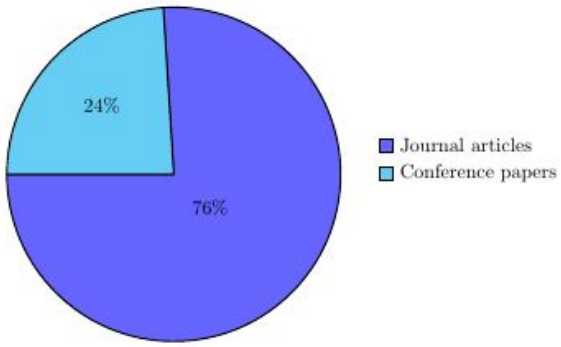


Figure 2.16: Distribution of journal articles and conference articles used in this methodological systematic analysis.

followed by CSFM accuracy, CCC, Specificity and F- score.

2.4.9 Tag cloud of common and frequent terms of reviewed studies

Figure 2.18 is a visual model that illustrates the highest frequency terms appearing in the title and abstract of the peer reviewed literature, where higher occurring terms appear in larger font size. From our cloud tag we can deduce that ML task of classification appears more, also we can can report that verbal autopsy is common within the health field through text or data that forms part of our narratives.

2.4.10 Models

The section below discusses various models in this methodological systematic analysis. Table 2.6 and Figure 2.19 show the distribution of statistical and machine learning models applied in this study. In the table we have the following abbreviations; VA algorithms (VA), Logistic Regression (LR), Random Forests (RF), Decision Tree (DT), K- nearest neighbor (k-nn), Support Vector Machine (SVM), Artificial Neural Network (ANN), Naive Bayes Classifier (NBC), K-means clustering (k-means), Rule-based (RB), Ensemble Classifier (EC), Convolutional neural network (CNN), Gated Recurrent neural network (GRN). This methodological systematic analysis suggests that out of 36 studies, conventional VA algorithms are the most widely used technique, followed by Naive Bayes Classifier, RF and SVM and k-nn.

2.4.11 Machine learning framework guidelines for VA

Our proposed guiding framework for VA researchers and practitioners is based on the results of this methodological systematic analysis. First and foremost, due to the nature of VA datasets, which may be balanced or imbalanced, we recommend performing exploratory data analysis and data preprocessing as the initial steps in data

Table 2.5: Evaluation metrics techniques applied in primary studies. Legend: Specificity (Sp), Precision (Pr), Recall (Re), F-score (Fs), Accuracy (Ac), CSFM (CS), CCC, PCCC, Kappa (Ka), Case Similarity measure (CSM), Top Cause (TC), Top3CoD (T3), Area Under Curve (AUC)

	Evaluation metrics of primary studies												
Study	Sp	Pr	Re	Ac	CSFM	CCC	PCCC	Ka	CSM	Fs	TC	T3	AUC
[1]		✓	✓		✓		✓						
[4]	✓	✓	✓		✓								
[2]					✓								
[28]	✓	✓			✓	✓							
[29]		✓											
[11]		✓	✓										
[12]		✓	✓							✓			
[15]		✓	✓	✓						✓			
[16]					✓	✓							
[17]	✓	✓			✓		✓						
[18]		✓	✓	✓						✓			
[30]					✓						✓		
[31]	✓	✓	✓		✓			✓					
[66]					✓	✓					✓	✓	
[67]					✓	✓							
[68]					✓	✓							
[69]		✓			✓		✓						
[70]					✓	✓		✓					
[71]	✓			✓		✓							
[8]		✓	✓	✓					✓			✓	
[44]		✓	✓	✓						✓			✓
[45]	✓	✓	✓		✓	✓		✓					
[19]	✓		✓	✓		✓							
[48]						✓		✓					
[54]			✓				✓		✓				
[20]				✓									
[21]		✓	✓							✓			
[41]		✓	✓		✓					✓			
[57]		✓	✓							✓			
[63]			✓	✓							✓		
[72]		✓	✓							✓			
[77]			✓										
[87]	✓	✓	✓		✓	✓		✓					
[93]		✓	✓										
[52]													✓

Table 2.6: Models applied in primary studies.

	Models applied in of primary studies												
Study	VA	LR	RF	DT	k-nn	SVM	ANN	NBC	k-means	RB	EC	CNN	GRN
[1]		✓	✓	✓		✓	✓	✓					
[2]	✓												
[28]	✓		✓										
[65]	✓												
[29]						✓					✓		
[11]			✓			✓		✓					
[12]			✓			✓		✓					
[15]			✓	✓	✓	✓		✓		✓			
[16]	✓	✓	✓										
[17]	✓							✓					
[18]			✓	✓	✓	✓		✓			✓		
[30]	✓												
[31]	✓							✓					
[66]	✓												
[67]	✓												
[68]	✓												
[69]	✓		✓										
[43]					✓			✓		✓			
[70]	✓												
[71]	✓												
[8]		✓	✓	✓	✓	✓	✓	✓				✓	
[44]			✓	✓	✓	✓		✓					
[45]	✓												
[19]		✓					✓						
[47]		✓											
[48]	✓												
[54]								✓			✓		
[20]		✓		✓		✓		✓					
[21]									✓			✓	✓
[41]													✓
[57]					✓				✓	✓			
[63]					✓				✓	✓			
[72]						✓							
[77]	✓												
[87]	✓												
[90]	✓												
[52]		✓											

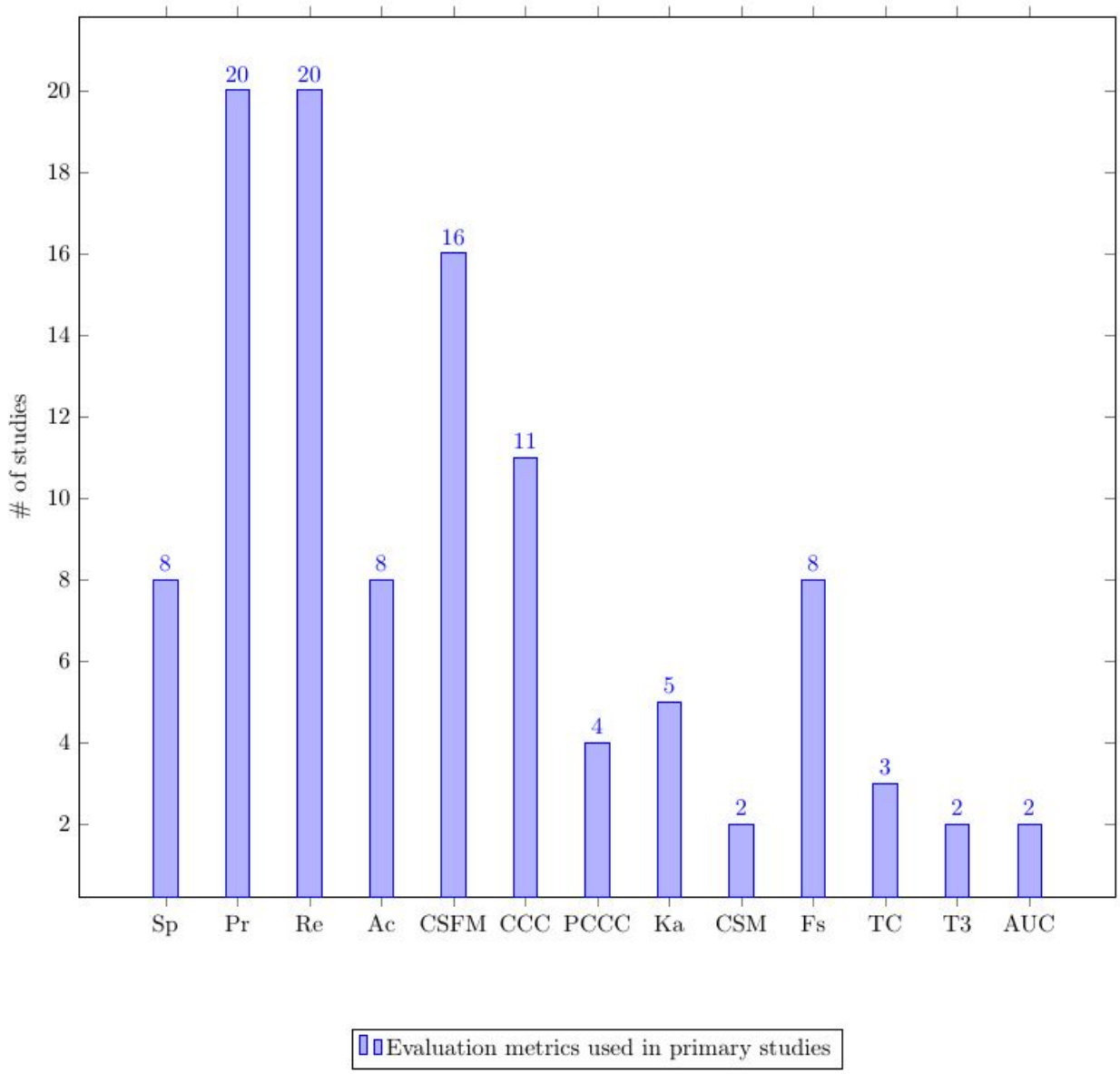


Figure 2.17: Distribution of evaluation metrics from primary studies reviewed.

cleaning. After successful data cleaning, we suggest conducting well-structured and logical feature selection and engineering to identify the most discriminative and representative features for our vector space. To address class imbalance in the VA domain, and depending on the dataset's type and characteristics, we propose applying techniques such as under-sampling, over-sampling, re-sampling or hybrid methods to balance the class distribution. Moreover, it is crucial to ensure that the applied techniques do not compromise the integrity of the data, particularly in cases of rare causes of death. It is also imperative to evaluate different data balancing methods and their impact on model performance. This evaluation allows for the selection of the optimal approach while maintaining the clinical relevance of predictions. Notably, caution should be exercised to avoid eliminating or overly simplifying minority classes, as this could result in the loss of critical information related to less common causes of death during class balancing.

The next vital step involves establishing baseline models. Start with simple models such as logistic regression and decision trees as benchmarks to evaluate the effectiveness of more complex models. This approach helps in understanding the basic patterns in the data before moving on to advanced techniques. Additionally, compare these baseline models against more sophisticated machine learning models, such as Random Forests,

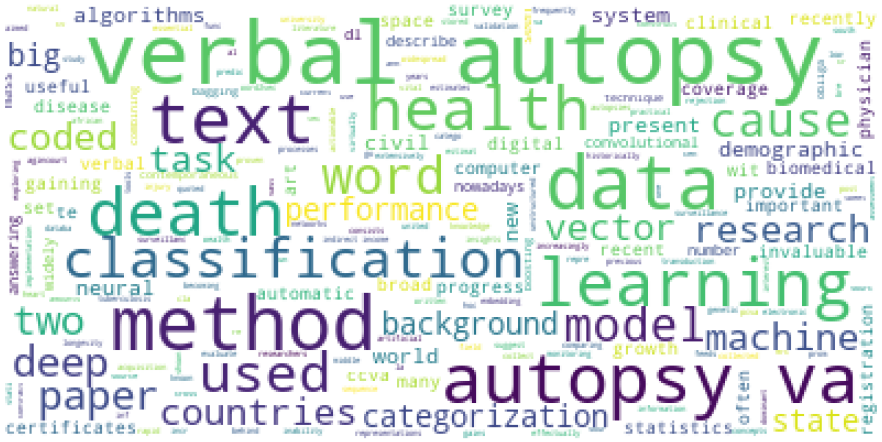


Figure 2.18: Tag cloud of this methodological systematic analysis

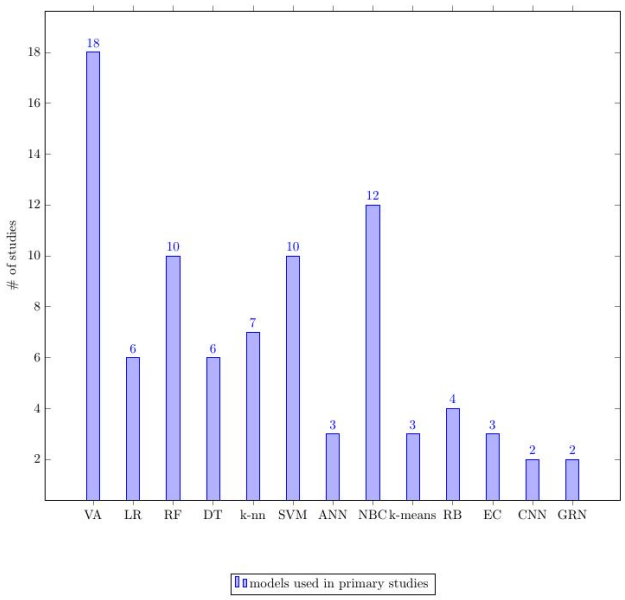


Figure 2.19: Distribution of statistical and ML models used in primary studies.

Gradient Boosting Machines, SVM and Neural Networks. During this process, ensure that each model is properly tuned and cross-validated to prevent overfitting and provide a fair comparison. It is also important to compare ensemble model performance with individual models to determine the best approach for VA data. Periodically re-evaluate the benchmark models with updated datasets to ensure they remain relevant and effective as new data becomes available. However, this systematic methodological analysis shows that SVM and ensemble classifiers generally perform better than other models. After this process, we recommend applying evaluation metrics to assess the performance of the approaches and choose the optimal one for classification.

This analysis also showed that the most common feature extraction techniques are bag of words and n-grams, while the most frequently used feature value representation schemes are Binary Representation (BR) and Term Frequency (TF). Furthermore, our investigation revealed that the most common feature selection or dimensionality reduction techniques include expert-driven feature selection by physicians, information gain, and Pearson correlation.

Moreover, the few studies that have used the emerging trend of advanced machine learning known as deep learning have proved that distributed vector representations (Word2vec, Sentence2vec, Doc2vec and Glove) show promising results. CNN improves VA performance through use of character embeddings. New models exist in deep learning such as the Recursive Neural Network and Transformer, which have shown promising results in other domains and can prove useful to explore with, in our VA field. Performance evaluation metrics such as Precision, Recall, Accuracy, AUC and F-score can be applied to evaluate models.
Below Figure 2.20 is the proposed visual representation of the guiding framework.

2.5 Discussion

Our study conducted a comprehensive methodological analysis of publications that are in the VA field in terms of statistical and machine learning application to determine cause of death. Subsequently, several main sub fields related to ML in VA were presented. We further gave an insight on the application of statistical, ML and advanced deep learning approaches in VA, which has not been done at this stage. Therefore from this detailed methodological systematic analysis, one can be able to identify various statistical and ML applications in VA and thus also identify opportunities and challenges in the VA domain. Most of the studies carried out in VA have been using mostly supervised learning techniques with a few reports on unsupervised learning approaches. Moreover, there is little or few studies that apply deep learning as a ML approach in VA. Deep learning model application can improve results within the field of VA, such as better knowledge representation systems for VA, thus enforcing standardisation and a shared data repository for further research.

Generally all ML studies in VA have been trained on sparse high dimensional data and high training cost of samples, thus applying alternative approaches can effectively handle increased training data samples and improve efficiency. There is also a limitation in terms of having publicly available training datasets, which can be used by researchers to train and improve models. This should be one key area to be looked into within the VA community.

Future VA studies should incorporate and utilise well structured and logical feature selection and feature engineering techniques such as filter, wrapper and embedded methods, as they have shown promising and excellent results in other fields. These methods make it possible to attain relevant features that are discriminative, thereby addressing issues of noise and high hyper parameter space amongst many drawbacks [72].

Literature to date has presented various studies that have applied model evaluation in terms of performance for various statistical and ML frameworks in VA. Yet little evidence exists that shows a consensus on optimal best performing algorithm. Our methodological systematic analysis investigated commonly applied statistical and ML approaches including deep learning approaches and we found out that the SVM as the most optimum classifier and we deduced that hybrid models are better than single models classifiers. However, there is

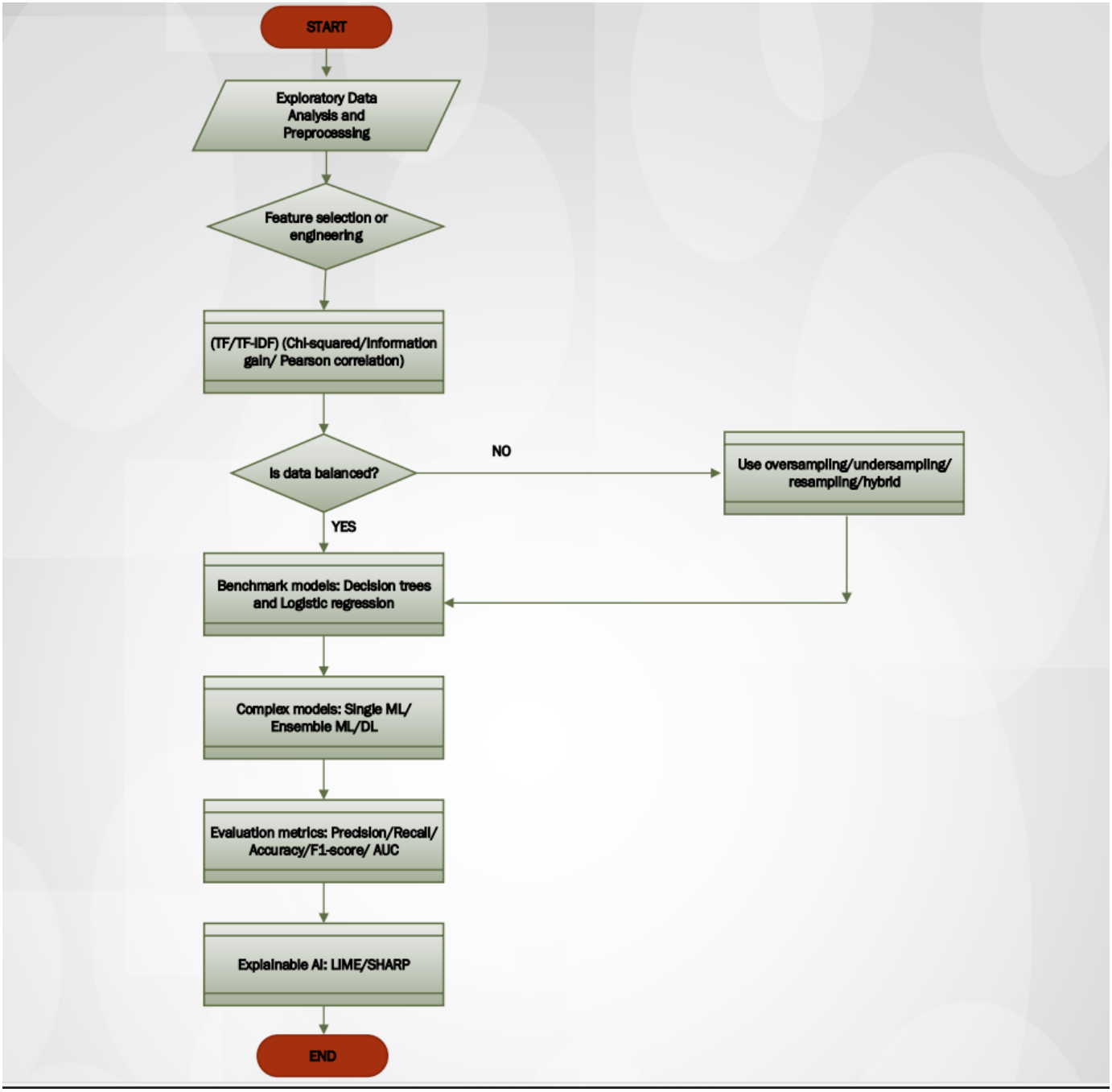


Figure 2.20: Proposed guidelines for the VA machine learning framework

less application of deep learning models in VA, even though the application of the CNN from literature showed improvements to classification accuracy as well as better knowledge representation using character embeddings in the VA domain. This was better in comparison to conventional statistical and classical machine learning models.

Our study further flags limitations such as lack of standardised VA questionnaires, scarcity and cost of experts to perform VAs, various medical ontologies resulting in heterogeneous datasets, relying on expert's diagnosis as the gold standard, scarce availability of publicly available biomedical datasets, various ways of acquiring data, lack of effective preprocessing procedures, few data instances, high and complex feature space that id highly dimensional amongst many. The proposed guiding framework for VA practitioners and researchers in this study suggests performing robust and logical exploratory data analysis and pre-processing for data

cleaning, applying feature selection and feature engineering in the most effective way, performing data balancing, employing benchmark models and applying evaluation metrics for model performance. This survey pointed to emerging directions in VA such as the applications of deep learning models in VA and explainable AI.

Prospectively the VA community should come up with strategies of effective data balancing of datasets. They should seek to apply emerging techniques apart from common over-sampling and under-sampling. Newer novice approaches like linear dependence approach and wavelet data transformation should be employed in VA. Upcoming work should consider doing more effective exploratory data analysis before model development. This gives the researcher the opportunity to fully understand the feature space in terms of complexity, feature distribution, association and correlation. As a result of this one can be able to identify key important and relevant features that are the best predictors of our outcome variable. In other words we can esily narrow down our feature space without losing any key information and this improves model performance. Future research should introduce and apply standard VA questionnaires which are key input to the VA process. Moreover, there should be standardisation of medical ontologies and incorporation of homogeneous datasets. Data acquisition and pre-processing techniques should be standardised and be performed in a robust and logical manner.

There is scanty evidence from past studies in employing effective exploratory data analysis. There is great need to incorporate such in order to to better understand distributions of features and address model complexity to allow efficiency in model development in VAs. As a result of exploratory analysis, one can easily acquire knowledge prior to model development such as; identifying the correlation between the predictors and outcome feature and predictors that are more important in predicting the target feature. Apart from this it helps in the reduction of the feature space, hence addressing complexity and improve on performance of algorithms. Furthermore, there is need to make available public datasets in order to train models for improved performance. If resources permit, there should be a team of experts that specifically deal with VAs.

The future of the VA community should focus on transparency, to justify and make non-transparent models explainable. One example of a transparent model is Local Interpretable Model-Agnostic Explanations (LIME). LIME can explain any classifier prediction through learning an interpretable model around a prediction. Transparency is key in that it builds trust in humans in terms of predictions, such that if there is need for interventions or any form of action then it can be taken. Transparency gives the leverage for an expert to accept or reject a model.

Most VA studies have employed supervised machine learning approaches. There is need for the VA research community to consider other forms of machine learning such as; unsupervised learning (which do not need a labelled dataset, hence performs classification or clustering based on unlabelled points) , semi-supervised machine learning (makes use of few labelled data points to make predictions), transfer learning (relevant where the input data set has few labelled features to effectively train a model) and reinforcement learning (uses environment interaction to learn through past exposures to accurately classify VA narratives using a a reward system).
To get better improved results going forward it is vital for all researchers and stakeholders that deal with VA studies to collaborate effectively and build strategic partnership that seek to address the confines within the VA domain with particular emphasis on tools, approaches, data sources among many and put in place effective strategies that can address the problems at hand.

2.6 Conclusion

In this study, we reviewed in detail traditional statistical and ML approaches to determine cause of deaths from VAs, identified limitations of each of these techniques, proposed machine learning framework guidelines for determining causes of death from VAs and pointed to emerging future directions. Most VA models use the PCVA approach as a gold standard, even though it is expensive, inconsistent and inaccurate (subjective and prone to errors because of the VA narratives that are taken as input). In practice, as an alternative to PCVA, mostly statistical approaches such as LR and NB are employed because of simplicity and transparency. Nevertheless, in literature complex machine learning and emerging deep learning frameworks exist, that can be applied to improve cause of death classification. Most of the machine learning models reviewed apply the RF and SVM. However, incorporating ensemble classifiers can yield better results. To achieve optimum predictions in determining cause of deaths from VAs, models need to employ effective data exploration (fully understand datasets before modelling), data balancing, pre-processing (improve knowledge representation of medical datasets) and robust emerging feature engineering techniques (extract relevant features and narrow down the feature space, thus improving model performance). Additionally, other types of machine learning like transfer learning and reinforcement learning need to be explored further as they have shown promising results in other domains. Since VAs are key inputs to the civil registration, there is need to build transparent models for cause of death classification.

The significant contribution of VA data in informing interventions, health priorities, monitoring and evaluation programmes, is of importance. Therefore, it is imperative for the VA community to continue exploring and developing new strategies in line with the long term efforts of improving the civil registration systems. Advances in VA methods should seek to achieve consistency, comparability, and generalisation to other studies.

2.7 Declarations

2.7.1 Ethics approval and consent to participate

This is not applicable to this study.

2.7.2 Consent for publication

This is not applicable to this study.

2.7.3 Availability of data and materials

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

2.7.4 Competing interests

The authors declare that they have no financial or personal relationship(s) that may have inappropriately influenced them in writing this article.

2.7.5 Declaration of interest

The authors declare that they do not have any financial and personal relationships that could inappropriately influence this study.

2.7.6 Funding

This research was not grant funded.

2.7.7 Authors' contributions

All authors listed below have read and approved the manuscript.

MM is the main author and conducted the whole methodological systematic analysis process including the write up.

EM reviewed and contributed on statistical models.

CK reviewed and contributed on verbal autopsy theory.

TC reviewed and contributed on the machine learning concepts.

2.7.8 Corresponding authors

Michael T. Mapundu

University of the Witwatersrand

School of Public Health

Department of Biostatistics and Epidemiology

7 York Road, Parktown, Johannesburg, South Africa

Tel: +27 11 71-72627

Email: michael.mapundu@wits.ac.za

2.7.9 Acknowledgements

This work was supported by the Developing Excellence in Leadership, Training and Science (DELTAS) Africa Initiative Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) [Grant No.DEL-15-005]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS) Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [Grant No. 107754/Z/15/Z] and the United Kingdom government.

3 Performance evaluation of Machine Learning and Computer Coded Verbal Autopsy (CCVA) algorithms for cause of death determination: A comparative analysis of data from rural South Africa

This study was published in Frontiers in Public Health Journal (September 2022. doi:10.3389/fpubh.2022.990838)

Michael T. Mapundu[1]*, Chodziwadziwa W. Kabudula[1,2], Eustasius Musenge[1], Victor Olago [3], Turgay Celik[4,5].

[1]Division of Epidemiology and Biostatistics, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Parktown, Johannesburg, South Africa
[2]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), University of The Witwatersrand, Johannesburg, South Africa
[3]National Health Laboratory Service (NHLS), National Cancer Registry, Johannesburg, South Africa
[4]Wits Institute of Data Science, University of The Witwatersrand, Johannesburg, South Africa
[5]School of Electrical and Information Engineering, University of The Witwatersrand, Johannesburg, South Africa

3.1 Abstract

Computer Coded Verbal Autopsy (CCVA) algorithms are commonly used to determine the cause of death (CoD) from questionnaire responses extracted from verbal autopsies (VAs). However, they can only operate on structured data and cannot effectively harness information from unstructured VA narratives. Machine Learning (ML) algorithms have also been applied successfully in determining the CoD from VA narratives, allowing the use of auxiliary information that CCVA algorithms cannot directly utilise. However, most ML-based studies only use responses from the structured questionnaire, and the results lack generalisability and comparability across studies. We present a comparative performance evaluation of ML methods and CCVA algorithms on South African VA narratives data, using data from Agincourt Health and Demographic Surveillance Site (HDSS) with physicians' classifications as the gold standard. The data was collected from 1993 to 2015, and has 16338 cases. The random forest and extreme gradient boosting classifiers outperformed

the other classifiers on the combined dataset, attaining an accuracy of 96% respectively, with significant statistical differences in algorithmic performance ($p < 0.0001$). All our models attained Area Under Receiver Operating Characteristics (AUROC) of greater than 0.884. The InterVA CCVA attained 83% Cause Specific Mortality Fraction accuracy and an Overall Chance-Corrected Concordance of 0.36. We demonstrate that ML models could accurately determine the cause of death from VA narratives. Additionally, through mortality trend and pattern analysis, we discovered that in the first decade of the civil registration system in South Africa, the average life expectancy was approximately 50 years. However, in the second decade, the life expectancy significantly dropped, and the population was dying at a much younger average age of 40 years, mostly from the leading HIV related causes. Interestingly, in the third decade, we see a gradual improvement in life expectancy, possibly attributed to effective health intervention programmes. Through a structure and semantic analysis of narratives where experts disagree, we also demonstrate the most frequent terms of traditional healer consultations and visits. The comparative approach also makes this study a baseline that can be used for future research enforcing generalization and comparability. Future work will entail exploring with deep learning models for CoD classification.

Keyword: Cause of Death, Machine Learning, Verbal Autopsy

3.2 Introduction

More than 65 percent of the population in the world lack high quality information on cause of death (CoD) since every year about sixty million deaths worldwide are not assigned a medically certified cause [1]. As such, most of the countries in the world fail to meet the United Nations 90 percent death registration coverage requirement, as deaths in many Low to Medium Income Countries (LMICs) are not captured in civil registration systems [2, 3]. On the contrary, the CoD information is vital for public health monitoring, informing critical health policies and priorities. Therefore, in the absence of clinically oriented sources, CoD information should be derived from alternative sources. Verbal Autopsy (VA) is the mostly used tool worldwide as an alternative source of CoD information. VA is common in LMICs and is a process that is used to determine CoD where deaths occur outside health facilities and is not certified by a medical practitioner [4]. These sentiments are supported by [5], who also report on the importance of the VA process in determining CoD in countries where there are no active civil registration systems. The VA process is conducted by a non-medical personnel who seek to elicit valuable information using both structured questions and an open narrative section with next of kin of the deceased about circumstances and events that led to death [1]. Two doctors are given the full set of responses, both from structured questions and open narratives for assessment and reach a consensus on the CoD and if not a third physician is consulted, a process known as Physician Coded Verbal Autopsy (PCVA). PCVA is the mostly used process for determining CoD. However, it is widely criticised because of lack of robustness, cost, time, inconsistencies and inaccuracies as it is subjective and prone to errors amongst many drawbacks [7]. This results in PCVAs mostly employed for training and validation of computational approaches. The surge of technological advances has availed a plethora of automated methods for determining CoD which are faster, efficient and cost effective [1]. Most of the research that reports on ML application in the VA domain mainly use the responses from the questionnaire as the classical dataset. As such, this affects comparability and generalisability. In this study we validate the performance of various ML techniques using various VA data types for determining CoD using a comparative analysis approach. We apply enhanced data standardisation and normalisation strategies to achieve optimum transparency and accuracy through addressing most model limitations and applying recommendations that are reported in [10, 8]. We assess the robustness of several classifiers including; random forest (RF), k-nearest neighbour (KNN),

decision tree (DT), support vector machine (SVM), logistic regression (LR), artificial neural network (ANN), Bayes Classifier (BC), bagging and eXtreme Gradient Boosting (XGBoost) as ensemble classifiers. We also validate our dataset using the common conventional Computer Coded Verbal Autopsy (CCVA) algorithm; InterVA.

3.2.1 Computer Coded Verbal Autopsy Algorithms

Previous studies report on the most commonly used VA algorithms also known as CCVA algorithms. These CCVA approaches use expert-driven rules to determine CoD from VAs [69, 67, 94, 30, 17]. The VA algorithms make use of the responses from the standardised structured World Health Organisation questionnaire that denote signs or symptoms based on the deceased health history prior to death. Most of these VA algorithms take input of VA data derived from real deaths, symptom-cause information (SCI) which is a repository of information about symptoms that are related to each probable CoD. Additionally, they make use of logic that entails a logical algorithm that combines the SCI and VA data to identify cause-specific mortality fractions (CSMF), so as to assign a specific CoD.

The InterVA algorithm uses the Bayes Rule, which is a function that computes conditional probabilities of a sample when given evidence and SCI as conditional probabilities of symptoms given a specific CoD using CSMFs. This approach is reported in the work of [66, 28, 17, 45].

These VA approaches have been widely criticised in terms of their credibility and reliability. The study of [70] reports on evaluation of VA expert algorithms, and deduces that population level accuracy is similar to that of ML approaches with CSMF in the range of $57\% - 96\%$. Similar findings are also presented in the work of [95] who did a study where they validated data derived algorithms against the gold standard of physician review using various disease categories based on the CSMF. [28] argues that, there is little evidence to justify the CCVA as a possible replacement of the gold standard which is the PCVA. Therefore, there is need for further investigations and research with large datasets to train and test models on CoD classification.

Little research exists in the VA domain on the application of ML to determine CoD from VA narratives. These ML algorithms make use automated computer programs that can take input of data to learn new trends and patterns from complex data by applying optimisation techniques for VA classification [35].

3.2.2 Machine Learning in VA

Most ML model predictors commonly use only responses from the standardised questionnaire, attaining Sensitivity scores around 60% for individual CoD classification, using various numbers of CoD categories. On the contrary, the work of [1] demonstrates that the VA narratives have valuable rich information that can be used for CoD determination. ML can avail real-time results that is similar to that of physicians/experts [40]. Alternative complex ML approaches exist in literature and can be used as substitutes for the PCVA and CCVA algorithms as approaches of determining CoD.

[96] applied the Bayesian hierarchical factor regression models to infer CoD using VA narratives and report an improvement in model performance on inferring CoD and CSMF. However, they used thirty four disease categories. [97] applied six different ML algorithms (SVM, ANN, KNN, DT, C5.0 and gradient boosting). Their results report the SVM as the best classifier with Accuracy of more than 80%. However, they used six disease categories. Similar results are reported in the study of [15], with SVM attaining a Precision of 78.1%, Recall of 78.3%, F-score of 78.2% and overall Accuracy of 78.25% for 16 disease categories. Their study used text classification techniques to predict CoD from forensic autopsy reports. Other studies of [12, 18, 63] also

found the similar results and deduce that feature extraction approaches are grossly affected by variations in words as well as word combinations.

The work of [47] used the LR model to determine the completion rate of VA and factors associated with undetermined CoD. They report a completion rate of 83% to 89%. They ascertain that 94% of deaths submitted to physicians were assigned a specific cause and on the other hand 31% were labelled as undetermined. [46] reports various common diseases that lead to death using CSMF and LR classifier and they achieved 80% Specificity. [19] applied ANN classify CoD from VAs and achieved a Sensitivity of 45.3%. They concluded that more explorations are needed with large datasets and large training samples to improve results of the ANN. The study of [16] used the RF classifier to assign CoD categories and affirmed that the RF algorithm performed better if not as the PCVA approach. Additionally, they point out that the RF classifier was better than PCVA on overall chance concordance and CSMF accuracy for both adults and children.

Related work that has also used VA data for cause of death determination is also reported elsewhere in [73]. They conclude that using word occurrences produced better results as compared to word occurrence features and suggest using large datasets in order to improve model performance. Their sentiments are echoed in the work of [20, 54]. Additionally, [44, 18, 8, 15] has done vast work in the VA domain and argue that, uni-grams are better feature extraction techniques, Term Frequency (TF) and TF-IDF are better feature representation schemes, Chi-squared is a better dimensionality reduction approach. They recommend employing effective data cleaning strategies and feature engineering techniques to get improved performance.

Despite the reported results in literature, both CCVA algorithms and ML models applied to VA data to determine CoD, suffer from challenges and limitations as they lack concrete evidence where there is limited expert diagnosis and cannot be fully utilised to inform health priorities [2]. Most of the CCVA approaches use statistical concepts and scores to determine CoD [65, 69]. Moreover, these approaches are affected in terms of optimal performance because of their dependency on sample size, age group, causes of death, and characteristics of the sample [70, 4, 71, 17, 65]. Other issues that affect VA data quality, emanate from having interviewers being untrained, incompetent and unqualified to appropriately elicit relevant and appropriate symptoms on causes of death. Additionally, language barriers call for the need for the interviewer and interviewee to speak the same language so as to derive best results. [4] recommended incorporating fully trained multiple translators. The other downside is the length of the recall period which can create a bias in the collected VA data. The heterogeneity of various autopsies in terms of the non-intersecting dialects of the English language (terms being in the native language) compromises data quality as most of these approaches tend to omit such autopsies in their model prediction, yet they might entail valuable information.

All the discussed challenges and limitations affect the VA data quality that are taken as input to the CCVA and ML approaches. Therefore, we can deduce that there is a great need addressing these challenges in order to remove room for any bias and mis-interpretations of the models, thereby enforcing generalisability and comparability. This study demonstrates the robust assessment of ML approaches and CCVA algorithms in determining CoD, thus availing a baseline ML framework that can be used for comparability and generalisation across all VA dataset types.

3.3 Methods

3.3.1 Study design

This is a retrospective cross-sectional study that uses secondary data analysis. All the cleaned VA datasets, model performance and classification results of various tasks are pushed from a Python Jupyter Notebook

environment and housed within a PostgreSQL Version 4.2 object-relational database management system.

3.3.2 Population

This study uses VA narratives data from the study area of the Agincourt Health and Demographic Surveillance System (HDSS). The HDSS came into existence in 1992 and is located in the rural Sub-district of Bushbuckridge under Ehlanzeni District, in Mpumalanga Province, in north-eastern South Africa. The study area covers approximately 420km^2. According to Agincourt fact sheet of 2019, the population was at 116 247 individuals residing in 28 villages with 22 716 households, with males being 55 961, females being 60 280, children under 5 years being 11 724 and school going children with ages from $5 - 19$ being 35 928 [98].

3.3.3 Data Source

The source of our data for this study is the Agincourt HDSS. It is a surveillance site that specifically provides evidence based health monitoring that seeks to strengthen health priorities, practice and inform policy. The VA narratives data is for the period of 1993 to 2015. However, physician diagnosis was done from 1993 to 2010, and this target variable of the doctors diagnosis is enough for model training and prediction.

In this study, we used three types of datasets. The responses from the standard questionnaire, narratives and a combination of the responses and the narratives. The whole dataset had 287 columns/features and 16338 records/observations. For the responses only, we took all features that had responses from the standard questionnaire as our predictors and the CoD assigned by physicians using International Classification of Diseases-10 (ICD-10) code for each record in the dataset as our target variable. Ultimately, we had 231 predictors (all symptoms, age at death and gender) and 1 target variable, and all our features were in English. The predictions using the narratives were done using age at death, gender and the narrative feature and 1 target variable.

For the combined VA dataset we used 232 predictors and 1 target variable. We only added the VA narrative feature to the responses dataset in order to have our combined dataset. We further created twelve CoD categories with corresponding labels, class distribution with number of samples for each class before and after data balancing for our training dataset, as shown in Table 3.1. The CoD categories were derived based on InterVA user guide and literature studies of [94, 12, 99, 1].

Table 3.1: Twelve disease classes and number of data samples before and after data balancing.

Class labels and corresponding number of samples			
Disease category	Label	Samples before data balancing	Samples after data balancing
HIV/TB	0	3388	3388
Other infectious	1	964	3388
Metabolic	2	242	3388
Cardiovascular	3	140	3388
Indeterminate	4	1468	3388
Maternal and Neonatal	5	121	3388
Abdominal	6	117	3388
Neoplasms	7	93	3388
External causes	8	89	3388
Neurological	9	57	3388
Respiratory	10	46	3388
Other NCD	11	21	3388

Fig 3.1 illustrates the logical steps that we follow for this study experiments. We first do data acquisition of our VA narratives as a comma separated value text file (csv), followed by data exploration and cleaning. Additionally, we do feature engineering, data balancing and feed our data to our models for training, validation and testing. Lastly we do CoD classification.

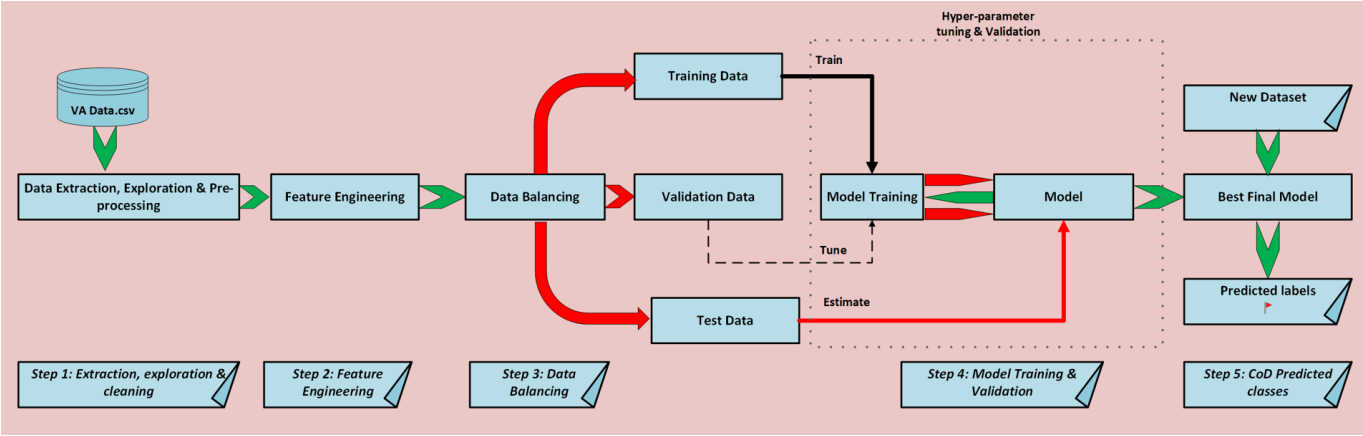


Figure 3.1: Schematic diagram of ML process followed.

3.3.4 Data Pre-processing and Encoding

For the questionnaire responses dataset, we cleaned and replaced all nulls with zeros, implying that there was no symptom assigned for a missing value in a record. All symptoms that had a 'Y' were encoded as a 1 meaning that the record had a present symptom value. On the other hand, all features that had a 'N', were encoded as a 0 meaning that those records had no symptoms present. In order to normalise and standardise the narrative feature used with the combined dataset, we did pre-processing in order to remain with only relevant data. Data was first imported in comma separated value format, followed by pre-processing. The pre-processing stage entailed converting all text to lowercase, removing all punctuation, spaces, numbers and special characters. Tokenisation was done by splitting a document (seen as a string) into tokens. Stopword removal was then applied to do away with insignificant words using the NLTK library of English stopwords. We applied normalisation using the Python spacy package, a process known as lemmatization. Lemmatization uses a dictionary of known word forms and considers the role of a word in a sentence with the aim of extracting some normal form of a word. Lastly we applied feature engineering to determine the most representative features, as we then aimed at retaining only relevant words in the vector space by applying a weighting scheme [14]. All categorical data was encoded using the one-hot encoding technique to create numeric vectors. This was followed by concatenating the narratives and the questionnaire responses datasets using horizontal stacking which were pushed to our models for training, validation and testing.

3.3.5 Feature Engineering

We did feature engineering in order to derive new input features from existing ones. This process was done in three phases namely;feature extraction, feature selection and feature value representation. Feature extraction was applied in order to get only relevant and useful features from textual data using n-gram models. The n-grams are a set of words that are sequential as they make use of continuous number of items such as characters or words from a given sequence of narratives. n-gram models can be of the form; a) n = 1 (unigram), b) when n = 2 (bigram), c) n = 3 (trigram) and d) hybrid-grams (mixture of unigram, bigram, and trigram) [8, 15]. This was followed by the feature value representation stage employing the TF-IDF approach. In

this phase we sought to create a numeric vector of features, were each feature will have a corresponding numeric value that can be used for model learning. TF-IDF considers a feature important if the feature occurs frequently in the VA narratives belonging to one class and less frequently available in narratives belonging to another class. Lastly we applied feature selection in order to attain the most useful subset of features from the narratives. This was achieved using Singular Value Decomposition (SVD) as a selection approach to reduce the dimensionality of our feature space, thus removing noise in our dataset. This dimensionality reduction technique, creates a matrix that only has relevant information producing an exact representation of data in a low dimensional space without any loss of data [100, 101].

3.3.6 Data Balancing and feature scaling

We applied data balancing to the training set to address data imbalance where one or more classes are less represented than the other classes, meaning that the majority classes have more samples as compared to other minority classes. As such, this creates bias in that minority classes as they will have fewer data points that can cause large misclassification errors. The ratio of majority against minority class was $1:160$. In order to address the issue of data imbalance we explored with various techniques (under sampling, over sampling, threshold and class weight). We attained optimal results when using the Random Over Sampling Examples (ROSE) and Synthetic Minority Oversampling Technique (SMOTE). After our experiments, we chose SMOTE as our best choice for our dataset. This possibly suggests that our dataset was well suited for SMOTE as a data balancing technique. Moreover, our balanced datasets behaved better than imbalanced datasets. SMOTE was applied by generating artificial samples for the minority class, through interpolation between the positive instances that lie together. This approach addresses the issue of over-fitting caused by the general oversampling approach that replicates existing positive cases [8]. We ended up having 3388 samples per each class. We did feature scaling using the Python Standard scaler library in order to get all our features within the same range as the target variable. After data balancing and feature scaling we fed the data into our twelve models for training and validation.

3.3.7 Machine Learning Models for CoD Prediction

We specifically applied supervised ML techniques to predict our target variable given input data. We aimed at predicting the related CoD by taking input of; questionnaire responses only, narratives only and combined questionnaire responses and narratives and then we fed into nine classifiers (SVM, DT, XGBoost, KNN, RF, Bagging, LR, BC and ANN). These ML approaches are reported elsewhere [102, 103, 51, 15, 100, 101, 54, 47, 44, 104, 1, 62, 36, 14, 97]. Using the questionnaire responses only, we created a feature space made up of binary responses as predictors and our target variable was a categorical ICD-10 code for CoD. Similarly we did the same for the narratives only dataset. For the combined dataset only, we added the narrative column to the list of our predictors.

3.3.8 Model Training, Validation and Testing

In this study we perform multi-class classification, where we generated individual prediction models for each of the twelve disease categories. Data was split into 70% training, 20% validation and 10% testing on unseen or new data, for all our nine models. We evaluated model performance by assessing the robustness of the nine classifiers by applying 10-fold cross-validation supplemented by the GridSearch algorithm. The k-fold cross-validation (k=10 in our study) is advantageous in that, it uses all observations for both training and validation, with each observation used for validation exactly once. On the contrary, this approach has a disadvantage of having to define the number of folds manually. In order to address the limitations of

the k-fold cross validation technique, we also used the automated GridSearch approach that eliminates the random setting of parameters and chooses optimum parameters automatically for a specific model.

In order to attain a better estimate of the generalization performance, we used 10-fold cross validation to evaluate the performance of each parameter combination, instead of using a single split into a training and validation set. Firstly, we specified the parameters for searching stored in a dictionary. GridSearch cross validation function then performed all the necessary model fits. All dictionary keys were the names of parameters that we wanted to tune, and the values, were the parameter settings that we wanted to test out. Applying cross validation, we managed to choose the optimal parameters that gave us the best model performance based on accuracy on the test set or unseen data.

We used optimisation parameters such as; cost complexity pruning and tuning parameter alpha through k-fold validation (tree based models). Moreover, we also used the Mean Squared Error (MSE) and Cross Entropy Error (CEE), Minkowski and Gini as cost functions to compute the minimal cost error between our predictor and the response using the k-fold cross validation approach to optimise model performance. These cost functions are described in [100]. Additionally we also employ L_1 and L_2 regularisation approaches to further optimise some of our models. L_1 regularisation involves eliminating features that are not useful for model prediction by setting some weights close to zero. On the contrary, L_2 regularisation tends to penalise large weights more and small weights less [101]. Table 3.2 depicts some of the model hyperparameters used in our models.

Table 3.2: Model optimal hyperparameters.

	Selected hyperparameters
Model name	Hyperparameters
XGBoost	L1, max_depth=10, objective=multi:softmax, learning_rate =0.1, alpha=0
RF	gini, max_depth =10, n_estimators=100, min_samples_leaf=1
ANN	relu, alpha=0.0001, solver=adam
KNN	minkowski, n_neighbors=5, p=2
SVM	gamma=scale, kernel=rbf, C=1.0
Bagging	KNN, max_samples, max_features
DT	gini, min_samples_split=2, min_samples_leaf=1,
LR	$L2, C = 1.0$
BC	alpha=1.0, fit_prior=True, class_prior=None

Legend: *XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbour; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier.*

3.4 CCVA algorithms

We followed the same preprocessing steps of our dataset and fed into our commonly applied CCVA algorithm InterVA. The data preprocessing steps entailed; de-duplication based on the identifier field (ID), dropping observations with peculiar IDs, filtering out observations with recorded age at death above 110 years and any observation with year of death before 1992 and after 2016. All records with unspecified sex were dropped from the raw dataset. All modelling for the InterVA was done in R. Libraries such as; knitr were used for dynamic report generation, lubridate was used for date and time functions, tidyr for organising and tidying of data, tidyverse for loading core packages, ggplot for plotting graphs, readxl for reading our excel raw data, InterVA for our CCVA algorithm. In order to determine the most probable CoD, we used the InterVA libraries

for analysis in our R-statistical analysis software guided by the work of [105, 30]. Since InterVA and InSilico are correlated, we decided to only validate the InterVA algorithm for comparability with ML approaches.

3.5 Identification of contradicting cases and best model predictors

In order to identify contradicting cases, where physicians were not agreeing on the diagnosis, we extracted a separate dataset. We used simple text mining techniques known as n-gram models for identifying of the contradicting cases and best features for our models (see section 3.3.5).

3.5.1 ML Techniques Model Evaluation

Performance evaluation of classifiers are evaluated using various metrics and we report the metrics based on studies by [44, 15]. We validated our results using one versus all with Accuracy, Precision, Recall , F-score and AUROC as our metrics for evaluation.

Accuracy denotes all classes with classified results that have been predicted correctly in fraction terms. Precision also known as the Positive Predictive Value (PPV) defines the proportion of VA narratives correctly predicted as positive to the total of positively predicted VA narratives. Recall also known as Sensitivity or True Positive Rate (TPR) defines the proportion of VA narrative correctly predicted as positive to all VA narratives in the actual positive category. F-measure computes the average or harmonic mean of Precision and Recall.

True Positives (TP) and True Negatives (TN) represent number of outcomes in which our prediction model correctly classifies positive and negative cases respectively. In our case TP denotes predicted positive VA narratives with a particular disease category from the twelve classes and are actually positive and TN denote predicted negative VA narratives with a particular disease category from the twelve classes and are actually negative. Conversely, False Positives (FP) and False Negatives (FN) denote the number of outcomes where our models incorrectly predicts the positive and negative classes respectively. As in our case the FP imply predicted positive VA narratives with a particular disease category from the twelve classes but are actually negative and FN depicts the predicted negative VA narratives with a particular disease category from the twelve classes but are actually positive.

The AUROC visualises the TPR against the false positive rate (FPR). Area under the ROC curve applies the principle of plotting a curve specific to a machine learning algorithm were the classifier is evaluated relative to a weighting on the area under the curve. Good performance of the algorithm is given a weight of close to 1, thus graph is AUROC closer to upper left corner and the poor performance of an algorithm is given a weight of 0.5 and below. Specificity computes the ratio of negative VA narratives that are correctly predicted as negative.

3.5.2 CCVA Techniques Model Evaluation

We explicitly validated the InterVA algorithm using CSMF accuracy and Overall Chance-Corrected Concordance (CCC). CCC computes the accuracy of individual cause assignment and ranges from 0 to 1 and the lower the CCC, the larger the error type on the accuracy of the underlying cause [66, 87]. On the other hand, CSMF accuracy defines accuracy as having a value between 0 and 1. This metric assumes the worst possible case for predicting CSMF and assigns a weight on the least possible CSMF value that matches total absolute error [106].

3.6 Statistical Analysis

We applied statistical tests for comparing the performance of our nine algorithms. We computed the variance of our models using descriptive statistics such as mean and standard deviation based on the results of our AUROC. Moreover, we computed some tests using 10-cross validation using the mean and standard deviation. Furthermore, we conducted some non-parametric tests since our data distribution was non-normal using the Kruskal-Wallis test, to test if the model's mean are different or the same. For the Kruskal-Wallis test, we considered $p < 0.005$ statistically significant. We applied the pairwise model comparisons using McNemar statistical tests, in order to be able to state objectively whether one model performs better than the other [107]. Since we did eight different tests, we used the Bonferroni corrected p-value of 0.0065, derived from $0.05/8$, where the denominator is the total number of tests. We used Python version 5.2.2 and STATA version 17 SE edition for all these statistical tests.

3.7 Results

In this section we present the results attained from CCVA algorithms and various classification techniques employed to determine CoD from various VA datasets (using only narratives as predictors, using questionnaire responses only and results of the combined features).

3.7.1 Performance Evaluation of ML Classifiers

We validated our results using one versus all with Precision, Recall, Accuracy, F1-score and AUROC. We report on Precision, Recall, Accuracy, F-score measure and AUROC for our nine classifiers in the CoD categorisation of the twelve disease classes for narratives only in Table 3.3, questionnaire responses only in Table 3.4, combined questionnaire responses and narratives in Table 3.5.

Results from only the VA Narrative Predictors

The XGBoost and RF classifier outperformed all the other classifiers with a Precision of 96%, Recall of 96%, F1-score of 96% and Accuracy of 96% respectively. The least performing classifier was the statistical BC classifier with an Accuracy of 71%. Overall, our nine models had an AUROCMIA (Area Under Receiver Operating Characteristics Micro Average) and AUROCMAA (Area Under Receiver Operating Characteristics Macro Average) between $0.910-0.998$ and $0.910-0.996$ respectively. Table 3.3 shows the detailed performance evaluation results of our nine models using VA narratives only.

Results from using questionnaire responses only as predictors

The ANN and XGBoost outperformed all the other classifiers when using questionnaire responses from the standardised questionnaire attaining a Precision, Recall, F1-score and Accuracy of 100% respectively. It was followed by Bagging our ensemble classifier and KNN that both recorded a Precision, Recall, F1-score and Accuracy of 98% respectively. Our statistical classifiers the LR and BC were on the lower ranking of our evaluation recording an Accuracy in the range of $74\% - 83\%$ respectively. All of our models attained highest AUROCMIAs within the range of 0.869 and 1 respectively. Our nine models XGBoost, RF, ANN, Bagging, SVM, LR, DT and KNN recording high scores and the BC a bit lower AUROCMIA of 0.869. Additionally, the same nine models attained highest AUROCMAAs within the range of 0.976 and 1 respectively. On the other hand the BC achieved an AUROCMAA score of 0.884. Table 3.4 shows the detailed performance evaluation results of our nine models using questionnaire responses only.

Table 3.3: Comparison of nine ML models using narratives only.

	Model evaluation					
Model name	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUROCMIA	AUROCMAA
XGBoost	96	96	96	96	0.927	0.906
RF	96	96	96	96	0.998	0.996
ANN	94	94	94	94	0.982	0.964
KNN	93	93	93	92	0.989	0.987
SVM	92	92	92	92	0.917	0.917
Bagging	91	91	91	91	0.997	0.995
DT	85	84	85	84	0.910	0.910
LR	82	82	82	82	0.977	0.959
BC	71	75	71	72	0.921	0.920

Legend: *XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbour; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier; AUROCMIA, Area Under Receiver Operating Characteristics Micro Average; AUROCMAA, Area Under Receiver Operating Characteristics Macro Average.*

Table 3.4: Comparison of nine ML models using questionnaire responses only.

	Model evaluation					
Model name	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUROCMIA	AUROCMAA
XGBoost	100	100	100	100	1	1
ANN	99	99	99	99	1	1
Bagging	98	98	98	98	0.998	0.998
KNN	98	98	98	98	0.997	0.997
RF	97	97	97	97	0.999	0.998
DT	97	97	97	97	0.976	0.976
SVM	94	94	94	94	0.990	0.988
LR	83	83	83	83	0.990	0.980
BC	74	77	74	75	0.869	0.884

Legend: *XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbour; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier; AUROCMIA, Area Under Receiver Operating Characteristics Micro Average; AUROCMAA, Area Under Receiver Operating Characteristics Macro Average.*

Results from using combined narratives and questionnaire responses

The XGBoost, ANN and the RF classifier outperformed all the other classifiers with a Precision of 96%, Recall of 96%, F1-score of 96% and Accuracy of 96% respectively. On the contrary BC and SVM where the least performing classifiers with Accuracy in the range $68\%-72\%$. All of our models attained highest AUROCMIAs within the range of 0.910 and 0.998 respectively. The RF, XGBoost, ANN, Bagging and KNN recording high scores and the rest a bit lower scores. Additionally, our models attained highest AUROCMAAs within the range of 0.907 and 0.996 respectively. Similarly the RF, XGBoost, ANN, Bagging and KNN recording high scores and the rest a bit lower comparable scores. However, the BC attained the lowest AUROCMIAs of 0.869 and 0.884 respectively. Table 3.5 shows the detailed performance evaluation results of our nine models using combined questionnaire responses and narratives. Additionally, Fig 4.4 shows the model validation done using AUROC.

We report on the performance validation of our nine algorithms using descriptive statistics such as the mean

Table 3.5: Comparison of nine ML models using combined narratives and questionnaire responses.

	Model evaluation					
Model name	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUROCMIA	AUROCMAA
XGBoost	96	96	96	96	0.994	0.990
RF	96	96	96	96	0.998	0.996
ANN	96	95	96	95	0.995	0.991
Bagging	93	92	93	92	0.994	0.994
KNN	91	91	91	90	0.982	0.981
DT	87	87	87	87	0.928	0.928
LR	76	76	76	76	0.985	0.973
BC	72	73	72	73	0.910	0.907
SVM	68	68	68	66	0.969	0.958

Legend: *XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbour; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier; AUROCMIA, Area Under Receiver Operating Characteristics Micro Average; AUROCMAA, Area Under Receiver Operating Characteristics Macro Average.*

and standard deviation based on the Micro and Macro averages of our AUROC reported in table 3.5. We report on 0.010282 and 0.010105 variance for AUROCMAA and AUROCMIA for our dataset, respectively. Table 3.6 shows the mean and standard deviation scores for each model throughout the 10-fold cross-validation training of the algorithms. The rank sums for each model column, depicts the Kruskal-Wallis test conducted. The test has revealed that the mean observation was not the same $(Chi = 85.383, p = 0.0001)$ across the nine models. This therefore implies that, there was a statistically significant difference in mean observation between the nine models. We also report a p-value greater than the significance level of 0.05, hence we fail to reject the null hypothesis and conclude that the nine model observations are not normally distributed. All model variances are very low or insignificant, implying that our dataset had a low degree of spread. Therefore, we can confidently state that our models were consistent in making predictions, thus even if different training data were used, they could still make a good estimate of the target variable. Additionally, we can infer that our sampled data points were very close to where our nine models predicted they would be.

The results of the McNemar tests on validating the performance of our nine models, suggest good performance on the XGBoost and RF classifiers. The pairwise tests on XGBoost and RF suggest that there is significance difference between the classifiers $(p < 0.0001)$, which is smaller than our significance threshold $(\alpha = 0.0065)$. Therefore, we reject our null hypothesis. We discovered that the XGBoost got 868 predictions right that RF got wrong. On the contrary, RF got 555 predictions correct that XGBoost got wrong. As such, based on this $1.5 : 1$ ratio, we may conclude that XGBoost performed substantially better than RF. Additionally, we performed comparative pairwise tests on all our models (LR, KNN, DT, SVM, ANN, BC, Bagging) and our best classifiers XGBoost and RF. Based on the tests, we can objectively reject our null hypothesis and and state that there is significant difference between our two best classifiers and the other seven classifiers in terms of model performance $(p < 0.0001)$ smaller than our significance threshold $(\alpha = 0.0065)$.

3.8 CCVA algorithm evaluation using CSMFs

This extract highlights how the InterVA and InSilico algorithms were evaluated using CSMFs. We also present CSMF and CCC as evaluation metrics for the InterVA algorithm.

Table 3.6: Statistical tests of our nine models.

Model scores			
Model name	Mean	Standard deviation	Rank sum
XGBoost	0.9622614	0.003209	836.00
RF	0.9566394	0.0030548	735.50
ANN	0.9530553	0.0025771	663.50
Bagging	0.9216445	2.91e+07	585.00
KNN	0.9015075	0.0033769	447.00
DT	0.8671503	0.003984	255.00
LR	0.7509405	0.0124037	155.00
BC	0.698092	0.0081906	55.00
SVM	0.6783361	0.0054433	50.00

Legend: *XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbour; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier.*

Fig 3.3 presents the twelve leading causes of death over time as determined by the InterVA algorithm using only one CoD. We observe that between the years 1993 and 2015, HIV/AIDS was the leading CoD across the population (CSMF=0.2739). This was closely followed by Pulmonary Tuberculosis (CSMF=0.1987) and Other Infectious/parasitic diseases (CSMF=0.1385). These three causes alone accounted for up to 61% of all deaths in the population during this period. The InterVA algorithm performance using CSMF accuracy and CCC, attained values of 83% and 0.36 respectively.

3.9 Trend and pattern analysis using ML and CCVA approaches

3.9.1 CCVA algorithms

This section presents mortality trend and pattern analysis using conventional CCVA algorithms based on gender (Fig 3.4A), age (Fig 3.4B) and population over time (Fig 3.4C), using data from structured questions. The visualisations are given in Fig 3.4.

We investigated the average age at death for the twelve leading causes of death. We discovered that, both males and females were more likely to die from any disease at an average age of 40 years (mean=40, median=39, IQR=36, SD=26), despite the sex. We notice more women deaths from HIV and circulatory diseases. On the contrary, we notice more male deaths from other infectious diseases, tuberculosis and external causes (see Fig 3.4A). However, these difference were not statistically significant.

Fig 3.4B depicts percentages of mortality trends across all age groups. To determine mortality across age groups, the data was grouped into five bins "0-4", "5-14", "15-49", "50-64", "65+". We significantly notice a declining trend in number of deaths among the persons aged between 0-4 years over time. In the earlier years of the Agincourt HDSS, there appears a declining trend in the number of deaths among individuals 65 years and above. However, this pattern is reversed and the mortality in the same age category is gradually increasing from the mid 2000's. Similarly, we also notice an almost comparable trend on the 50-64 age group to that of the 65+ group but the trend is gentle and stable. Among the 5-14 and 15-49 age groups, number of deaths is appearing to be constant over time.

Fig 3.4C shows mortality trends based on gender over time. A total of 16063 observations was collected, and composed of 52% male (n = 8354) and 48% female (n = 7709). We observe a gentle but steady increase in

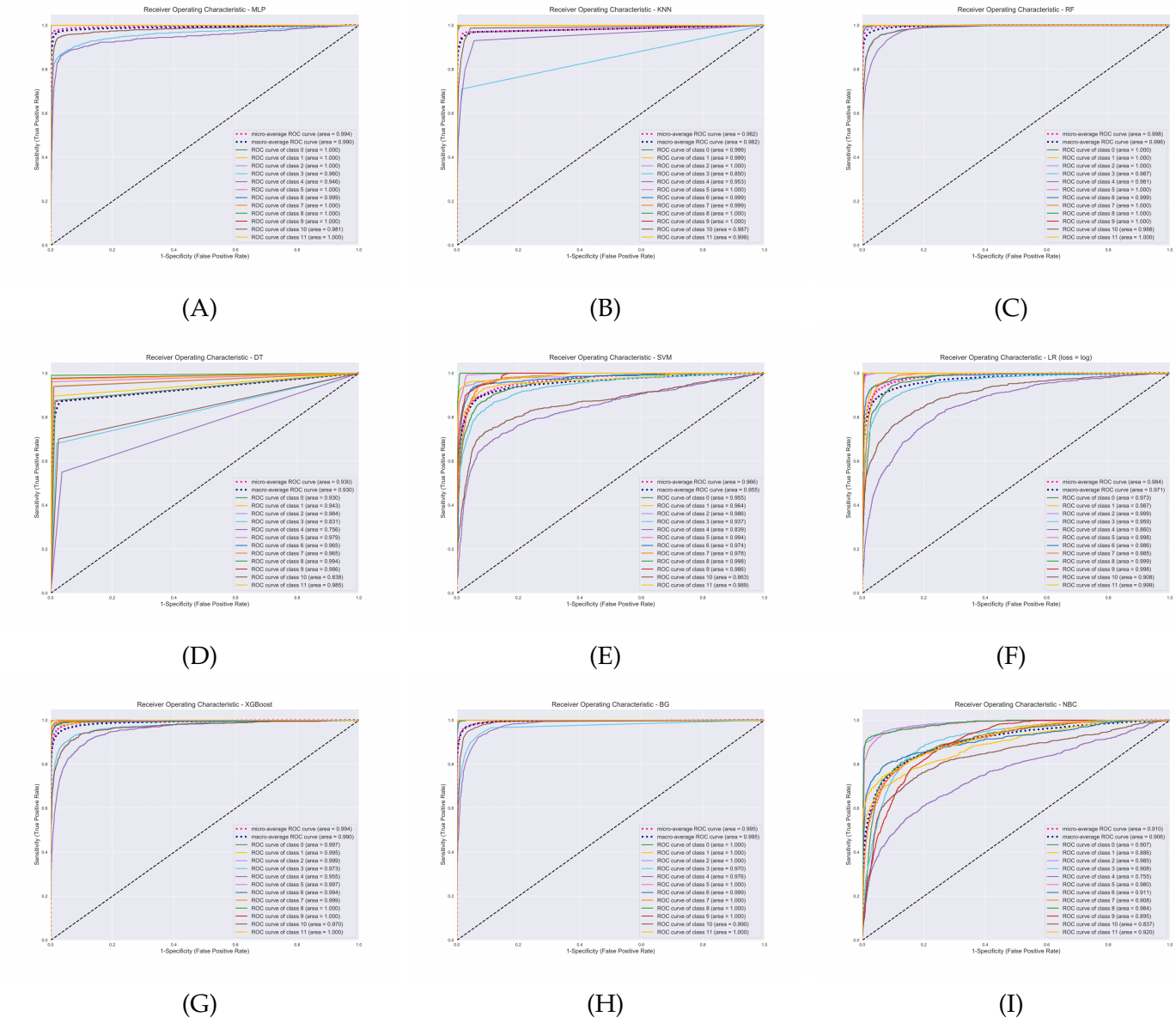


Figure 3.2: AUROC of our nine classifiers using combined questionnaire responses and narratives. **(A)** AUROC for ANN. **(B)** AUROC for KNN.**(C)** AUROC for RF. **(D)** AUROC for DT. **(E)** AUROC for SVM.**(F)** AUROC for LR. **(G)** AUROC for XGBOOST. **(H)** AUROC for BG.**(I)** AUROC for BC.

Legend: *XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbour; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier.*

mortality between the years 1993 up to 2000. This pattern rapidly accelerates among males between 2001 and 2008 before gradually declining.

3.9.2 ML techniques

In this section we present results of our trend and pattern analysis using ML approaches of mortality based on gender, age and population over time using narrative data combined with structured questions. We start by looking at the general distribution of our population based on gender, as depicted in 3.5A, age groups Fig 3.5B and 3.5C. All these graphs are depicted in Fig 3.5 . We observe that there were more male deaths than female deaths. Most of the deaths were within the 15-49 and 65+ age groups.

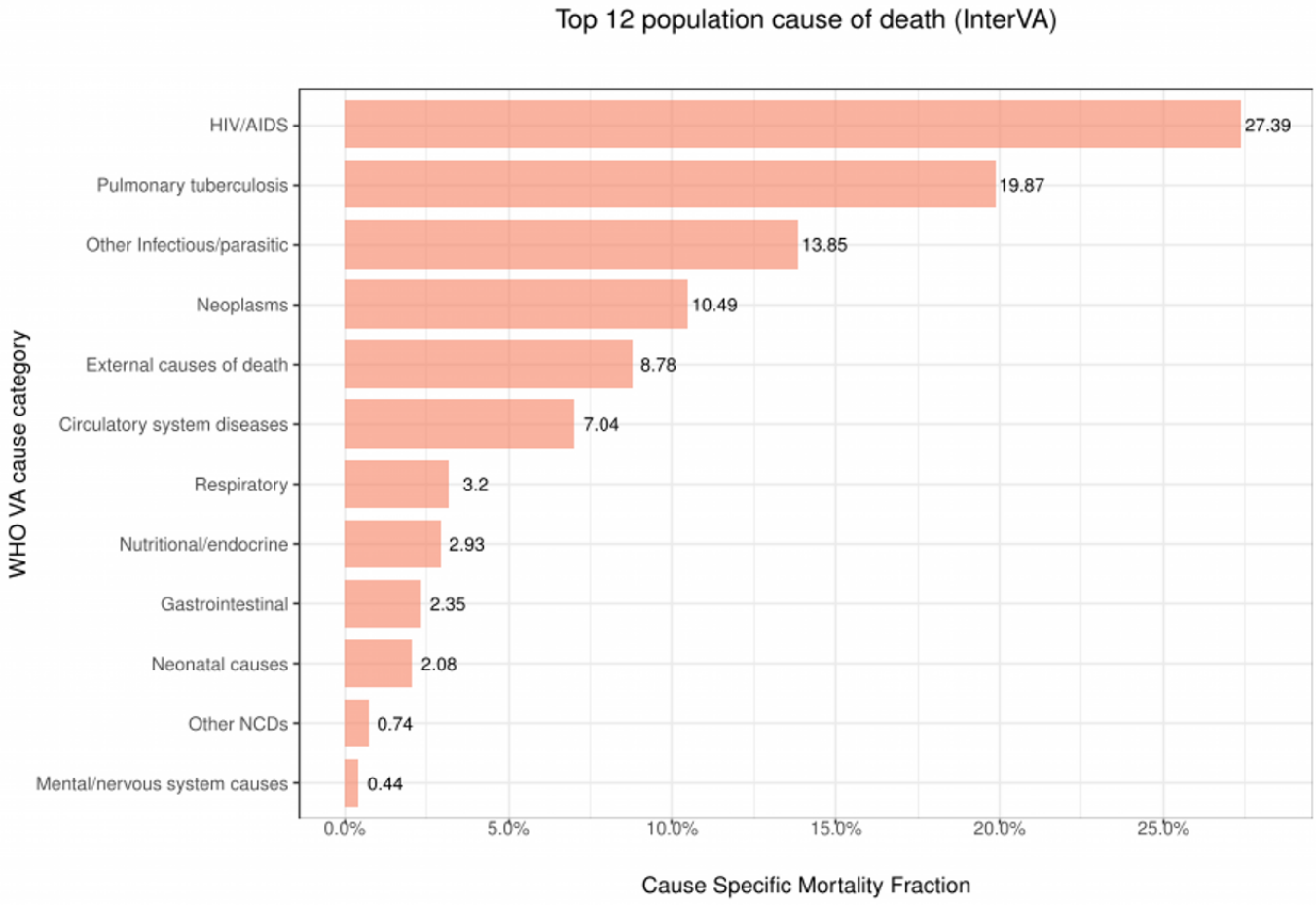
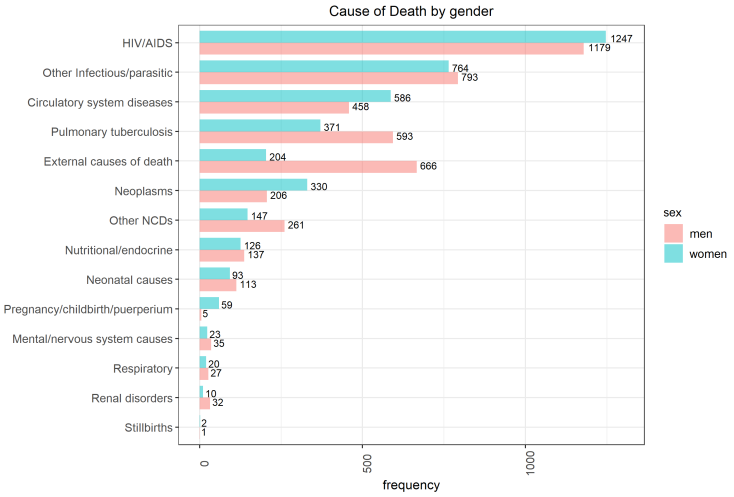
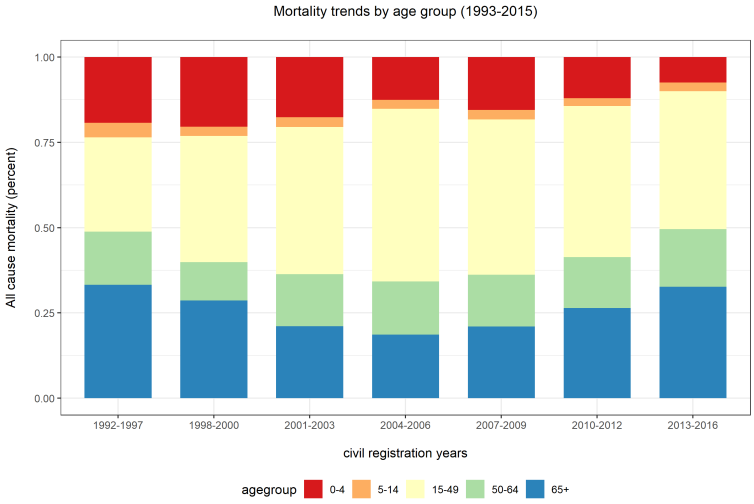


Figure 3.3: Top twelve CoD diseases



(A)

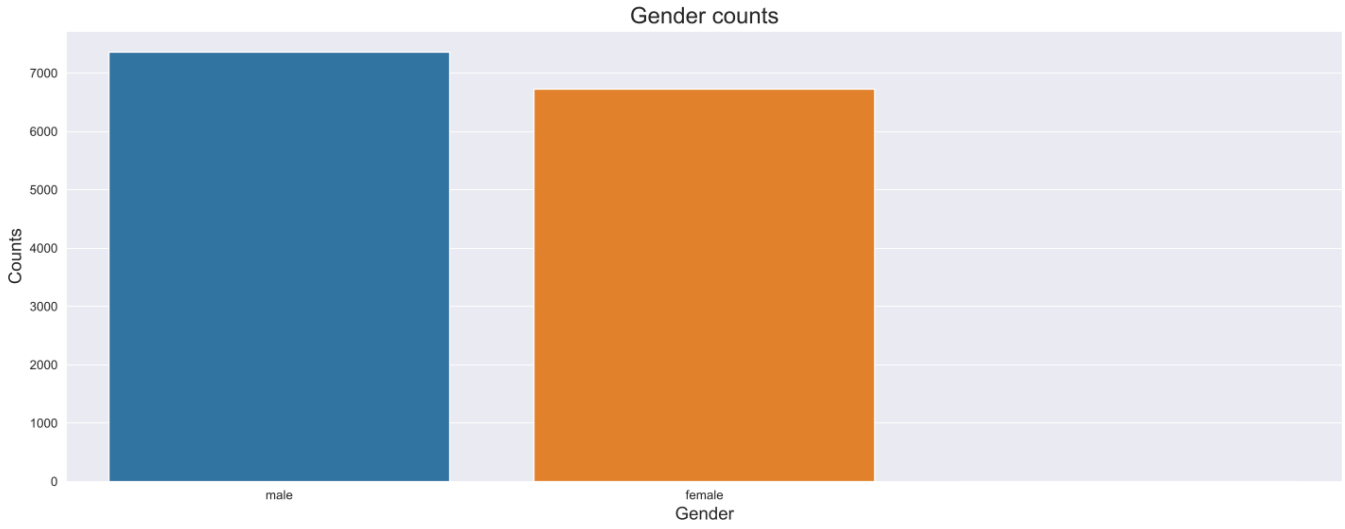


(B)

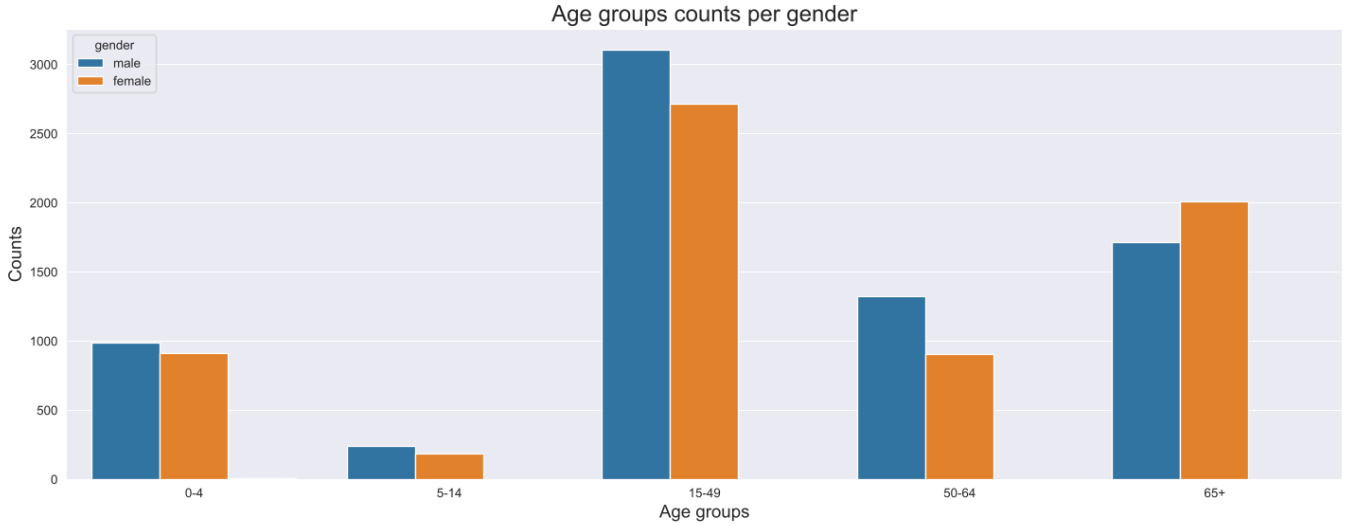


(C)

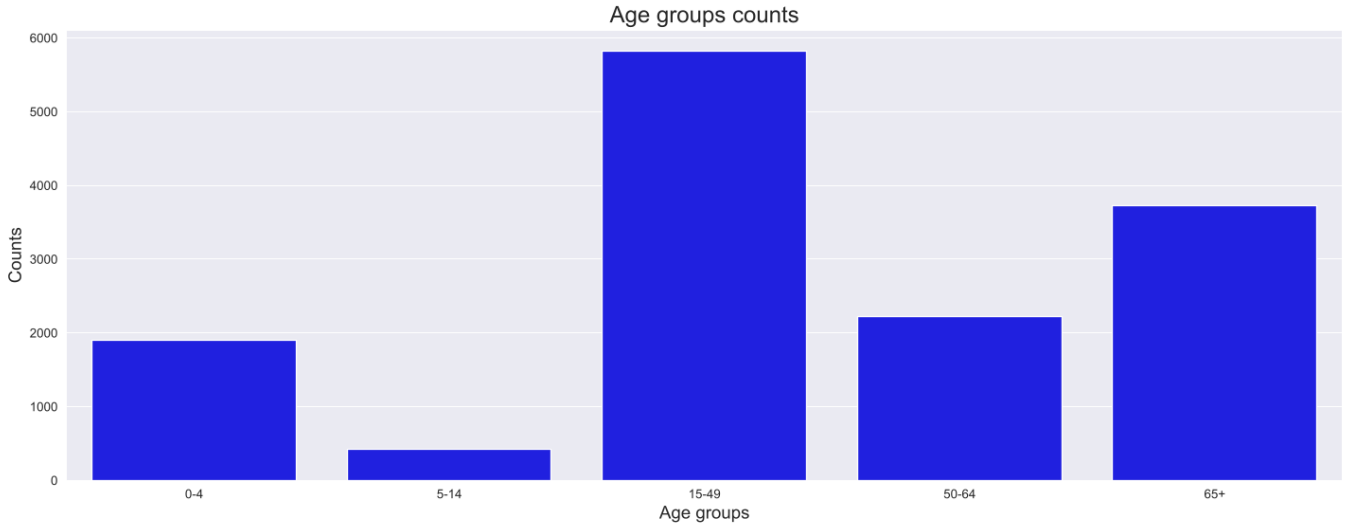
Figure 3.4: CCVA mortality trends based on age, population and gender. **(A)** Cause of death by sex. **(B)** Percentage of deaths by age group.**(C)** Yearly mortality trends by gender.



(A)



(B)



(C)

Figure 3.5: Gender and age group counts graphs. **(A)** Gender count. **(B)** Age group count per gender. **(C)** Age group count.

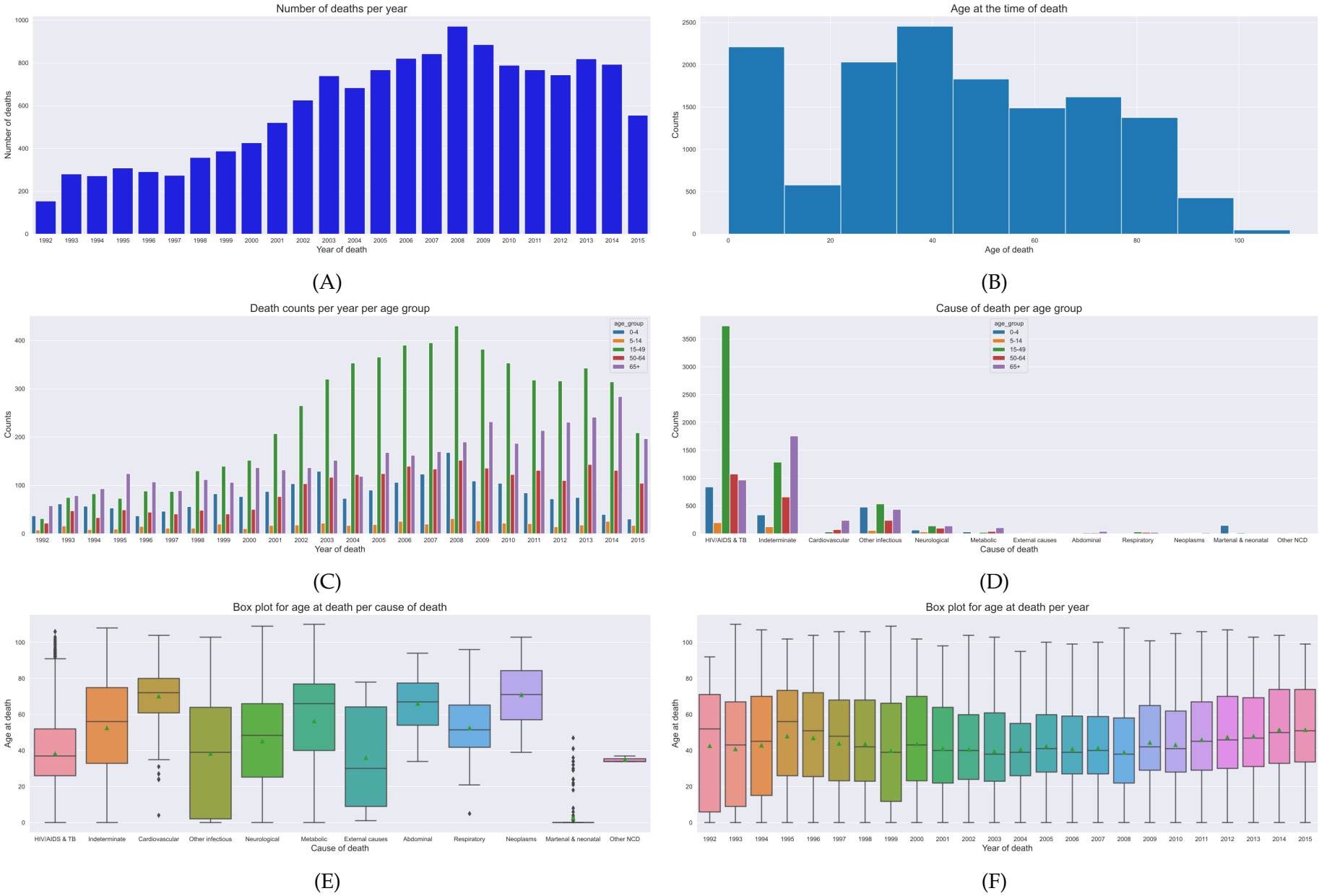
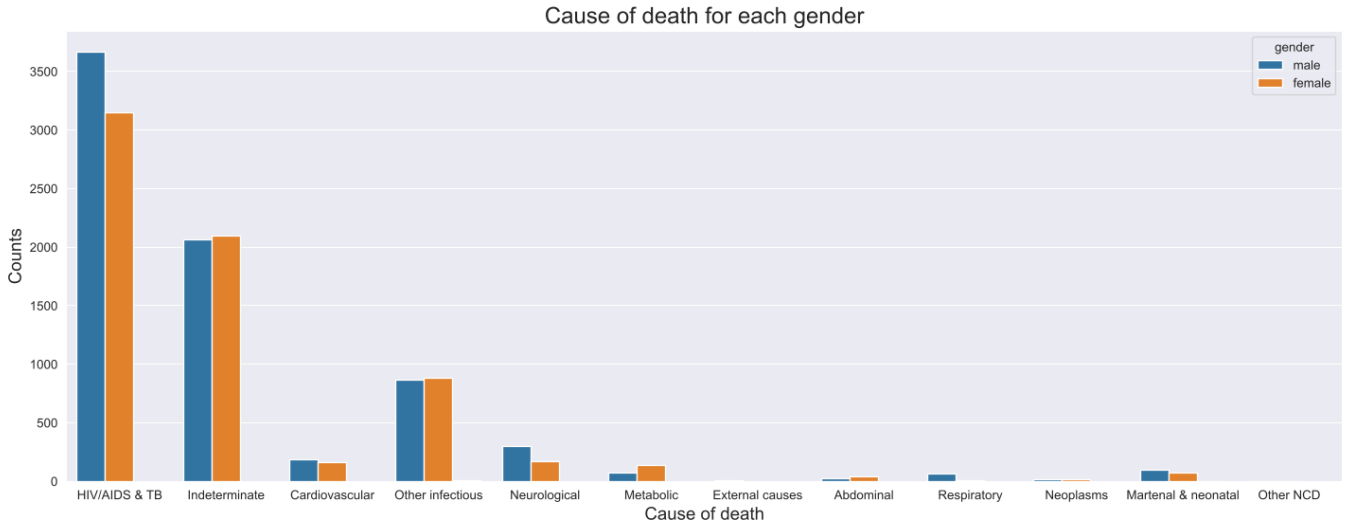
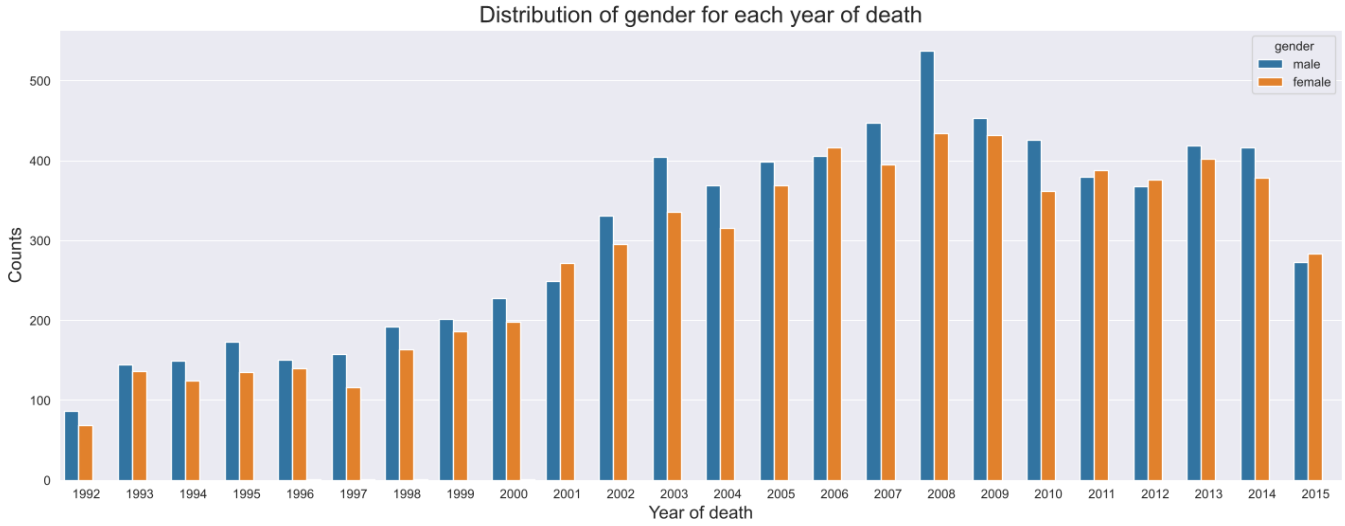


Figure 3.6: Mortality trends across age groups. **(A)** Number of deaths over time. **(B)** Age at death count. **(C)** Yearly death count across age groups. **(D)** Age group CoD count. **(E)** CoD and age death. **(F)** Age at death per year.



(A)



(B)

Figure 3.7: Yearly CoD based on gender.**(A)** CoD based on gender. **(B)** Yearly CoD based on gender.

We analysed our mortality trend and pattern based on age groups as in Fig 3.5. We observe that most deaths are within the 15-49 and 65+ age groups. The 65+ age group had more deaths recorded in the 1990's with a gradual increase till 2014. We also discovered that the 15-49 age group trend sharply increases till 2008 and then steadily goes down till 2015. We notice a constant trend for the 5-14 year age group over time. There is a high number of death from HIV causes affecting mostly the 15-49 age group. We also notice that most deaths appear to be common in the 0-10 year age group and 30-80 years age groups. Conversely we notice fewer deaths for 80+ years.

Fig 3.6E and 3.6F depict our boxplots on CoD and age at death over time and age at death per year. On average, the population died of HIV/AIDS or tuberculosis which where the leading CoD at a median age of 38 years. The plots depict an average death age of 66 years succumbed from cardiovascular, neoplasm, metabolic or abdominal diseases. Worth taking note of is the death from other infectious disease causes that show a dissimilar trend across all age groups. Additionally, on average, most of the cases died of metabolic causes at an elderly age of 65 years. Other NCDs causes of death were more prevalent in the 30-35 year old age group and neonatal and maternal causes in their first year (shown by the narrow IQRs). CoD from

neurological and respiratory causes show a mortality trend and pattern that illustrates an average age at death of 50 years. We observe that there were more male deaths than the females, despite the cause. There is a gradual up-trend from 1992 (less than 100 deaths) to 2008 (almost 500 deaths) and a steady decline in number of deaths from 2009 (see Fig 3.7A and 3.7B). Fig 3.6D illustrates that between the years 1993 and 1997, the average life expectancy was approximately 50 years. However, from 1998-2010, the life expectancy significantly dropped and population was dying at a much younger age of 40 years on average. From the year 2011, we see a gentle improvement in the life expectancy.

3.9.3 Analysis of contradicting cases

The extract details analysis of structure and semantics of features in cases where the doctors are in disagreement. We discovered that approximately 16% of the observations in the VA dataset denote cases where the experts are in disagreement. Further analysis of the structure and semantics availed insights that in most cases the narratives entail information related to traditional healers visits and consultations. Additionally, we deduced that also this is a result of cases where the captured information entails imminent loss of weight, vomiting and having a fever leading to a unexplainable sudden death. Fig 3.8 below shows our n-gram model of what was mined from the contradicting narratives.

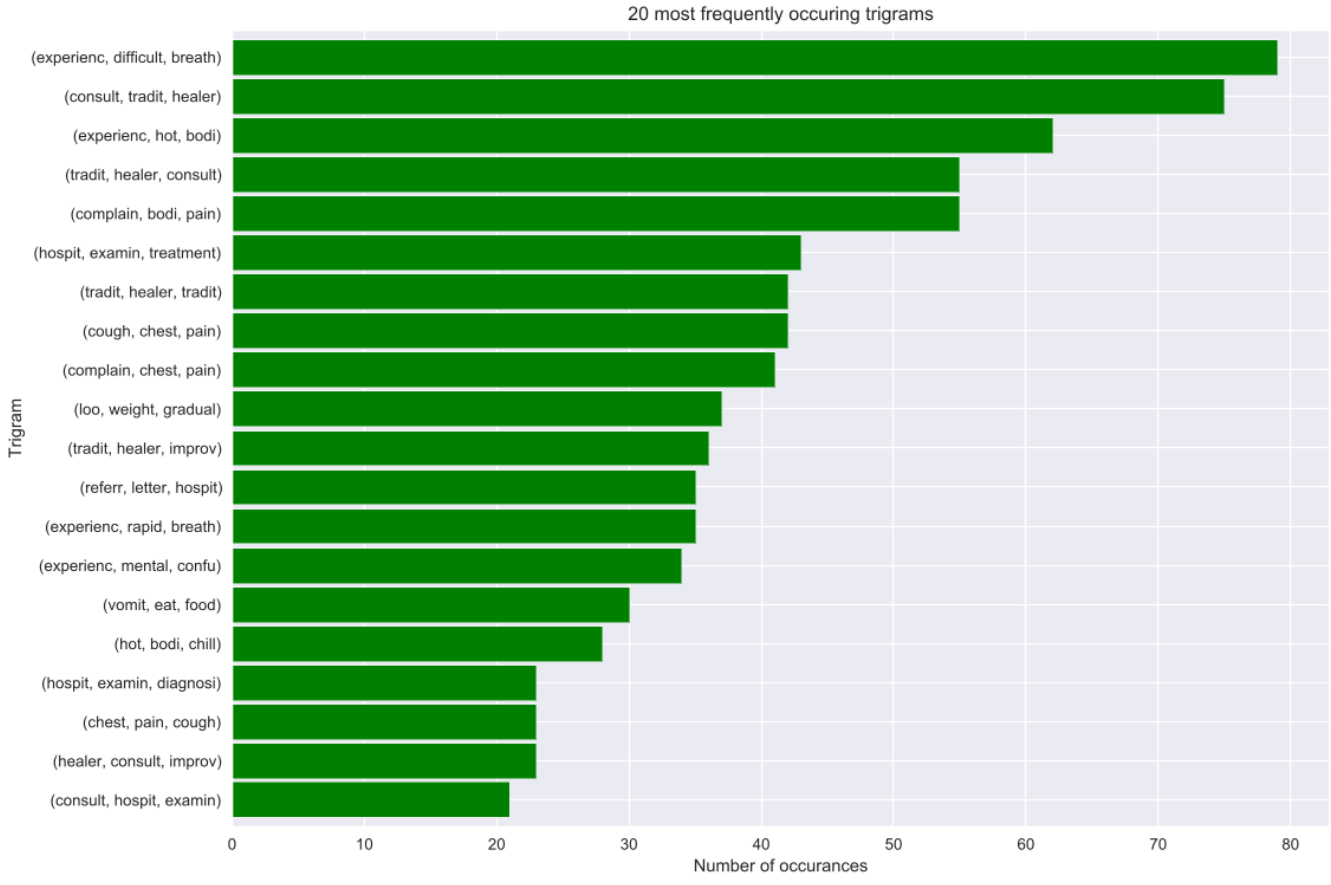


Figure 3.8: Tri-gram model showing frequently occurring contradicting cases.

3.9.4 Model best predictors

This part discusses how the most important narrative features that where identified and are the best predictors of our models. We chose the bi-grams as they show an evenly distributed frequency analysis of features (see Fig 3.9). As n increases the features start having more or less the same frequency.

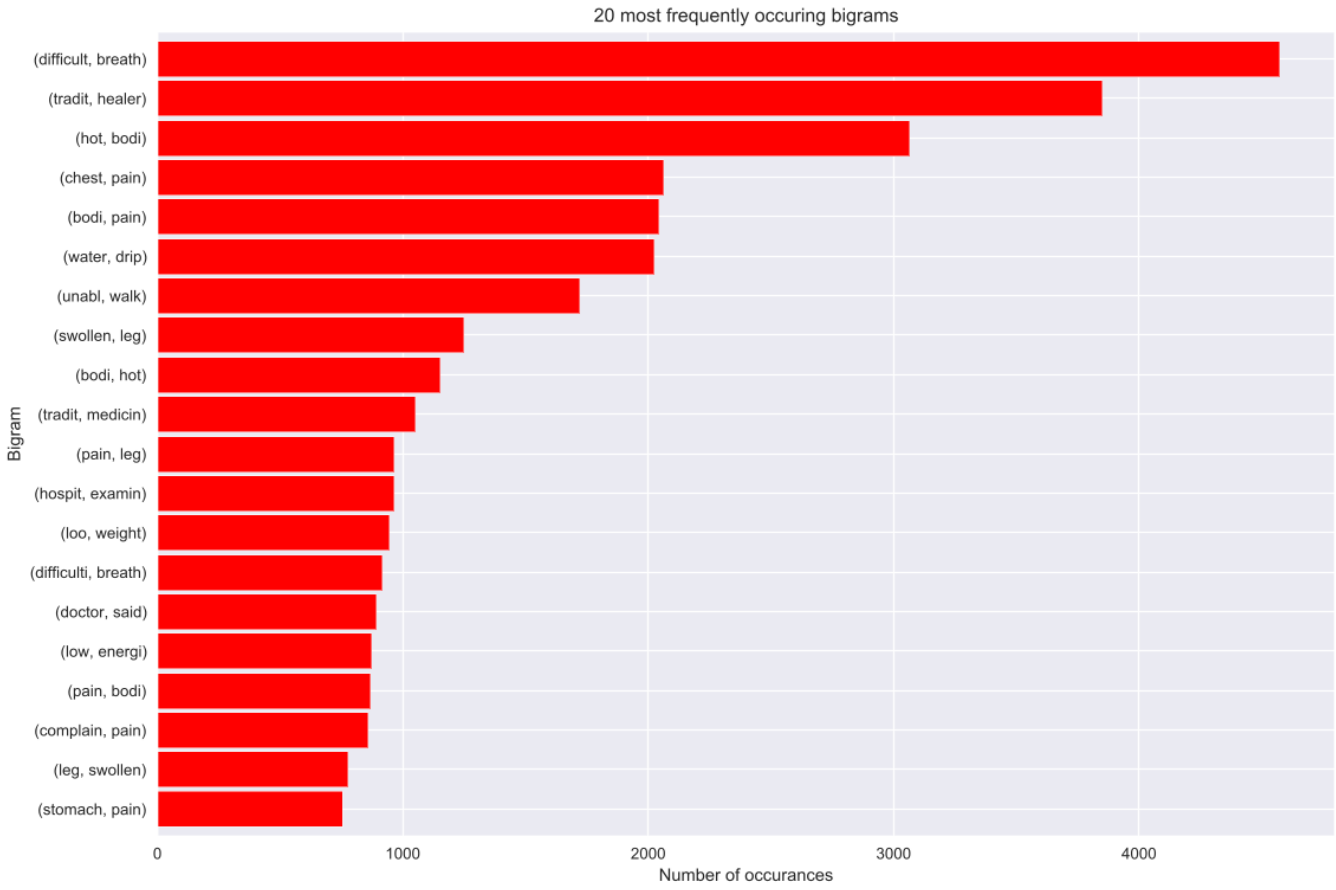


Figure 3.9: Bi-gram model of our best model predictors

3.10 Discussion

The process of determining causes of death using VAs still remains a manual task and suffers from many drawbacks. This negatively affects the VA reporting process, despite it being vital for strengthening health priorities and informing civil registration systems. Therefore, under such circumstances, there is a great need for innovative novel automated approaches to address these problems thereof.

In this study we explore with various VA data types, despite most studies in literature reporting results based on the classical dataset for CoD determination using ML approaches. Our aim being to investigate if the narratives can improve or enhance model prediction if added to the responses from the structured questionnaire. Our deductions suggest that the VA narratives have vital valuable information that should be used in model prediction. Consequently, we identify the best model predictors from the narratives. We further do a mortality pattern and trend analysis based on age, population and gender over time. We also do a structure and semantic analysis of narratives in cases where the experts agree and also disagree. To add to our findings we also investigate the best features that contribute to our models from the narratives.

Generally, the results of all our ML models used in this study, demonstrate that our models exhibited consistent superior performance on all datasets. This further reinforces the notion that, ML approaches can be used as alternatives for conventional approaches for CoD determination using VAs. Ensemble classifiers (XGBoost, bagging), tree based models (DT, RF), ANN and KNN performed exceptionally well on all datasets. Our results of the combined dataset do not exhibit a consistent model performance, as most models slightly drop in model performance. This can be attributed to the fact that the combined dataset creates high dimensionality of the feature space and this triggers model complexity with too many noisy data points. The CCVA approach,

InterVA, attained a CSFM accuracy of of 83% and CCC of 36%.

Our CCVA approaches and ML techniques produced similar mortality trends and patterns based on age, population and gender. Interestingly, we observed that in the first decade of the civil registration system, the average life expectancy was approximately 50 years. However, in the second decade, the life expectancy significantly dropped and population was dying at a much younger average age of 40 years. This suggests CoD mostly from the leading HIV and tuberculosis related causes. Interestingly, in the third decade, we see a gradual improvement in the life expectancy, possibly attributed to implementation of effective health intervention programmes. We notice that, cardiovascular, neoplasm, neurological, respiratory, and metabolic CoD mainly affected the elderly. We observe that other infectious disease and external causes affected the population disproportionately across all age groups, with the latter having an average age at death of 30 years. Despite the expected CoD from neonatal and maternal causes, we can also infer that those with HIV had a lower life expectancy as compared to the other CoDs. Of interest, is that most undetermined causes of deaths are found within the 65+ age groups. This suggests that as the elderly population grow older their health state deteriorates and they succumb to many symptoms that can lead to untimely hard to explain deaths. Other NCDs causes of death were more prevalent the younger age groups. We also discovered that, sudden deaths are common in the elderly, suggesting symptoms, such as imminent loss of weight, vomiting and having a fever leading to an unexplainable premature death. Generally we notice more male deaths than females.

We therefore propose that optimal model performance should be set at 80% accuracy. This threshold for optimal model performance was selected based on the specific requirements and context of our study. We recognize that the ideal performance threshold can vary depending on the application, and therefore, it should be determined on a case-by-case basis. Therefore, different thresholds may be more suitable for other scenarios depending on their unique objectives and constraints. In cases where the ML model fails to reach a threshold value of 80% accuracy in terms of performance, we propose expert's intervention for further exploration and assessment. Conversely, in cases where the experts are failing or do not reach a consensus we recommend the help of the machine to make predictions. Most of these cases where the machine can assist, entail narratives where the interviewee details most content about the deceased circumstances and events that led to death based on traditional healer visits and consultation. Interestingly, we still found out that traditional healer consultations are a common practice in the population as they occurred frequently in our model best predictors. This cements the notion that, most people in the HDSS seek traditional ways for their terminal illnesses, rather than western means. This finding opened exciting avenues for future work, which will focus on sequential text modelling with the aim of fully understanding treatment sequences for terminal illnesses. Nevertheless, in cases where the physicians were in agreement, these narrations about traditional healer's consultations where supplemented by enough symptoms that made it possible for the experts to give a proper diagnosis. We also discovered that our model best predictors entail matching symptoms with those in the responses to the structured questionnaire.

The results of this study, consistent with a number of studies that used VA data to determine CoD, suggest that ML approaches can accurately classify CoD from VA narratives. However, in most cases statistical approaches and CCVA approaches are always outperformed by ML approaches [1, 44, 18, 15, 8, 17, 69]. Therefore, it is imperative for future research studies to incorporate effective data handling strategies [8]. This study adds to the existing body of literature, suggesting that automated approaches can be used as alternatives to PCVA in a cost effective way, producing real-time results that are consistent, accurate and error free, thus strengthening health priorities. As such, VA processes are still key in capturing civil registration data where death occurs outside health facilities, up until a point when deaths start to take place in areas where it can be documented.

Given these complexities, there is a great need for novel automated approaches that can be used as alternatives [97].

The strength of this study lies in the application of various ML and CCVA algorithms on various VA data types. Moreover, our sample size was large and representative of deaths that occurred at Agincourt HDSS that were captured in a standard way. Moreover, our mortality trend and pattern analysis gave us valuable insights into our HDSS and this can be used to inform policy and practice. This enforces generalisation and comparability across studies. On the contrary, this study had limitations of data quality described in section 3.2.2.

3.11 Conclusion

In general this study demonstrates that ML techniques can be used as alternatives in determining CoD from VA narratives producing results comparable to physician diagnosis. Our findings should be used to inform policy and practice and enforce effective health interventions programmes and resource prioritisation to reduce the mortality rate and prolong the life expectancy. As such, they can help close the gap in civil registration systems. Our comparative analysis of the ML models on various VA datasets enforce comparability and generalisation, thus availing a baseline study for future research. Future work will entail exploring with deep learning methods and employ novel techniques like transfer learning to determine CoD.

3.12 Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

3.13 Author Contributions

All authors contributed to the study conception and design. MTM, VO did all algorithm experiments with the help of TC. MTM and all the other authors drafted and critically revised the work. The first draft of the manuscript was written by MTM and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

3.14 Funding

No funding was received for conducting this study. The authors have no financial or proprietary interests in any material discussed in this article.

3.15 Acknowledgments

This work was supported by the Developing Excellence in Leadership, Training and Science (DELTAS) Africa Initiative Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) [Grant No.DEL-15-005]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS) Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [Grant No. 107754/Z/15/Z] and the United Kingdom government.

3.16 Data Availability Statement

The datasets [GENERATED/ANALYZED] for this study can be made available upon request by contacting the corresponding author.

4 Text mining of verbal autopsy narratives to extract mortality causes and most prevalent diseases using Natural Language Processing

This study is in review in PlosOne journal
Michael T. Mapundu[1]*, Chodziwadziwa W. Kabudula[1,2], Eustasius Musenge[1], Victor Olago [3], Turgay Celik[4,5].

[1]Division of Epidemiology and Biostatistics, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Parktown, Johannesburg, South Africa
[2]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), University of The Witwatersrand, Johannesburg, South Africa
[3]National Health Laboratory Service (NHLS), National Cancer Registry, Johannesburg, South Africa
[4]Wits Institute of Data Science, University of The Witwatersrand, Johannesburg, South Africa
[5]School of Electrical and Information Engineering, University of The Witwatersrand, Johannesburg, South Africa

4.1 Abstract

Verbal autopsy (VA) narratives play a crucial role in understanding and documenting the causes of mortality, especially in regions lacking robust medical infrastructure. In this study, we propose a comprehensive approach to extract mortality causes and identify prevalent diseases from VA narratives utilizing advanced text mining techniques, so as to better understand the underlying health issues leading to mortality. Our methodology integrates n-gram-based language processing, Latent Dirichlet Allocation (LDA), and BERTopic, offering a multi-faceted analysis to enhance the accuracy and depth of information extraction. This is a retrospective study that uses secondary data analysis. We used data from the Agincourt Health and Demographic Surveillance Site (HDSS), which had 16338 observations collected between 1993 and 2015. Our text mining steps entailed data acquisition, pre-processing, feature extraction, topic segmentation, and discovered knowledge. The results suggest that the HDSS population may have died from mortality causes such as *vomiting, chest/stomach pain, fever, coughing, loss of weight, low energy, headache*. Additionally, we discovered that the most prevalent diseases entailed human immunodeficiency virus (HIV), tuberculosis (TB), diarrhoea, cancer, neurological disorders, malaria, diabetes, high blood pressure, chronic ailments (kidney, heart, lung, liver), maternal and accident related deaths. This study is relevant in that it avails valuable insights regarding mortality causes and most prevalent diseases using novel text mining approaches. These results can be integrated in the diagnosis

pipeline for ease of human annotation and interpretation. As such, this will help with effective informed intervention programmes that can improve primary health care systems and chronic based delivery, thus increasing life expectancy.

Keyword: Verbal autopsy, Machine learning, Algorithms, Natural Language Processing, Text mining

4.2 Introduction

The United Nations expects all countries in the world to meet their 90 percent death registration coverage requirement. However, this is not the case as most deaths that occur in Low to Medium Income Countries (LMICs) are not captured in civil registration systems [2, 3]. As such, this accounts for approximately 65 percent of the world population that do not have a medically certified cause of death (CoD) on a yearly basis [1]. Nevertheless, the CoD information is important as it is used to improve civil registration systems, public health monitoring, critical health policies and priorities [4]. In most instances were CoD cannot be derived from clinically based sources, the VA process is used as a substitute tool. The VA process is conducted by non-medical staff who seek to elicit information from the next of kin of the deceased, regarding circumstances and events that may have led to death [1]. This information is then captured and stored in textual format and is also known as VA narratives. These narratives contain rich valuable information that is used to supplement responses from the standard World Health Organisation CoD questionnaire by physicians, when giving a CoD diagnosis [1]. However, in most cases this information is encapsulated in the textual format, usually unstructured and is stored in large volumes. Consequently, it is difficult to manage, takes time to process and suffers from word polysemy, amongst many challenges. This therefore makes the analysis and harnessing of meaningful information difficult, even though it is vital in identifying mortality causes.

As such, the application of text mining (TM) which employs natural language processing (NLP) in an automated fashion can be beneficial to extract meaningful information regarding mortality causes and most prevalent diseases from VA narratives. There is scanty literature on the application of unsupervised TM approaches within the VA domain. Only a few TM studies that employ supervised learning, have been applied to date within the VA domain to determine CoD [8, 12, 108]. Mujtaba et al. [8] argues that the TM approaches used for cause of death classification, require large datasets which are labelled for training the machine learning models. This makes the process tedious, time consuming, expensive and prone to human error as the experts are the ones responsible for labelling the datasets [8, 7].

This study seeks to address this gap by reducing the need for human intervention, by implementing unsupervised TM approaches. The implementation of TM approaches in VA to determine mortality causes and identify the most prevalent disease has several advantages as reported in [8, 12]. Moreover, several studies within the healthcare domain also appraise on the benefits of applying TM as cited in [109, 110, 111, 112, 113, 114, 115, 116].

Firstly, TM can help in efficient analysis, through extracting vital information from large volumes of unstructured narratives in a quick, cheaper and accurate way [8]. Ultimately, this reduces time and effort required for manual analysis. Moreover, the TM techniques exhibit the ability to transform unstructured data into actionable insights that can help improve health outcomes, inform public health policies and interventions. The TM mathematical approaches are reported elsewhere [100]. Secondly, implementing TM techniques can result in improved accuracy as one can identify patterns and trends in the data that might be missed by human analysts, leading to more accurate and reliable results [112, 115]. Thirdly, there is consistency as one can categorise similar cases together, ensuring a consistent way of CoD categorisation across different VA records.

Additionally, there is some form of standardisation on the CoD categorisation, that ensures the application of the same criteria across different VA records. Interestingly, TM approaches can extract additional variables that can be useful in the CoD diagnosis [112]. Moreover, the semantic and structure analysis through sequential modelling, can even give physicians better insights on the most probable CoD as it entails sequences of events that might have led one to succumb to death [117]. As such, this might help in cases where there are fewer symptoms or unknown CoD to help one to get to a CoD diagnosis.

In this study we present novel TM techniques such as n-gram natural language processing, LDA and BERTopic as we seek to investigate mortality causes and the most prevalent diseases that led to death in rural north-east South Africa. The TM approaches will assist in automating the process of analysing VA data, in a fast and accurate manner. The findings of this study can be integrated in the diagnosis pipeline for ease of human annotation and interpretation. By incorporating modelling frameworks, researchers can potentially improve the granularity and precision of identified topics, leading to a more nuanced understanding of the underlying themes present in large VA text corpora. Ultimately, the discovered knowledge from the unstructured narratives will enable the design, implementation and sustainment of tailored health interventions programmes rather than adhoc generalizations. This will strengthen health priorities, improve life expectancy, VA reporting and decision making, thus informing policy and practice.

4.3 Literature review

4.3.1 Related work

There is scanty literature that reports on the application of TM techniques specifically in the VA domain. TM is an approach that seeks to close the gap between textual information and representation of the text in a structured fashion [115]. Leskovec et al. [101] argues that the mining of data entails use of very robust hardware, programming languages that employ efficient algorithms, and strive to solve various problems in domain specific areas. TM employs various techniques that span from NLP, data mining, machine learning (ML) and management of knowledge. TM can be used for information retrieval, information extraction, classification of documents and named entity recognition (NER), amongst many applications [114].

One notable study in the VA domain that has applied TM is that of Mujtaba et al. [8], who did a systematic literature review on clinical text classification trends using 72 articles. They point out that the free-text clinical report is very valuable for classification problems. Nonetheless, one can only derive meaningful information from such reports after effective data transformation, to generate useful trends and patterns. Similar work on text classification is reported in [73, 74, 72, 78].

Despite the little research that reports on application of TM methods in the VA domain, these approaches have been employed extensively in other health fields as reported in [118, 119, 120, 121, 121, 113, 122, 123]. These previous studies report that TM techniques make it possible to extract meaningful information from implicit information. As such, there is great need to use novel TM tools to extract and discover valuable information. Consequently, TM techniques, can be used as alternatives to manual processes. However, they raise issues of noise and redundancy that affect the quality of data. Moreover, in most cases unstructured textual data exhibits high levels of noise, sparsity, makes use of varying vocabulary terms, misspelled terms with many grammatical errors and entails native terms [8, 121, 124].

Recent developments in technology have triggered an increase in the application of electronic public health repositories, making most textual reports readily available for additional use, specifically in biomedicine to extract useful information [8, 123]. However, despite these availed opportunities, there is little to none

regarding VA literature that has applied these novel TM techniques for knowledge discovery. On the contrary, VA narratives entail rich valuable information, which can be utilised through NLP, and can be integrated in intelligent models to improve public health decision making processes [115, 1].

4.3.2 N-gram language processing

Literature reports on n-gram language processing being a feature extraction approach, that aims at attaining only relevant and useful features from textual data. The n-grams are a set of words that are sequential as they make use of continuous number of items such as characters or words from a given sequence of narratives [8, 15]. Another study by Danso et al. [12] reports on a comparative study of machine learning methods for VA text classification and they deduce that the bag of words using n-grams can result in better model performance with high accuracies.

Lucini et al. [114] did a study using TM techniques as they sought to predict hospital admissions using early medical records from the emergency department. They report that these tools make it possible to elicit valuable information, thereby improving clinical decision making processes. Similar findings are also reported in Kim and Chung [116], who conducted a study on information extraction from big health data and they suggest that word clouds can be effective visualisation tools that can be able to depict keywords or concepts in documents. Furthermore, they argue that, the more a term appears in the corpus repeatedly, the more important it is. Similar work which emphasises the importance of n-grams as a TM approach in gaining valuable insights is alluded in the studies of [125, 126, 127]. However, they point out the challenges of having too many insignificant words in the corpus.

4.3.3 Topic modelling

Topic modelling using LDA

Topic modelling entails using probabilistic algorithms to extract useful and common topics from a set of documents (in our case we aim to extract most prevalent diseases). Topic modelling is an unsupervised learning approach based on statistical principles, that seeks extract topics from a set of documents stored in given corpus. A single topic is made up of a collection of words, with a corresponding weight. A single word can belong to various topics and documents can be made up of several topics [128].

One commonly used topic segmentation method is the LDA, which has been applied in domains such as classification of genomic sequences, social networks, modelling of health data, extracting latent topics from clinical reports, amongst many uses. This approach has been reported to attain high precision and performance scores, especially when used with clinical reports [128, 129, 130, 110, 131]. LDA is generative in nature and can easily pick out hidden groups in data, using patterns of co-occurrences. Such patterns cannot easily be detected by conventional approaches, and as researchers, we may not even be aware of these associations in the data [125, 132]. Noteworthy, is that the LDA approach often struggles to capture the intricate nuances of language and context, and its implementation requires some programming expertise [133, 110].

Topic modelling using BERTopic

BERT, being a contextualized language model, takes into account the surrounding words when representing each word in a sentence. This bidirectional understanding enables BERT to create more accurate and contextually rich word embeddings. In the context of topic modelling, these embeddings can be utilized to derive topic representations that better reflect the semantic relationships between words [133].

The work of Scarpino et al. [134] supports this notion, as they report that BERTopic outperforms LDA approach in clustering in their study that investigated topic modelling techniques to extract meaningful insights in Italian long COVID narrations. The study of Silva et al. [135], reports on the successful application of BERTopic approach as an unsupervised natural language processing in the identification of patients with suspected COVID-19 infection. Similar successful application of BERTopic as a topic modelling approach in various health spaces is reported in [136, 137, 138, 29, 139, 140, 141]. As such, the application of BERT in topic modelling holds the promise of enhancing the sophistication and contextual relevance of topic extraction, thereby contributing to more accurate and meaningful insights derived from textual data.

The text mining process entailed the following steps; data acquisition, pre-processing, feature extraction, topic segmentation, and discovered knowledge. These techniques are reported elsewhere [112, 118, 111, 115, 130, 129, 142, 131].

4.4 Materials and methods

In this study, we apply TM techniques to extract mortality causes and identify the most prevalent diseases using VA data. This is a retrospective cross-sectional study that uses secondary data analysis through TM methods. The VA data was obtained from the study area of the Agincourt Health and Demographic Surveillance System (HDSS). This study has been approved by the relevant ethics committee in South Africa (University of the Witwatersrand Faculty of Health Sciences, Human Research Ethics Committee (Medical), and (approval ref. no: M1911132). It was conducted in accordance to the Declaration of Helsinki. Informed participant consent was waived by University of the Witwatersrand Faculty of Health Sciences, Human Research Ethics Committee (Medical). We got permission from the Agincourt Health and Demographic Surveillance Site to use their data for research purposes from 29 November 2019 to 29 November 2024. The inclusion criteria was all VA autopsy data for the period between 1993 and 2015.

The HDSS was established in 1992 and is situated in the rural Sub-district of Bushbuckridge under Ehlanzeni District, in Mpumalanga Province, in north-eastern South Africa. The geographical coverage is approximately 420km^2. As of 2019, the population was at 116 247 individuals residing in 28 villages with 22 716 households, with males being 55 961, females being 60 280, children under 5 years being 11 724 and school going children with ages from $5-19$ being 35 928 [98].

The Agincourt HDSS is a surveillance site that serves as an evidence based health monitoring site that seeks to strengthens health priorities, practice and inform policy. The data was collected between 1993 and 2015. For this study, we used only textual data from VA narratives, denoting one feature with 16,338 cases/observations, all of which were in English.. However, for our semantic and structure analysis we compare our extracted symptoms from the VA narratives to that of the responses from the structured questionnaire. The responses from the questionnaire dataset entailed 227 symptoms. All symptoms that had a 'Y' were encoded as a 1 meaning that the record had a present symptom value. On the other hand, all features that had a 'N', were encoded as a 0 meaning that those records had no symptoms present. All our features were in English. This comparative analysis was done to investigate if mined narratives exhibit similar and/or have additional features.

In this study, we first do data acquisition of our VA narratives as a comma separated value text file (csv), followed by preprocessing, then keyword extraction using n-gram natural language processing, followed by topic modelling using LDA and BERTopic and lastly identifying the relevant topics through knowledge discovery. We follow the TM steps as depicted in Fig 4.1 below.

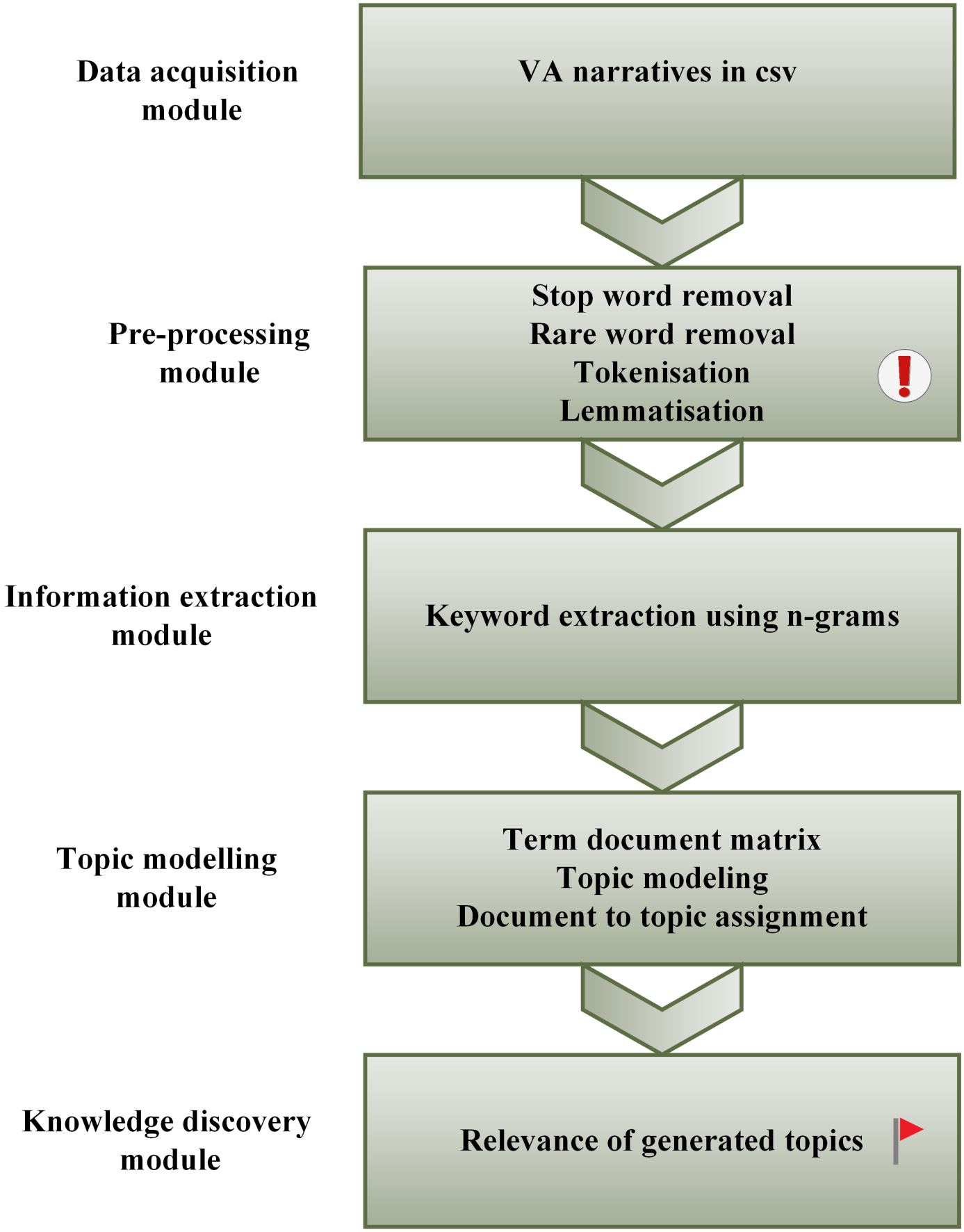


Figure 4.1: Text mining steps followed in this study

4.4.1 Data pre-processing

We performed pre-processing as part of data cleaning and normalisation, as we sought to do away with irrelevant data. In most cases input of textual data requires pre-processing and transformation before being mined [143]. After data acquisition, in csv format, we started by cleaning the unstructured narrative data by converting all text to lowercase, removed all punctuation, spaces, numbers and special characters. We applied stopword removal in order to remove insignificant words using the NLTK library of English stopwords. We performed tokenisation which is breaking up of text into tokens, of words or sentences. Lemmatization was applied by employing the Python Spacy package. This process uses a dictionary of known word forms and considers the role of a word in a sentence with the aim of extracting some normal form of a word (See Fig 4.1, Pre-processing module).

4.4.2 Information extraction

After the pre-processing phase, we now had our tokens as words, and at this stage features could now be defined. The features were extracted as tokens of words, using n-gram natural language processing. The n-grams were seen as tokens of adjacent words denoted as; $n = 1$ a unigram such as, *'pain'*, $n = 2$ a bigram, such as *'chest pain'*, $n = 3$ a trigram, such as *'high blood pressure'* and $n = 4$ a quodgram, such as *'blood result hiv positive'* [15]. The bigger the n-gram word frequencies, the more contextual information that we extracted. These n-grams models gave us an understanding of our narrative corpus and denoted the frequently occurring terms, also known as symptoms or mortality causes (See Fig 4.1, Information extraction module).

Additionally, we also performed keyword extraction defined as a representation of a document's content using a word or a sequence of words. These words provide meaning within a sentence's context. As such, one can discover hidden patterns and correlations using TM approaches in an automated fashion from large volumes of unstructured data using keywords [108]. Recent novelties in TM, now allow the extraction of keywords from documents without the assistance of experts, a process known as expert driven.

Keyword extraction is a form of text analysis that seeks to pull out relevant keywords automatically from unstructured text. These keywords are a good descriptive summary of the unstructured text. In this study we use Term Frequency with Inverse Document Frequency (TF-IDF). This approach mainly captures the most frequently occurring word(s) in one document and less less frequently available in other documents [108] (See Fig 4.1, Information extraction module).

4.4.3 Topic modelling using LDA

The LDA components that we used entail; *a) Term document matrix*, that represents a document as a bag of words. It is also a easier way of representing the corpus, known as a word document matrix. The word document matrix is then taken as input and fed in topic modelling [8] [144], *b)Topic modelling* using LDA. This approach is statistical in nature and is used for clustering text documents. LDA makes the assumption that each document is made up of several topics and on the other hand each topic has various words belonging to it. The representations are done through document-topic w distributions and topic-word distribution v. As such, w and v are considered as dirichlet distributions. In the process the topics are discovered since they are not predefined, and *c) Document to topic assignment* is the most useful process of the LDA algorithm [144].

LDA algorithm

The LDA process follows an iterative manner starting with a random sample. This implies that the LDA process seeks to maximise the probability of a document belonging to a particular topic, and the probability of a word belonging to a certain topic iteratively. Once the algorithm converges, it generates final collection of words that are representative of a particular topic. It then computes the topics based on probability, and the topic with the highest weighting will denote the document's most dominant topic (See Fig 1, Topic modelling module). It is key to find an optimal value relative to the topic distribution per document using the hyper-parameter α. A high value of α may lead to a homogeneous distribution of topics. On the contrary, a low value of α may hinder a fair topical percentage distribution and any inferences thereof [144]. The LDA approach we use is reported in [110, 108, 144].

We applied the pyLDAvis package in Python in order to discover important information about our topics. We used a coherence score of 19 topics which we derived from the exhaustive search of the optimal number of topics and α (See Fig 2). The inter topical distance graph depicts topics as circles, and the distance between the circles denotes the relationship between topics (See Fig 6). The relationship is generated through dimensionality reduction using principal component analysis which employs probability distributions. The goal of this process is to try and have optimal topics that are distinct and do not overlap. To achieve minimal overlap we have to optimise our model by fine tuning model parameters such as α and relevancy (method for ranking terms within topics) to attain terms that are most important and exclusively belong to a particular topic. This overcomes the issue of having a biased representation of a topic, where highly ranked terms maybe frequent across the whole corpus [145]. In this study we follow the LDA methodological algorithm and approach reported in [144].

4.4.4 Topic modelling using BERTopic

This present study implemented BERTopic in four steps namely; 1) converting documents to their embedding representation using a pre-trained language model, 2) using uniform manifold approximation and projection (UMAP) for dimensionality reduction, 3) applying the hierarchical density-based spatial clustering of applications with noise (HDBSCAN) algorithm to cluster VA narratives into categories that have a similar meaning, and 4) generating classes through a class-based term frequency with inverse document frequency (c-TF-IDF) weighting scheme based on importance and relevance of a word, thus denoting the most representative words for each topic [134].

BERTopic algorithm

We used the BERTopic library in Python in order to discover important information about our topics. The BERtopic algorithm starts by creating document embeddings. Documents are embedded to create representations in a vector form that is semantically related. This implies that documents that entail the same topic are semantically similar. Document embedding is done using sentence-BERT framework, thus enabling users to convert text in sentences and paragraphs to dense vector representations using pre-trained language models. As such, these embeddings can be applied to perform clustering of similar documents. However, the embeddings are not used in generating topics [146]. After creating the document embeddings, the algorithm then performs dimensionality reduction using UMAP. This results in a more representative vector space which preserves local and global features. The reduced document embeddings are then clustered into similar groups using HDBSCAN. During the clustering process, HDBSCAN also clusters noise as outliers. Ultimately, this avoids clustering unrelated documents and improves the process of topic assignment. Lastly, topic assignments are

done using c-TF-IDF, and are based on the documents in a cluster, where one topic will be a cluster [146, 138, 133, 147].

Optimisation of LDA and BERTopic algorithms

We carried out exploratory topic modelling through an exhaustive search of the number of topics. In order to attain an optimal number of topics *k* we fine tuned α our hyper-parameter for the LDA algorithm. Additionally, we explored with number of topics from $k = 5$ to $k = 30$, and we looked at values of α that had a uniform distribution with values in the range $0-1$, and these where $(0.49, 0.504, 0.505, 0.508, 0.50, 0.48)$. After experimenting with $(k = 12, k = 19, k = 23, k = 29)$ topics, we discovered most coherent topics using 19 topics, as most of the words in a topic were associated. It should be noted that even if we generated high coherence scores for $(k = 12, \alpha = 0.54)$, $(k = 23, \alpha = 0.525)$, and $(k = 29, \alpha = 0.53)$, we settled for $(k = 19, \alpha = 0.505)$, since an increase in topic numbers results in difficulties in terms of human interpretability. Moreover, a large α value leads to similar generation of words in topics. Therefore, we chose lower values that follow a normal distribution and had most diverse representative terms [144]. Fig 4.2 shows the exhaustive search process that was done to identify the optimal number of topics.

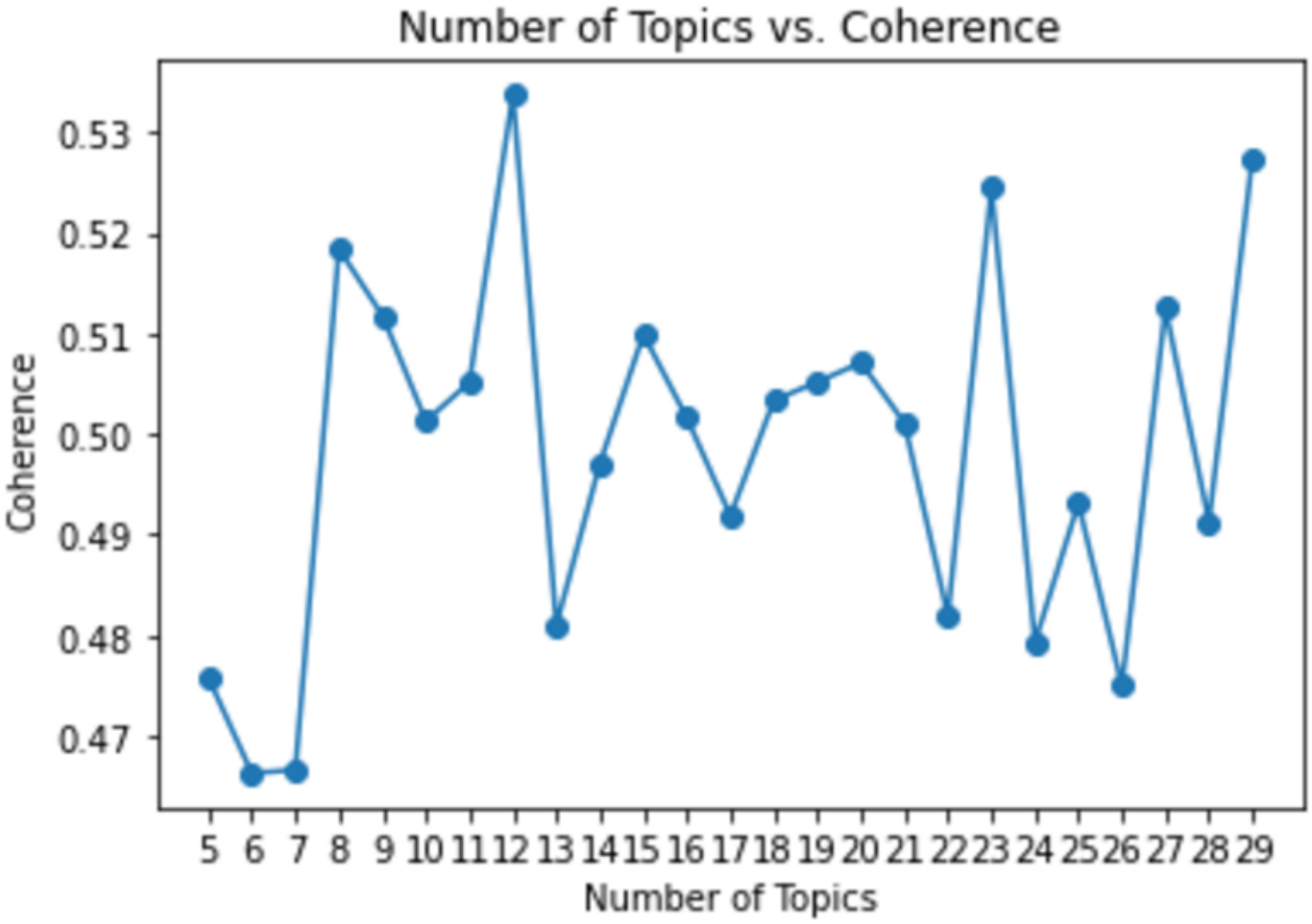


Figure 4.2: Hyper-parameter optimization for number of topics.

The implementation of the BERTopic algorithm entailed optimisation through the use of UMAP to improve topic assignment. BERTopic model evaluation was done using c_v (it is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity), and c-umass (it takes into consideration the document co-occurrence

counts, one-preceding segmentation, and a logarithmic conditional probability as a confirmation measure) [148]. Both measures seek to determine the degree of significance of words in a topic and the level of interpretability, thus assessing topic quality from a human perspective. The c_v measure ranges between 0 and 1. The higher the c_v score, the more understandable and coherent a topic is to a human. On the contrary, the closer the c-umass score to 0 the better [148]. The BERTopic attained a c_v of 0.66 and c-umass of -0.32. The mathematical approach is reported in the study of [140, 133].

4.4.5 Knowledge discovery

Zhu et al. [118] defines knowledge discovery as identifying meaningful information from large volumes of unstructured text. This knowledge can entail implicit or explicit facts, information or descriptions which relates to the contextual understanding of a specific domain. Furthermore, this discovered knowledge can serve as extra information or data that can be further utilised to discover more interesting patterns. In our context we employ knowledge discovery through data mining of the VA narratives in order to find answers to mortality causes and the most prevalent diseases that might have attributed to high mortality numbers within the HDSS. Additionally, further deductions and inferences through integration of our biomedical VA narrative data with other sources of data can give us further insights on our investigations (See Fig 1, Knowledge discovery module).

Buenano-Fernandez et al. [144] points out that at this phase we have to identify relevant topics by adding topic labels through a synthesis of the content generated by the topic assignment algorithms. However, the LDA approach has a challenge of interpretability, as the generated topics are not easily interpretable by humans because the initial assignment of words is random. Therefore, it is of paramount importance to have human interaction from beginning to the end of the process, in order to attain an optimal final model. As such, we engaged with a subject matter expert to assist with the topic labels. The expert synthesised the generated tokens and allocated a probable disease category. Similar steps and consultations were done with experts in order to get the best probable cause of death on the BERTopic algorithm results.

4.5 Data analysis

In this study we performed information extraction through word frequencies (n-grams) and keyword extraction using TF-IDF to identify the most frequent terms in the corpus, for our data analysis phase. We removed 2247 cases that had missing textual narratives. We further chucked out terms that had a high and low sparsity respectively, thus they were infrequently occurring in the corpus. We ended up with a 14079 documents for our analysis. Additionally, we also employed inter-topic distance measure through principal component analysis and applying HBSCAN to understand how our topics are related and clustered in our corpus. Moreover, we also performed structure and semantic analysis of random twenty (20) narratives in order to investigate if the narratives have same or additional variables as compared to the symptoms/variables from the responses to the structured questionnaire.

4.6 Statistical packages

All statistical analyses were done using Python 3.11.5 using Jupyter Notebook platform. The following libraries were used; *pandas* and *numpy* for data manipulation and analysis, *nltk* for preprocessing, *matplotlib* for generating visualisation, *sklearn* for feature extraction *bertopic*, *gensim* and*spacy* for topic modelling.

4.7 Results

The extracts below describe the results attained from the n-gram natural language processing approach and topic modelling.

4.7.1 Term frequencies using word cloud

Fig 4.3 depicts the the word cloud of terms that we extracted from our VA corpus. The word cloud captures a visual illustration and representation of word frequency. The more bold and large the text is, the more important it is.

4.7.2 Term frequencies using TF-IDF

Table 4.1 depicts the top 8 word frequencies using TF-IDF, a process known as keyword extraction. We found out that terms like *swollen, pain, hot, weak, diarrhoea, headache, stomach, vomiting, swollen, stomach, mental confusion, hiv positive, high blood pressure, cough, and chest pain* had high weightings in that listing order signifying their importance in corpus. Additionally, terms like *chest pain, low energy, diarrhoea vomiting* also had high weightings in the corpus. This suggests that these terms had high frequency scores, denoting importance in a document but less common in other collections of documents. Similarly, we can deduce that these term's importance was high in given documents, but very rare in other documents.

Table 4.1: Most frequently occurring terms.

Terms and corresponding frequency weighting	
Term	TF-IDF
pain	0.0926
swollen	0.0925
weak	0.0763
diarrhoea	0.066
vomit	0.058
hot	0.057
cough	0.055
chest pain	0.053

4.7.3 Term frequencies using n-grams

We illustrate our word frequencies in Fig 4.4 using n-grams. Fig 4a shows a visual representation of the most common symptoms from our corpus using unigrams. In a similar fashion, we managed to discover the same frequent terms as in our TF-IDF approach and the word cloud. We can see that the symptoms such as *pain, diarrhoea, swollen, hot, cough, weak, blood, stomach, chest, headache and vomiting* formed part of the most occurring uni-grams. Fig 4b denotes symptoms using bi-grams. We can see that *(chest, pain), (mental, confusion), (swollen, stomach), (diarrhoea, vomiting), (diarrhoea, diarrhoea), (high, blood),(stomach, swollen),(blood, pressure),(stomach, pain) and (low, energy)* were the most frequently occurring bi-grams. Fig 4c shows that *(chest, pain, cough), (blood, pressure, high), (pain, chest, pain), (cough, chest, pain) and (high, blood, pressure)* were the most occurring tri-grams. Fig 4d shows that, *(diarrhoea, vomit, diarrhoea, vomit),(vomit, diarrhoea, vomit, diarrhoea), (chest, pain, stomach, pain), (cough, chest, pain, hot), (pain, cough, chest, pain), (chest, pain, low, energy), (blood, result, hiv, positive), (cough, chest, pain, cough), (high, blood, pressure, diabetes), (chest, pain, chest, pain), (high,*

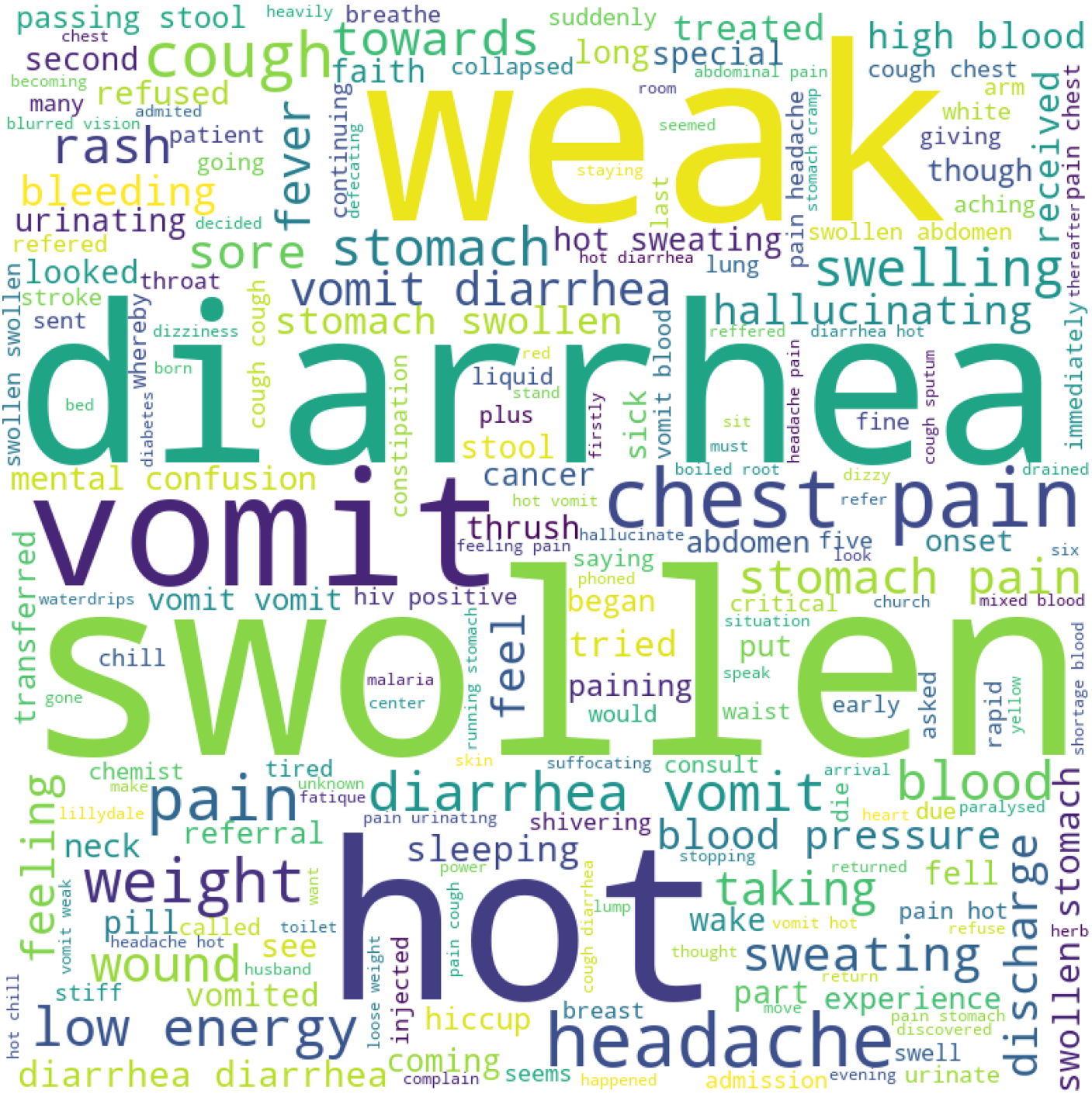


Figure 4.3: Word frequencies using word cloud

blood, pressure, high) as the most occurring quodgrams. Based on these word frequencies, one can possibly deduce that, these symptoms were the most common mortality causes of our HDSS population.

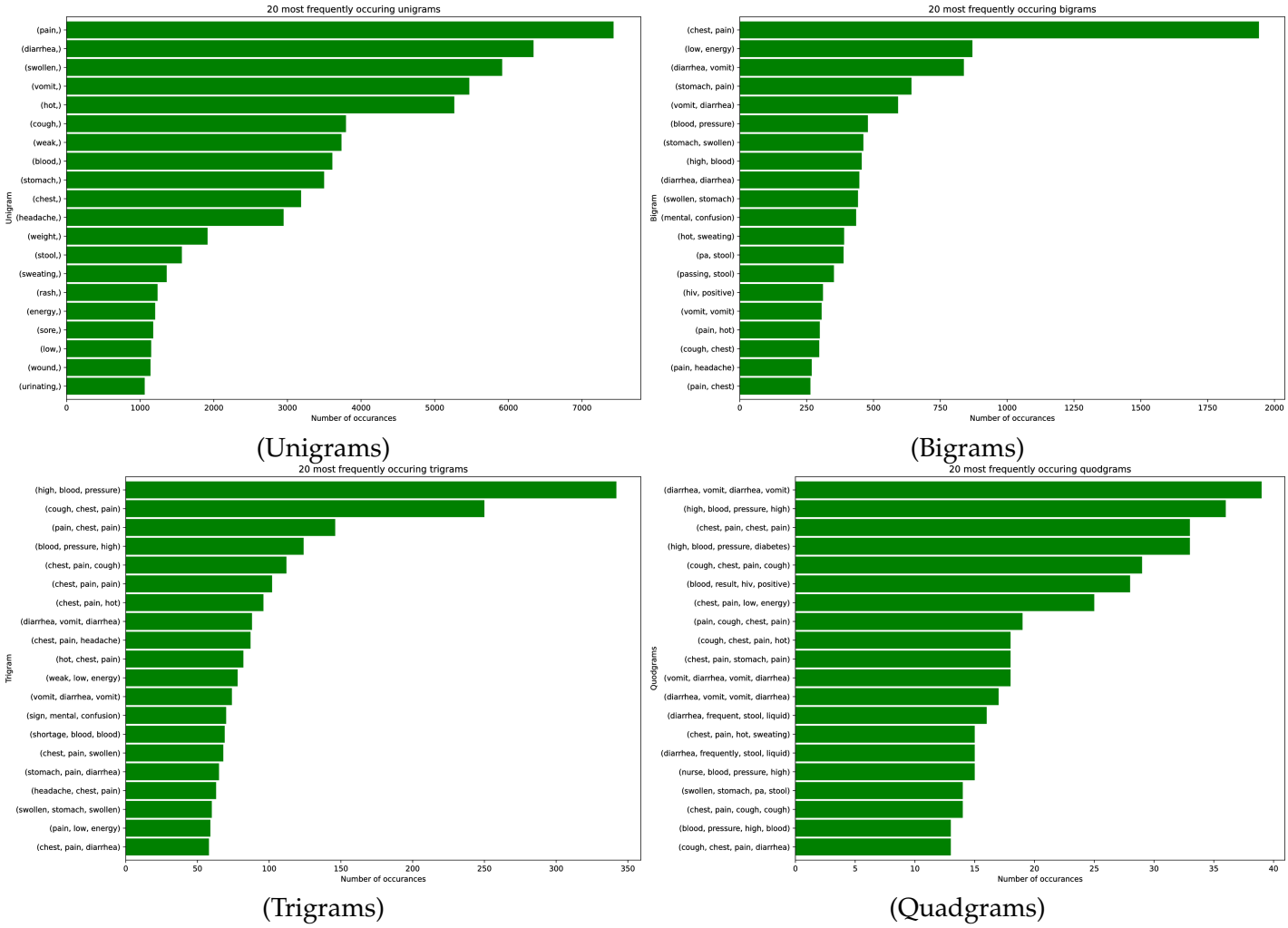


Figure 4.4: Most prevalent mortality causes using n-grams language processing.

4.7.4 Semantic and structure analysis of additional and similar variables

This section discusses results of the structure and semantics analysis of twenty random cases that we extracted to investigate if the narratives had additional information or exhibit the same symptoms as in the responses from the structured questionnaire. Interestingly, we note that the narratives entail the same symptoms as in the responses from the structured questionnaire, and further analyses gave us insights that they also entail additional variables that can be used to help with disease diagnosis especially in cases where the cause of death is unknown or there is a disagreement between physicians. Table 4.2 below is a comparison of symptoms from the responses from the structured questionnaire and symptoms the extracted textual data. Symptoms highlighted in the colour yellow denote similar symptoms and orange colour additional symptoms. We used anonymous record identifiers. Fig 4.5 is a VA narrative example that shows how text was preprocessed to generate clean processed symptoms/tokens that were used for the comparative analysis with responses from the structured questionnaire.

We can see that from Table 4.2, we managed to elicit similar symptoms from the textual data as compared to the responses from the structured questionnaire (see A1, A2, A8, A10, A16, A17). We also note that the extracted textual narratives also exhibit additional symptoms that can be easily interpreted to assist physicians in reaching a diagnosis (A5, A8, A9, A11, A14, A15, A17). On average from our random sample of cases,

UNPROCESSED VA TEXT NARRATIVE
Deceased started to ill while at working place, He came home while experiencing cough with chest pain, difficult in breathing, tiredness and blood vision. The after visited Belfast clinic to get treatment but no improvement. Afterwards deceased complained of stomach pain. Then after experienced diarrhea. He was given traditional medicine but did not change. Afterwards he vomiting worms and diarrhea continued. He continued using traditional medicine and the condition remains the same. Three days before death deceased sneezed a thing like a worm. He died at home and he also experienced hot body. It was examined that his chest and throat developed wounds. Treatment given but no change. His lower lip also had rash that at time chapping and a lot of blood will comes out. After treatment that lip became healed He was taken to traditional healer, but condition unchanged. He was taken Tintswalo hospital, where he was admitted Oxygen supplier was given but he finally passed away on the third day at hospital. A week before death he complained about body pain. At the beginning deceased also had cough and complained of headache during the night only throughout the illness. A month before death he experienced hiccup which continued until death but recurrent, he skips days not defecating When defecate the stool were hard then after yellowish and black few days before death. Deceased also developed ring worms on both checks but healed before death
PROCESSED VA TEXT NARRATIVE
['cough', cough',' chest',' pain',' tiredness',' blood',' vision',' stomach',' pain',' ' vomit',' worms','diarrhea',' sneezed',' worm',' hot',' chest',' throat',' ' lip',' rash',' chapping',' blood',' lip',' pain',' cough',' headache',' hiccup'," defecating',' defecate',' stool',' yellowish',' ring','worms']

Figure 4.5: An example of a processed VA narrative

Table 4.2: Comparison of symptoms

RecID: Symptoms from questionnaire	Extracted symptoms from textual data
A1: fever, cough ,breath, chest pain, whooping	cough , whooping , cough, screaming, vomited , hot , swollen
A2: cancer , fever, sweating, breath low, yellow, abdomen pain, urine , hair, weight loss	stomach, vaginal bleeding , womb, cancer treatment, painful, stomach bleeding, urinating , swollen breast
A3: hypertension,asthma,low breath, weight loss, swollen legs, alcohol intake	passing stools , swollen
A4: hypertension, skin, urine, weight loss, swollen legs, excessive drinking	develop, sore toe , toes infection , wounds , grow infection, smell , worms , swollen
A5: fever, breath, chest pain , weight loss, excessive drinking, sweating	carrying , heavy box , twisted , pains , toes , affected , toes, swell , pain , blood, clot , removed, swell, huge, lump bursts , holes lump , chest , armpits , groin , throat , lumps, busted, puss coming
A6: TB, epilepsy, fever, cough , breath low, chest pain, diarrhoea , convulsion, vomiting, stiff neck	swollen , diarrhoea , stools , cough , go, toilet
A7: nil	pain , treated, looking better, weak , admitted, drips injection, worse
A8: hiv , fever, cough, low breath, yellow, alcohol, diarrhoea , vomit, weight loss , abdomen , swelling	sweats , lose weight , vomit , weak , hiv , water drips, diarrhoea , hallucinating , swollen stomach
A9: sweating, cough , low breath, urine, weight loss, paralysed, alcohol	cough , heavily , rot , worms , rotting , spread , thigh , buttocks , waist , extent , intestines , visible , abdomen
A10: fever , cough , vomit , weight loss, alcohol	cough , signs, vomits , low , energy , cough critical , tried, natural, fever
A11: hypertension, fever, low breath	fell , forehead , improve, speak, tears, rolled eyes , mixed blood , accident , urinate
A12: fever , cough , low breath, abdomen , swelling , headache, injury	initially, swollen ears , swollen , abdomen , critical , puss , ears , fever , cough
A13: nil	swollen , wounds , top, feet , wounds , burst , puss , swollen, feet, cuts , razor blade , rubbed
A14: fever, low breath, chest pain, diarrhoea , vomit , weight loss , hair,eyes sunken	committing, thinner , diarrhoea , oral, dehydration , solution, vomit
A15: chronic, fever, skin, weight loss, swollen legs , alcohol, smoking	pain , toe , long, kneel, cut , flesh wound , toe, swollen toe , infected
A16: fever, cough, diarrhoea , blood diarrhoea, vomit , weight loss	diarrhoea , vomit , weight loss , oral, dehydration , thrush
A17:hiv,stroke,low breath, weight loss, alcohol,chest pain	fell , weak , touch, hold, feeling, pains , arv, years, clear, pass , stools , giving, arv, sweats , minutes

the retrieval rate of additional is approximately $n > 2$, per every case where n is the number of terms. Moreover, we discovered that the text-extraction algorithm misses a symptom that is in the responses from the questionnaire cases $\approx 2\%$ of the times. This can be attributed to the fact that some of the narratives entail shallow content and are less informative than others.

4.7.5 Topic modelling results using LDA approach

In this section we illustrate through graphs the results of our LDA process. Table 4.3 shows the results of the common 19 topics that we mined from our VA narratives (the most prevalent diseases). Additional visualisation is also given on the inter topic distance map in Fig 6. Basically, we noted that the most prevalent diseases were mainly communicable and non communicable. The highlights being *diarrhoea, TB, HIV, malaria, chronic ailments, neurological disorders, asthma, accidents and maternal related deaths*. We also had topic 7 and 12 which had fewer symptoms to relate to any disease and we categorised them as unknown. Expectedly, we note that maternal and accident related deaths which are topic 19 and 15 respectively are distinct and far from the other clusters. This is attributed to the fact that they have very dissimilar symptoms from the other topics. We also discovered that topic 14 and 16 which are mental confusion and heart ailment respectively, are on their own cluster. This suggests that the two topics share some common similar symptoms. Another interesting observation is the cluster with 5, 7, and 12 which are topics for chronic liver ailments and unknown topics respectively. This grouping depicts similar symptoms for unknown which are also present in the liver complications. This suggests that cases with such symptoms in the unknown can have symptoms that lead to liver complications.

One interesting discovery is that topic 2 on HIV is central to most of the clusters such as TB, diarrhoea, malaria, cancer, epilepsy, diabetes, hypertension, and chronic ailments. One possible reason for this is that most of these disease are related to HIV and most cases who have these complicated diseases also have HIV, or might be at risk of having HIV, as evidenced by the similar symptoms. Furthermore, we also observe that jaundice/ malaria with skin infection is related to neurological disorders as evidenced by the close cluster embedding of topic 6 in 11. This suggests that some cases had a long untreated jaundice which might have led to brain damage, a condition known as kernicterus. Of particular interest is cluster diarrhoea with respiratory tract infections, malaria and chronic heart ailments (topic 1, 10 and 18). This suggests that these diseases share similar symptoms and if one suffers from the earlier diseases, it can lead to heart chronic ailments. Another finding is that cluster 8 and 9 on breast cancer and chronic ailments respectively have similar characteristic features. This suggests that if one has cancer they might also have complications of liver, lungs, and kidneys.

Fig 4.6 depicts the relationship and association between topics using the inter topic distance measure through principal component analysis. This process basically illustrates and models the visual representation of similar topics being in close proximity (one cluster) and dissimilar topics distant to each other. The bigger the topic circle the more tokens it uses from the corpus for topic generation. As such, we can see that topic 1, topic 2, topic 3 and topic 4 use 8.8%, 8.6%, 8.2% and 8.1% of the tokens in the corpus respectively. Topics that are overlapping imply having some similar characteristics symptom wise, hence they fall within the same cluster. We can see that topic *15* and *19* are distinct topics. Noticeably, we can also see clusters *(14 and 16)*, *(5, 7, 12)*, *(3,4)*, *(2, 3, 13)*,*(9,8,17)*, *(11,6)*, *(18,1,10)* having similar characteristics.

4.7.6 Topic modelling results using BERTopic approach

This extract uses figures to depict the results of our BERTopic model in topic assignment. Initially, the model generated 154 topics, identified 6265 documents as outliers. UMAP was applied for dimensionality reduction

Table 4.3: Most prevalent diseases identified using LDA

Topic number and name	Tokens
T1: Diarrhoea/respiratory tract infection	cough body diarrhoea weight hot sputum rapid pains vomit headache rash chills
T2: HIV	diarrhoea vomit hiv hot weak positive refuse arv blood cd count stopped
T3: Tuberculosis	cough pain chest sputum hot lungs weight results sharp spit blood lungs
T4: Cervical cancer	bleeding blood cancer urinate swollen urine wound urinating coming vaginal nasal puss
T5:Chronic liver ailment	paining stools pass pains liver weak running constipation stomach admitted shortness tired
T6:Neurological disorders	diarrhoea looked mental confusion hence vomit weight encountered abscess suddenly improve though
T7:Unknown	pains chest headache feeling hot hallucinating abdominal migrant faith breathe energy low
T8: Breast cancer/infection	wound swelling swollen feet sore puss lump burst ointment wounds breast sent hole hip infected
T9: Chronic ailments (liver,lung, kidneys)	swollen abdomen liver discharge diarrhoea vomit yellow drained lungs kidneys lungs swell
T10: Malaria	received malaria breast oral groaning feeding diarrhoea vomit toilet problems dizziness dehydration
T11: Jaundice/malaria/skin infection	rash sores vomit itchy yellow diarrhoea thrush hot eyes sunken weak sweating
T12: Unknown	panado syrup lillydale years liquid problems toilet die go oral whenever going
T13: Diabetes/Hypertension	high blood pressure stroke collapsed diabetes low hot energy shivering paralysed hours
T14: Mental confusion	feels starts refuse injected admission confused weak hiccups care mental makes sick
T15: Accidents	car accident shot road broken truck driving taxi fell collapsed paralyzed hanged
T16: Septic wound/heart ailment	wound received swelling began whilst swollen stomach cramps herbs hot bigger sore
T17: Epilepsy	stiff tongue teeth bent gone onset epilepsy jaws pills uncontrolled neck sudden
T18: Chronic ailments (heart)	heart fast swelling failure beating collapse urinate pass penis catheter bladder paralysis
T19: Maternal and Asthma	crying hours born room pregnant breast delivered delivery asthma breastfeeding cried incubator

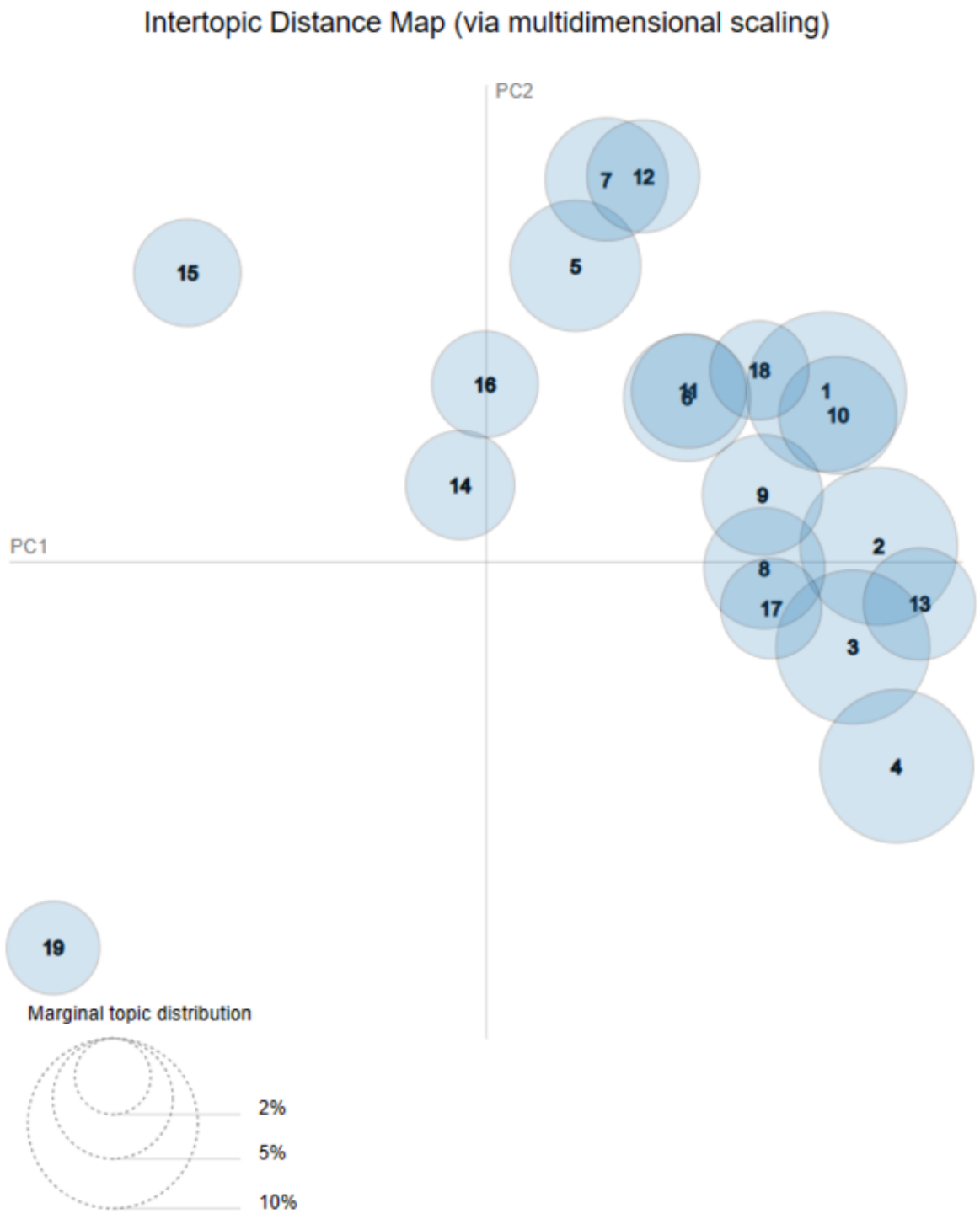


Figure 4.6: Inter-topic distance map using multi-dimensional scaling
Legend: T1:Diarrhoea/Respiratory tract infection, T2:HIV, T3:TB, T4:Cervical cancer, T5:Chronic liver ailment, T6:Neurological disorder, T7:Unknown, T8:Breast cancer, T9:Chronic ailments, T10:Malaria, T11:Jaundice/malaria/skin infection, T12:Unknown, T13:Diabetes, T14:Mental confusion, T15:Accidents, T16:Heart ailment/septic wound, T17:Epilepsy, T18:Heart chronic ailments, T19:Maternal and Asthma

and we were left with 30 topics. Fig 4.7 shows the generated thirty topics and corresponding top words with scores for each topic. Topic 0 is about HIV that has top word scores of *hiv, positive, arv, count*. Topic 1 shows skin related diseases. Topic 2 is TB as evidenced by symptoms with high word scores such as *sputum, cough, smoking, chest, lung*. Topic 3 suggests septic wound diseases which might be attributed to other diseases such as diabetes. Topic 4 is unspecified. Topic 5 is a sexually transmitted diseases as evidenced by the genital related symptoms. Topics 6 and 8 are unspecified. Topic 7 and 14 denotes symptoms that might be related to diarrhoea. Topic 9 is about abdominal swelling complications. Topic 10 is about malaria. Topics 11 *shot, gun, bullet, police, shooter*, 17 *stabbed, knife, fight*, 19 *car, accident, road, injured, driving*, and 20 *fire, paraffin, burned, burnt* are accident related causes of death. Topic 12 is maternal related death as highlighted by these high word weightings *infant, feeding, breast, born, milk*. Topics 13 shows symptoms that might be related to other diseases such as HIV. Topic 15 suggests diabetes and hypertension related causes of death as denoted by these terms *diabetes, toe, high, pressure, cut*. Topic 16 suggests ulcers and TB related symptoms. Topic 18 might be related chronic liver ailments and jaundice like disease. This is shown by high word terms like *yellow, eye, liver, urine, yellowish, substance*. Topic 21 is related to stroke and paralysis possibly because of diabetes and hypertension causes, with terms such as *stroke, paralysed, collapsed, paralysis, stiff*. Topic 22 talks to cervical cancer prone in women as shown by these terms *vaginal, bleeding, cancer, cervix, cervical*. Topic 23 shows symptoms that are related to diseases such as malaria and HIV. Topic 24 depicts symptoms that are TB related. Topic 25 is a result of epilepsy evidenced by top word scores such as *epilepsy, seizure, tongue, uncontrolled, stiff, jaw*. Topic 26 exhibits hiccup symptoms coupled with pain and low energy. This suggests possible diseases such as pneumonia, bowel diseases, pancreatitis, pregnancy, bladder irritation, liver cancer or hepatitis. Topic 27 show symptoms related to diabetes such as consciousness, sweating, hot body and sunken eyes. Topic 28 is related to psychiatric conditions because of high weighting on terms like hallucination. Topic 29 is unknown.

Fig 4.8 shows the inter-topic distance map with 5 dominant clusters. Cluster 1 is made up of topics (5, 18, 22). These are correlated due to genital mortality causes both in men and women. This cluster suggests sexually transmitted diseases and infections that ultimately lead to liver and chronic ailments. Cluster 2 is made up of topics (3, 9, 15). This cluster denotes diabetes and high blood pressure causes of death. These chronic diseases can lead to amputation and abdominal complications. Cluster 3 shows correlation of topics (11, 19, 20) that are accident and injury related causes of death. Cluster 4 shows the dominant HIV as topic 0 and closely correlated to topics (1, 4, 7, 8, 12, 13, 14, 23, 27, 29). This suggests that most of the HIV diseases are related to diarrhoea, and symptoms such as *low energy, no appetite, headaches, skin infections, vomiting, loss of weight, sweating*. Interestingly, we observe that similar symptoms are correlated to maternal deaths. Noteworthy, is the striking number of migrant workers who exhibit same symptoms and also succumbed to death from HIV and maternal related diseases. Cluster 5 is made up of topics (2, 6, 16, 21, 24, 26, 28). Topic 2 which is related to TB and lung diseases is closely correlated to topic 16 that talks to chest pains. Noteworthy, we observe epileptic seizures being correlated to loss of consciousness, paralysis and collapsing.

Fig 4.9 shows a heatmap similarity matrix with distinct topics, generated from BERTopic document embeddings. This was a result of applying the cosine similarity to document embeddings to get the relationship and distinction between topics in a quantifiable way. Noteworthy, the similarity score on the topics on HIV, skin related diseases, migrant workers, malaria, TB, chest pain, hallucination, septic wounds,diabetes and hypertension, and epilepsy are very high amongst others as depicted in Fig 4.9.

The results illustrated in Fig 4.8 are also visually depicted in our hierarchical clustering in Fig 4.10. For example, as observed from the data, the closely clustered linked topics in this group are *11, 17, 19, 20*. These are related to accident and injury mortality causes. This suggests that there were more deaths from road accidents, gender based violence and fire accidents. There is need to raise awareness to avoid the surge of

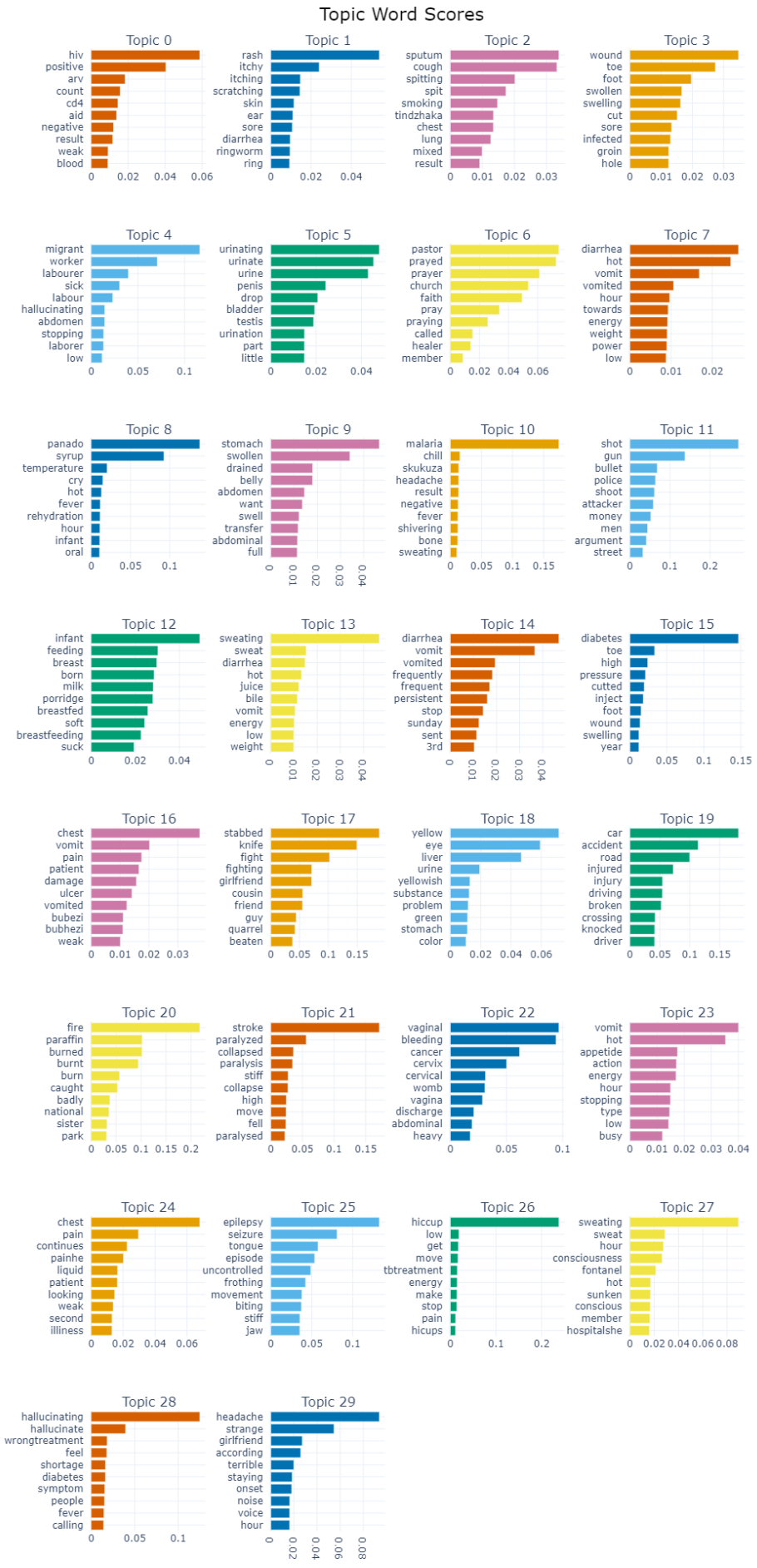


Figure 4.7: Topics generated using BERTopic model

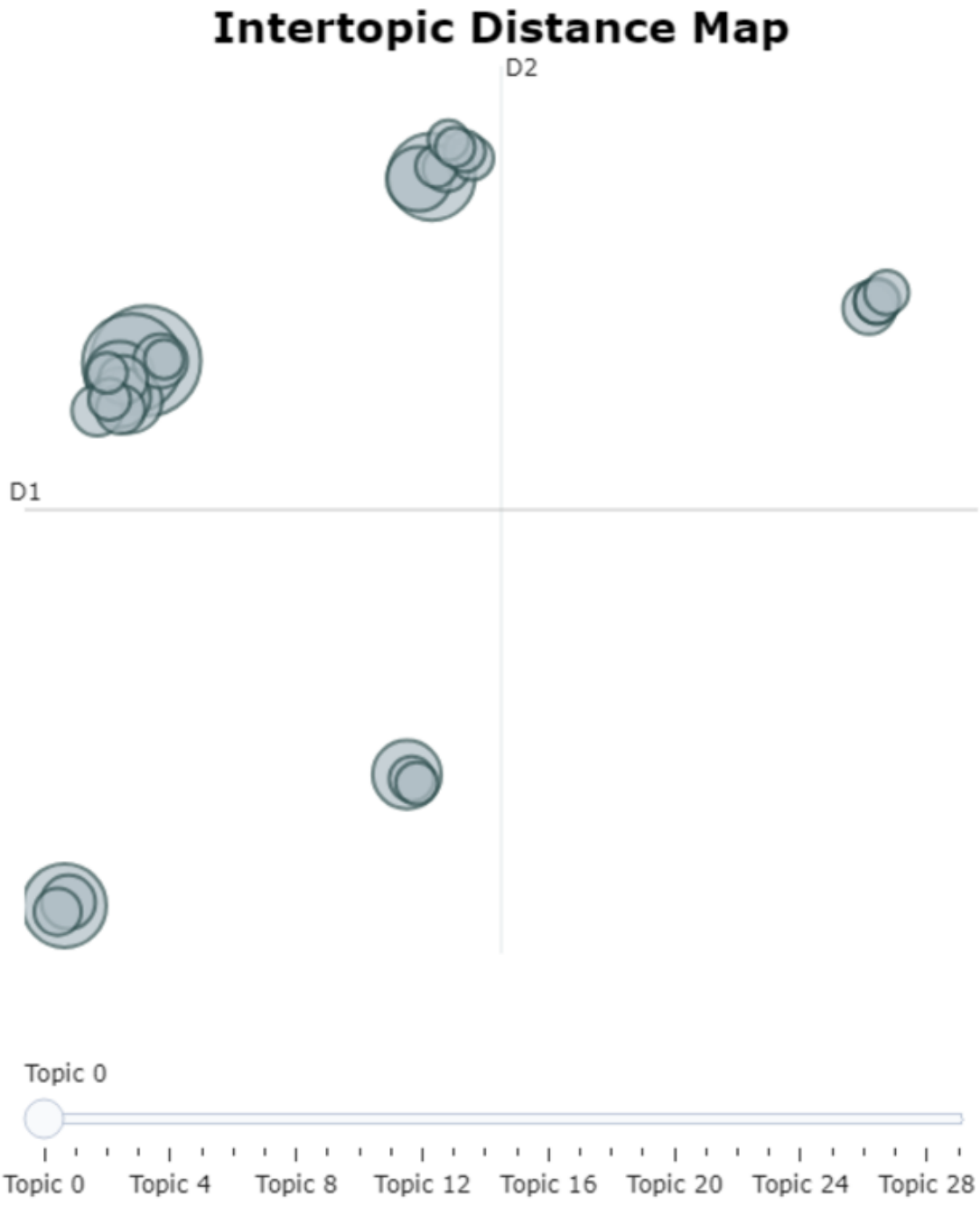


Figure 4.8: BERTopic inter-topic distance map

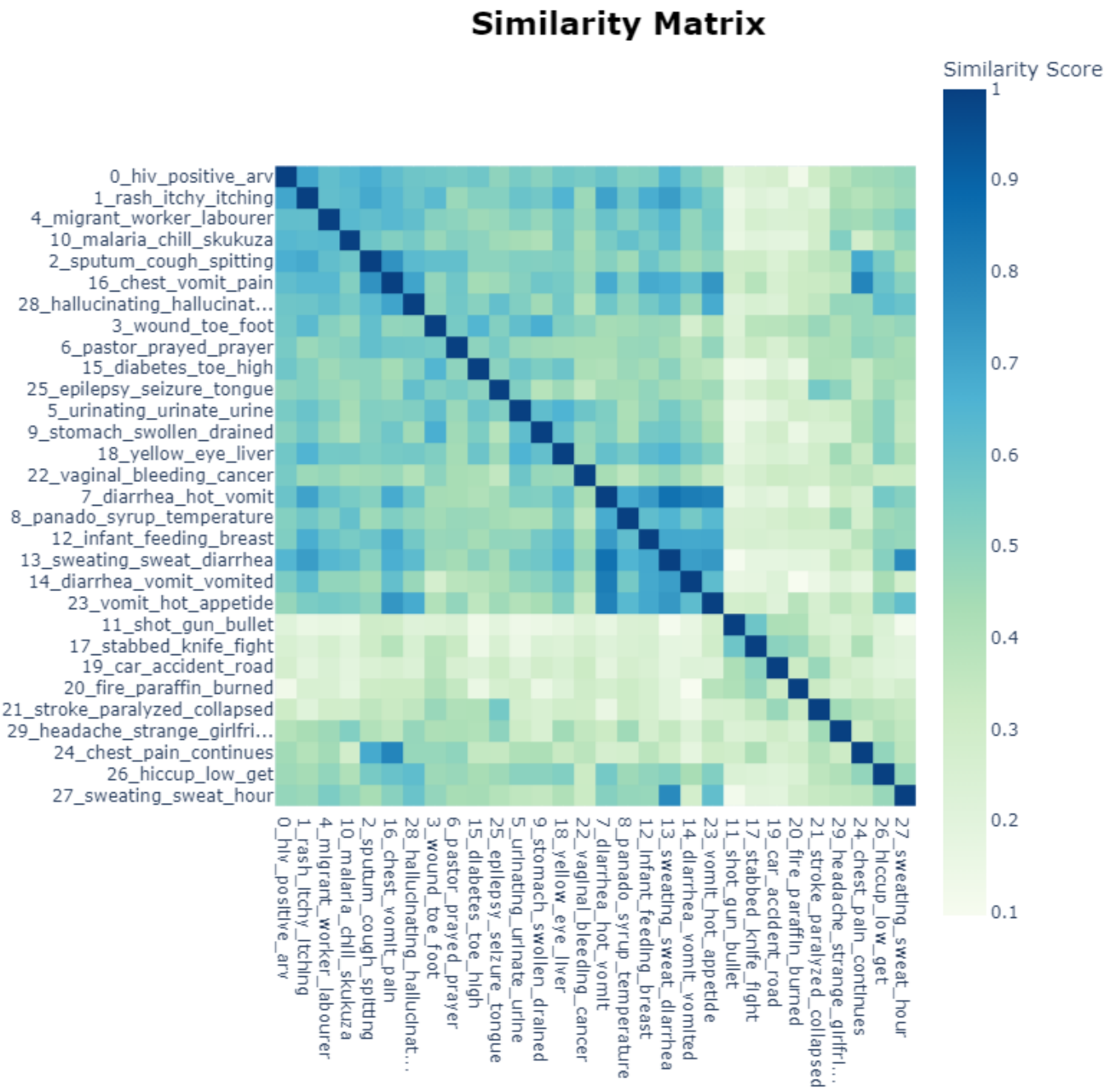


Figure 4.9: Heatmap similarity matrix

such cases. Interestingly, we also note that HIV, diabetes, hypertension, TB, chronic lung and liver ailments and cervical cancer are closely clustered. This might be attributed to the fact that TB leads to lung and cancer complications. Furthermore, most HIV patients are prone to secondary diseases, that eventually worsen their condition and give rise to more deaths. Noticeably, we observe that patients with liver cirrhosis often present with urinary complaints as shown in the closely clustered topics 5 and 18. Noteworthy, is the diarrhoea cluster exhibiting typical symptoms such as *sweating, low energy, weight loss, hot body, sunken eyes*. As expected in topic 12, we note an isolated cluster on maternal related death at birth. As observed, we see that epilepsy seizures are closely linked to stroke and paralysis. As such, there is need to invest in more care workers to look after epileptic patients. Interestingly, we note that there is a high correlation on diabetic cases that result in hallucinations. Therefore, there is great need to invest in health awareness programmes, improved primary health care and early screening processes to try and curb the rise of these diseases.

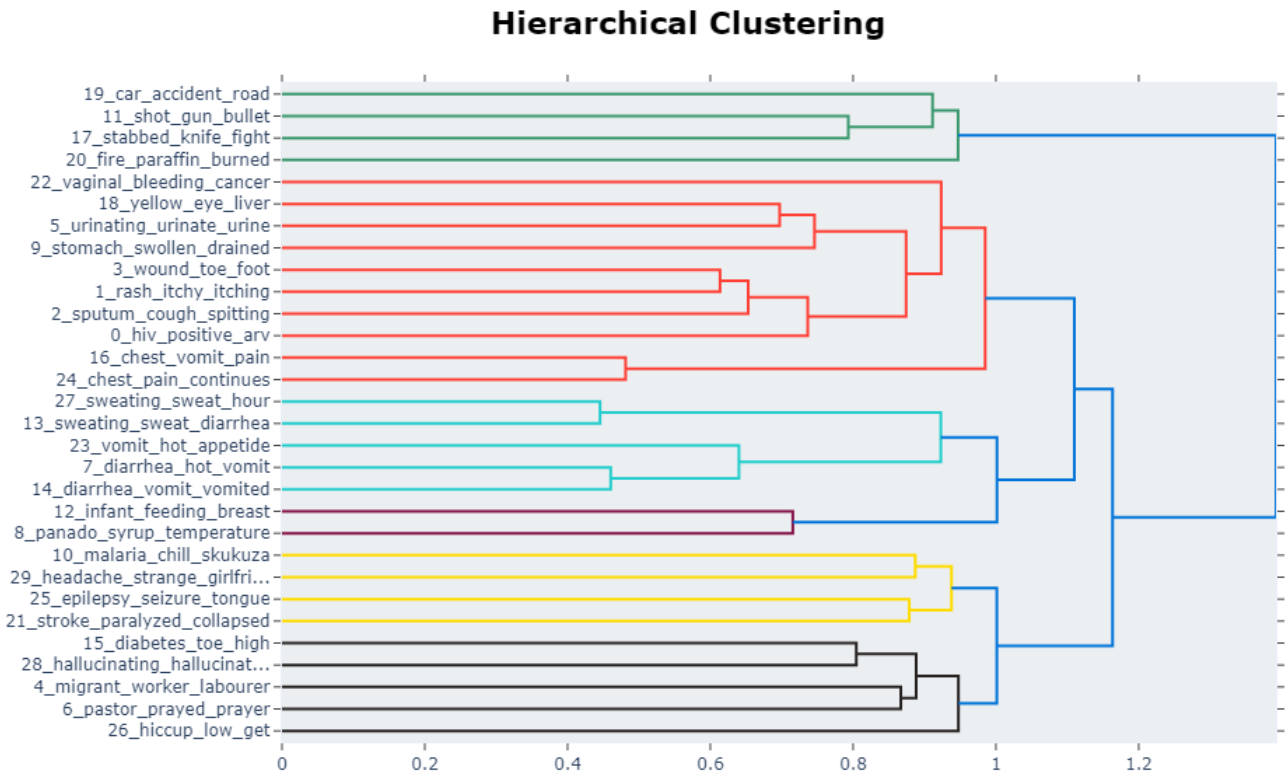


Figure 4.10: Hierarchical clusters of similar topics

4.7.7 Results from the knowledge discovery process: Identified relevant topics

Our analysis suggests that most of the population in the HDSS could have died from diseases such as diarrhoea, TB, HIV, cancer, maternal, accident and injury and chronic ailments. Our LDA topic clustering shows that accident (T15) and maternal related deaths (T19) were distinct and distant from other clusters. A similar pattern and trend is also noticeable on T4 which is cervical cancer. We notice accident related deaths from the HDSS, but mostly intriguing are suicidal deaths. Of interest is cluster T1, T10 and T18. It shows that cases that succumbed to diarrhoea, respiratory tract infections, malaria and heart ailments related deaths, exhibited mostly similar symptoms. Another paramount discovery is cluster T2, T3 and T13. This cluster depicts that subjects who died from HIV also had other conditions related to TB and diabetes. Furthermore,

we can also note that HIV related deaths were also associated with cancers and chronic ailments (T8, T9). T17 shows epilepsy which is also a neurological disorder. This disease is clustered mostly with T8, which is breast cancer. Additionally, we can also see that it shares some similar characteristics with HIV and TB. We can also deduce that cancer of the breast or breast infections have some common symptoms with chronic ailments of the liver, lung and kidneys as shown on the T8 and T9 cluster.

Cluster T5,T7 and T12 show that some deaths from liver ailments can lead to complications from not easy to identify symptoms which are in the unknown topics. One important finding is that HIV related cases share common symptoms with TB related deaths. From the HIV topic, through token analysis, we notice that some of the subjects were defaulting from taking their medication *stopped, refused*. Interestingly, we notice that some cases were affected by hypertension and diabetes, and this led to stroke and paralysis. This finding implies that diabetes and high blood pressure could be indicative of a form of heart metabolic syndrome. Moreover, hallucinations and mental confusion can happen with very high blood sugar levels (T7). We also discovered some cases of jaundice related deaths possibly prone in the new born babies. This calls for better controlled and effective immunisation measures in order to curb such diseases. Similar findings are also exhibited on our BERTopic analysis, as we observed from the discussions in previous paragraphs.

4.8 Discussion

The present study's novelty was to apply text mining techniques specifically n-gram natural language processing, LDA and BERT in topic modelling to investigate mortality causes and most prevalent diseases, using VA narrative data. As such, we managed to explore the relevancy of text mining techniques in order to discover mortality causes and the most prevalent diseases. The rich textual narratives availed valuable insights by extracting implicit knowledge on symptoms and cause of death. This is consistent with previous findings by [1]. Moreover, this study, findings are correlated with results from our previous work reported in [117, 149].

We first present the most relevant terms that were the main causes of mortality and identify the main topical diseases that were prevalent in the HDSS population. Additionally, we also managed to perform topic modelling, thus identifying interesting patterns and trends based on disease symptom similarities and differences. The generated associated disease clusters provide a statistical potential on discovering semantic level associations. Of interest, are findings from semantic and structure analysis of random cases, that show that the narratives exhibit similar and in some instances have additional symptoms as compared to those in the responses to the structured questionnaire. These findings can therefore be used to guide the health community and close the civil registration system gap. Our work also provides a methodological framework for text mining and topic clustering. Therefore, robust automated unsupervised systems that analyse VA narratives and avail interesting insights for public health improvement are vital for the community and world at large.

The findings on the n-gram natural language processing suggest that the HDSS population could have died, mostly from mortality causes such as *diarrhoea, fever, headaches, vomiting, pain, blood pressure, low energy, hot/swollen body*, amongst many. Interestingly, using topic modelling to identify most prevalent diseases, we discovered that communicable diseases such as HIV and tuberculosis were more prevalent in the HDSS. Interestingly, non communicable diseases such as diabetes, chronic ailments (kidney, heart, lung, liver), diarrhoea, cancer, respiratory tract infections, neurological disorders, hypertension and mental health ailments, were also very common in the HDSS. Expectedly, we notice distinct topics for maternal and accident and injury related deaths. This suggests that these topics have least relationships with other topics as they have dissimilar characteristics/ symptoms. A very striking observation is that neurological disorders is embedded in jaundice/ malaria/ wound with skin infection. This can be attributed to the fact that some cases had a long untreated

jaundice which might have led to brain damage, a condition known as kernicterus. A similar pattern and trend is exhibited in malaria and diarrhoea and tract respiratory infections. Our topical distance map shows that HIV is central to most clusters (TB, cancer, epilepsy, diabetes, hypertension, chronic ailments), thus implying that most of the diseases exhibit symptoms as in HIV. Consequently, mortality from HIV can also be a result of complications from the earlier diseases. Noteworthy, is the fact that most people who suffer from uncontrolled hypertension and diabetes are prone to stroke and paralysis. The above findings are consistent with our initial work reported in [117, 26].

On the contrary, we note two unknown diseases, that have less informative symptoms to get to a cause of death diagnosis. However, we can see that these topics are also in the same cluster with chronic related liver ailments. Therefore, in such cases the doctors can merge the fewer symptoms in the unknowns and the ones in the known cause of death topic to probably get to a diagnosis as the symptoms are related and associated.

Our structure and semantic analyses can further provide insights that can help the public health community to extract additional implicit knowledge and symptom associations about cause of death from the large VA data. This can prove beneficial especially in cases where the physicians disagree on the cause of death and also where there is limited diagnosis information (symptoms)on the responses from the structured questionnaire. Interestingly, interpretability of CoD is made easier using our word sequence modelling approach that identifies semantics and relations in text. As such, this will save time, is cost effective and improves the diagnosis turnaround time. Of paramount importance is that, this present study methodological approach improves knowledge representation of VA textual data which supports multi disciplinary research across disciplines (See Fig 5).

Our initial study also reports on interesting findings on the most sought after treatment method when the HDSS population is faced with terminal illnesses [117]. Even though this was not the aim of this study, we notice a similar trend and pattern in the current study, as people mostly seek the traditional ways. Noteworthy, we observe diseases such as HIV, that are affecting immigrants in the HDSS. This suggests that that immigrants are also spreading such diseases. Consequently, there is great need for further investigations as this might lead to identifying subjects who default from medication and western ways of getting medical help. As such, this can possibly suggest some of the premature deaths. Informed by the study of Kabudula et al. [150], socio-economic status also plays a big role in morbidity and mortality, hence there is a great need to try address this challenge in order to improve public health.

Literature reports on the threat of non-communicable diseases and chronic diseases. Furthermore, they suggest effective implementation of primary health care systems as well as scaling up chronic based delivery, in order to curb the high mortality numbers [151]. This present study, also supports the notion raised by these authors in an effort to curb high mortality numbers.

It is imperative to improve population health, through effective intervention and support programmes, and early screening processes that are well monitored and controlled through initiatives by relevant authorities. This can prove beneficial especially for diseases such as cervical cancer.

This present study will lead to improved public health planning through the use of the findings ton inform public health strategies by identifying major health issues contributing to mortality, allowing for targeted interventions and resource allocation. Moreover, the results can be used to develop evidence-based policies addressing the specific health challenges prevalent in a given population, leading to more effective healthcare delivery. The application of topic modelling to VA data represents an innovative approach in health research, showcasing the potential of NLP in extracting valuable insights from unstructured medical narratives.

This study was limited in complexity of the VA narratives used. The analysis entailed some terms in topics which were of little significance. This is consistent with the findings of Hacking et al. [127]. Moreover, topic segmentation has a challenge of human interpretation of discovered topics. The narratives were highly unstructured, and entailed high levels of noise, sparsity, had varying technical vocabulary terms, contain misspelled terms, had many grammatical errors, and exhibited issues around word polysemy. Nevertheless, the findings of this study still stand despite all these limitations.

4.9 Future work

Informed by the findings of this study, specifically the semantic and structure analysis, our future direction will seek to identify under-reported morbidity occurrences at the Agincourt HDSS, by effectively linking HDSS population data and NHLS public health data registries. This can further give us insights on other factors that might be leading to higher mortality numbers.

4.10 Conclusion

The application of the unsupervised TM approaches in the current study creates a synergistic framework that leverages the strengths of each method, providing a comprehensive and accurate analysis of mortality causes and prevalent diseases within the VA narratives. The results obtained from this study contribute to the advancement of public health research by offering a sophisticated and interpretable methodology for extracting valuable insights from verbal autopsy data. As such, they can be included in the pipeline for identifying mortality causes and most prevalent diseases alongside human annotation, and interpretation. The increase in textual digital repositories can be fully utilised using automated TM approaches that can extract implicit knowledge on cause of death and avail valuable insights, which could not have been possible through manual processes.

4.11 Acknowledgements

We acknowledge the University of the Witwatersrand, Faculty of Health Sciences for providing high computing infrastructure resources.

5 Explainable Stacked Ensemble Deep Learning (SEDL) Framework to Determine Cause of Death from Verbal Autopsies

This study was published in Machine Learning and Knowledge Extraction journal (October 2023, Mach. Learn. Knowl. Extr. 2023, 5(4), 1570-1588. https://doi.org/10.3390/make5040079)

Michael T. Mapundu[1]*, Chodziwadziwa W. Kabudula[1,2], Eustasius Musenge[1], Victor Olago [3], Turgay Celik[4,5].

[1]Division of Epidemiology and Biostatistics, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Parktown, Johannesburg, South Africa
[2]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), University of The Witwatersrand, Johannesburg, South Africa
[3]National Health Laboratory Service (NHLS), National Cancer Registry, Johannesburg, South Africa
[4]Wits Institute of Data Science, University of The Witwatersrand, Johannesburg, South Africa
[5]School of Electrical and Information Engineering, University of The Witwatersrand, Johannesburg, South Africa

5.1 Abstract

Verbal autopsies (VA) are commonly used in Low to Medium Income Countries (LMIC) to determine cause of death (CoD) where death occurs outside clinical settings, with the most commonly used international gold standard being physician medical certification. Interviewers elicit information from relatives of the deceased, regarding circumstances and events that might have led to death. This information is stored in textual format as VA narratives. The narratives entail detailed information that can be used to determine CoD. However, this approach, still remains a manual task that is costly, inconsistent, time consuming and subjective (prone to errors) amongst many drawbacks. As such, this negatively affects the VA reporting process, despite it being vital for strengthening health priorities and informing civil registration systems. Therefore, this study seeks to close this gap by applying novel deep learning (DL) interpretable approaches for reviewing VA narratives and generate CoD prediction in a timely, easily interpretable, cost effective and error-free way. We validate our DL models using optimisation and performance accuracy machine learning (ML) curves as a function of training samples. We report on validation with training set accuracy [LSTM=76.11%, CNN=76.35%, SEDL=82.1%], validation accuracy [LSTM=67.05%, CNN=66.16%, SEDL=82%], and test set accuracy [LSTM=67%, CNN=66.2%, SEDL=82%] for our models. Furthermore, we also present Local Interpretable

Model-agnostic Explanations (LIME) for ease of interpretability of the results, thereby building trust in the use of machines in healthcare. We presented robust deep learning methods to determine CoD from VAs, with the Stacked Ensemble Deep Learning (SEDL) approaches performing optimally, and better than Long Short Term Memory (LSTM) and Convolutional Neural Network (CNN). Our empirical results suggest that ensemble DL methods may be integrated in the CoD process to help experts get to a diagnosis. Ultimately, this will reduce the turnaround time needed by physicians to go through the narratives in order to be able to give an appropriate diagnosis, cut costs and minimise errors. This study was limited by the number of samples needed for training our models, and high levels of lexical variability of words used in our textual information.

Keywords: Cause of Death, CNN, Deep Learning, LIME, LSTM, Machine Learning, NLP, SEDL, Verbal Autopsy

5.2 Introduction

More than 65 percent of the world population lack high quality information on cause of death (CoD), accounting to two thirds of the of sixty million deaths worldwide, that are not assigned a medically certified cause [2, 3, 1]. This information is vital for informing civil registration systems, policies and strengthening of health priorities. Most of these deaths are common in Low to Medium Income Countries (LMIC) as they occur outside health facilities [4]. As such, the verbal autopsy (VA) tool is used to determine CoD. VA is a process that entails non medical personnel interviewing next of kin of the deceased using a structured questionnaire where they seek to elicit information on circumstances and events that could have led to CoD. The collected information known as VA narratives are then housed as text and given to two medical doctors who determine the probable CoD and if they do not agree a third physician is consulted, a process known as Physician Coded Verbal Autopsy (PCVA) [1]. Detailed information on the VA process is reported in [152]. Even though the PCVA is the only gold standard used in CoD decisions, this process is expensive, time consuming, inconsistent and subjective (prone to errors in information elicitation) amongst many drawbacks.

This has led to alternative novel approaches that use probabilities, tariff scores and statistical operations to get to the CoD [65, 69]. These approaches are known as Computer Coded Verbal Autopsy (CCVA). Recent studies have mainly focussed on statistical approaches that seek to investigate CoD determination using VA narratives by employing CCVA approaches [69, 67, 94, 30, 17]. Nevertheless, these approaches perform poorly because of the unstructured narratives that they take as input for CoD classification. Therefore, the CCVA approaches cannot be applied to guide health priorities as they fail to avail enough evidence where there is limited expert diagnosis [2]. On the other hand there have been efforts to apply machine learning (ML) approaches, which have shown better results than the CCVA approach. Considerable research was done by applying various shallow ML algorithms which are trained on very high dimensional (where there are many predictors and outcome variables) and sparse features (predictors and outcome variables are not within the same scale) to get meaningful information from the VA narratives [19, 1, 16, 95, 47, 63, 57, 44, 18, 15]. Most of these ML techniques perform poorly because of high dimensionality data caused by lack of feature scaling, missing effective feature engineering techniques, ineffective preprocessing strategies and lack of data balancing.

Advances in technology specifically in ML and natural language processing (NLP) have availed new opportunities and proficiencies in processing and classification of textual narratives. In this study we present novel robust DL architectures that use NLP, specifically tailored to effectively process VA narratives and predict probable CoD.

Furthermore, we aim to assess the potential impact of the DL methods in healthcare and mortality assessment. In so doing we will be able to enhance CoD prediction accuracy by leveraging the advanced capabilities of ensemble DL approaches as compared with existing baseline methods reported in [117]. Our DL methods will further give us insights into how it arrived at a CoD prediction, a step crucial for enforcing transparency and trustworthiness in the use of AI in disease diagnosis in the public health space. As such, this will allow us to explore the potential application of the SEDL in the real world, opening avenues for assessing its applicability beyond the study's context, such as different populations, regions, and datasets. Consequently, discovering new knowledge, presenting valuable insights and semantic relationships between the VA narratives and their corresponding probable CoD.

Even though DL techniques are more accurate in terms of prediction, they have a limitation in terms of complexity and model interpretability (understanding how the model performed in the way it did). Despite these challenges, DL has shown promising results because of it's efficiency, superior performance, integrated feature learning and effective capabilities of attaining end to end learning from complex and multi-modality data [9]. Conversely, conventional statistical and ML techniques require firstly to perform feature engineering to attain effective and robust features representative of the data, and then build predictive models. There is scanty literature, especially on applying advanced ML techniques known as DL models as approaches to solving VA problems.

This study enhances transparency, trustworthiness and accountability in the use of AI-driven healthcare applications. Integrating such an explainable system in CoD diagnosis will minimise diagnosis errors, reduce bias and improve diagnosis turnaround time and cut costs, amongst many benefits. As such, it will enforce effective collaborations on use of AI systems in aiding human interpretation and annotation of disease diagnosis, thus improving the quality of VA reporting and decision making, which is key in informing civil registration systems and strengthening of health priorities.

5.3 Deep Learning in VA

DL functionalities such as automated feature extraction and engineering, has triggered the application of automated diagnosis systems, something that was initially impossible with conventional ML approaches [153]. DL has produced good results and has already been applied in the health domain as in the work of [154, 155, 156, 157]. Other studies by [158, 159, 157] also showed positive results of DL approaches. [160] point out that the DL approaches can improve ML results in various fields such as speech recognition, drug discovery, object detection amongst many. Moreover, DL models with optimal hidden layers have been developed to reveal information not easily detectable with traditional statistical and ML models. These DL architectures can achieve learning of data representation with varying levels of data abstraction when even computational techniques use few processing layers [8]. Feature engineering in DL to extract, represent and select features is done using an end to end system, that learns in an automated fashion from training data. Therefore, these advanced approaches can be relevant in VA data with high-dimensional sparse data.

Initially the Convolutional Neural Network (CNN) architecture was used for image recognition using multiple data arrays. Nevertheless, further inquiries with this architecture showed better results when used for text classification using one dimensional data structures and character levels [161, 159]. Various researchers have applied the basic Recurrent Neural Network (RNN), with some investigators modifying the approach for sentiment analysis producing good results [162, 163, 164, 165, 166, 167]. Other work by [168, 160] uses a fusion of CNN and RNN and attain good results using a hierarchical implementation.

[21] investigated a multi task learning model for CoD classification. They used the CNN and Linear Discriminant Analysis (LDA) for topic segmentation. The predicted key phrase clusters outperformed the LDA and CNN models in extracting keywords from the VA data. They employed a CNN feed forward network that took input of word embeddings, using 10 fold cross validation for optimisation to generate distinct LDA topics. They attained a precision of 0.779, recall 0.778 and F1-score of 0.774. However, they used a small dataset and had word clusters that were more frequent and longer than others creating non-representative features.

[41] also explored character embeddings on VA data to try and establish if their approach can improve CoD classification. They used four datasets with varying disease categories and deduce that character information improves accuracy when used with smaller datasets as the models can handle unknown words and different forms of spelling. They report a significant improvement using models such as CNN. However, their study only used small datasets of less than 1000 cases which varied in size.

[169] used DL interpretable methods to extract CoD from VAs by applying logistic regression, Bayes classifier, random forest ensemble and two variants of RNN DL architectures. Their experiments used three datasets but high scores were attained only for DL architecture on the neonatal dataset with a precision of 63.2%, accuracy of 63.0%, recall of 63.0%, F-score of 61.3% .

[170] experimented on the prediction of death status on treatment course in SARS-COV-2 patients using DL and ML methods and reported an accuracy of 97.15 % on DL approaches and 92.15% for random forest, 93.4% for k-nearest neighbour and 99.7% for the ensemble classifier XGBoost.

[171] did an analysis of railway accident narratives using DL distributed vector representations and RNN and CNN. They achieved a high accuracy 75% and F-score of 0.65 using word2vec as a distributed vector representation of words and CNN.

5.3.1 Ensemble DL

Apart from conventional DL approaches discussed above, there has been studies that applied what are known as ensemble DL approaches (combination of several individual models) in a quest to attain better generalisation performance. Ensemble DL make use of an averaging or voting process of the combined models to get to a final model prediction.

[153] did a study where they applied SEDL using the CNN for pediatric pneumonia diagnosis by using chest X-ray images. They combined conventional traditional ML approaches with DL approaches. They reported an accuracy of 98.3%, precision of 99.29%, recall of 98.36%, F1-score of 98.83% and AUC of 98.24%. They concluded that their stacked ensemble models attained optimal model performance and the findings can be used to assist clinicians in the diagnosis process. Similar work is reported in [172], which point out the robustness of ensemble classifiers on attaining high model performance. However, they highlight the importance of addressing class imbalance for improved results. [173] did a study on application of ensemble DL for COVID-19 case detection using chest X-ray images. They attained an accuracy of 95% and reported that ensemble classifiers can yield better performance. Nevertheless, they point out the need to have more data points for model improvement.

[174] did a review on the various types of ensemble DL approaches, and reported that ensemble approaches are better in terms of model performance. However, they argue that these DL models are difficult to train with smaller training samples, they are complex in nature and selecting models to include in the ensemble architecture is also a challenge. The authors further point out that there is a need for researchers to investigate

how best to find a strategy to define the number of base learners to be included in the architecture, find a criterion for model selection and combine different fusion strategies.

[153] argued that stacking is one effective ensemble DL approach, which entails creating a stack of predictions from base classifiers, also known as individual classifiers. These predictions on base classifiers are used as features, which are taken as input for training the meta learner, which is the final classifier. This results in model improvement as stacking combines the strengths of the base classifiers.

[175] reported on the application of ensemble DL models for heart disease classification using data from Mexico. They report on accuracy and F1-scores within the range of 91% and 96%. They conclude that ensemble DL can attain high model performance, which can be used for real-time reporting and diagnosis in the health space.

5.3.2 Explainable AI

Although DL methods have shown promising results in various domains, they are difficult to interpret. In other words, it is difficult to tell how a model got to a final prediction, due to the complexity. This is the reason why DL models are known as a black box [176, 177, 178]. As such, this has given rise to the field of Explainable Artificial Intelligence (XAI), which strives to ease DL model interpretability. [177] pointed out that XAI approaches can be categorised into intrinsic and post hoc. The authors elaborated that intrinsic approaches are easy to understand in as how they got to the decision making process. On the contrary, post hoc approaches help us understand the context of the input data that leads to a classification decision in any classifier. They usually use visual explanations, local explanations, and explanation by simplification, amongst many functionalities.

[179] argued that XAI can be applied in the medical field to aid clinicians in getting to a diagnosis in a way that is transparent, understandable and explainable. This can be effectively implemented only if ML models are able to justify how they arrived at a decision. Additionally, the authors further elaborate that if XAI is employed properly it will increase the chances of trusting and implementing such strategies within the medical domain. The authors applied XAI to aid in human decision support within the medical space. They applied three different strategies and conclude that if properly implemented and improved, these approaches may prove key in the decision making process for clinicians. [176] also supports the above described notion and argues that most of the DL model predictions are black boxes and they lack interpretability on how the model got to a decision. As such, this makes it difficult for the health experts to be able to understand and trust the results generated by ML models, hence the limited use within the public health space. In order to build trust in model predictions, there is a need to incorporate explainability and interpretation of results generated by the models. This may position DL models as an encouraging choice in its application and usage in healthcare to improve healthcare outcomes.

In this study we used Local Interpretable Model-agnostic Explanations (LIME), because it eases interpretability and provides key meaningful information for decision making [178]. The XAI approach gets to a decision based on validating model classifiers behaviour in close proximity with the cases to be explained based on the local models such as logistic regression or decision trees [179]. [176] pointed out that XAI can be thought of as features that describe how a particular complex model derived its predictions. Therefore, it makes it easier for a user to understand how it got to a decision. XAI highlights the most important and relevant features that it used to get to a prediction. It takes data input from the model, whilst observing any prediction changes in its response. Moreover, it generates a prediction explanation of a single case and not the entire dataset using unseen or test data. It uses individual feature prediction probabilities to explain how it arrived at a prediction

[176]. As such, the results generated by XAI approaches should be visualised, thus making generalisations of findings easier, specifically in clinical settings. Therefore, this will make the DL approaches more reliable and will thus improve such application acceptance levels in healthcare.

This present study investigated the application of SEDL architectures to determine CoDs from VA narratives. There is limited research in the application of XAI on SEDL architectures, specifically in the VA domain. One notable study of [169] used XAI methods, however they did not employ stacked ensemble approaches. The contribution of the current study is to extend current XAI research and its application in healthcare, in order to build trustworthy and easy to understand DL models that are cost effective, timely, error free, consistent and accurate in order to improve VA reporting.

In this study we apply LSTM, CNN and SEDL to determine easily interpretable CoD from VA narratives, thus making it easier to understand how the model arrived to its predicted outcome. The XAI DL approaches can therefore be effectively integrated into the process of identifying mortality causes, alongside human annotation, and interpretation. This can prove very beneficial for clinicians and other interested researchers.

5.4 Materials and Methods

This section will describe the methods used in this study.

5.4.1 Study Design

This study is a retrospective cross-sectional study that uses secondary data analysis. We standardised and normalised our dataset within the Python Spyder environment and created data frames for cleaned VA narratives, model performance and classified VA categories which are stored within a PostgreSQL version 4.2 Relational Database Management System.

5.4.2 Study Population

In this study we used VA data from the Agincourt Health and Demographic Surveillance System (HDSS). The Agincourt HDSS study area came into existence in 1992 and is located in rural north-eastern South Africa. Specifically it is situated in the rural sub-district of Bushbuckridge under the Ehlanzeni District, in the Mpumalanga Province. The Agincourt study area covers approximately 420km^2. According to the Agincourt fact sheet of 2019, the population comprised 116 247 individuals residing in 28 villages with 22 716 households, with 55 961 males, 60 280 females, 11 724 children under 5 years and 928 school going children aged from $5-19$ [98].

5.4.3 Data source and description

This study used the Agincourt HDSS dataset for the period of 1993 to 2015. Our DL models specifically used input from VA narratives to predict CoD in twelve disease categories which were described in the literature [94, 12, 99, 1]. The dataset had 287 columns/features and 16338 records/observations. We therefore, used all the symptoms and the VA narrative column as our predictors X and the corresponding doctors diagnosis International Classification of Diseases 10th Revision (ICD-10) code as our target variable Y with 16338 records.

Table 5.1 depicts the data labelling as in this study for the twelve disease categories.

5.4.4 Pre-processing

Pre-processing addressed data that were incomplete, noisy and inconsistent, through data cleaning, munging, transformation and reduction [180]. Data cleaning was also done to simplify text by removing meaningless text that was deemed irrelevant for the models. This was achieved by converting text to lower case, removing all punctuation marks, spaces, numbers and special characters. Spelling correction was applied using the TextBlob Python library. Tokenisation was employed to create string tokens by splitting of a document. Stopword removal was done to remove irrelevant words using the NLTK library of English stopwords. Lemmatization was applied to convert all possible word variations into the root form known as the lemma. All nulls were dropped to remove bias in the modelling. Feature scaling was done by using the Python StandardScaler to set the magnitude of our predictors X and response variables Y to be within the same range. To remove bias after exploratory data analysis we applied data balancing using the Synthetic Minority Oversampling Technique (SMOTE) for our training dataset, as we had imbalanced classes.

5.4.5 Word Embedding and representation

In this study we used the Global Vectors (GloVe) distributed vector word embedding and representation since we aimed to address the issue of complex and high data dimensionality, resulting in a distributed representation of words in low dimensional space. Distributed vectors mark the beginning of a data processing layer in a DL model and imply that sentences with similar meanings tend to occur in similar context. The novel GloVe was applied instead of term frequency (TF) and Term Frequency with Inverse Document Frequency (TF-IDF), because GloVe is robust in discovering relationships as compared with the traditional TF which has common words dominating the vector space and TF-IDF fails to find similarity between words. The GloVe is described in detail in [81, 83].

5.4.6 Text Classification with Deep Learning architectures

In this study we employed the Long Term Short Memory (LSTM), CNN and SEDL architectures as our text classification techniques. These approaches are able to deal with the challenge of large vocabularies that are made up of unknown words or out of vocabulary words (words that appear a few times in the test set but are not in the training corpus of words), unlike distributed vector representations. In the literature, they are known as character embeddings because each word is considered as no more than a composition of individual letters [22].

Long Term Short Memory

LSTM is a variant of RNN which uses a feed-back loop between layers and thus is dynamic [100]. RNNs are used to process information that is sequential and apply same operation on each sequential instance. RNNs use input that relies on output of prior steps or previous computational output [101]. They make use of memory from previous computations to effectively process the current tasks. RNNs are made up of three layers input, hidden layers and output [100]. However, RNNs have issues of learning long distance associations known as the vanishing gradient (where weighted activation function inputs of a neural network increase or decrease whilst the derivative function approaches zero). Consequently, this ultimately affects accuracy and makes it difficult for the architecture to learn and hyper tune parameters for optimum results. RNNs are described elsewhere in [22].

One variant of RNN which overcomes the vanishing gradient issue is the LSTM. The LSTM has three gates namely; forget, input and output. It is distinctly efficient as compared with other RNN models in that it

can remove information from memory, can save selected information to memory and can focus only on instant important aspects that are relevant [101]. The forget gate differentiates itself from other basic RNN architectures and makes it possible to back propagate an infinite number of times. Furthermore, it calculates the hidden layer by averaging the three layers [22]. It calculates the hidden state by combining the three gates [22]. Figure 5.1 depicts the LSTM architecture and the conventional RNN.

Convolutional Neural Network

Convolutional Neural Network (CNN) was developed as a feature extraction function to extract high level features from n-grams, and thus, these features are used for various NLP tasks [22, 84, 85]. The CNN uses the weight constraint, meaning that a single convolutional layer must have the same weights on inputs for all nodes in that layer. As such, this makes the CNN more efficient than training a basic neural network [101]. The CNN is made up of an input layer, convolutional layers, pooling layers, fully connected layers and softmax function for classification. The functionality of the CNN is described in [22]. Figure 5.2 illustrates the pictorial representation of the CNN architecture.

5.4.7 Stacked ensemble DL models

Figure 5.3 depicts ensemble DL methods using stacking. We employ various base learners ($B_1 - B_n$) that go through several iterations ($Level_1 - Level_n$) and feed into the meta learner (M_{output}) for final model prediction. This study employs stacking ideas formulated in the work of [181]. The dataset is first split into n equal parts, and take each n^{th} fold cross-validation set for testing and the remainder for training the model. We created pairs of train-test sub-datasets to attain various model predictions that were used as input to our meta model for the final prediction. As such, this approach improved model performance and reduces room for bias in results [174].

5.4.8 Explainable AI using LIME

Figure 5.4 is a visual representation of a general schematic diagram of the implementation of XAI using post hoc explanations by a simplification approach LIME. We first feed VA narratives as input into the black box, SEDL models. This is followed by explaining the model output using the LIME component to provide CoD diagnosis and explanations to experts. In this study we follow the XAI mathematical approach (LIME) reported on in [179, 177].

5.5 Experiments

5.5.1 Deep Learning models

This section details the distributed word vector representations, DL architectures and Python libraries used in this study. The Global vectors (GloVe) were used in order to attain a vectorised representation of words using Keras Python library, and we re-trained it with our own corpus. We generated a 128 dimension vector for each word using a window size of 5. The input VA narratives were padded to be of the same size of 250 words for all narratives. As we experimented it was discovered that changing the dimensions to higher values would not have any significant effect. CNN architecture was implemented which consisted of four 1D convolutional layers (128, 256, 256 and 512) with corresponding maximum pooling and drop out layers. The kernel size of the convolutional and max pooling layers was set at 3. We had a flattening layer for generating all input of the convolutional layers into a single long feature vector that was connected to the fully-connected layer. The

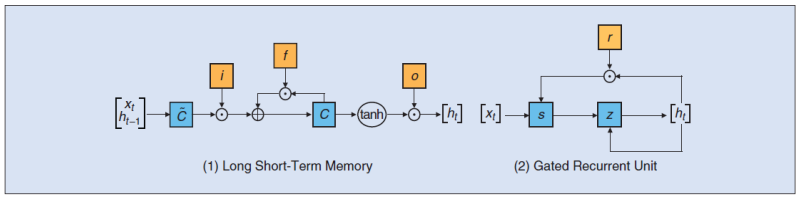


Figure 5.1: LSTM deep learning framework, Source: [22].

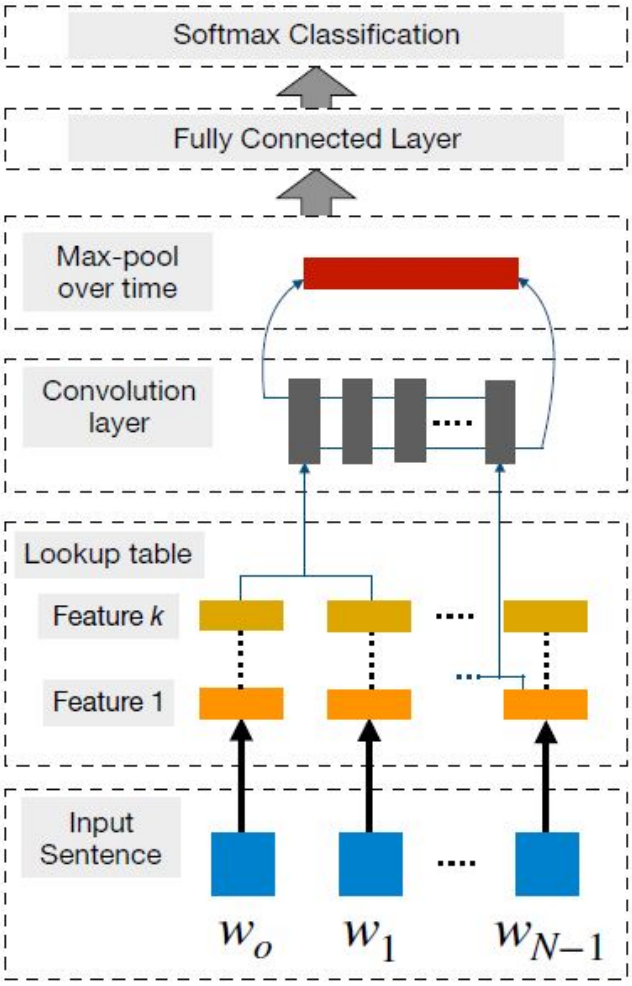


Figure 5.2: CNN deep learning framework, Source: [22].

Table 5.1: Labelling of the twelve disease classes.

Disease category	Class label	Number of samples
HIV/TB	0	3388
Other infectious	1	964
Metabolic	2	242
Cardiovascular	3	140
Indeterminate	4	1468
Maternal and Neonatal	5	121
Abdominal	6	117
Neoplasms	7	93
Neurological	9	57
Respiratory	10	46
Other NCD	11	21

[1] Legend: HIV- Human Immunodeficiency Virus, TB- Tuberculosis, NCD- Noncommunicable diseases.

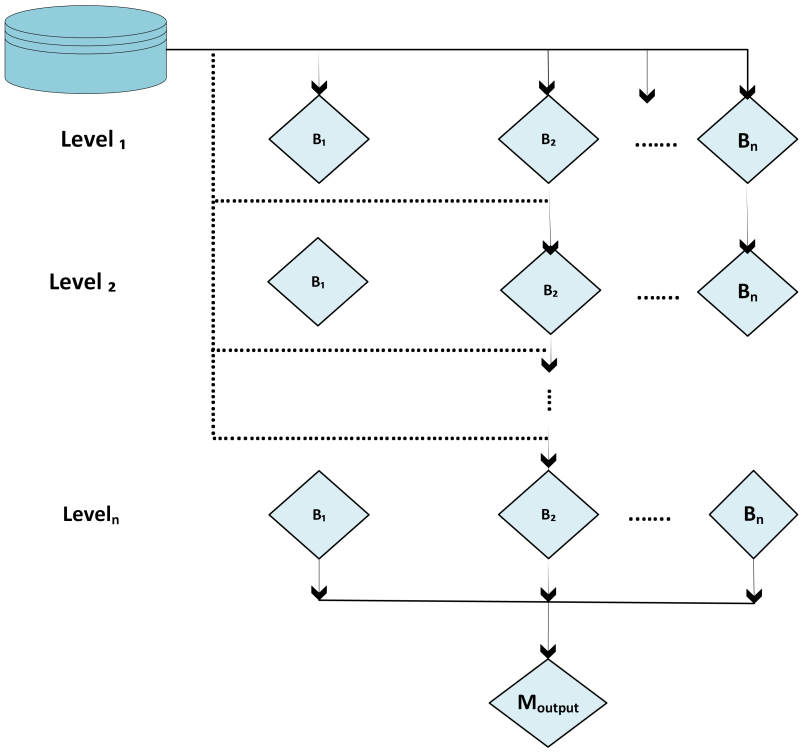


Figure 5.3: Stacking of our models, Source: [174].

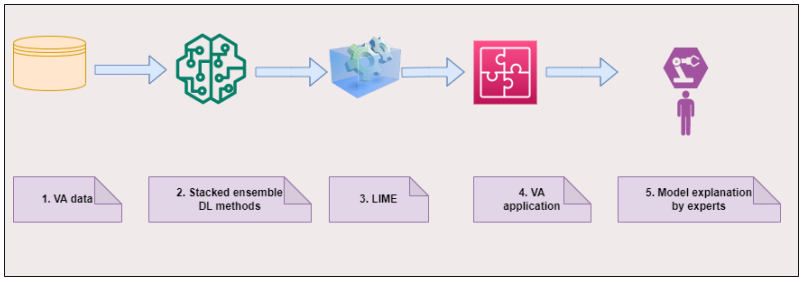


Figure 5.4: General pipeline of XAI implementation using VA data.

fully connected layer was set to 128 units and also used a dropout layer. All the layers used the rectified linear unit (ReLU) activation function except the classification layer which used softmax. For optimisation, cross entropy for loss was used and adam was used as the optimiser. The model was set to early stopping if there was no further improvement in training.

The RNN implementation using the LSTM variant included two Gated Recurrent layers with 256 and 512 nodes respectively. The fully connected layer had 12 nodes at the end and also has a drop out layer with the rate set at 0.002. The classification layer used the softmax as the activation function and we optimised using cross entropy for loss and set our optimiser as adam. We also set our model to early stopping if there was no effective further training.

We fine-tuned batch size and number of epochs as part of batch normalisation where we sought to normalise the distribution of each input dimension of our dataset [143]. Furthermore, we tuned the optimization algorithm used to train the network, each with default parameters. Additionally, we optimised the learning rate which controls how much to update the weight at the end of each batch and the momentum controls how much to let the previous update influence the current weight update. Optimisation was also applied to the selection of network weight initialization by evaluating all of the available techniques. Moreover, we fine-tuned the activation function which controls the non-linearity of individual neurons and when to fire. Drop-out rate fused with weight constraint were optimised by tuning the drop-out rate for regularisation in an effort to limit overfitting and improve the model's ability to generalise. We further tuned the number of neurons in the hidden layer as a larger network requires more training and at least the batch size and number of epochs should ideally be optimised with the number of neurons.

Our conventional DL models failed to converge, as they failed to attain our optimal model performance set on 80 % in our initial work reported in [117]. Therefore, we explored with ensemble DL methods, specifically employing SEDL architectures. A sequential model from the Keras ML library was used. It had one global average pooling layer and four dense layers with 1024, 512, 128, and 64 units respectively. Additionally, to address model complexity and avoid overfitting, we added drop-out layers and employed regularisation using Lasso regularisation. We used adam as our optimiser, set patience to 3, used 500 epochs and a batch size of 128. This formed the basis of our base learners which were made up of 100 members.

5.6 DL Model evaluation

In this study we applied optimisation and performance learning curves to evaluate our DL models, specifically using training and validation datasets to visualise the cross entropy loss and model performance. This was done in various various iterations. We used the optimisation learning curves that depict a training loss and validation loss graph over time. The training loss depicts the extent to which a model is fitting the training data, and on the other hand the validation loss indicates how well the model fits new unseen data. Performance learning curves which use accuracy as a metric to evaluate models were employed. Training accuracy indicates how well the model fits the training data, whilst validation accuracy denotes the extent to which a model fits new unseen data. Learning curves are described elsewhere in the studies of [182, 101, 183].

We also report on the accuracy of our models. Accuracy denotes all classes with classified results that have been predicted correctly in fraction terms [44, 15]. True Positives (TP) and True Negatives (TN) represent the number of outcomes in which our prediction model correctly classifies positive and negative cases respectively. In our study, TP denotes predicted positive VA narratives with a particular disease category from the twelve classes, and are actually positive while TN denotes predicted negative VA narratives with a particular disease

category from the twelve classes and are actually negative. Conversely, False Positives (FP) and False Negatives (FN) denote the number of outcomes where our models incorrectly predicted the positive and negative classes. In our case the FP imply predicted positive VA narratives with a particular disease category from the twelve classes but are actually negative and FN depicts the predicted negative VA narratives with a particular disease category from the twelve classes but are actually positive. The mathematical approach of accuracy is given by;

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.1}$$

5.7 Results

In this section we discuss the results attained from our DL models in the classification of VA narratives into twelve disease categories.

We removed 2247 VA narratives that had null values, and were left with 14091 cases of textual information that were fed into our DL models. We also dropped 5170 target outcome values, that were also missing. The longest narrative had 715 words, thus 3949 characters. On the contrary, the shortest narrative had 7 words, thus 55 characters. Generally our narratives had a mean sentence length of 99.76 words.

Twelve disease classes with 3388 samples each for our training dataset, were created after data balancing using SMOTE. This implied that the majority of classes had more data points compared with other minority classes which were less represented in terms of data samples. Models that lack proper implementation of these data handling strategies, usually suffer from model bias and can cause huge misclassification errors and misinterpretations.

Each VA narrative was taken as a single short document which is made up of a sequence of words. These sequence of words are unigrams (n=1) and bigrams (n=2) which were taken as input by our DL architectures for predicting the probable CoD into any of the twelve disease categories (target class). We sought to predict values of our target features for the test set as unseen data. These text sequences were converted into numeric vector representations of word embeddings using Global vectors (Glove) where our models applied mathematical operations for prediction.

5.7.1 Model performance

Table 5.2 below presents our performance evaluation metrics for the three DL architectures (LSTM, CNN and SEDL) attained by employing optimisation and performance ML curves. The SEDL outperformed LSTM and CNN in the disease classification. Our ML curve depicts SEDL achieving a training accuracy of 82.1%, as compared with LSTM at 76.11% and CNN at 76.35%. In terms of validation accuracy SEDL achieved 82%, compared with LSTM and CNN that attained 67.05% and 66.16% respectively. A test accuracy of 82%, 67% and 66.2% for SEDL, LSTM and CNN respectively. In terms of test loss, SEDL attained 1.15%, LSTM 11.95% and CNN 12.64%. SEDL, LSTM and CNN are respectively approximately 82%, 67% and 66.2% accurate. This implies that all three models predicted the class correctly for 82%, 67% and 66.2% of the samples in the test dataset.

Figures 5.5 and 5.6 depict the DL model performances using learning curves. Our DL models were able to generalise well on new unseen data. This implies that, our models are able to make accurate predictions on new data illuminating the same characteristics of the training data. In order to be able to generalise with our models and avoid overfitting and underfitting, we reduced model complexity by using dropout (20%). We also used early stopping (monitor=validation loss), patience (3), and minimum delta (0.001%) to force the

Table 5.2: Performance evaluation of models.

Model	Training Accuracy (%)	Validation Accuracy (%)	Test Accuracy (%)	Test loss (%)
LSTM	76.11	67.05	67	11.95
CNN	76.35	66.16	66.2	12.64
SEDL	82.1	82	82	1.15

model to stop when there was no further model improvement. The small variance between our validation and test accuracies denotes a good fit of our models.

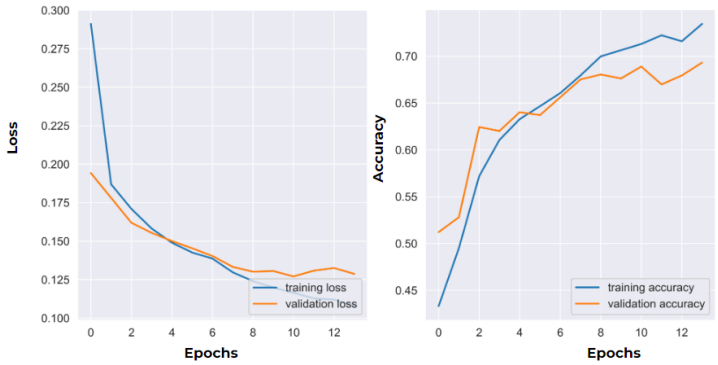


Figure 5.5: Learning curve depicting performance evaluation of CNN

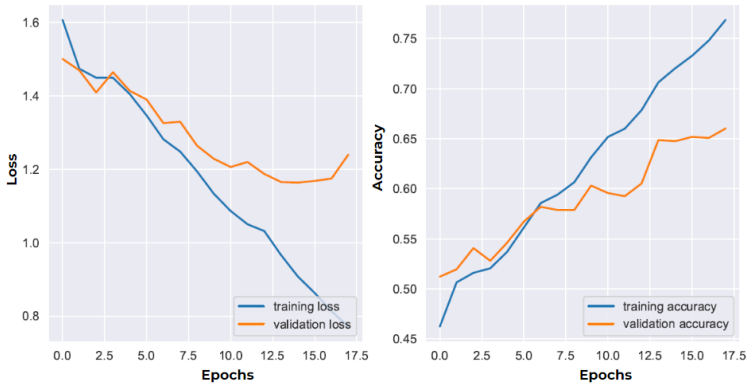


Figure 5.6: Learning curve depicting performance evaluation of LSTM

Figures 5.7 and 5.8 show a learning curve performance evaluation of the stacked ensemble model. The variance between the train and test curve is minimal and thus shows that we did not have any under-fitting or over-fitting in our ensemble models.

5.7.2 Model explainability

This extract discusses model interpretability of our SEDL architectures using LIME. Generally, we attained different CoD predictions and interpretability from our models. However, further insight was derived from the results attained by using LIME to ease explainability.

Figures 5.9 and 5.10 show text narratives of cases diagnosed with metabolic, maternal and neonatal respectively, and were correctly predicted by our SEDL model. The case of a patient who succumbed from metabolic diabetes disease shows that the respondent raised symptoms such as *("high sugar" and "sugar diabetic")*. These

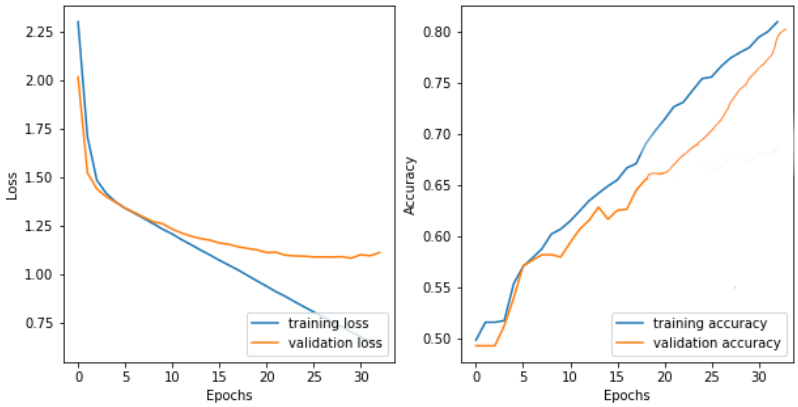


Figure 5.7: Learning curve depicting SEDL performance evaluation during model training

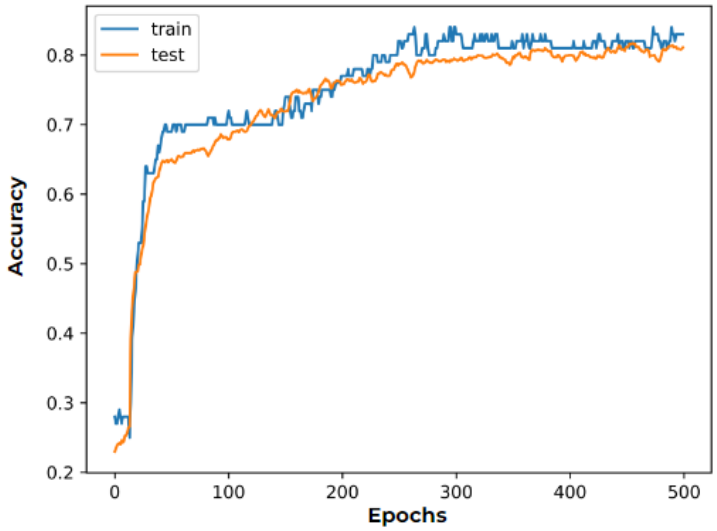


Figure 5.8: Learning curve depicting SEDL performance evaluation

terms also constituted high probabilities in our XAI LIME model. Figure 5.9 depicts another correct prediction by our model. Interestingly, LIME manages to arrive at a prediction based on the patient's death description based on the keywords *(" baby", "incubator" , "mother", "delivery" and "birth")*, which all had high probabilities. It is noteworthy that in certain instances, our models got to some CoDs by inferring symptoms, treatment and laboratory tests done. Moreover, such cases had more correct predictions given by our models.

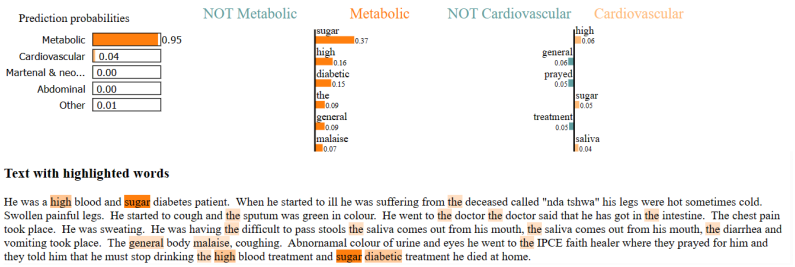


Figure 5.9: Example of LIME correct prediction

Figures 5.11 and 5.12 show cases which were incorrectly predicted by our models as HIV and TB being the CoDs whilst the gold standard were non-communicable diseases and respiratory causes respectively. On the contrary, we found that in certain cases the models had challenges in NLP proficiencies, since some extracts of the narratives had irrelevant historical information of patients which were insignificant to achieve a proper

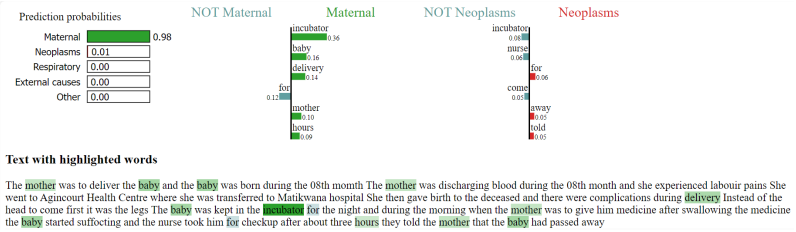


Figure 5.10: Example of LIME correct prediction

CoD. As such, the correct predictions were achieved by paying attention to relevant descriptions that led to death, rather than historical patient information. This is evident in the textual extract given in Figure 5.12.

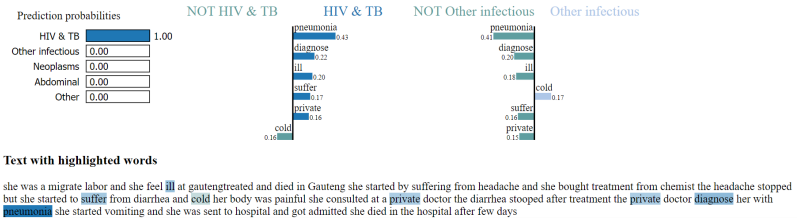


Figure 5.11: Example of LIME incorrect prediction

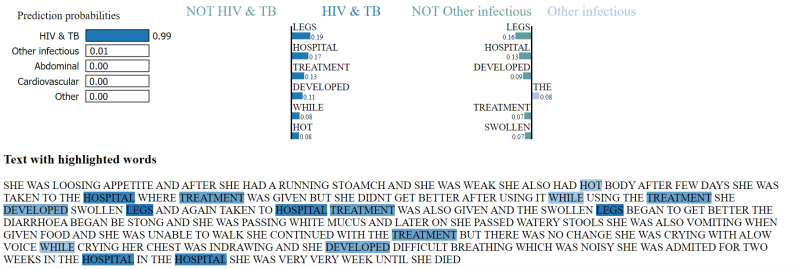


Figure 5.12: Example of LIME incorrect prediction

We designed an experiment that sought to elicit mean opinion scores (MOS) from doctors, based on evaluation of LIME results and model predictions. The goal of this was to assess if the LIME explanations made sense or not. We randomly selected 20 LIME predictions and gave it to 10 doctors for review and assessment. The MOS were denoted by the following terms and corresponding weights;

1. Bad
2. Poor
3. Fair
4. Good
5. Excellent

This calculated MOS value represents the average quality perception of the LIME predictions. We attained a high MOS of 4.21 which implies that the LIME predictions made sense.

5.8 Discussion

Determining CoD from VAs using narratives only, largely remains a manual task that is tedious, time consuming, prone to errors and costly. Specifically, the turn-around time to get a proper diagnosis on CoD from physicians

after reviewing the VA narratives is still a challenge. Despite efforts to improve the VA processes, the VA elicitation process suffers from many drawbacks and still lags behind in the determination of CoD from VAs. This ultimately affects VA reporting as it does not happen in real time, even though it is key in informing civil registration systems and strengthen health priorities.

Generally, we employed novel DL architectures specifically using only VA narratives, with the aim of reducing the turnaround time needed by physicians to give an appropriate diagnosis after reviewing the narratives. Whilst literature mainly uses the structured responses to the VA questionnaire for information extraction, we found the VA narratives having rich and valuable information [1]. Comparison with related work that usually uses the closed ended responses to the structured questionnaires to build predictive models, we have shown that using the VA narratives only produces comparative results which are in close agreement with physician determined CoD. Additionally, to obtain improved information from the VA narratives we need to consider using new novel technological ways such as using tablets for recording the VA process with relatives of the deceased and transcribing it in an automated fashion. Consequently, doing away with summarising the interview, may produce much more reliable results with an error percentage below 6% [169].

In this study we use the LSTM and CNN to determine CoD using VA narratives only. Consistent with previous studies, we discovered that the DL approaches do not always produce the best performance as expected, as they require vast amounts of data and computing resources. Therefore, we faced challenges of few training data samples, as well as high prevalence of out of vocabulary words with high lexical variability of terms (words with broad meaning in the medical domain) [169]. Even though our DL models underperformed, the LSTM produced better results than the CNN. Nevertheless, the LSTM under performance signifies that our model was not effective for sequential modelling of narratives and could not address the vanishing gradient issue. This implies that during training of our model, using gradient methods, our gradient was getting smaller, thus preventing our models to train and learn effectively as weights remained unchanged or not updated. Consequently, this issue might have negatively affected the LSTM's capability of using its forget gate functionality (to forget previous computations by removing it from memory, ability to save only useful information to memory and focus only on components of memory that are recently relevant), possibly suggest such performance.

It follows that the CNN, also performed moderately. The performance of this model suggests that, during training, it did not effectively constraint input weights to be the same for all nodes in each convolutional layer. Furthermore, we noted that we did not have a good number of parameters at each layer in order to attain a balanced efficient training of data points in this model. Moreover, this model did not fully utilise its ability to take combined results from previous layers as input to other layers to attain a cutting edge experience to actually learn more abstract and complex relationships.

As stated by [8], DL approaches require vast amounts of training data in order to perform optimally in classification, thus in cases where there is limited data points it does not achieve the best results as it needs to learn various feature weights and select the discriminative ones for model prediction. Conversely, our initial work reported in [117] proved that conventional approaches are better in CoD determination from VAs as compared with basic DL techniques. In a quest to prove the capabilities of DL approaches, we explored SEDL models.

Our SEDL architectures attained optimal results as compared with conventional DL approaches, as they surpassed the 80% model performance threshold set in our initial work reported in [117]. This can be attributed to the fact that the stacked ensemble classifier managed to effectively improve performance during iterative

model training, and by combining all capabilities of all base classifiers. Our empirical findings suggest that these SEDL architectures can be effectively used for CoD determination using narratives only.

DL models are a black box and difficult to implement as an alternative for the physician CoD gold standard. Additionally, the public health domain still lacks trust in implementing machines for disease diagnosis and prognosis, since this sensitive field requires more accurate models that are easily interpretable. However, the findings of this study position our SEDL framework as an encouraging alternative to physician diagnosis. This is consistent with the findings reported in the study of [169]. This is due to automated techniques enforce a reduction in human error and inconsistencies, thus enforcing standardisation. Additionally, ML models are scalable and can be used to analyse and interpret large datasets, something not possible through human processes. Moreover, automated models are time saving (reduce time needed to review the narratives), cost effective (less costly than the physician), free from human error and enforces consistency in CoD determination (the machine can improve CoD diagnosis and prognosis through model training).

The process of eliciting information from relatives of the deceased remains a huge challenge of the VA process, in terms of interpretability because of the subjective nature of recall bias. Our findings can result in improved efficiency by automating the VA analysis process, thus reducing the time and effort required for human analysis. Therefore, this improves VA reporting and decision making.

Interestingly, it should be noted that clinical data is not always found in large volumes, thus in cases where the datasets are limited or small in terms of training samples, it is better to use conventional ML approaches. A comparative analysis of the results attained in this study and our previous study where we used conventional ML approaches [117], we found that traditional ML approaches are better in terms of performance when using VA narratives for CoD classification. Nevertheless, the novelty of this study is in the use of SEDL frameworks that are easily interpretable using LIME. Moreover, the evaluation done by the physicians presents this study as a promising accurate and cost effective alternative. This further builds trust in the application of machines in the public health space.

This study was limited in the number of training samples which are needed by DL models to improve performance. Moreover, model performance was also affected by high lexical variability of terms in the medical domain. The study was also limited by the length of the recall period which could have created a bias in the collected VA data. In addition, some of the data collectors might have had inadequate training or inconsistencies in interviewing techniques which could have introduced bias into the data collection process. The VA data that we used did not entail native terms, which thus may have impacted the generalisability of the findings and limited a chance in model

5.9 Conclusions

This study showed that our SEDL framework may be included in the process for determining mortality causes, alongside human annotation, and interpretation. As such, the findings of this study may be used as a baseline of building trust in the implementation of machines in disease prognosis and diagnosis in the public health field. Ultimately, this will reduce turnaround time needed for CoD determination, cut costs, strengthen civil registration systems and health priorities as well as inform policy and practice. Future work will entail application of record linkage to identify under reported morbidity occurrences by effectively linking HDSS population data and NHLS public health data registries.

5.10 Author contributions

Conceptualization, M.M., C.K., T.C. and E.M.; M.M., C.K., E.M. and T.C. methodology, M.M., and V.O.; software, M.M. and V.O.; validation, M.M., T.C. and V.O.; formal analysis, M.M. and V.O.; investigation, M.M.; data curation, M.M. and C.K.; writing—original draft preparation, M.M.; writing—review and editing, M.M.; visualization, M.M. and V.O.; supervision, C.K., T.C. and E.M.; project administration, M.M. All authors have read and agreed to the published version of the manuscript.

5.11 Funding

This research received no external funding.

5.12 Ethics

The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of University of the Witwatersrand Faculty of Health Sciences, Human Research Ethics Committee (Medical), (protocol code M1911132 and approved 23 September 2020). Patient consent was waived due to REASON (This study is a retrospective study that used secondary data analysis. We got permission from the Agincourt Health and Demographic Surveillance Site to use their data for research purposes and maintained confidentiality through anonymity.

5.13 Data availability

The data presented in this study are available on request from the corresponding author. The data are not publicly available due to [due to privacy and ethical restrictions].

5.14 Acknowledgments

This work was supported by the Developing Excellence in Leadership, Training and Science (DELTAS) Africa Initiative Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) [Grant No.DEL-15-005]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS) Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [Grant No. 107754/Z/15/Z] and the United Kingdom government. Furthermore, we received support from the University of the Witwatersrand, Faculty of Health Sciences Seed funding.

5.15 Conflicts of interest

The authors declare that they have no competing interests.

5.16 Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CCVA	Computer Coded Verbal Autopsy
CoD	Cause of Death
CNN	Convolutional Neural Network
DL	Deep Learning
XAI	Explainable Artificial Intelligence
LIME	Local Interpretable Model-agnostic Explanations
LSTM	Long Short Term Memory
ML	Machine Learning
NLP	Natural Language Processing
PCVA	Physician Coded Verbal Autopsy
RNN	Recurrent Neural Network
SEDL	Stacked Ensemble Deep Learning
VA	Verbal Autopsy

6 Summary and conclusion

6.1 Introduction

This chapter gives a conclusive synthesis of all the findings presented in prior chapters. This study explored the intricacies of health dynamics within rural South African communities using robust advanced computational techniques, specifically for determining CoD using VA narratives. We started by applying CCVA, conventional statistical and ML techniques, secondly we employed TM techniques to investigate most prevalent mortality diseases. We then used SEDL to encourage the use of machines in the public health space. Our last chapters, are still ongoing and will form part of future work. The sections below will highlight the implications of our findings, contextualize them within existing literature, acknowledge limitations, and suggest potential avenues for future research.

The application of computational approaches on various datasets has shed light on multifaceted aspects of morbidity and mortality patterns, offering insights that have both immediate and far-reaching implications for public health policies and strategies. Chapter 2 has shed light on conventional ML approaches being most commonly used as compared to DL methods. Moreover, we note that most models do not incorporate effective feature engineering approaches and lack efficient exploratory data analysis to fully understand features before modelling. Henceforth, this leads to poor model performance. It should be noted that ensemble approaches yield better results compared to single models. In line with our chapter 5 findings, we discovered that there is need to utilise more transparent DL models, thus building trust in the use of machines for disease diagnosis and prognosis.

Chapter 3 generally, demonstrated that ML techniques can be used as alternatives in determining CoD from VA narratives producing results comparable to physician diagnosis. Specifically, ensemble classifiers produced best results compared to statistical, CCVA and single algorithms. Interestingly, we discovered a higher life expectancy (50 years) in the first decade of the civil registration system in South Africa. However, we witnessed a drop in in following years, as the population was dying at a much younger average age of 40 years, mostly from the leading HIV related causes. Noteworthy, in the third decade, we see a gradual improvement in life expectancy, possibly attributed to effective health intervention programmes. Moreover, we also discovered that people in the HDSS seek traditional ways when faced with terminal illnesses as compared to western ways. The comparative approach also makes this study a baseline that can be used for future research enforcing generalization and comparability.

We employ n-grams and LDA as our text mining techniques to investigate the most prevalent diseases in chapter 4. We found out that the HDSS population mainly succumbed to death due to mortality causes such as *vomiting/diarrhoea, chest/stomach pain, fever, coughing and high blood pressure*. Additionally, we discovered that the most prevalent diseases entailed *human immunodeficiency virus (HIV), tuberculosis (TB), neurological disorders, malaria, diabetes, high blood pressure, chronic ailments (kidney, heart, lung, liver), maternal and accident related deaths*.

In chapter 5, we presented robust DL methods to determine CoD from verbal autopsies, with the stacked ensemble deep learning approaches performing optimally, and better than LSTM and CNN. Our empirical results suggest that both conventional ML and ensemble deep learning methods, could be included in a pipeline for identifying mortality causes, alongside human annotation, and interpretation. Ultimately, this will reduce the turnaround time needed by physicians to go through the narratives in order to able to give an appropriate diagnosis, thus cutting costs and minimise errors, among many benefits.

Our robust predictive models, developed through conventional ML, DL algorithms and statistical modelling, have provided a compelling depiction of disease prevalence and mortality causes in rural South Africa. Additionally, the automated techniques have enabled a more granular analysis, revealing specific associations that might have been overlooked using traditional methods. The results highlighted possible impact of socio-economic factors on health outcomes, with distinct correlations between disease risk factors and disease burden. Interestingly, we note that mortality rates might be influenced by limited access to healthcare resources. These include lack of effective screening tests, primary healthcare systems, and chronic based delivery. Consequently, this creates distinct health disparities, thus there is a crucial call for targeted interventions to improve healthcare accessibility in marginalized communities.

There are various implications of our research findings, generated by our computational predictive models for the identification of disease hotspots and related mortality rates. Generally, they offer a tool for optimizing resource allocation, enabling health stakeholders to strategically distribute resources to areas with elevated health risks. The prevalence of communicable and non-communicable diseases underscores the necessity of integrated healthcare strategies. As such, socio-economic determinants and their effect on rural setting health outcomes, should be taken into account in policy formulation. Interestingly, the integration of computational insights with cultural beliefs and socio-economic context has availed more nuanced understanding of health dynamics, such as the need to understand patient care seeking patterns when faced with terminal illnesses. This can further make us understand the pivotal influencers in shaping health-seeking behaviours and health outcomes. Therefore, this kind of contextualization provides a foundation for designing interventions that resonate with the communities' unique needs, ultimately contributing to more effective and culturally sensitive public health initiatives.

6.2 Strengths of this study

This thesis on computational approaches to characterize morbidity and mortality in rural South Africa holds significant strengths due to the potential to address pressing healthcare challenges and contribute to scientific knowledge. The highlights of this study's strengths, span from, long period covered by the study, robustness and rigour of the computational approaches which used vast datasets, model performance comparability of various statistical, ML and DL approaches, as well as the comprehensive application of various models. This study managed to give us data driven insights by leveraging the use of diverse large datasets to identify morbidity and mortality patterns, trends, risk factors, and potential correlations. Something that could not have been possible through manual conventional approaches.

The use of computational approaches provided a more accurate approach of identifying mortality trends, thus proving very vital in resource constrained regions that can benefit from targeted interventions to inform policy and practice. Moreover, the application and deployment of computational models can act as early warning systems or syndromic surveillance systems for disease outbreaks, that can be used to predict the at-risk populations, especially in rural areas with limited healthcare. The automated approaches can therefore prove very helpful for timely resource allocation and implementation of preventative measures. The findings of

this study can be used to guide policy and strengthen health priorities by providing evidence based solutions into the most significant health challenges in rural South Africa. Interestingly, our computational models are scalable, and can be used to address topical health issues in larger geographical areas.

The use of a large dataset over a long period of time, is essential in having a deep understanding on the evolving nature of various disease and healthcare needs through trends over time. As such, the computational approaches are resource efficient as make it possible to get results in a short time frame as compared to conventional approaches that are time consuming and resource intensive. Consequently, this promotes innovativene and encourages the use of machines in public health. Therefore, this research supports an interdisciplinary approach, that can lead to innovative solutions that tackle complex health challenges from multiple angles. Ultimately, this study enforces broader implications for similar rural and under-resourced populations worldwide, thus the methods used in this study could be adapted to address health disparities in other regions.

6.3 Limitations of this study

This study had several limitations, that span from the nature of the datasets being heterogeneous, having lexical variability, incomplete and unstructured among many drawbacks. Moreover, this study used VA data, which may suffer from reporting biases, long recall periods and incomplete records. Additionally, the computational algorithms employed, while effective, do not account for nuanced factors that influence health outcomes. We also had a limitation that we used a corpus which entailed English terms only and we dropped the native terms. Perhaps, adding a translation module might have resulted in better improved model performance. Future research could explore machine learning techniques to integrate socio-economic and cultural variables, allowing for a more comprehensive analysis.

6.4 Future Research Directions

While our study has made substantial strides in characterizing morbidity and mortality patterns in rural South Africa using computational approaches, several avenues for future research remain open. Firstly, we aim to investigate the morbidity dimension of this study through application of record linkage on HDSS and NHLS registries, to investigate under-reported morbidity and mortality. Furthermore, we will also apply unsupervised learning approaches to understand mortality clusters based on household characteristics and other factors. The clustering approach on a longitudinal dataset will assist in tracking changes in morbidity and mortality patterns over time. Consequently, this could shed more light into the effectiveness of interventions. Moreover, applying a more generalised approach of our models and findings to other regions globally can present a comparative analyses across different rural regions. Therefore, this has the potential of revealing region-specific determinants and solutions. Integrating qualitative research could delve into the underlying cultural and social determinants that drive health behaviours, offering a more comprehensive picture of health dynamics. Lastly, there is need to investigate if incorporating and integrating a translation module to cater for non-english terms to be part of the corpus can improve model performance.

6.5 Conclusion

In conclusion, our computational analysis of morbidity and mortality patterns in rural South Africa provides valuable insights into the complex interplay of disease burdens, socio-economic factors, and cultural influences. By addressing the limitations and embracing the interdisciplinary nature of healthcare, we can drive positive

change and ultimately contribute to improved health outcomes in these communities. The integration of computational approaches with localized context has proven to be a potent tool in characterizing morbidity and mortality patterns in rural South Africa. Our study demonstrates the potential for data-driven strategies to drive evidence-based interventions, ultimately improving the health outcomes of vulnerable populations. The insights garnered from our study have direct policy and practical implications. Our findings can inform the design of healthcare initiatives that address the evolving health landscape of rural South Africa. Policy makers should prioritize investments in healthcare infrastructure, access to clean water, and health education. Moreover, the computational techniques we employed can be adapted for real-time monitoring of health trends, enabling swift responses to emerging health crises. As we move forward, the collaboration between computational researchers, public health experts, and local communities will be instrumental in addressing the intricate health challenges faced by rural South African regions and the world at large.

A Appendices

A.1 Equations

A.1.1 Feature reduction techniques

Pearson correlation

The Pearson correlation is given by:

$$r(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})}} \tag{A.1}$$

.

Chi-squared

$$x^2 = \sum \left(\frac{(O_i - E_i)^2}{E_i} \right) \tag{A.2}$$

.

Information gain

The following equation as defined by [44] denotes information gain: Gain(D,A)=I-$\sum_{i\varepsilon values(A)} \left(\frac{(t_i)}{(S)} \right) I_i$, where D is the dataset, A is the attribute, I is the number of information before split and $\sum_{i\varepsilon values(A)} \left(\frac{(t_i)}{(S)} \right) I_i$ is the sum of information after split where I is the information node for I and t is the number of objects in node i.

Principal component analysis

[49] reports that PCA aims at defining a low dimension space that represents a dataset with minimal variation. Given n observations that exist in p varying dimensional space are defined by PCA linear approach of the p features. Therefore the principal components of set features $\{X_1, X_2, ..., X_p\}$ is the normalised linear combination of features

$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + ...\phi_{p1}X_p$ with highest variance in the normalised form denoted by:

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1 \tag{A.3}$$

,

where $\phi_{11}, ..., \phi_{p1}$ are loadings of the first principal defined in vector $\phi_1 = (\phi_{11}\phi_{21}...\phi_{p1})^T$.

The PCA aims at maximising the following given A.3:

$$maximise \quad \phi_{11}, ..., \phi_{p1} \left\{ \ \frac{1}{n} \sum_{i=1}^{n} (\sum_{j=1}^{p} \phi_{j1} x_{ij})^2 \ \right\}$$

A.1.2 Statistical equations

Logistic regression

$$ln \left(\frac{P(w = 1|X)}{P(w = 0|X)} \right) \tag{A.5}$$

.

Naive Bayes

The Naive Bayes rule classifier assumes that features in a dataset are mutually exclusive and categorises for class w_i **if** $\forall j \neq i$:

$$P\left(w_i|x\right) \geq P\left(w_j|x\right) \tag{A.6}$$

,

where $\mathrm{P}(w_i|x) = \left(\frac{p(x|w_i)P(w_i)}{p(x)} \right)$, and $p\left(x\right) = \sum_{j=1}^{2} p(x|w_j)P(w_j)$.

A.1.3 Machine Learning equations

k-nearest neighbour

The Euclidean distance is calculated as follows:

$$\sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{A.7}$$

.

Random Forests

The voting of prediction is given by the following:

$$f = \frac{1}{B} \sum_{b=1}^{B} f_b(x) \tag{A.8}$$

.

The diagram below illustrates the RF approach:

Support Vector Machine

The SVM approach aims at maximising the margin $\|e\| = \sqrt{\sum_{i=1}^{N} e_i^2}$.The SVM principles can be thought of a hybrid of linear and non-linear and is derived from the structural risk minimisation principle.SVM employs a maximal margin hyperplane where it computes the maximum linear distance between separated classes in

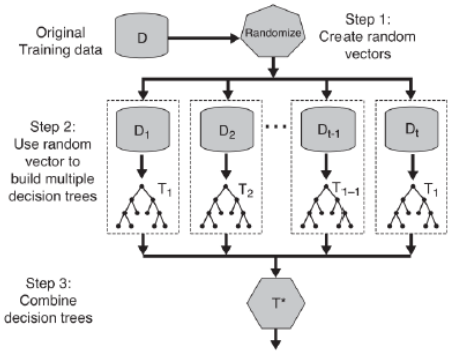


Figure A.1: Random Forests

feature space. In non-linear cases the SVM has the ability to use kernels which can map non-linearity between classes, categories and feature space. Classification of the correct class employs the following equation:

$$e = \left\{ \begin{array}{l} 1, \\ \quad \text{if } b + e^T x^k \geq +1 \\ -1, \\ \quad \text{if } b + e^T x^k \leq -1 \end{array} \right\}$$

Assume we are given a set S of points $X_i \in R^n$ with $i = 1, 2, ..., N$. Each point X_i, belongs to either of two classes and thus is given a label $Y_i \in \{-1, 1\}$. Hyperplane divides S leaving all the points of the same class on the same side. The classification is performed by constructing an N dimensional hyperplane that optimally separates the data into two categories. Learning in this respect is finding the maximum margin separating the hyperplane between two classes of points. All training points have associated coefficients with them. The coefficients express the strength with which the points are embedded in the final decision function for any given test points. By using kernel function K such that;

$$x_1, x_2 \in XK(x1, x2) = \langle \phi(x1) \cdot \phi(x2) \rangle$$

.

A.1.4 Artificial Neural Networks

The ANN is made up of the input, hidden and output layers. It applies the error function at the output layer to generate updated weights using back propagation. The model is as below:

$$error function = (w_{actual} - w_{predicted})^2 \tag{A.11}$$

A.1.5 k-means clustering

k-means clustering is of the opinion that: $C_1 \cup C_2 \cup ... \cup C_k = \{1, ..., n\}$ denotes an observation falling into at least a cluster and $C_k \cap C_{k'} = \emptyset$ for all $k \neq \quad k'$ implying that none of the given observations belong to one class.

We aim at minimising the following problem:

$$minimise \quad c_1, ..., c_k \left\{ \ \sum_{k=1}^{K} W(C_k) \ \right\}$$

,

where the euclidean distance that minimises the rate at which observations from various classes differ from each other is given by:

$$minimise \quad c_1, ..., c_k \left\{ \ \sum_{k=1}^{K} \frac{1}{|C_K|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \ \right\}$$

.

Boosting

The equation below is an AdaBoost classifier that denotes a class where α_t is the weight of the classifier given by $h_t(x)$. The classifier is given by the following:

$$w = sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \tag{A.14}$$

.

Bagging

The equation below denotes the bagging classifier, that takes K bootstraps from datasets and and creates a classifier for every bootstrap using the the majority vote to K classifiers:

$$w = \arg\max_{y \varepsilon Y} \sum_{k=1}^{K} 1(y = h_k(x)) \tag{A.15}$$

,

where

$$1(y = h_k(x)) = \begin{cases} 1, & \text{if } y = h_k(x) \\ 0, & \text{if } y \neq h_k(x) \end{cases} \tag{A.16}$$

.

A.1.6 Deep Learning equations

A.1.7 Distributional vectors

Diagram below adapted from [22] denotes distributional vectors where $D << V$, where V is size of Vocabulary.

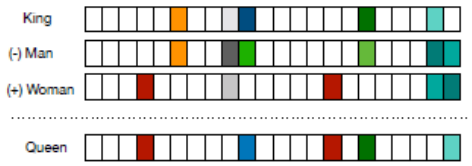


Figure A.2: Distributional vectors represented by a D-dimensional vector

A.1.8 Neural Language Model

Diagram below adapted from [184] as cited in [22] graphically represents the Neural Language Model where C(i) is the ith word embedding.

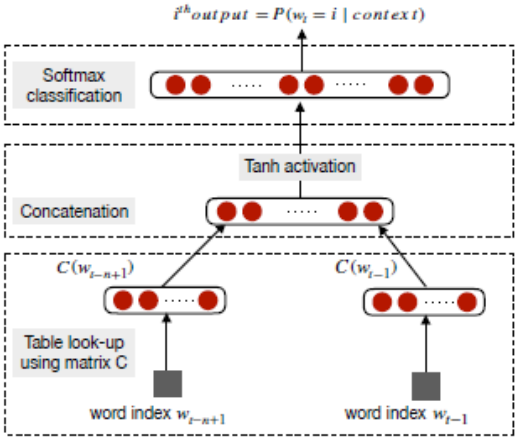


Figure A.3: Neural Language Model

A.2 Word embeddings

Word2vec

The Word2vec makes use of the input, projection and output layers [81]. The input layer takes one hot word vector of V neurons and corresponding to hidden layer N neurons.The softmax is output that contains all words in the vocabulary. Each word in the vocabulary is denoted by two learned vectors v_c and v_w where the connected layers are given the following weight matrix[22]:

$$W \epsilon R^{V \times N} \tag{A.17}$$

, and

$$W' \epsilon R^{H \times V} \tag{A.18}$$

The model suggests that the k_{th} word in the vocabulary will have corresponding context and target representations denoted by:

$$v_c = W_{(k,.)} \quad and \quad v_w = W'_{(.,k)} \tag{A.19}$$

.

[22] defines w_i with a given context of word c as input denoted by:

$$P(\frac{w_i}{c}) = y_i = \frac{e^{u_i}}{\sum_{i=1}^{V} e^{u_i}} \tag{A.20}$$

,

where

$$u_i = V_{w_i}^T . V_c \tag{A.21}$$

.

The objective function is the log likelihood and finding its gradient where the parameters $\theta = V_w, V_c$ are learned through the objective function [22] and is denoted by:

$$l(\theta) = \sum_{w \epsilon Vocabulary} log(P(\frac{w}{c})) \tag{A.22}$$

, which is equivalent to:

$$\frac{\partial l(\theta)}{\partial V_w} = V_c(1 - P(\frac{w}{c})) \tag{A.23}$$

, where the CBOW model takes the one-hot vector of word context as input in the form:

$$h = W^T(x_1 + x_2 + ... + x_c) \tag{A.24}$$

.

Below is the diagrammatic representation of Word2vec:

CBOW model

The CBOW model [**p.~4**, 22] is depicted below:

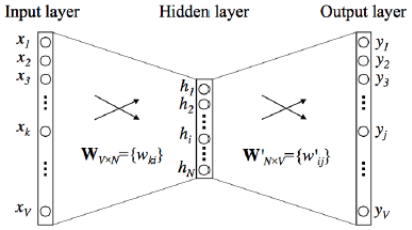


Figure A.4: Model for CBOW

Sentence2vec

Sentence2vec represents each sentence in a document as a vector, unlike Word2vec which represents each word as a vector [83]. Their sentiments are augmented by the mathematical model that; let all the vocabulary in the corpus be $V = w_1, w_2...w_n$ where, n is vocabulary size V .The vector representation of sentence S will be given by vector $S = f_1, f_2...f_n$ where f_i is the extracted word feature of w_i in sentence S. This implies that we can derive f_i by computing the frequency of word w_i in sentence S. We can further deduce that f_i is the TF-IDF.

The sentence vectorisation applies the following formula:

$$\overrightarrow{V}_d = \frac{\sum_{i=0}^{n} \overrightarrow{v}_i}{n} \tag{A.25}$$

.

Global Vectors

This approach first constructs the co-occurrence matrix X from a training corpus where,
X_{ij} denotes frequency of word i co-occurring with word j

$X_{ij} = \sum_{k}^{V} X_{ik}$ denotes total occurrences of word i in the corpus and V is corpus size. w

Secondly is the process of vectorisation by factorising X. By this process we apply the following equation:

$$F(w_i - w_j, \tilde{w}_k) = \frac{P_i k}{P_j k} \tag{A.26}$$

,

where w_i, w_j and w_k are three words in a vector, $P_i k = X_{ik}/X_i$ is probability k occurring in context of the word i.Thereby w are word vectors and $\tilde{w}_k$ are context vectors. To maintain linearity and avoid mixing dimensions, we apply vector differences and dot products of our parameters in above equation to derive:

$$F((w_i - w_j)^T \tilde{w}k) = \frac{P_i k}{P_j k} \tag{A.27}$$

,

Resolving the symmetry and maintain same mathematical model we exchange $w \rightarrow \tilde{w}$ and $X \rightarrow X^T$ to derive:

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = log(X_{ik}) \tag{A.28}$$

The least squares objective function for weighting given by $f(x)$ will be denoted by:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - log(X_{ij}))^2 \tag{A.29}$$

.

A.2.1 Character embeddings

Convolutional Neural Networks (CNN)

This approach takes sequence input $s_1, s_2 ... s_n$ of n words and transforms to a vector $w_{s_1}, w_{s_2}, ... w_{s_n}$ applying the lookup table to each of the words [22]. The equation below denotes CNN sentence modelling where for each sentence the word embedding is given by $w_i \epsilon R^d$ and represents the i^{th} word, where d is the word embedding dimension. This implies that given a sentence with n words we can denote it using the embedding matrix $W \epsilon R^{n \times d}$ and if $w_{i:i+j}$ is the concatenation of vectors $w_i, w_{i+1}, ... w_j$. Convulsion takes place at the input embedding layer and applies a filter $k \epsilon R^{hd}$ to a window of h to generate a new feature. Below we have a new feature c_i that is produced using the window of words $w_{i:i+h-1}$.

$$c_i = f(w_{i:i+h-1}.k^T + b) \tag{A.30}$$

, where $b \epsilon R$ is the bias term and f is the non-linear activation function and k is the filter function applied to all possible windows using same weights to create the feature map given by:

$$c = [c_1, c_2, ... c_{n-h+1}] \tag{A.31}$$

.

Below are two figures illustrating the CNN framework for word and text prediction respectively as cited by [22] in [75]:

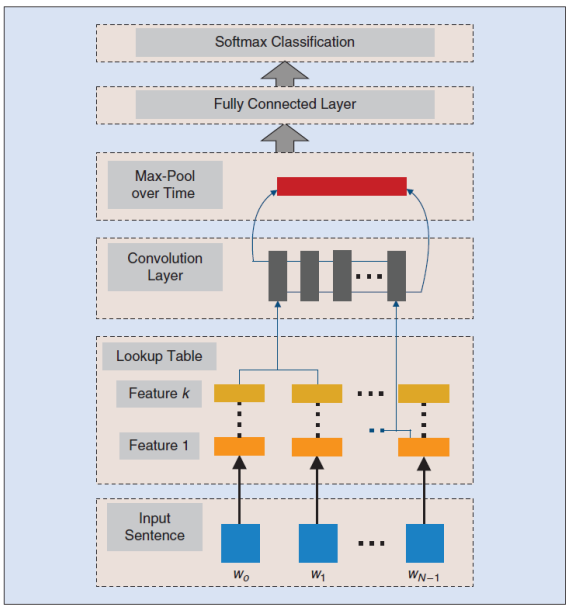


Figure A.5: CNN framework for word prediction

Simple Recurrent Neural Networks

The mathematical model is given by the following equation:

$$s_t = f(Ux_t + Ws_{t-1}) \tag{A.32}$$

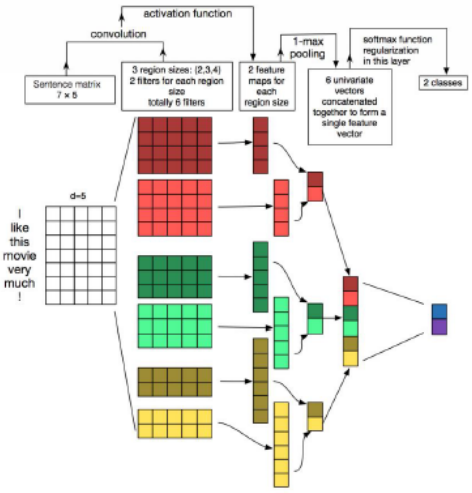


Figure A.6: CNN modelling on text

, where x_t is the input to network at time step t and s_t is the hidden step at same time step, thus s_t is computed based on current input and previous time step's hidden state, function f assumed to be a non-linear transformation and x_t is made up specifically of one hot embeddings [22].

Recurrent Neural Networks

Below is the RNN model as cited by [22] in [75]:

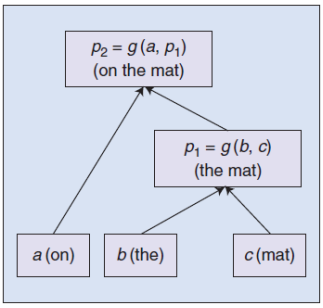


Figure A.7: Simple RNN network

A.2.2 Long Short-Term Memory RNN

It calculates the hidden state by combining the three gates and is denoted by the following equations:

$$x = \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} \tag{A.33}$$

,

$$f_t = \sigma(W_f.x + b_f) \tag{A.34}$$

,

$$i_t = \sigma(W_i.x + b_i) \tag{A.35}$$

,

$$o_t = \sigma(W_o.x + b_o) \tag{A.36}$$

,

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_c.X + b_c) \tag{A.37}$$

,

to derive a one hot word vector in the form:

$$h_t = o_t \odot tanh(c_t) \tag{A.38}$$

.

Below is the LSTM and GRU model as cited by [22] :

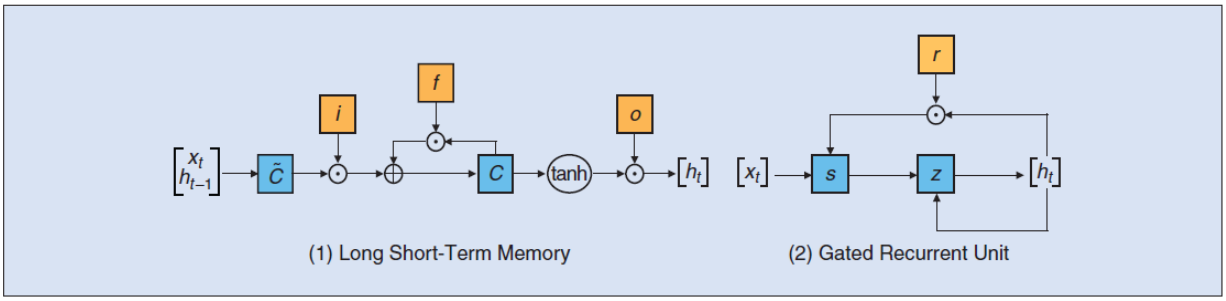


Figure A.8: Illustration of an LSTM and GRU gate

Gated RNN

The Gated recurrent unit is modelled in the following equations:

$$z = \sigma(U_z.x_t + W_z.h_{t-1}) \tag{A.39}$$

,

$$r = \sigma(U_r.x_t + W_r.h_{t-1}) \tag{A.40}$$

,

$$s_t = tanh(U_z.x_t + W_s.(h_{t-1} \odot r)) \tag{A.41}$$

, to get the one hot word vector:

$$h_t = (1 - z) \odot s_t + z \odot h_{t-1} \tag{A.42}$$

.

Matrix Vector Recursive Neural Network

The mathematical model is denoted by:

$$p_1 = tanh(W \begin{bmatrix} Cb \\ Bc \end{bmatrix}), P_2 = tanh(W_M \begin{bmatrix} B \\ C \end{bmatrix}) \tag{A.43}$$

, where $b, c, p_1 \epsilon R^D, B, C, P_1 \epsilon R^{D \times D} and \quad W_M \epsilon R^{D \times 2D}$.

Recursive Neural Tensor Network

The network compositional function g on word and phrases representations (b, c or a, p_1 to produce a higher level phrase($p_1 \quad or \quad p_2$), thus representation of all nodes take the same form. Moreover the compositional function g where each node is a d -dimensional vector and $W \epsilon R^{D \times 2D}$ is denoted by the following equation:

$$p_1 = tanh(\begin{bmatrix} b \\ c \end{bmatrix}^T V^{[1:D]} \begin{bmatrix} b \\ c \end{bmatrix} + W \begin{bmatrix} b \\ c \end{bmatrix}) \tag{A.44}$$

,

where $V \varepsilon R^{2D \times 2D \times 2D}$ is a tensor that denotes several bilinear states.

A.2.3 Transformer architecture

Below is the transformer architecture as cited by [92] :

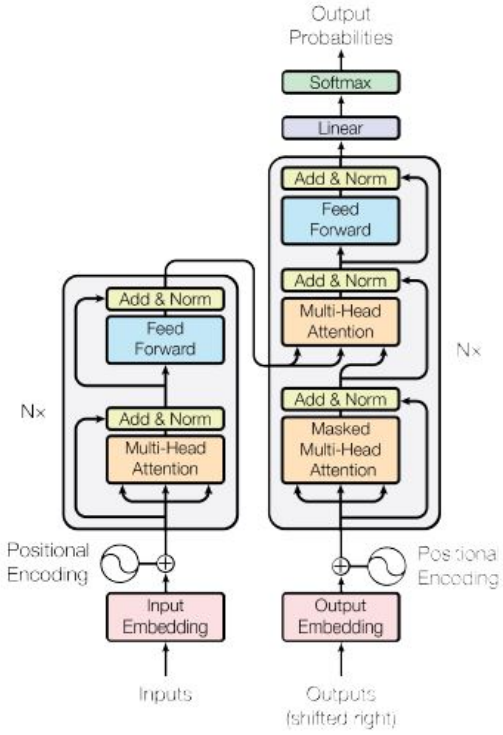


Figure A.9: Illustration of the Transformer Architecture

[92] illustrates the transformer deep learning architecture and uses a decoder and encoder that are connected layers with structured attention. The encoder takes symbol representations in a sequence $(x_1, ..., x_n)$ to continuous sequenced representations $(z_1, ..., z_n)$, thus decoder uses z sequenced outputs $(y_1, ..., y_m)$ features one at a time. Each phase the model takes previous input to generate the next symbols. The encoder has 6 stacks of similar layers which have 2 sub layers apiece for attention and and feed forward functionality. The decoder is made up of 6 stacked similar layers and 2 sub layers apiece. It then has a third sub layer which functions as multi head attention over output from the encoder. The attention has mappings to requests and uses keys with values that correspond to output. It creates vectors for requests/ queries, keys, values and output. The end result is a calculated weight of all values using a compatibility function of the request and the related

key. It applies the scaled dot product that takes input as queries of dimension d_k and values d_v. It applies the softmax function to obtain weight by taking the dot product of the query and keys and divide each by $\sqrt{d_k}$. The mathematical matrix output model is denoted below:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{A.45}$$

, where attention is a function on query sets that are computed at once and loaded to matrix Q with keys and values loaded to matrices K and V.

The multi-head attention attends to varying representations at differing positions in space by permitting model in a combined way. The multi-head is denoted by:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{A.46}$$

, where

$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$,

where the projections are matrices with parameters:

$W_i^Q \in R^{d_model \times d_k}, W_i^K \in R^{d_model \times d_k}, W_i^V \in R^{d_model \times d_k}$ and $W^O \in R^{hd_v \times d_model}$.

The feed forward network of the decoder and encoder is denoted by:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{A.47}$$

.

A.2.4 Performance Metrics

Macro precision

The macro precision will be given by the following mathematical model:

$$Precision_M = \frac{\sum_{i=1}^{c} \frac{TP_i}{TP_i + FP_i}}{C} \tag{A.48}$$

,

where

TP_i denotes true positives (instances correctly predicted as required),

FP_i denotes false positives (instances incorrectly predicted as required)

C denotes positive cases.

Macro recall

The mathematical model is given by:

$$Recall_M = \frac{\sum_{i=1}^{c} \frac{TP_i}{TP_i+FN_i}}{C} \tag{A.49}$$

,

where

TP_i denotes true positives(instances correctly predicted as required),

FN_i denotes false negatives (instances incorrectly predicted as not required) and C denotes positive cases.

Macro F-measure (F-measure

The mathematical model is given by:

$$F - measure_M = \frac{(\beta^2 + 1)Recall_M \times Precision_M}{\beta^2(Recall_M + Precision_M)} \tag{A.50}$$

.

Overall accuracy

The mathematical model is given by:

$$Accuracy_Avg = \frac{\sum_{i=1}^{c} \frac{TP_i+TN_i}{TP_i+FN_i+TN_i+Fp_i}}{C} \tag{A.51}$$

,

where

TP_i denotes true positives(instances correctly predicted as required),

TN_i denotes true negatives(instances incorrectly predicted as not required),

FN_i denotes false negatives (instances incorrectly predicted as not required) and C denotes positive cases.

Area under curve (AUC)

The mathematical model below denotes the AUC:

$$AUC = \frac{s_0 - n_0(n_0 + 1)/2}{n_0 n_1} \tag{A.52}$$

, where n_0 and $n_1$denote positive and negative examples and s_0=$\sum r_i$ where r_i is the rank of the i_{th} positive example in the ranked list.

Overall chance-corrected concordance (CCC)

This metric is denoted by the following equation:

$$CCC_j = \frac{\frac{TP_j}{TP_j+TN_j} - \frac{1}{N}}{1 - \frac{1}{N}} \tag{A.53}$$

, where TP_j is the number of true positives correctly assigned or predicted for cause$j$, and TN_jdenotes true negatives for cause j,thus it implies number of cases where the cause assigned is not j whilst the true cause is j [66].

Cause Specific Mortality Fractions (CSMF) accuracy

The mathematical model is denoted by the following equation:

$$ACC_{csmf} = 1 - \frac{\sum_{c=1}^{C} \mid CSMF_c^{true} - CSMF_c^{pred} \mid}{2(1 - minCSMF^{true})} \tag{A.54}$$

.

Top cause accuracy

The model is given by the following equation:

$$ACC_1 = \frac{\#ofcorrectCODbeingfirstcauseassignment}{N} \tag{A.55}$$

.

Top 3 cause accuracy

The model is of the form:

$$ACC_3 = \frac{\#ofCCODwithinthefirst3causeassignments}{N} \tag{A.56}$$

, where CCOD denotes correct cause of death.

A.2.5 Validation

k-fold cross validation Error estimator

If we have a dataset S_n that is partitioned into k folds randomly of similar size $P = \{P_1, ..., P_k\}$ and supposing that n is a multiple of k. Let $T_i = S_n$ be complement of data set P_i ,then algorithm $A(.)$ induces a classifier from $T_i, \psi = A(T_i)$, and predicts estimation error with P_i. Therefore the k-vc error estimate will be given by :

$$\bar{\epsilon}_k(S_n, P) = \frac{1}{n} \sum_{i=1}^{k} \sum_{(x,c)\epsilon P_i} 1(c, \psi_i(x)), \tag{A.57}$$

,

where $1(i, j) = 1$ iff $i \neq j$ and zero [50].

k-cross fold validation for a single algorithm [89] assumes that, given a dataset D that is divided into disjoint folds $F_1, F_2, ..., F_k$ and if $|D| = n$ and $|F_j|$ is the instance number in D and F_j, thus $n = km$. Moreover let r_j be number of correct predictions on F_j by algorithm A, and let p be prediction accuracy of A on all data points in set D . This implies that [89, 88];

Theorem 1: If $r_j \geq 5$ and $m - r_j \geq 5$ for $j = 1, 2, ..., k$, then the sampling distribution of the mean accuracy $\overline{p} = \frac{\sum_{j=1}^{k} p_j}{k}$ can be estimated by a normal distribution with mean p and variance $\frac{p(1-p)}{n}$ where $p_j = \frac{r_j}{m}$ for $j = 1, 2, ..., k$.

Proof: The above applies if every pair of instances in D are independent , and $F_i \cap F_j = \emptyset$ for any $i \neq j$. [89] argues that if the above is true it follows that:

$$E(\overline{p}) = \frac{\sum_{j=1}^{k} E(p_j)}{k} = p \tag{A.58}$$

,

and

$$Var(\overline{p}) = \frac{\sum_{j=1}^{k} Var(p_j)}{k^2} = \frac{p(1-p)}{n} \tag{A.59}$$

.

k-cross validation for comparing two algorithms assumes that; let p_A and p_B be prediction accuracies of classes A and B on data set D that will be divided into folds $F_1, F_2, ...F_k$, and A and B are trained by instances in D/F_j and tested by F_j in the jth iteration. Assuming that $|D| = n$ and $|F_j| = m$ for $j = 1, 2, ..., k$.. Let r_{ij} be number of instances in F_j correctly classified by by algorithm i for A and B and $j = 1, 2, ...k$. To assume appropriateness of our estimator $p_A = p_B$ we compute

$$\overline{d} = \frac{\sum_{j=1}^{k} d_j}{k} \tag{A.60}$$

, where

$$d_j = \frac{(r_{Aj} - r_{Bj})}{m} \tag{A.61}$$

, for

$$j = 1, 2, ...k \tag{A.62}$$

.

B Appendices

frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 27 September 2022
DOI 10.3389/fpubh.2022.990838

Check for updates

OPEN ACCESS

EDITED BY
Yi-Ju Tseng,
National Yang Ming Chiao Tung University, Taiwan

REVIEWED BY
Shibiao Wan,
St. Jude Children's Research Hospital, United States
Jin Cho,
University of Tennessee at Chattanooga, United States
Igor Saveljic,
University of Kragujevac, Serbia

*CORRESPONDENCE
Michael T. Mapundu
michael.mapundu@wits.ac.za

SPECIALTY SECTION
This article was submitted to Digital Public Health, a section of the journal Frontiers in Public Health

RECEIVED 10 July 2022
ACCEPTED 31 August 2022
PUBLISHED 27 September 2022

CITATION
Mapundu MT, Kabudula CW, Musenge E, Olago V and Celik T (2022) Performance evaluation of machine learning and Computer Coded Verbal Autopsy (CCVA) algorithms for cause of death determination: A comparative analysis of data from rural South Africa. *Front. Public Health* 10:990838. doi: 10.3389/fpubh.2022.990838

COPYRIGHT
© 2022 Mapundu, Kabudula, Musenge, Olago and Celik. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Performance evaluation of machine learning and Computer Coded Verbal Autopsy (CCVA) algorithms for cause of death determination: A comparative analysis of data from rural South Africa

Michael T. Mapundu[1]*, Chodziwadziwa W. Kabudula[1,2], Eustasius Musenge[1], Victor Olago[3] and Turgay Celik[4,5]

[1]Department of Epidemiology and Biostatistics, School of Public Health, University of the Witwatersrand, Johannesburg, South Africa, [2]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), University of the Witwatersrand, Johannesburg, South Africa, [3]National Health Laboratory Service (NHLS), National Cancer Registry, Johannesburg, South Africa, [4]Wits Institute of Data Science, University of the Witwatersrand, Johannesburg, South Africa, [5]School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa

Computer Coded Verbal Autopsy (CCVA) algorithms are commonly used to determine the cause of death (CoD) from questionnaire responses extracted from verbal autopsies (VAs). However, they can only operate on structured data and cannot effectively harness information from unstructured VA narratives. Machine Learning (ML) algorithms have also been applied successfully in determining the CoD from VA narratives, allowing the use of auxiliary information that CCVA algorithms cannot directly utilize. However, most ML-based studies only use responses from the structured questionnaire, and the results lack generalisability and comparability across studies. We present a comparative performance evaluation of ML methods and CCVA algorithms on South African VA narratives data, using data from Agincourt Health and Demographic Surveillance Site (HDSS) with physicians' classifications as the gold standard. The data were collected from 1993 to 2015 and have 16,338 cases. The random forest and extreme gradient boosting classifiers outperformed the other classifiers on the combined dataset, attaining accuracy of 96% respectively, with significant statistical differences in algorithmic performance ($p < 0.0001$). All our models attained Area Under Receiver Operating Characteristics (AUROC) of greater than 0.884. The InterVA CCVA attained 83% Cause Specific Mortality Fraction accuracy and an Overall Chance-Corrected Concordance of 0.36. We demonstrate that ML models could accurately determine the cause of death from VA narratives. Additionally, through mortality trends and pattern analysis, we discovered that in the first decade of the civil registration system in South Africa, the average life expectancy was approximately 50 years. However, in the second decade, life expectancy significantly dropped, and the population was dying at a

much younger average age of 40 years, mostly from the leading HIV related causes. Interestingly, in the third decade, we see a gradual improvement in life expectancy, possibly attributed to effective health intervention programmes. Through a structure and semantic analysis of narratives where experts disagree, we also demonstrate the most frequent terms of traditional healer consultations and visits. The comparative approach also makes this study a baseline that can be used for future research enforcing generalization and comparability. Future study will entail exploring deep learning models for CoD classification.

KEYWORDS

cause of death, machine learning, Verbal Autopsy, CCVA, algorithms

1. Introduction

More than 65% of the population in the world lacks high quality information on the cause of death (CoD) since every year about sixty million deaths worldwide are not assigned a medically certified cause (1). As such, most of the countries in the world fail to meet the United Nations 90% death registration coverage requirement, as deaths in many Low to Medium Income Countries (LMICs) are not captured in civil registration systems (2, 3). On the contrary, the CoD information is vital for public health monitoring, informing critical health policies and priorities. Therefore, in the absence of clinically oriented sources, CoD information should be derived from alternative sources. Verbal Autopsy (VA) is the most used tool worldwide as an alternative source of CoD information. VA is common in LMICs and is a process that is used to determine CoD where deaths occur outside health facilities and is not certified by a medical practitioner (4). These sentiments are supported by Mapoma et al. (5), who also reports on the importance of the VA process in determining CoD in countries where there are no active civil registration systems. The VA process is conducted by non-medical personnel who seek to elicit valuable information using both structured questions and an open narrative section with the next of kin of the deceased about circumstances and events that led to death (1). Two doctors are given the full set of responses, both from structured questions and open narratives for assessment and to reach a consensus on the CoD and if not a third physician is consulted, a process known as Physician Coded Verbal Autopsy (PCVA). PCVA is the most used process for determining CoD. However, it is widely criticized because of its lack of robustness, cost, time, inconsistencies, and inaccuracies as it is subjective and prone to errors among many drawbacks (6). This results in PCVAs mostly employed for the training and validation of computational approaches. The surge of technological advances has availed a plethora of automated methods for determining CoD which are faster, efficient, and cost effective (1). Most of the research that reports on ML applications in the VA domain mainly uses the responses from the questionnaire as the classical dataset. As such, this affects comparability and generalisability. In this study, we validate the performance of various ML techniques using various VA data types for determining CoD using a comparative analysis approach. We apply enhanced data standardization and normalization strategies to achieve optimum transparency and accuracy through addressing most model limitations and applying recommendations that are reported in Reeves and Quigley (7) and Mujtaba et al. (8). We assess the robustness of several classifiers including; random forest (RF), k-nearest neighbor (KNN), decision tree (DT), support vector machine (SVM), logistic regression (LR), artificial neural network (ANN), Bayes Classifier (BC), bagging and eXtreme Gradient Boosting (XGBoost) as ensemble classifiers. We also validate our dataset using the common conventional Computer Coded Verbal Autopsy (CCVA) algorithm; InterVA.

1.1. Computer Coded Verbal Autopsy algorithms

Previous studies report on the most commonly used VA algorithms also known as CCVA algorithms. These CCVA approaches use expert-driven rules to determine CoD from VAs (9–13). The VA algorithms make use of the responses from the standardized structured World Health Organization questionnaire that denote signs or symptoms based on the deceased health history prior to death. Most of these VA algorithms take input from VA data derived from real deaths, and symptom-cause information (SCI) which is a repository of information about symptoms that are related to each probable CoD. Additionally, they make use of logic that entails a logical algorithm that combines the SCI and VA data to identify cause-specific mortality fractions (CSMF), so as to assign a specific CoD.

The InterVA uses the Bayes rule to compute the probability of cause of death, given the availability of indicators such as SCI from the VAs. This approach is reported in the study of Clark

et al. (14), Leitao et al. (15), Miasnikof et al. (13), and Murray et al. (16).

These VA approaches have been widely criticized in terms of their credibility and reliability. The study of Kalter et al. (17) reports on the evaluation of VA expert algorithms and deduces that population level accuracy is similar to that of ML approaches with CSMF in the range of 57−96%. Similar findings are also presented in the study of Quigley et al. (18) who did a study where they validated data derived algorithms against the gold standard of physician review using various disease categories based on the CSMF. Leitao et al. (15) argues that, there is little evidence to justify the CCVA as a possible replacement of the gold standard which is the PCVA. Therefore, there is a need for further investigations and research with large datasets to train and test models on CoD classification.

Little research exists in the VA domain on the application of ML to determine CoD from VA narratives. These ML algorithms make use of automated computer programs that can take input of data to learn new trends and patterns from complex data by applying optimization techniques for VA classification (19).

1.2. Machine learning in VA

Most ML model predictors commonly use only responses from the standardized questionnaire, attaining Sensitivity scores of around 60%for individual CoD classification, using various numbers of CoD categories. On the contrary, the study of Jeblee et al. (1) demonstrates that the VA narratives have valuable rich information that can be used for CoD determination. ML can avail real-time results that are similar to that of physicians/experts (20). Alternative complex ML approaches exist in the literature and can be used as substitutes for the PCVA and CCVA algorithms as approaches to determining CoD.

Moran et al. (21) applied the Bayesian hierarchical factor regression models to infer CoD using VA narratives and report an improvement in model performance on inferring CoD and CSMF. However, they used thirty-four disease categories. Idicula-Thomas et al. (22) applied six different ML algorithms (SVM, ANN, KNN, DT, C5.0, and gradient boosting). Their results report the SVM as the best classifier with an Accuracy of more than 80%. However, they used six disease categories. Similar results are reported in the study of Mujtaba et al. (23), with SVM attaining a Precision of 78.1%, Recall of 78.3%, F-score of 78.2%, and overall Accuracy of 78.25%for 16 disease categories. Their study used text classification techniques to predict CoD from forensic autopsy reports. Other studies by Danso et al. (24), Mujtaba et al. (25), and Koopman et al. (26) also found similar results and deduce that feature extraction approaches are grossly affected by variations in words and word combinations.

The study of Mwanyangala et al. (27) used the LR model to determine the completion rate of VA and factors associated with undetermined CoD. They report a completion rate of 83–89%. They ascertain that 94% of deaths submitted to physicians were assigned a specific cause, and on the other hand, 31% were labeled as undetermined. Quigley et al. (28) reports various common diseases that lead to death using CSMF and LR classifier and they achieved 80% Specificity. Boulle et al. (29) applied ANN to classify CoD from VAs and achieved a Sensitivity of 45.3%. They concluded that more explorations are needed with large datasets and large training samples to improve the results of the ANN. The study of Flaxman et al. (30) used the RF classifier to assign CoD categories and affirmed that the RF algorithm performed better if not as the PCVA approach. Additionally, they point out that the RF classifier was better than PCVA on overall chance concordance and CSMF accuracy for both adults and children.

Related work that has also used VA data for cause of death determination is also reported elsewhere in Danso et al. (31). They conclude that using word occurrences produced better results as compared to word occurrence features and suggest using large datasets in order to improve model performance. Their sentiments are echoed in the study of Pestian et al. (32) and Murtaza et al. (33). Additionally, Mujtaba et al. (8, 23, 25, 34) have done vast work in the VA domain and argue that unigrams are better feature extraction techniques, Term Frequency (TF) and TF-IDF are better feature representation schemes, and Chi-squared is a better dimensionality reduction approach. They recommend employing effective data cleaning strategies and feature engineering techniques to get improved performance.

Despite the reported results in the literature, both CCVA algorithms and ML models applied to VA data to determine CoD, suffer from challenges and limitations as they lack concrete evidence where there is a limited expert diagnosis and cannot be fully utilized to inform health priorities (2). Most of the CCVA approaches use statistical concepts and scores to determine CoD (9, 35). Moreover, these approaches are affected in terms of optimal performance because of their dependency on sample size, age group, causes of death, and characteristics of the sample (4, 13, 17, 35, 36). Other issues that affect VA data quality, emanate from having interviewers being untrained, incompetent, and unqualified to appropriately elicit relevant and appropriate symptoms on causes of death. Additionally, language barriers call for the need for the interviewer and interviewee to speak the same language so as to derive the best results. Soleman et al. (4) recommended incorporating fully trained multiple translators. The other downside is the length of the recall period which can create a bias in the collected VA data. The heterogeneity of various autopsies in terms of the non-intersecting dialects of the English language (terms being in the native language) compromises data quality as most of these approaches tend to omit such autopsies in their model prediction, yet they might entail valuable information.

All the discussed challenges and limitations affect the VA data quality that is taken as input to the CCVA and ML approaches. Therefore, we can deduce that there is a great need to address these challenges in order to remove room for any bias and misinterpretations of the models, thereby enforcing generalisability and comparability. This study demonstrates the robust assessment of ML approaches and CCVA algorithms in determining CoD, thus availing a baseline ML framework that can be used for comparability and generalization across all VA dataset types.

2. Methods

2.1. Study design

This is a retrospective cross-sectional study that uses secondary data analysis. All the cleaned VA datasets, model performance, and classification results of various tasks are pushed from a Python Jupyter Notebook environment and housed within a PostgreSQL Version 4.2 object-relational database management system.

2.2. Population

This study uses VA narrative data from the study area of the Agincourt Health and Demographic Surveillance System (HDSS). The HDSS came into existence in 1992 and is located in the rural Sub-district of Bushbuckridge under Ehlanzeni District, in Mpumalanga Province, in north-eastern South Africa. The study area covers approximately 420 km^2. According to the Agincourt fact sheet of 2019, the population was at 1,16,247 individuals residing in 28 villages with 22,716 households, with men being 55 961, women being 60,280, children under 5 years being 11,724, and school going children with ages from 5 — 19being 35,928 (37).

2.3. Data source

The source of our data for this study is the Agincourt HDSS. It is a surveillance site that specifically provides evidence based health monitoring that seeks to strengthen health priorities, practice and inform policy. The VA narratives data is from 1993 to 2015. However, physician diagnosis was done from 1993 to 2010, and this target variable of the doctors' diagnosis is enough for model training and prediction.

In this study, we used three types of datasets such as the responses from the standard questionnaire, narratives, and a combination of the responses and the narratives. The whole dataset had 287 columns/features and 16,338 records/observations. For the responses only, we took all features that had responses from the standard questionnaire as our predictors and the CoD assigned by physicians using the International Classification of Diseases-10 (ICD-10) code for each record in the dataset as our target variable. Ultimately, we had 231 predictors (all symptoms, age at death, and gender) and 1 target variable, and all our features were in English. The predictions using the narratives were done using age at death, gender, the narrative feature, and 1 target variable.

TABLE 1 Twelve disease classes and the number of data samples before and after data balancing.

Class labels and corresponding number of samples			
Disease category	**Label**	**Samples before data balancing**	**Samples after data balancing**
HIV/TB	0	3,388	3,388
Other infectious	1	964	3,388
Metabolic	2	242	3,388
Cardiovascular	3	140	3,388
Indeterminate	4	1,468	3,388
Maternal and Neonatal	5	121	3,388
Abdominal	6	117	3,388
Neoplasms	7	93	3,388
External causes	8	89	3,388
Neurological	9	57	3,388
Respiratory	10	46	3,388
Other NCD	11	21	3,388

For the combined VA dataset, we used 232 predictors and 1 target variable. We only added the VA narrative feature to the responses dataset in order to have our combined dataset. We further created 12 CoD categories with corresponding labels, and class distribution with the number of samples for each class before and after data balancing for our training dataset, as shown in Table 1. The CoD categories were derived based on the InterVA user guide and literature studies of Byass et al. (11), Danso et al. (24), King and Lu (38), and Jeblee et al. (1).

Figure 1 illustrates the logical steps that we follow for this study's experiments. We first do data acquisition of our VA narratives as a comma separated value text file (csv), followed by data exploration and cleaning. Additionally, we do feature engineering and data balancing and feed our data to our models for training, validation, and testing. Finally, we do CoD classification.

2.4. Data pre-processing and encoding

For the questionnaire responses dataset, we cleaned and replaced all nulls with zeros, implying that there was no symptom assigned for a missing value in a record. All symptoms

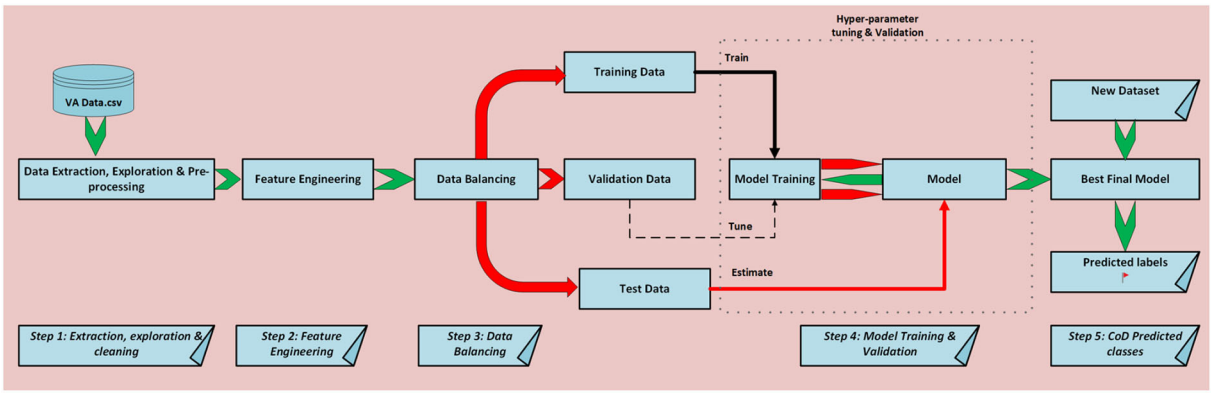


FIGURE 1
Schematic diagram of ML process followed.

that had a 'Y' were encoded as a 1 meaning that the record had a present symptom value. On the other hand, all symptoms that had an 'N', were encoded as a 0 meaning that those records had no symptoms present. In order to normalize and standardize the narrative feature used with the combined dataset, we pre-processed in order to retain with only relevant data. Data were first imported in comma separated value format, followed by pre-processing. The pre-processing stage entailed converting all text to lowercase and removing all punctuation, spaces, numbers, and special characters. Tokenisation was done by splitting a document (seen as a string) into tokens. Stopword removal was then applied to do away with insignificant words using the NLTK library of English stopwords. We applied normalization using the Python spacy package, a process known as lemmatization. Lemmatization uses a dictionary of known word forms and considers the role of a word in a sentence with the aim of extracting some normal form of a word. Finally, we applied feature engineering to determine the most representative features, as we then aimed at retaining only relevant words in the vector space by applying a weighting scheme (39). All categorical data was encoded using the one-hot encoding technique to create numeric vectors. This was followed by concatenating the narratives and the questionnaire response datasets using horizontal stacking which was pushed to our models for training, validation, and testing.

2.5. Feature engineering

We did feature engineering in order to derive new input features from existing ones. This process was done in three phases namely; feature extraction, feature selection, and feature value representation. Feature extraction was applied in order to get only relevant and useful features from textual data using n-gram models. The n-grams are a set of words that are sequential as they make use of the continuous number of items such as characters or words from a given sequence of narratives. n-gram models can be of the form; a) $n = 1$ (unigram), b) when $n = 2$ (bigram), c) $n = 3$ (trigram), and d) hybrid-grams (mixture of unigram, bigram, and trigram) (8, 23). This was followed by the feature value representation stage employing the TF-IDF approach. In this phase, we sought to create a numeric vector of features, where each feature will have a corresponding numeric value that can be used for model learning. TF-IDF considers a feature important if it occurs frequently in the VA narratives belonging to one class and less frequently available in narratives belonging to another class. Finally, we applied feature selection in order to attain the most useful subset of features from the narratives. This was achieved using Singular Value Decomposition (SVD) as a selection approach to reduce the dimensionality of our feature space, thus removing noise in our dataset. This dimensionality reduction technique creates a matrix that only has relevant information producing an exact representation of data in a low dimensional space without any loss of data (40, 41).

2.6. Data balancing and feature scaling

We applied data balancing to the training set to address data imbalance where one or more classes are less represented than the other classes, meaning that the majority classes have more samples as compared to other minority classes. As such, this creates a bias in the minority classes as they will have fewer data points that can cause large misclassification errors. The ratio of the majority against the minority class was 1 : 160. In order to address the issue of data imbalance, we explored various techniques (under sampling, over sampling, threshold, and class weight). We attained optimal results when using the Random Over Sampling Examples (ROSE) and Synthetic Minority Oversampling Technique (SMOTE). After our experiments, we chose SMOTE as the best choice for our dataset. This

possibly suggests that our dataset was well suited for SMOTE as a data balancing technique. Moreover, our balanced datasets behaved better than imbalanced datasets. SMOTE was applied by generating artificial samples for the minority class, through interpolation between the positive instances that lie together. This approach addresses the issue of over-fitting caused by the general oversampling approach that replicates existing positive cases (8). We ended up having 3, 388samples per class. We did feature scaling using the Python Standard scaler library in order to get all our features within the same range as the target variable. After data balancing and feature scaling, we fed the data into our 12 models for training and validation.

2.7. Machine learning models for CoD prediction

We specifically applied supervised ML techniques to predict our target variable given input data. We aimed at predicting the related CoD by taking input of; questionnaire responses only, narratives only, and combined questionnaire responses and narratives. The input was then fed into nine classifiers (SVM, DT, XGBoost, KNN, RF, Bagging, LR, BC, and ANN). These ML approaches are reported elsewhere (1, 22, 23, 27, 33, 34, 39–47). Using the questionnaire responses only, we created a feature space made up of binary responses as predictors and our target variable was a categorical ICD-10 code for CoD. Similarly, we did the same for the narratives only dataset. For the combined dataset only, we added the narrative column to the list of our predictors.

2.8. Model training, validation, and testing

In this study, we perform multi-class classification, where we generated individual prediction models for each of the 12 disease categories. Data were split into 70% training, 20% validation, and 10% testing on unseen or new data, for all our nine models. We evaluated model performance by assessing the robustness of the nine classifiers by applying 10-fold cross-validation supplemented by the GridSearch algorithm. The k-fold cross-validation (k=10 in our study) is advantageous in that, it uses all observations for both training and validation, with each observation used for validation exactly once. On the contrary, this approach has the disadvantage of having to define the number of folds manually. In order to address the limitations of the k-fold cross-validation technique, we also used the automated GridSearch approach that eliminates the random setting of parameters and chooses optimum parameters automatically for a specific model.

In order to attain a better estimate of the generalization performance, we used 10-fold cross-validation to evaluate the performance of each parameter combination, instead of using a single split into a training and validation set. First, we specified the parameters for searching stored in a dictionary. GridSearch cross validation function then performed all the necessary model fits. All dictionary keys were the names of the parameters that we wanted to tune, and the values were the parameter settings that we wanted to test out. Applying cross-validation, we managed to choose the optimal parameters that gave us the best model performance based on the accuracy of the test set or unseen data.

We used optimisation parameters such as; cost complexity pruning and tuning parameter alpha through k-fold validation (tree based models). Moreover, we also used the Mean Squared Error (MSE) and Cross Entropy Error (CEE), Minkowski and Gini as cost functions to compute the minimal cost error between our predictor and the response using the k-fold cross-validation approach to optimize model performance. These cost functions are described in Zaki and Meira (40). Additionally, we also employ $L_1$and L_2regularization approaches to further optimize some of our models. L_1regularization involves eliminating features that are not useful for model prediction by setting some weights close to zero. On the contrary, L_2regularization tends to penalize large weights more and small weights less (41). Table 2 depicts some of the model hyperparameters used in our models.

TABLE 2 Model optimal hyperparameters.

Selected hyperparameters	
Model name	**Hyperparameters**
XGBoost	L1, max_depth=10, objective=multi:softmax, learning_rate =0.1, alpha=0
RF	gini, max_depth =10, n_estimators=100, min_samples_leaf=1
ANN	relu, alpha=0.0001, solver=adam
KNN	minkowski, n_neighbors=5, p=2
SVM	gamma=scale, kernel=rbf, C=1.0
Bagging	KNN, max_samples, max_features
DT	gini, min_samples_split=2, min_samples_leaf=1,
LR	$L2, C = 1.0$
BC	alpha=1.0, fit_prior=True, class_prior=None

XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbor; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier.

3. CCVA algorithms

We followed the same preprocessing steps of our dataset and fed it into our commonly applied CCVA algorithm InterVA. The data preprocessing steps entailed de-duplication based on the identifier field (ID), dropping observations with peculiar IDs, filtering out observations with recorded age at death above 110 years, and any observation with the year of death before 1992

and after 2016. All records with unspecified sex were dropped from the raw dataset. All modeling for the InterVA was done in R. Libraries such as knitr were used for dynamic report generation, lubridate was used for date and time functions, tidyr for organizing and tidying of data, tidyverse for loading core packages, ggplot for plotting graphs, readxl for reading our excel raw data, and InterVA for our CCVA algorithm. In order to determine the most probable CoD, we used the InterVA libraries for analysis in our R-statistical analysis software guided by the study of Li et al. (48) and McCormic et al. (12). Since InterVA and InSilico are correlated, we decided to only validate the InterVA algorithm for comparability with ML approaches.

4. Identification of contradicting cases and best model predictors

In order to identify contradicting cases, where physicians were not agreeing on the diagnosis, we extracted a separate dataset. We used simple text mining techniques known as n-gram models for identifying the contradicting cases and best features for our models (refer to Section 2.5).

4.1. ML techniques model evaluation

Performance evaluation of classifiers is evaluated using various metrics and we report the metrics based on studies by Mujtaba et al. (23, 34). We validated our results using one vs. all with Accuracy, Precision, Recall, F-score, and AUROC as our metrics for evaluation.

Accuracy denotes all classes with classified results that have been predicted correctly in fraction terms. Precision also known as the Positive Predictive Value (PPV) defines the proportion of VA narratives correctly predicted as positive to the total of positively predicted VA narratives. Recall also known as Sensitivity or True Positive Rate (TPR) defines the proportion of VA narratives correctly predicted as positive to all VA narratives in the actual positive category. F-measure computes the average or harmonic mean of Precision and Recall.

True Positives (TP) and True Negatives (TN) represent the number of outcomes in which our prediction model correctly classifies positive and negative cases, respectively. In our case, TP denotes predicted positive VA narratives with a particular disease category from the 12 classes and are actually positive and TN denotes predicted negative VA narratives with a particular disease category from the 12 classes and are actually negative. Conversely, False Positives (FP), and False Negatives (FN) denote the number of outcomes where our models incorrectly predict the positive and negative classes, respectively. As in our case, the FP implies predicted positive VA narratives with a particular disease category from the 12 classes but are actually negative and FN depicts the predicted negative VA narratives with a particular disease category from the 12 classes but are actually positive.

The AUROC visualizes the TPR against the false positive rate (FPR). The area under the ROC curve applies the principle of plotting a curve specific to a machine learning algorithm where the classifier is evaluated relative to a weighting on the area under the curve. Good performance of the algorithm is given a weight of close to 1, thus graph is AUROC closer to the upper left corner and the poor performance of an algorithm is given a weight of 0.5 and below. Specificity computes the ratio of negative VA narratives that are correctly predicted as negative.

4.2. CCVA techniques model evaluation

We explicitly validated the InterVA algorithm using CSMF accuracy and Overall Chance-Corrected Concordance (CCC). CCC computes the accuracy of individual cause assignment and ranges from 0 to 1 and the lower the CCC, the larger the error type on the accuracy of the underlying cause (14, 49). On the other hand, CSMF accuracy defines accuracy as having a value between 0and 1. This metric assumes the worst possible case for predicting CSMF and assigns a weight on the least possible CSMF value that matches the total absolute error (50).

5. Statistical analysis

We applied statistical tests for comparing the performance of our nine algorithms. We computed the variance of our models using descriptive statistics such as mean and standard deviation based on the results of our AUROC. Moreover, we computed some tests using 10-fold cross-validation using the mean and standard deviation. Furthermore, we conducted some non-parametric tests since our data distribution was non-normal using the Kruskal-Wallis test, to test if the model's mean is different or the same. For the Kruskal-Wallis test, we considered $p < 0.005$statistically significant. We applied the pairwise model comparisons using McNemar statistical tests, in order to be able to state objectively whether one model performs better than the other (51). Since we did eight different tests, we used the Bonferroni corrected p-value of 0.0065, derived from 0.05/8, where the denominator is the total number of tests. We used Python version 5.2.2 and STATA version 17 SE edition for all these statistical tests.

6. Results

In this section, we present the results attained from CCVA algorithms and various classification techniques employed to determine CoD from various VA datasets (using only narratives as predictors, using questionnaire responses only, and results of the combined features).

TABLE 3 Comparison of nine ML models using narratives only.

	Model evaluation					
Model name	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**AUROCMIA**	**AUROCMAA**
XGBoost	96	96	96	96	0.927	0.906
RF	96	96	96	96	0.998	0.996
ANN	94	94	94	94	0.982	0.964
KNN	93	93	93	92	0.989	0.987
SVM	92	92	92	92	0.917	0.917
Bagging	91	91	91	91	0.997	0.995
DT	85	84	85	84	0.910	0.910
LR	82	82	82	82	0.977	0.959
BC	71	75	71	72	0.921	0.920

XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbor; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier; AUROCMIA, Area Under Receiver Operating Characteristics Micro Average; AUROCMAA, Area Under Receiver Operating Characteristics Macro Average.

6.1. Performance evaluation of ML classifiers

We validated our results using one vs. all with Precision, Recall, Accuracy, F1-score, and AUROC. We report on Precision, Recall, Accuracy, F-score measure, and AUROC for our nine classifiers in the CoD categorization of the 12 disease classes for narratives only in Table 3, questionnaire responses only in Table 4, combined questionnaire responses and narratives in Table 5.

6.1.1. Results from only the VA narrative predictors

The XGBoost and RF classifier outperformed all the other classifiers with a Precision of 96%, Recall of 96%, F1-score of 96%, and Accuracy of 96%, respectively. The least performing classifier was the statistical BC classifier with an Accuracy of 71%. Overall, our nine models had an AUROCMIA (Area Under Receiver Operating Characteristics Micro Average) and AUROCMAA (Area Under Receiver Operating Characteristics Macro Average) between 0.910 − 0.998and 0.910 − 0.996, respectively. Table 3 shows the detailed performance evaluation results of our nine models using VA narratives only.

6.1.2. Results from using questionnaire responses only as predictors

The ANN and XGBoost outperformed all the other classifiers when using questionnaire responses from the standardized questionnaire attaining a Precision, Recall, F1-score, and Accuracy of 100%, respectively. It was followed by Bagging our ensemble classifier and KNN both recorded a Precision, Recall, F1-score, and Accuracy of 98%, respectively. Our statistical classifiers LR and BC were on the lower ranking of our evaluation recording an Accuracy in the range of 74–83%, respectively. All of our models attained the highest AUROCMIAs within the range of 0.869and 1, respectively. Our nine models XGBoost, RF, ANN, Bagging, SVM, LR, DT, and KNN record high scores and the BC a bit lower AUROCMIA of 0.869. Additionally, the same nine models attained the highest AUROCMAAs within the range of 0.976and 1, respectively. On the other hand, the BC achieved an AUROCMAA score of 0.884. Table 4 shows the detailed performance evaluation results of our nine models using questionnaire responses only.

6.1.3. Results from using combined narratives and questionnaire responses

The XGBoost, ANN, and the RF classifier outperformed all the other classifiers with a Precision of 96%, Recall of 96%, F1-score of 96%, and Accuracy of 96%, respectively. On the contrary, BC and SVM were the least performing classifiers with Accuracy in the range of 68 − 72%. All of our models attained the highest AUROCMIAs within the range of 0.910and 0.998, respectively. The RF, XGBoost, ANN, Bagging, and KNN recorded high scores and the rest a bit lower scores. Additionally, our models attained the highest AUROCMAAs within the range of 0.907and 0.996, respectively. Similarly, the RF, XGBoost, ANN, Bagging, and KNN recorded high scores and the rest a bit lower comparable scores. However, the BC attained the lowest AUROCMIAs of 0.869and 0.884, respectively. Table 5 shows the detailed performance evaluation results of our nine models using combined questionnaire responses and narratives. Additionally, Figure 2 shows the model validation done using AUROC.

We report on the performance validation of our nine algorithms using descriptive statistics such as the mean and SD based on the Micro and Macro averages of our AUROC

TABLE 4 Comparison of nine ML models using questionnaire responses only.

	Model evaluation					
Model name	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**AUROCMIA**	**AUROCMAA**
XGBoost	100	100	100	100	1	1
ANN	99	99	99	99	1	1
Bagging	98	98	98	98	0.998	0.998
KNN	98	98	98	98	0.997	0.997
RF	97	97	97	97	0.999	0.998
DT	97	97	97	97	0.976	0.976
SVM	94	94	94	94	0.990	0.988
LR	83	83	83	83	0.990	0.980
BC	74	77	74	75	0.869	0.884

XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbor; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier; AUROCMIA, Area Under Receiver Operating Characteristics Micro Average; AUROCMAA, Area Under Receiver Operating Characteristics Macro Average.

TABLE 5 Comparison of nine ML models using combined narratives and questionnaire responses.

	Model evaluation					
Model name	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**AUROCMIA**	**AUROCMAA**
XGBoost	96	96	96	96	0.994	0.990
RF	96	96	96	96	0.998	0.996
ANN	96	95	96	95	0.995	0.991
Bagging	93	92	93	92	0.994	0.994
KNN	91	91	91	90	0.982	0.981
DT	87	87	87	87	0.928	0.928
LR	76	76	76	76	0.985	0.973
BC	72	73	72	73	0.910	0.907
SVM	68	68	68	66	0.969	0.958

XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbor; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier; AUROCMIA, Area Under Receiver Operating Characteristics Micro Average; AUROCMAA, Area Under Receiver Operating Characteristics Macro Average.

reported in Table 5. We report on 0.010282and 0.010105variance for AUROCMAA and AUROCMIA for our dataset, respectively. Table 6 shows the mean and standard deviation scores for each model throughout the 10-fold cross-validation training of the algorithms. The rank sums for each model column depict the Kruskal-Wallis test conducted. The test revealed that the mean observation was not the same ($Chi = 85.383, p = 0.0001$)across the nine models. This, therefore, implies that there was a statistically significant difference in mean observation between the nine models. We also report a p-value greater than the significance level of 0.05, hence, we fail to reject the null hypothesis and conclude that the nine model observations are not normally distributed. All model variances are very low or insignificant, implying that our dataset had a low degree of spread. Therefore, we can confidently state that our models were consistent in making predictions, thus even if different training data were used, they could still make a good estimate of the target variable. Additionally, we can infer that our sampled data points were very close to where our nine models predicted they would be.

The results of the McNemar tests on validating the performance of our nine models suggest good performance on the XGBoost and RF classifiers. The pairwise tests on XGBoost and RF suggest that there is a significant difference between the classifiers ($p < 0.0001$), which is smaller than our significance threshold ($\alpha = 0.0065$). Therefore, we reject our null hypothesis. We discovered that the XGBoost got 868 predictions right that RF got wrong. On the contrary, RF got 555predictions correct that XGBoost got wrong. As such, based on this 1.5 : 1ratio, we may conclude that XGBoost performed substantially better than RF. Additionally, we performed comparative pairwise tests on all our models (LR, KNN, DT, SVM, ANN, BC, and Bagging)

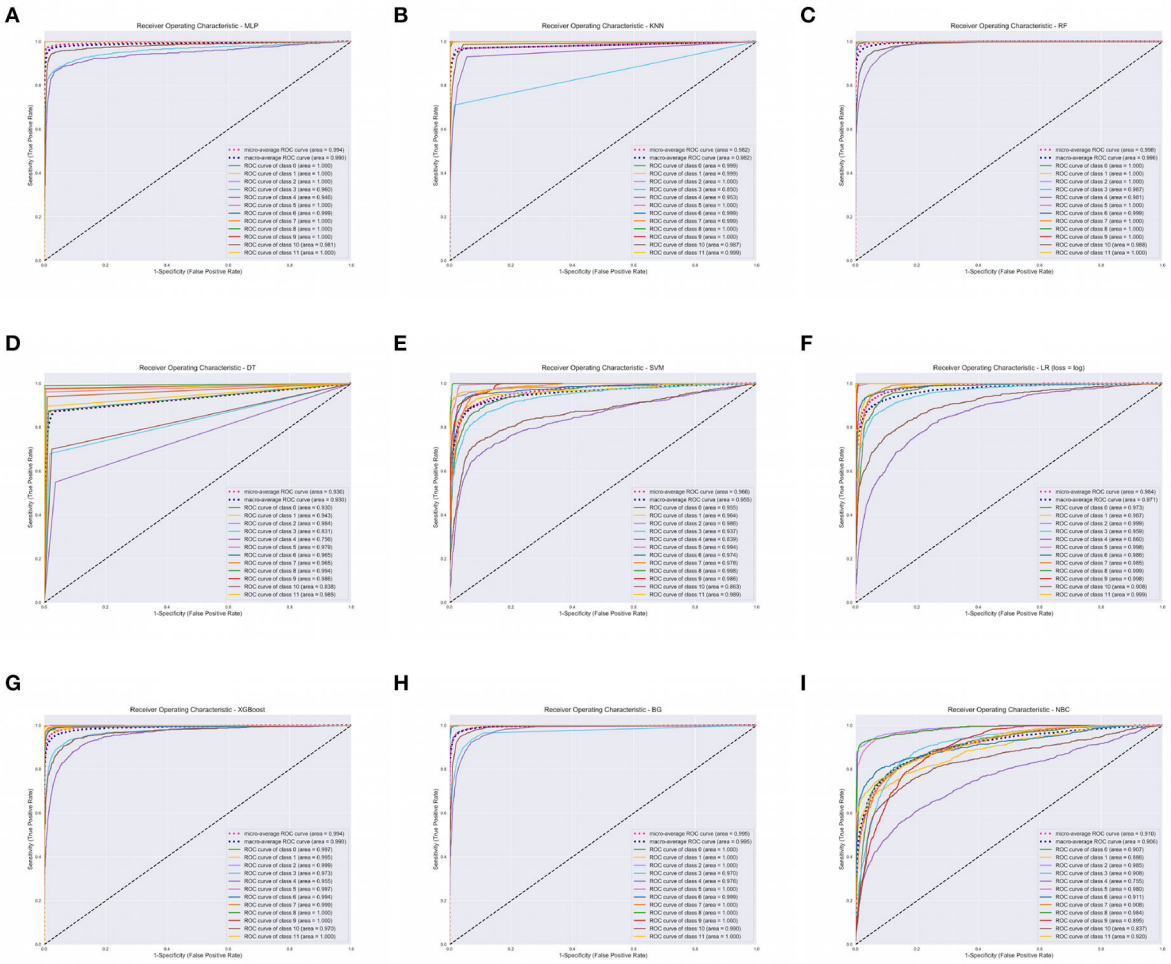


FIGURE 2
Area Under Receiver Operating Characteristics (AUROC) of our nine classifiers using combined questionnaire responses and narratives. **(A)** AUROC for ANN. **(B)** AUROC for KNN. **(C)** AUROC for RF. **(D)** AUROC for DT. **(E)** AUROC for SVM. **(F)** AUROC for LR. **(G)** AUROC for XGBOOST. **(H)** AUROC for BG. **(I)** AUROC for BC. XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbor; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier.

and our best classifiers XGBoost and RF. Based on the tests, we can objectively reject our null hypothesis and state that there is a significant difference between our two best classifiers and the other seven classifiers in terms of model performance ($p < 0.0001$)smaller than our significance threshold ($\alpha = 0.0065$).

7. CCVA algorithm evaluation using CSMFs

This extract highlights how the InterVA and InSilico algorithms were evaluated using CSMFs. We also present CSMF and CCC as evaluation metrics for the InterVA algorithm.

Figure 3 presents the 12 leading causes of death over time as determined by the InterVA algorithm using only one CoD. We observe that between the years 1993 and 2015, HIV/AIDS was the leading CoD across the population (CSMF=0.2739). This was closely followed by Pulmonary Tuberculosis (CSMF=0.1987) and Other Infectious/parasitic diseases (CSMF=0.1385). These three causes alone accounted for up to 61% of all deaths in the population during this period. The InterVA algorithm performance using CSMF accuracy and CCC attained values of 83% and 0.36, respectively.

8. Trend and pattern analysis using ML and CCVA approaches

8.1. CCVA algorithms

This section presents mortality trend and pattern analysis using conventional CCVA algorithms based on gender (Figure 4A), age (Figure 4B), and population over time

(Figure 4C), using data from structured questions. The visualizations are given in Figure 4.

We investigated the average age at death for the 12 leading causes of death. We discovered that both men and women were more likely to die from any disease at an average age of 40 years (mean=40, median=39, IQR=36, SD=26), despite the sex. We notice more women's deaths from HIV and circulatory diseases. On the contrary, we notice more male deaths from other infectious diseases, tuberculosis, and external causes (refer to Figure 4A). However, these differences were not statistically significant.

Figure 4B depicts percentages of mortality trends across all age groups. To determine mortality across age groups, the data were grouped into five bins "0–4," "5–14," "15–49," "50–64," and "65+." We significantly notice a declining trend in the number of deaths among persons aged between 0 and 4 years over time. In the earlier years of the Agincourt HDSS, there appears a declining trend in the number of deaths among individuals 65 years and above. However, this pattern is reversed and the mortality in the same age category is gradually increasing since the mid-2000s. Similarly, we also notice an almost comparable trend in the 50–64 age group to that of the 65+ group but the trend is gentle and stable. Among the 5–14 and 15–49 age groups, the number of deaths is appearing to be constant over time.

TABLE 6 Statistical tests of our nine models.

Model scores			
Model name	**Mean**	**Standard deviation**	**Rank sum**
XGBoost	0.9622614	0.003209	836.00
RF	0.9566394	0.0030548	735.50
ANN	0.9530553	0.0025771	663.50
Bagging	0.9216445	2.91e+07	585.00
KNN	0.9015075	0.0033769	447.00
DT	0.8671503	0.003984	255.00
LR	0.7509405	0.0124037	155.00
BC	0.698092	0.0081906	55.00
SVM	0.6783361	0.0054433	50.00

XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbor; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier.

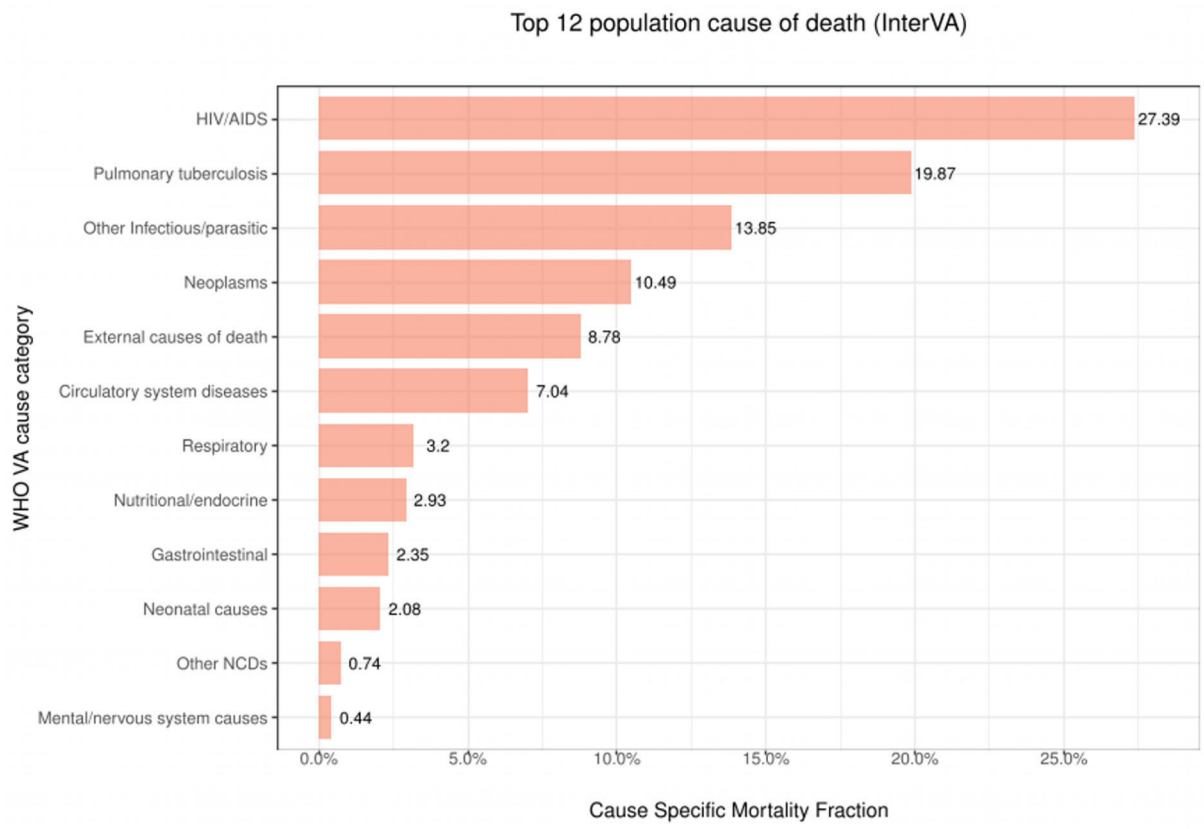


FIGURE 3
Top 12 CoD diseases.

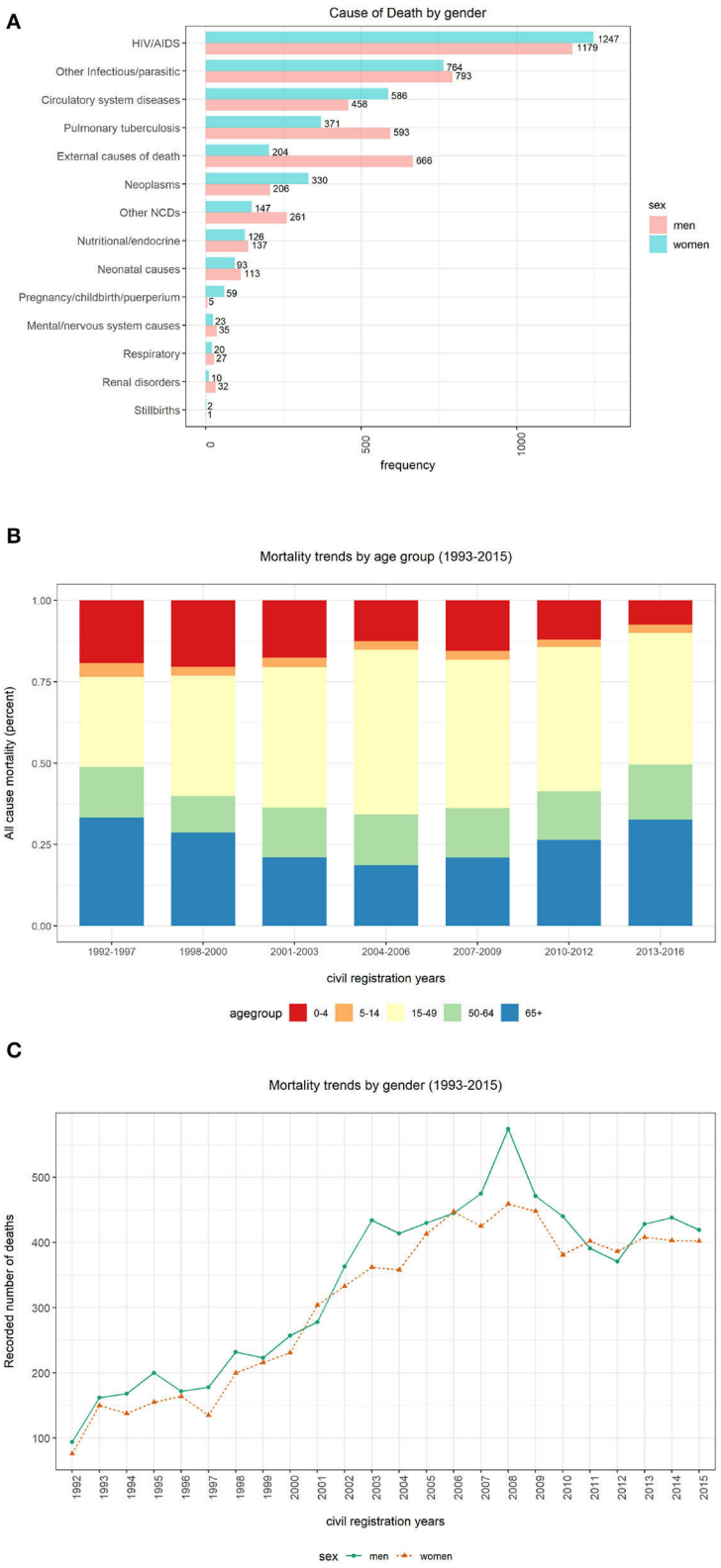


FIGURE 4
Computer Coded Verbal Autopsy (CCVA) mortality trends based on age, population, and gender. **(A)** Cause of death by sex. **(B)** Percentage of deaths by age group. **(C)** Yearly mortality trends by gender.

10.3389/fpubh.2022.990838

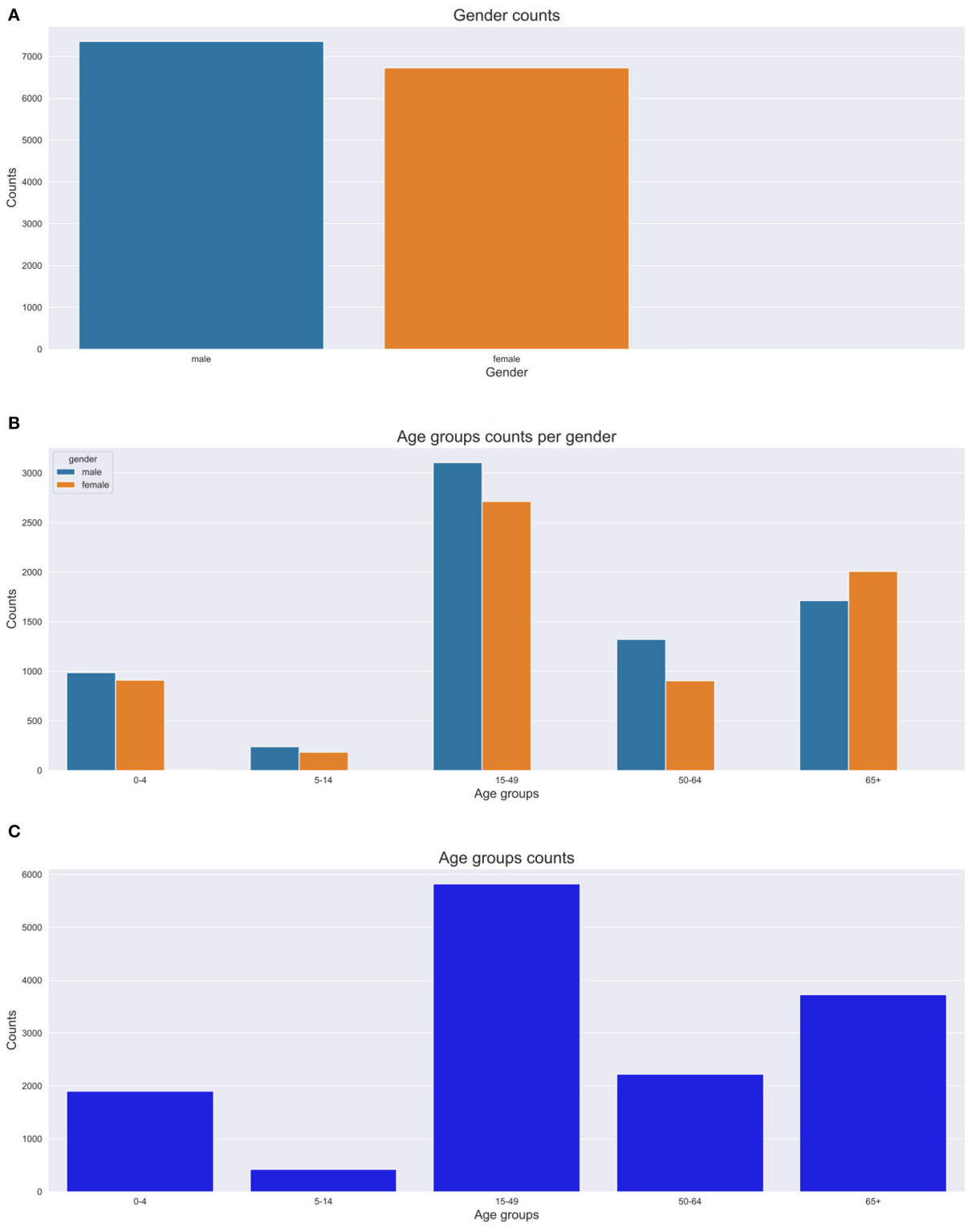


FIGURE 5
Gender and age group counts graphs. **(A)** Gender count. **(B)** Age group count per gender. **(C)** Age group count.

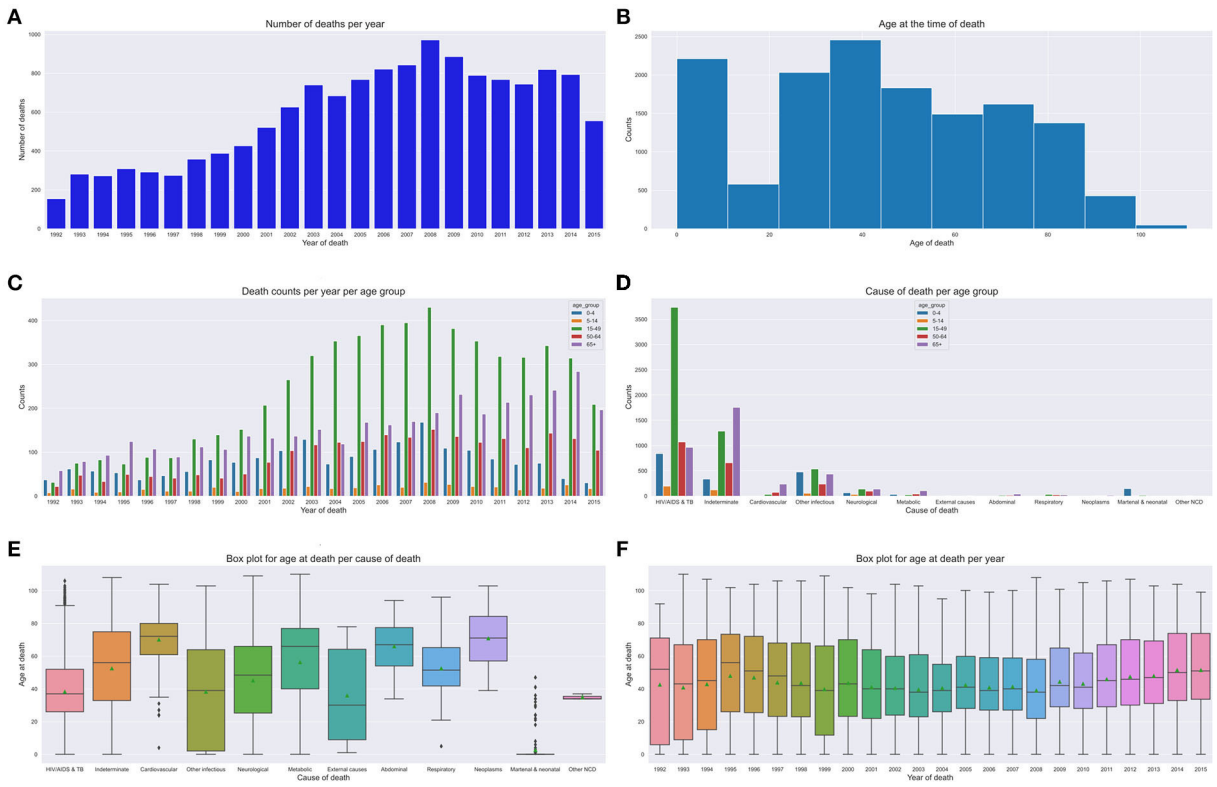


FIGURE 6
Mortality trends across age groups. **(A)** Number of deaths over time. **(B)** Age at death count. **(C)** Yearly death count across age groups. **(D)** Age group CoD count. **(E)** CoD and age death. **(F)** Age at death per year.

Figure 4C shows mortality trends based on gender over time. A total of 16,063 observations was collected, and composed of 52% men (n = 8,354) and 48% women (n = 7,709). We observe a gentle but steady increase in mortality between the years 1993 and 2000. This pattern rapidly accelerates among men between 2001 and 2008 before gradually declining.

8.2. ML techniques

Figures 6A–C depict the number of deaths over time, age at death count and yearly death count across age groups respectively. In this section, we present the results of our trend and pattern analysis using ML approaches to mortality based on gender, age, and population over time using narrative data combined with structured questions. We start by looking at the general distribution of our population based on gender, as depicted in Figure 5A, age groups (Figures 5B,C). All these graphs are depicted in Figure 5. We observe that there were more male deaths than female deaths. Most of the deaths were within the 15–49 and 65+ age groups.

We analyzed our mortality trend and pattern based on age groups as in Figure 5. We observe that most deaths are within the 15–49 and 65+ age groups. The 65+ age group had more deaths recorded in the 1990s with a gradual increase till 2014. We also discovered that the 15–49 age group trend sharply increases till 2008 and then steadily goes down till 2015. We notice a constant trend for the 5–14 year age group over time. There is a high number of deaths from HIV causes affecting mostly the 15–49 age group. We also notice that most deaths appear to be common in the 0–10 year age group and 30–80 years age groups. Conversely, we notice fewer deaths for 80+ years.

Figures 6E,F depict our boxplots on CoD and age at death over time and age at death per year. On average, the population died of HIV/AIDS or tuberculosis which was the leading CoD at a median age of 38 years. The plots depict an average death age of 66 years succumbed to cardiovascular, neoplasm, metabolic, or abdominal diseases. Worth taking note of is the death from other infectious disease causes that show a dissimilar trend across all age groups. Additionally, on average, most of the cases died of metabolic causes at an elderly age of 65 years. Other Non Communicable Diseases (NCDs') causes of death were more prevalent in the 30–35-year-old age group and neonatal and maternal causes in their first year (shown by the narrow IQRs). CoD from neurological and respiratory causes show a mortality

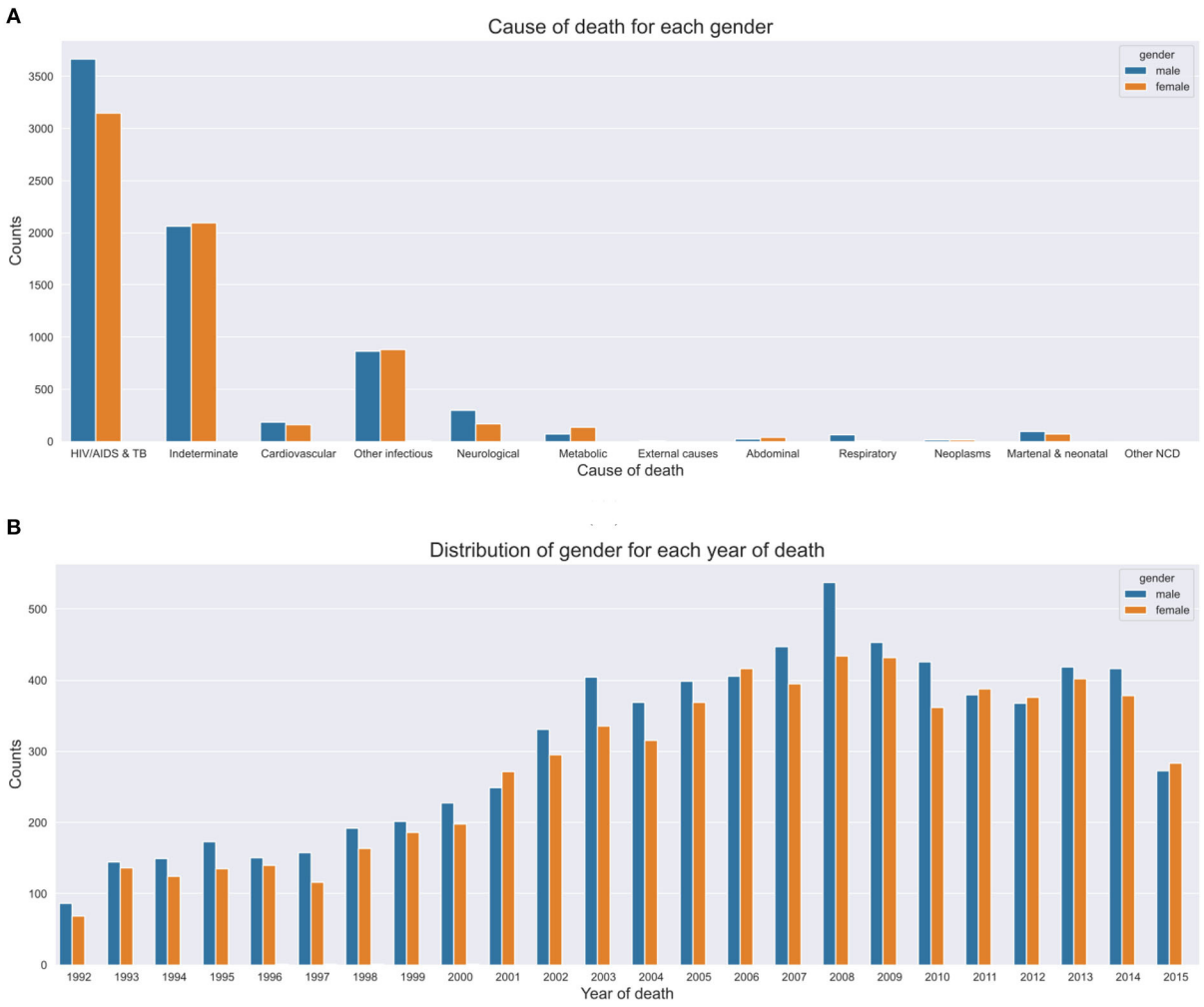


FIGURE 7
Yearly CoD based on gender. **(A)** CoD based on gender. **(B)** Yearly CoD based on gender.

trend and pattern that illustrates an average age at death of 50 years. We observe that there were more deaths in men than women, despite the cause. There is a gradual up-trend from 1992 (less than 100 deaths) to 2008 (almost 500 deaths) and a steady decline in the number of deaths from 2009 (refer to Figures 7A,B). Figure 6D illustrates that between the years 1993 and 1997, the average life expectancy was approximately 50 years. However, from 1998 to 2010, life expectancy significantly dropped and the population was dying at a much younger age of 40 years on average. From the year 2011, we see a gentle improvement in life expectancy.

8.3. Analysis of contradicting cases

The extract details an analysis of the structure and semantics of features in cases where the doctors are in disagreement. We discovered that approximately 16%of the observations in the VA dataset denote cases where the experts are in disagreement. Further analysis of the structure and semantics availed insights that in most cases the narratives entail information related to traditional healers' visits and consultations. Additionally, we deduced that also this is a result of cases where the captured information entails the imminent loss of weight, vomiting, and having a fever leading to an unexplainable sudden death. Figure 8 below shows our n-gram model of what was mined from the contradicting narratives.

8.4. Model best predictors

This part discusses how the most important narrative features where identified as the best predictors of our models. We chose the bi-grams as they show an evenly distributed

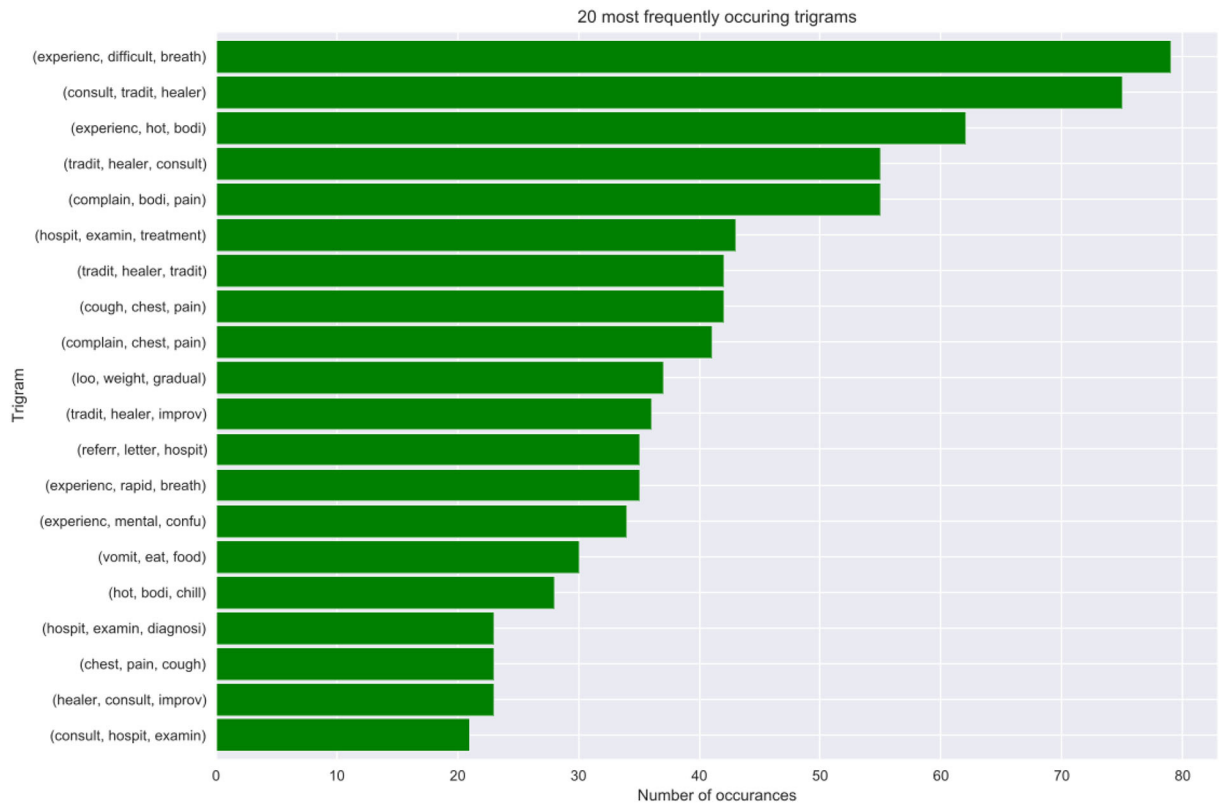


FIGURE 8
Tri-gram model showing frequently occurring contradicting cases.

frequency analysis of features (refer to Figure 9). As *n*increases the features start having more or less the same frequency.

9. Discussion

The process of determining causes of death using VAs still remains a manual task and suffers from many drawbacks (refer to Section 1). This negatively affects the VA reporting process, despite it being vital for strengthening health priorities and informing civil registration systems. Therefore, under such circumstances, there is a great need for innovative novel automated approaches to address these problems thereof.

In this study, we explore various VA data types, despite most studies in literature reporting results based on the classical dataset for CoD determination using ML approaches. Our aim is to investigate if the narratives can improve or enhance model prediction if they are added to the responses from the structured questionnaire. Our deductions suggest that the VA narratives have vital valuable information that should be used in model prediction. Consequently, we identify the best model predictors from the narratives. We further do a mortality pattern and trend analysis based on age, population, and gender over time. We also do a structure and semantic analysis of narratives in cases where the experts agree and also disagree. To add to our findings, we also investigate the best features that contribute to our models from the narratives.

Generally, the results of all our ML models used in this study, demonstrate that our models exhibited consistent superior performance on all datasets. This further reinforces the notion that ML approaches can be used as alternatives to conventional approaches for CoD determination using VAs. Ensemble classifiers (XGBoost, bagging), tree based models (DT, RF), ANN and KNN performed exceptionally well on all datasets. Our results of the combined dataset do not exhibit a consistent model performance, as most models slightly drop in model performance. This can be attributed to the fact that the combined dataset creates high dimensionality of the feature space and this triggers model complexity with too many noisy data points. The CCVA approach, InterVA, attained a CSFM accuracy of 83% and CCC of 36%.

Our CCVA approaches and ML techniques produced similar mortality trends and patterns based on age, population, and gender. Interestingly, we observed that in the first decade of the civil registration system, the average life expectancy was approximately 50 years. However, in the second decade,

10.3389/fpubh.2022.990838

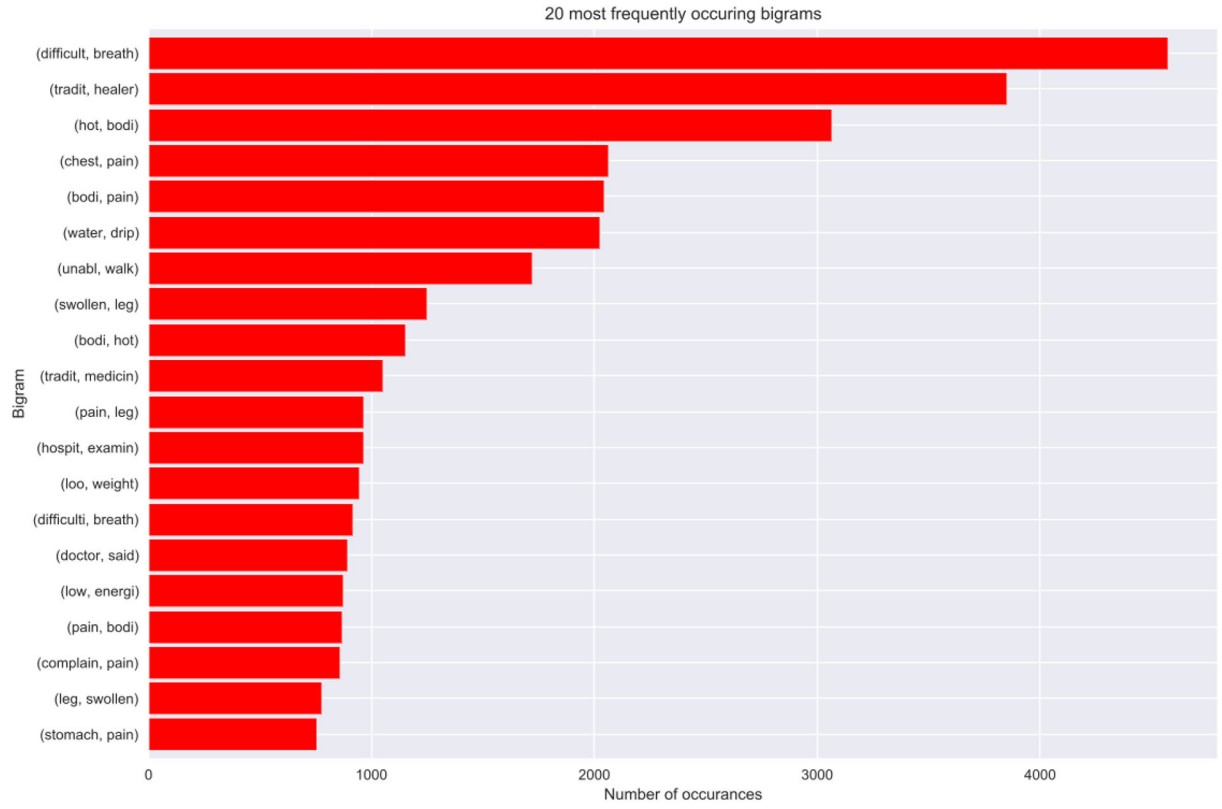


FIGURE 9
Bi-gram model of our best model predictors.

life expectancy significantly dropped and the population was succumbing to death at a slightly lower average age of 40 years. This suggests CoD mostly from the leading HIV and tuberculosis related causes. Interestingly, in the third decade, we see a gradual improvement in life expectancy, possibly attributed to the implementation of effective health intervention programmes. We notice that cardiovascular, neoplasm, neurological, respiratory, and metabolic CoD mainly affected the elderly. We observe that other infectious diseases and external causes affected the population disproportionately across all age groups, with the latter having an average age at death of 30 years. Despite the expected CoD from neonatal and maternal causes, we can also infer that those with HIV had a lower life expectancy as compared to the other CoDs. Of interest, is that most undetermined causes of death are found within the 65+ age group. This suggests that as the elderly population grows older, their health state deteriorates and they succumb to many symptoms that can lead to untimely hard to explain deaths. Other NCDs, causes of death were more prevalent in the younger age groups. We also discovered that sudden deaths are common in the elderly, suggesting symptoms, such as imminent loss of weight, vomiting, and having a fever leading to an unexplainable premature death. Generally, we notice more deaths in men than women.

We, therefore, propose that optimal model performance should be set at 80% accuracy. In cases where the ML model fails to reach a threshold value of 80% accuracy in terms of performance, we propose an expert's intervention for further exploration and assessment. Conversely, in cases where the experts are failing or do not reach a consensus, we recommend the help of the machine to make predictions. Most of these cases where the machine can assist, entail narratives where the interviewee details most content about the deceased circumstances and events that led to death based on traditional healer visits and consultation. Interestingly, we still found out that traditional healer consultations are a common practice in the population as they occurred frequently in our model best predictors. This cements the notion that most people in the HDSS seek traditional ways for their terminal illnesses, rather than western means. This finding opened exciting avenues for future study, which will focus on sequential text modeling with the aim of fully understanding treatment sequences for terminal illnesses. Nevertheless, in cases where the physicians were in agreement, these narrations about traditional healer's consultations were supplemented by enough

symptoms that made it possible for the experts to give a proper diagnosis. We also discovered that our model's best predictors entail matching symptoms with those in the responses to the structured questionnaire.

The results of this study, consistent with several studies that used VA data to determine CoD, suggest that ML approaches can accurately classify CoD from VA narratives. However, in most cases, statistical approaches and CCVA approaches are always outperformed by ML approaches (1, 8, 9, 13, 23, 25, 34). Therefore, it is imperative for future research studies to incorporate effective data handling strategies (8). This study adds to the existing body of literature, suggesting that automated approaches can be used as alternatives to PCVA in a cost effective way, producing real-time results that are consistent, accurate, and error free, thus strengthening health priorities. As such, VA processes are still key in capturing civil registration data where death occurs outside health facilities, up until a point when deaths start to take place in areas where it can be documented. Given these complexities, there is a great need for novel automated approaches that can be used as alternatives (22).

The strength of this study lies in the application of various ML and CCVA algorithms to various VA data types. Moreover, our sample size was large and representative of deaths that occurred at Agincourt HDSS that were captured in a standard way. Moreover, our mortality trend and pattern analysis gave us valuable insights into our HDSS and this can be used to inform policy and practice. This enforces generalization and comparability across studies. On the contrary, this study had limitations of data quality described in Section 1.2.

10. Conclusion

In general, this study demonstrates that ML techniques can be used as alternatives in determining CoD from VA narratives producing results comparable to physician diagnosis. Our findings should be used to inform policy and practice and enforce effective health intervention programmes and resource prioritization to reduce the mortality rate and prolong life expectancy. As such, they can help close the gap in civil registration systems. Our comparative analysis of the ML models on various VA datasets enforces comparability and generalization, thus availing a baseline study for future research. Future work will entail exploring deep learning methods and employing novel techniques such as transfer learning to determine CoD.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by University of the Witwatersrand, Faculty of Health Sciences Ethics Committee (Certificate No. M1911132). Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements. Written informed consent was not obtained from the individual(s), nor the minor(s)' legal guardian/next of kin, for the publication of any potentially no identifiable images or data are presented in the manuscript or data included in this article.

Author contributions

MM and VO did all algorithm experiments with the help of TC. MM and all the other authors drafted and critically revised the study. The first draft of the manuscript was written by MM and all the authors commented on previous versions of the manuscript. All the authors contributed to the study's conception and design, read, and approved the final manuscript.

Funding

This study was supported by the Developing Excellence in Leadership, Training and Science (DELTAS) Africa Initiative Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) (Grant No. DEL-15-005). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS) Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (Grant No. 107754/Z/15/Z) and the United Kingdom government.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Jeblee S, Gomes M, Jha P, Rudzicz F, Hirst G. Automatically determining cause of death from verbal autopsy narratives. *BMC Med Inform Decis Mak*. (2019) 19:127. doi: 10.1186/s12911-019-0841-9

2. Nichols EK, Byass P, Chandramohan D, Clark SJ, Flaxman AD, Jakob R, et al. The WHO 2016 verbal autopsy instrument: an international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0. *PLoS Med*. (2018) 15:e1002486. doi: 10.1371/journal.pmed.1002486

3. Thomas LM, D'mbruoso L, Balabanova D. Verbal autopsy in health policy and systems: a literature review. *BMJ Global Health*. (2018) 3:e000639. doi: 10.1136/bmjgh-2017-000639

4. Soleman N, Chandramohan D, Shibuya K. Verbal autopsy: current practices and challenges. *Bull World Health Organ*. (2006) 84:239–245. doi: 10.2471/BLT.05.027003

5. Mapoma CC, Munkombwe B, Mwango C, Bwalya BB, Kalindi A, Gona NP. Application of verbal autopsy in routine civil registration in Lusaka District of Zambia. *BMC Health Serv Res*. (2021) 21:408. doi: 10.1186/s12913-021-06427-y

6. Lozano R, Lopez AD, Atkinson C, Naghavi M, Flaxman AD, Murray CJ. Performance of physician-certified verbal autopsies: multisite validation study using clinical diagnostic gold standards. *Popul Health Metr*. (2011) 9:1–13. doi: 10.1186/1478-7954-9-32

7. Reeves BC, Quigley M. A review of data-derived methods for assigning causes of death from verbal autopsy data. *Int J Epidemiol*. (1997) 26:1080–9. doi: 10.1093/ije/26.5.1080

8. Mujtaba G, Shuib L, Idris N, Hoo WL, Raj RG, Khowaja K, et al. Clinical text classification research trends: systematic literature review and open issues. *Expert Syst Appl*. (2019) 116:494–520. doi: 10.1016/j.eswa.2018.09.034

9. Desai N, Aleksandrowicz L, Miasnikof P, Lu Y, Leitao J, Byass P, et al. Performance of four computer-coded verbal autopsy methods for cause of death assignment compared with physician coding on 24,000 deaths in low-and middle-income countries. *BMC Med*. (2014) 12:20. doi: 10.1186/1741-7015-12-20

10. James SL, Flaxman AD, Murray CJ. Performance of the Tariff Method: validation of a simple additive algorithm for analysis of verbal autopsies. *Popul Health Metr*. (2011) 9:31. doi: 10.1186/1478-7954-9-31

11. Byass P, Herbst K, Fottrell E, Ali MM, Odhiambo F, Amek N, et al. Comparing verbal autopsy cause of death findings as determined by physician coding and probabilistic modelling: a public health analysis of 54 000 deaths in Africa and Asia. *J Glob Health*. (2015) 5: 010402. doi: 10.7189/jogh.05.010402

12. McCormick TH, Li ZR, Calvert C, Crampin AC, Kahn K, Clark SJ. Probabilistic cause-of-death assignment using verbal autopsies. *J Am Stat Assoc*. (2016) 111:1036–49. doi: 10.1080/01621459.2016.1152191

13. Miasnikof P, Giannakeas V, Gomes M, Aleksandrowicz L, Shestopaloff AY, Alam D, et al. Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths. *BMC Med*. (2015) 13:286. doi: 10.1186/s12916-015-0521-2

14. Clark SJ, Li Z, McCormick TH. Quantifying the contributions of training data and algorithm logic to the performance of automated cause-assignment algorithms for verbal autopsy. *arXiv preprint arXiv:180307141*. (2018). doi: 10.48550/arXiv.1803.07141

15. Leitao J, Desai N, Aleksandrowicz L, Byass P, Miasnikof P, Tollman S, et al. Comparison of physician-certified verbal autopsy with computer-coded verbal autopsy for cause of death assignment in hospitalized patients in low- and middle-income countries: systematic review. *BMC Med*. (2014) 12:22. doi: 10.1186/1741-7015-12-22

16. Murray CJ, Lozano R, Flaxman AD, Serina P, Phillips D, Stewart A, et al. Using verbal autopsy to measure causes of death: the comparative performance of existing methods. *BMC Med*. (2014) 12:1–19. doi: 10.1186/1741-7015-12-5

17. Kalter HD, Perin J, Black RE. Validating hierarchical verbal autopsy expert algorithms in a large data set with known causes of death. *J Glob Health*. (2016) 6:010601. doi: 10.7189/jogh.06.010601

18. Quigley MA, Chandramohan D, Setel P, Binka F, Rodrigues LC. Validity of data-derived algorithms for ascertaining causes of adult death in two African sites using verbal autopsy. *Trop Med Int Health*. (2000) 5:33–9. doi: 10.1046/j.1365-3156.2000.00517.x

19. Nithya B, Ilango V. Predictive analytics in health care using machine learning tools and techniques. In: *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*. Madurai: IEEE (2017). p. 492–9.

20. Flaxman AD, Vos T. Machine learning in population health: opportunities and threats. *PLoS Med*. (2018) 15:e1002702. doi: 10.1371/journal.pmed.1002702

21. Moran KR, Turner EL, Dunson D, Herring AH. Bayesian hierarchical factor regression models to infer cause of death from verbal autopsy data. *J R Stat Soc Ser C Appl Stat*. (2019) 70:532–57. doi: 10.1111/rssc.12468

22. Idicula-Thomas S, Gawde U, Jha P. Comparison of machine learning algorithms applied to symptoms to determine infectious causes of death in children: national survey of 18,000 verbal autopsies in the Million Death Study in India. *BMC Public Health*. (2021) 21:1–11. doi: 10.1186/s12889-021-11829-y

23. Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K. Prediction of cause of death from forensic autopsy reports using text classification techniques: a comparative study. *J Forensic Leg Med*. (2018) 57:41–50. doi: 10.1016/j.jflm.2017.07.001

24. Danso S, Atwell E, Johnson O. A comparative study of machine learning methods for verbal autopsy text classification. *arXiv preprint arXiv:14024380*. (2014). doi: 10.48550/arXiv.1402.4380

25. Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K, Al-Garadi MA. Classification of forensic autopsy reports through conceptual graph-based document representation model. *J Biomed Inform*. (2018) 82:88–105. doi: 10.1016/j.jbi.2018.04.013

26. Koopman B, Karimi S, Nguyen A, McGuire R, Muscatello D, Kemp M, et al. Automatic classification of diseases from free-text death certificates for real-time surveillance. *BMC Med Inform Decis Mak*. (2015) 15:53. doi: 10.1186/s12911-015-0174-2

27. Mwanyangala MA, Urassa HM, Rutashobya JC, Mahutanga CC, Lutambi AM, Maliti DV, et al. Verbal autopsy completion rate and factors associated with undetermined cause of death in a rural resource-poor setting of Tanzania. *Popul Health Metr*. (2011) 9:41. doi: 10.1186/1478-7954-9-41

28. Quigley MA, Chandramohan D, Rodrigues LC. Diagnostic accuracy of physician review, expert algorithms and data-derived algorithms in adult verbal autopsies. *Int J Epidemiol*. (1999) 28:1081–7. doi: 10.1093/ije/28.6.1081

29. Boulle A, Chandramohan D, Weller P. A case study of using artificial neural networks for classifying cause of death from verbal autopsy. *Int J Epidemiol*. (2001) 30:515–20. doi: 10.1093/ije/30.3.515

30. Flaxman AD, Vahdatpour A, Green S, James SL, Murray CJ. Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Popul Health Metr*. (2011) 9:29. doi: 10.1186/1478-7954-9-29

31. Danso S, Johnson O, Ten Asbroek A, Soromekun S, Edmond K, Hurt C, et al. A semantically annotated Verbal Autopsy corpus for automatic analysis of cause of death. *ICAME J*. (2013) 37:37–69.

32. Pestian J, Nasrallah H, Matykiewicz P, Bennett A, Leenaars A. Suicide note classification using natural language processing: A content analysis. *Biomed Inform Insights*. (2010) 3:BII-S4706. doi: 10.4137/BII.S4706

33. Murtaza SS, Kolpak P, Bener A, Jha P. Automated verbal autopsy classification: using one-against-all ensemble method and Naïve Bayes classifier. *Gates Open Res*. (2018) 2:63. doi: 10.12688/gatesopenres.12891.1

34. Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K, Al-Garadi MA. Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. *PLoS ONE*. (2017) 12:e0170242. doi: 10.1371/journal.pone.0170242

35. Clark SJ. A guide to comparing the performance of VA algorithms. *arXiv preprint arXiv:180207807*. (2018). doi: 10.48550/arXiv.1802.07807

36. Chandramohan D, Setel P, Quigley M. Effect of misclassification of causes of death in verbal autopsy: can it be adjusted? *Int J Epidemiol*. (2001) 30:509–14. doi: 10.1093/ije/30.3.509

37. Kabudula CW, Tollman S, Mee P, Ngobeni S, Silaule B, Gómez-Olivé FX, et al. Two decades of mortality change in rural northeast South Africa. *Glob Health Action*. (2014) 7:25596. doi: 10.3402/gha.v7.25596

38. King G, Lu Y, et al. Verbal autopsy methods with multiple causes of death. *Stat Sci*. (2008) 23:78–91. doi: 10.1214/07-STS247

39. Korde V, Mahender CN. Text classification and classifiers: a survey. *Int J Artif Intell Appl*. (2012) 3:85. doi: 10.5121/ijaia.2012.3208

40. Zaki MJ, Meira Jr W. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge: Cambridge University Press (2019).

41. Leskovec J, Rajaraman A, Ullman JD. *Mining of Massive Data Sets*. Cambridge: Cambridge University Press (2020).

42. Pičulin M, Smole T, Žunkovič B, Kokalj E, Robnik-Šikonja M, Kukar M, et al. Disease progression of hypertrophic cardiomyopathy: modeling using machine learning. *JMIR Med Inform*. (2022) 10:e30483. doi: 10.2196/30483

43. Yang Y, Zheng J, Du Z, Li Y, Cai Y, et al. Accurate prediction of stroke for hypertensive patients based on medical big data and machine learning algorithms: retrospective study. *JMIR Med Inform*. (2021) 9:e30277. doi: 10.2196/30277

44. Fawagreh K, Gaber MM, Elyan E. Random forests: from early developments to recent advancements. *Syst Sci Control Eng*. (2014) 2:602–9. doi: 10.1080/21642583.2014.956265

45. Poole DL, Mackworth AK. *Artificial Intelligence: Foundations of Computational Agents*. Cambridge: Cambridge University Press (2010).

46. Byrne MD. Machine learning in health care. *J PeriAnesthesia Nurs*. (2017) 32:494–6. doi: 10.1016/j.jopan.2017.07.004

47. Iqbal Z, Ilyas R, Shahzad W, Inayat I. A comparative study of machine learning techniques used in non-clinical systems for continuous healthcare of independent livings. In: *2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*. Penang: IEEE (2018). p. 406–11.

48. Li ZR, McCormick TH, Clark SJ. InterVA4: an R package to analyze verbal autopsy data. In: *Center for Statistics and the Social Sciences, University of Washington*. Vienna, Austria: R Foundation for Statistical Computing (2014).

49. Murray CJ, Lozano R, Flaxman AD, Vahdatpour A, Lopez AD. Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. *Popul Health Metr*. (2011) 9:1–11. doi: 10.1186/1478-7954-9-28

50. Murray CJ, Lopez AD, Black R, Ahuja R, Ali SM, Baqui A, et al. Population health metrics research consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. *Popul Health Metr*. (2011) 9:27. doi: 10.1186/1478-7954-9-27

51. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput*. (1998) 10:1895–923. doi: 10.1162/089976698300017197

C Appendices

PLOS ONE

RESEARCH ARTICLE

Text mining of verbal autopsy narratives to extract mortality causes and most prevalent diseases using natural language processing

Michael Tonderai Mapundu[1☯]*, **Chodziwadziwa Whiteson Kabudula**[1,2☯], **Eustasius Musenge**[1☯], **Victor Olago**[3☯], **Turgay Celik**[4,5☯]

1 Department of Epidemiology and Biostatistics, School of Public Health, University of the Witwatersrand, Johannesburg, South Africa, **2** MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), Johannesburg, South Africa, **3** National Health Laboratory Service (NHLS), National Cancer Registry, Johannesburg, South Africa, **4** Wits Institute of Data Science, University of The Witwatersrand, Johannesburg, South Africa, **5** School of Electrical and Information Engineering, University of The Witwatersrand, Johannesburg, South Africa

☯ These authors contributed equally to this work.
* michael.mapundu@wits.ac.za

Check for updates

OPEN ACCESS

Citation: Mapundu MT, Kabudula CW, Musenge E, Olago V, Celik T (2024) Text mining of verbal autopsy narratives to extract mortality causes and most prevalent diseases using natural language processing. PLoS ONE 19(9): e0308452. https://doi.org/10.1371/journal.pone.0308452

Editor: Edison Arwanire Mworozi, King Caesar University, UGANDA

Received: May 20, 2023

Accepted: July 24, 2024

Published: September 19, 2024

Copyright: © 2024 Mapundu et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data cannot be shared publicly because of ethical and confidentiality issues as it is sensitive patient data. Data are available from the University of the Witwatersrand, South Africa Institutional Data Access / Ethics Committee (contact via Chairperson of the Medical HREC, Emeritus Professor Paul Ruff, email: paul.ruff@wits.ac.za) for researchers who meet the criteria for access to confidential data.

Abstract

Verbal autopsy (VA) narratives play a crucial role in understanding and documenting the causes of mortality, especially in regions lacking robust medical infrastructure. In this study, we propose a comprehensive approach to extract mortality causes and identify prevalent diseases from VA narratives utilizing advanced text mining techniques, so as to better understand the underlying health issues leading to mortality. Our methodology integrates n-gram-based language processing, Latent Dirichlet Allocation (LDA), and BERTopic, offering a multi-faceted analysis to enhance the accuracy and depth of information extraction. This is a retrospective study that uses secondary data analysis. We used data from the Agincourt Health and Demographic Surveillance Site (HDSS), which had 16338 observations collected between 1993 and 2015. Our text mining steps entailed data acquisition, pre-processing, feature extraction, topic segmentation, and discovered knowledge. The results suggest that the HDSS population may have died from mortality causes such as *vomiting, chest/stomach pain, fever, coughing, loss of weight, low energy, headache*. Additionally, we discovered that the most prevalent diseases entailed human immunodeficiency virus (HIV), tuberculosis (TB), diarrhoea, cancer, neurological disorders, malaria, diabetes, high blood pressure, chronic ailments (kidney, heart, lung, liver), maternal and accident related deaths. This study is relevant in that it avails valuable insights regarding mortality causes and most prevalent diseases using novel text mining approaches. These results can be integrated in the diagnosis pipeline for ease of human annotation and interpretation. As such, this will help with effective informed intervention programmes that can improve primary health care systems and chronic based delivery, thus increasing life expectancy.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

The United Nations expects all countries in the world to meet their 90 percent death registration coverage requirement. However, this is not the case as most deaths that occur in Low to Medium Income Countries (LMICs) are not captured in civil registration systems [1, 2]. As such, this accounts for approximately 65 percent of the world population that do not have a medically certified cause of death (CoD) on a yearly basis [3]. Nevertheless, the CoD information is important as it is used to improve civil registration systems, public health monitoring, critical health policies and priorities [4]. In most instances were CoD cannot be derived from clinically based sources, the VA process is used as a substitute tool. The VA process is conducted by non-medical staff who seek to elicit information from the next of kin of the deceased, regarding circumstances and events that may have led to death [3]. This information is then captured and stored in textual format and is also known as VA narratives. These narratives contain rich valuable information that is used to supplement responses from the standard World Health Organisation CoD questionnaire by physicians, when giving a CoD diagnosis [3]. However, in most cases this information is encapsulated in the textual format, usually unstructured and is stored in large volumes. Consequently, it is difficult to manage, takes time to process and suffers from word polysemy, amongst many challenges. This therefore makes the analysis and harnessing of meaningful information difficult, even though it is vital in identifying mortality causes.

As such, the application of text mining (TM) which employs natural language processing (NLP) in an automated fashion can be beneficial to extract meaningful information regarding mortality causes and most prevalent diseases from VA narratives. There is scanty literature on the application of unsupervised TM approaches within the VA domain. Only a few TM studies that employ supervised learning, have been applied to date within the VA domain to determine CoD [5–7]. Mujtaba et al. [5] argues that the TM approaches used for cause of death classification, require large datasets which are labelled for training the machine learning models. This makes the process tedious, time consuming, expensive and prone to human error as the experts are the ones responsible for labelling the datasets [5, 8].

This study seeks to address this gap by reducing the need for human intervention, by implementing unsupervised TM approaches. The implementation of TM approaches in VA to determine mortality causes and identify the most prevalent disease has several advantages as reported in [5, 6]. Moreover, several studies within the healthcare domain also appraise on the benefits of applying TM as cited in [9–16].

Firstly, TM can help in efficient analysis, through extracting vital information from large volumes of unstructured narratives in a quick, cheaper and accurate way [5]. Ultimately, this reduces time and effort required for manual analysis. Moreover, the TM techniques exhibit the ability to transform unstructured data into actionable insights that can help improve health outcomes, inform public health policies and interventions. The TM mathematical approaches are reported elsewhere [17]. Secondly, implementing TM techniques can result in improved accuracy as one can identify patterns and trends in the data that might be missed by human analysts, leading to more accurate and reliable results [12, 15]. Thirdly, there is consistency as one can categorise similar cases together, ensuring a consistent way of CoD categorisation across different VA records. Additionally, there is some form of standardisation on the CoD categorisation, that ensures the application of the same criteria across different VA records. Interestingly, TM approaches can extract additional variables that can be useful in the CoD diagnosis [12]. Moreover, the semantic and structure analysis through sequential modelling, can even give physicians better insights on the most probable CoD as it entails sequences of events that might have led one to succumb to death [18]. As such, this

might help in cases where there are fewer symptoms or unknown CoD to help one to get to a CoD diagnosis.

In this study we present novel TM techniques such as n-gram natural language processing, LDA and BERTopic as we seek to investigate mortality causes and the most prevalent diseases that led to death in rural north-east South Africa. The TM approaches will assist in automating the process of analysing VA data, in a fast and accurate manner. The findings of this study can be integrated in the diagnosis pipeline for ease of human annotation and interpretation. By incorporating modelling frameworks, researchers can potentially improve the granularity and precision of identified topics, leading to a more nuanced understanding of the underlying themes present in large VA text corpora. Ultimately, the discovered knowledge from the unstructured narratives will enable the design, implementation and sustainment of tailored health interventions programmes rather than adhoc generalizations. This will strengthen health priorities, improve life expectancy, VA reporting and decision making, thus informing policy and practice.

Literature review

Related work

There is scanty literature that reports on the application of TM techniques specifically in the VA domain. TM is an approach that seeks to close the gap between textual information and representation of the text in a structured fashion [15]. Leskovec et al. [19] argues that the mining of data entails use of very robust hardware, programming languages that employ efficient algorithms, and strive to solve various problems in domain specific areas. TM employs various techniques that span from NLP, data mining, machine learning (ML) and management of knowledge. TM can be used for information retrieval, information extraction, classification of documents and named entity recognition (NER), amongst many applications [14].

One notable study in the VA domain that has applied TM is that of Mujtaba et al. [5], who did a systematic literature review on clinical text classification trends using 72 articles. They point out that the free-text clinical report is very valuable for classification problems. Nonetheless, one can only derive meaningful information from such reports after effective data transformation, to generate useful trends and patterns. Similar work on text classification is reported in [20–23].

Despite the little research that reports on application of TM methods in the VA domain, these approaches have been employed extensively in other health fields as reported in [13, 24–29]. These previous studies report that TM techniques make it possible to extract meaningful information from implicit information. As such, there is great need to use novel TM tools to extract and discover valuable information. Consequently, TM techniques, can be used as alternatives to manual processes. However, they raise issues of noise and redundancy that affect the quality of data. Moreover, in most cases unstructured textual data exhibits high levels of noise, sparsity, makes use of varying vocabulary terms, misspelled terms with many grammatical errors and entails native terms [5, 27, 30].

Recent developments in technology have triggered an increase in the application of electronic public health repositories, making most textual reports readily available for additional use, specifically in biomedicine to extract useful information [5, 29]. However, despite these availed opportunities, there is little to none regarding VA literature that has applied these novel TM techniques for knowledge discovery. On the contrary, VA narratives entail rich valuable information, which can be utilised through NLP, and can be integrated in intelligent models to improve public health decision making processes [3, 15].

N-gram language processing

Literature reports on n-gram language processing being a feature extraction approach, that aims at attaining only relevant and useful features from textual data. The n-grams are a set of words that are sequential as they make use of continuous number of items such as characters or words from a given sequence of narratives [5, 31]. Another study by Danso et al. [6] reports on a comparative study of machine learning methods for VA text classification and they deduce that the bag of words using n-grams can result in better model performance with high accuracies.

Lucini et al. [14] did a study using TM techniques as they sought to predict hospital admissions using early medical records from the emergency department. They report that these tools make it possible to elicit valuable information, thereby improving clinical decision making processes. Similar findings are also reported in Kim and Chung [16], who conducted a study on information extraction from big health data and they suggest that word clouds can be effective visualisation tools that can be able to depict keywords or concepts in documents. Furthermore, they argue that, the more a term appears in the corpus repeatedly, the more important it is. Similar work which emphasises the importance of n-grams as a TM approach in gaining valuable insights is alluded in the studies of [32–34]. However, they point out the challenges of having too many insignificant words in the corpus.

Topic modelling

Topic modelling using LDA. Topic modelling entails using probabilistic algorithms to extract useful and common topics from a set of documents (in our case we aim to extract most prevalent diseases). Topic modelling is an unsupervised learning approach based on statistical principles, that seeks extract topics from a set of documents stored in given corpus. A single topic is made up of a collection of words, with a corresponding weight. A single word can belong to various topics and documents can be made up of several topics [35].

One commonly used topic segmentation method is the LDA, which has been applied in domains such as classification of genomic sequences, social networks, modelling of health data, extracting latent topics from clinical reports, amongst many uses. This approach has been reported to attain high precision and performance scores, especially when used with clinical reports [10, 35–38]. LDA is generative in nature and can easily pick out hidden groups in data, using patterns of co-occurrences. Such patterns cannot easily be detected by conventional approaches, let alone us as researchers may not be aware of such associations in their data [32, 39]. Noteworthy, is that the LDA approach often struggles to capture the intricate nuances of language and context, and its implementation requires some programming expertise [10, 40].

Topic modelling using BERTopic. BERT, being a contextualized language model, takes into account the surrounding words when representing each word in a sentence. This bidirectional understanding enables BERT to create more accurate and contextually rich word embeddings. In the context of topic modelling, these embeddings can be utilized to derive topic representations that better reflect the semantic relationships between words [40].

The work of Scarpino et al. [41] supports this notion, as they report that BERTopic outperforms LDA approach in clustering in their study that investigated topic modelling techniques to extract meaningful insights in Italian long COVID narrations. The study of Silva et al. [42], reports on the successful application of BERTopic approach as an unsupervised natural language processing in the identification of patients with suspected COVID-19 infection. Similar successful application of BERTopic as a topic modelling approach in various health spaces is reported in [43–49]. As such, the application of BERT in topic modelling holds the promise of

enhancing the sophistication and contextual relevance of topic extraction, thereby contributing to more accurate and meaningful insights derived from textual data.

The text mining process entailed the following steps; data acquisition, pre-processing, feature extraction, topic segmentation, and discovered knowledge. These techniques are reported elsewhere [11, 12, 15, 24, 36–38, 50].

Materials and methods

In this study, we apply TM techniques to extract mortality causes and identify the most prevalent diseases using VA data. This is a retrospective cross-sectional study that uses secondary data analysis through TM methods. The VA data was obtained from the study area of the Agincourt Health and Demographic Surveillance System (HDSS). This study has been approved by the relevant ethics committee in South Africa (University of the Witwatersrand Faculty of Health Sciences, Human Research Ethics Committee (Medical), and (approval ref. no: M1911132). It was conducted in accordance to the Declaration of Helsinki. Informed participant consent was waived by University of the Witwatersrand Faculty of Health Sciences, Human Research Ethics Committee (Medical). We got permission from the Agincourt Health and Demographic Surveillance Site to use their data for research purposes from 29 November 2019 to 29 November 2024. The inclusion criteria was all VA autopsy data for the period between 1993 and 2015.

The HDSS was established in 1992 and is situated in the rural Sub-district of Bushbuckridge under Ehlanzeni District, in Mpumalanga Province, in north-eastern South Africa. The geographical coverage is approximately 420km^2. As of 2019, the population was at 116 247 individuals residing in 28 villages with 22 716 households, with males being 55 961, females being 60 280, children under 5 years being 11 724 and school going children with ages from 5–19 being 35 928 [51].

The Agincourt HDSS is a surveillance site that serves as an evidence based health monitoring site that seeks to strengthens health priorities, practice and inform policy. The data was collected between 1993 and 2015. For this study, we only use textual data as VA narratives denoting one feature with 16338 cases / observations, which where in English. However, for our semantic and structure analysis we compare our extracted symptoms from the VA narratives to that of the responses from the structured questionnaire. The responses from the questionnaire dataset entailed 227 symptoms. All symptoms that had a 'Y' were encoded as a 1 meaning that the record had a present symptom value. On the other hand, all features that had a 'N', were encoded as a 0 meaning that those records had no symptoms present. All our features were in English. This comparative analysis was done to investigate if mined narratives exhibit similar and/or have additional features.

In this study, we first do data acquisition of our VA narratives as a comma separated value text file (csv), followed by preprocessing, then keyword the extraction using n-gram natural language processing, followed by topic modelling using LDA and BERTopic and lastly identifying the relevant topics through knowledge discovery. We follow the TM steps as depicted in Fig 1 below.

Data pre-processing

We performed pre-processing as part of data cleaning and normalisation, as we sought to do away with irrelevant data. In most cases input of textual data requires pre-processing and transformation before being mined [52]. After data acquisition, in csv format, we started by cleaning the unstructured narrative data by converting all text to lowercase, removed all punctuation, spaces, numbers and special characters. We applied stopword removal in order to

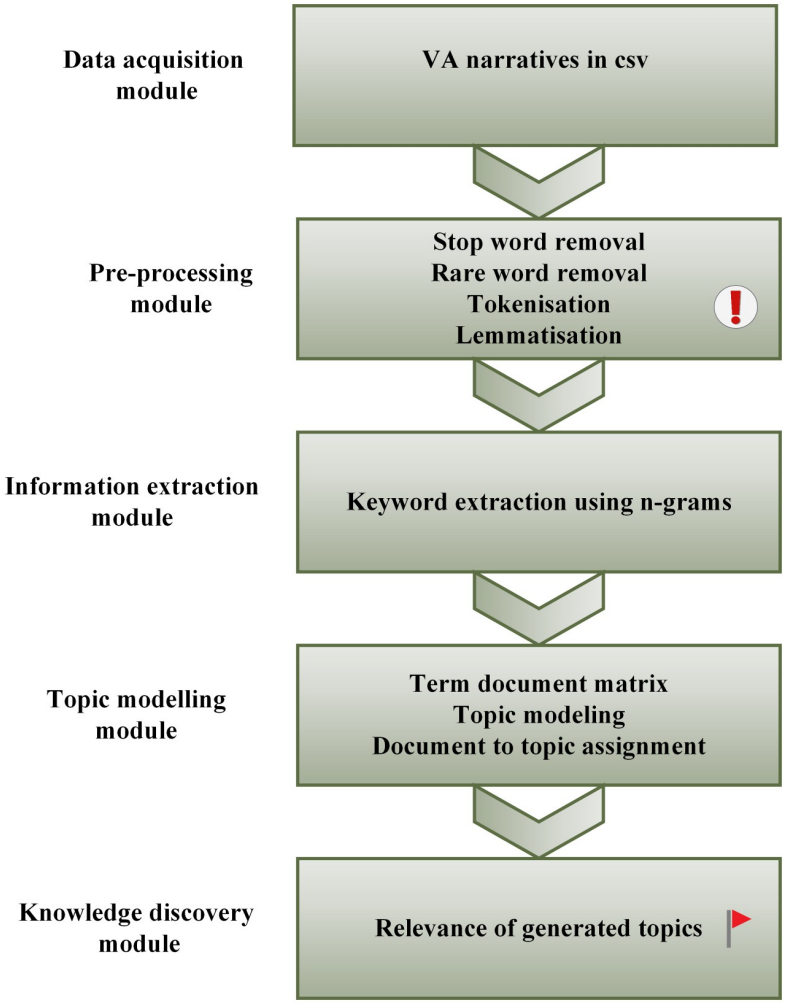


Fig 1. Text mining steps followed in this study.

https://doi.org/10.1371/journal.pone.0308452.g001

remove insignificant words using the NLTK library of English stopwords. We performed tokenisation which is breaking up of text into tokens, of words or sentences. Lemmatization was applied by employing the Python Spacy package. This process uses a dictionary of known word forms and considers the role of a word in a sentence with the aim of extracting some normal form of a word (See Fig 1, Pre-processing module).

Information extraction

After the pre-processing phase, we now had our tokens as words, and at this stage features could now be defined. The features were extracted as tokens of words, using n-gram natural language processing. The n-grams were seen as tokens of adjacent words denoted as; $n = 1$ a unigram such as, *'pain'*, $n = 2$ a bigram, such as *'chest pain'*, $n = 3$ a trigram, such as *'high blood pressure'* and $n = 4$ a quodgram, such as *'blood result hiv positive'* [31]. The bigger the n-gram word frequencies, the more contextual information that we extracted. These n-grams models

gave us an understanding of our narrative corpus and denoted the frequently occurring terms, also known as symptoms or mortality causes (See Fig 1, Information extraction module).

Additionally, we also performed keyword extraction defined as a representation of a document's content using a word or a sequence of words. These words provide meaning within a sentence's context. As such, one can discover hidden patterns and correlations using TM approaches in an automated fashion from large volumes of unstructured data using keywords [7]. Recent novelties in TM, now allow the extraction of keywords from documents without the assistance of experts, a process known as expert driven.

Keyword extraction is a form of text analysis that seeks to pull out relevant keywords automatically from unstructured text. These keywords are a good descriptive summary of the unstructured text. In this study we use Term Frequency with Inverse Document Frequency (TF-IDF). This approach mainly captures the most frequently occurring word(s) in one document and less less frequently available in other documents [7] (See Fig 1, Information extraction module).

Topic modelling using LDA

The LDA components that we used entail; *a) Term document matrix*, that represents a document as a bag of words. It is also a easier way of representing the corpus, known as a word document matrix. The word document matrix is then taken as input and fed in topic modelling [5, 53], *b)Topic modelling* using LDA. This approach is statistical in nature and is used for clustering text documents. LDA makes the assumption that each document is made up several topics and on the other hand each topic has various words belonging to it. The representations are done through document-topic w distributions and topic-word distribution v. As such, w and v are considered as dirichlet distributions. In the process the topics are discovered since they are not predefined, and *c) Document to topic assignment* is the most useful process of the LDA algorithm [53].

LDA algorithm. The LDA process follows an iterative manner starting with a random sample. This implies that the LDA process seeks to maximise the probability of a document belonging to a particular topic, and the probability of a word belonging to a certain topic iteratively. Once the algorithm converges, it generates final collection of words that are representative of a particular topic. It then computes the topics based on probability, and the topic with the highest weighting will denote the document's most dominant topic (See Fig 1, Topic modelling module). It is key to find an optimal value relative to the topic distribution per document using the hyper-parameter α. A high value of α may lead to a homogeneous distribution of topics. On the contrary, a low value of α may hinder a fair topical percentage distribution and any inferences thereof [53]. The LDA approach we use is reported in [7, 10, 53].

We applied the pyLDAvis package in Python in order to discover important information about our topics. We used a coherence score of 19 topics which we derived from the exhaustive search of the optimal number of topics and α (See Fig 2). The inter topical distance graph depicts topics as circles, and the distance between the circles denotes the relationship between topics (See Fig 6). The relationship is generated through dimensionality reduction using principal component analysis which employs probability distributions. The goal of this process is to try and have optimal topics that are distinct and do not overlap. To achieve minimal overlap we have to optimise our model by fine tuning model parameters such as α and relevancy (method for ranking terms within topics) to attain terms that are most important and exclusively belong to a particular topic. This overcomes the issue of having a biased representation of a topic where highly ranked terms maybe frequent across the whole corpus [54]. In this study we follow the LDA methodological algorithm and approach reported in [53].

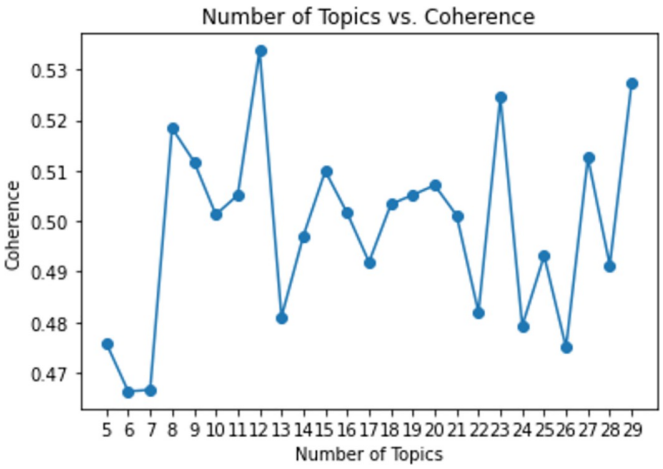


Fig 2. Hyper-parameter optimization for number of topics.

https://doi.org/10.1371/journal.pone.0308452.g002

Topic modelling using BERTopic

This present study implemented BERTopic in four steps namely; 1) converting documents to their embedding representation using a pre-trained language model, 2) using uniform manifold approximation and projection (UMAP) for dimensionality reduction, 3) applying the hierarchical density-based spatial clustering of applications with noise (HDBSCAN) algorithm to cluster VA narratives into categories that have a similar meaning, and 4) generating classes through a class-based term frequency with inverse document frequency (c-TF-IDF) weighting scheme based on importance and relevance of a word, thus denoting the most representative words for each topic [41].

BERTopic algorithm. We used the BERTopic library in Python in order to discover important information about our topics. The BERtopic algorithm starts by creating document embeddings. Documents are embedded to create representations in a vector form that is semantically related. This implies that documents that entail the same topic are semantically similar. Document embedding is done using sentence-BERT framework, thus enabling users to convert text in sentences and paragraphs to dense vector representations using pre-trained language models. As such, these embeddings can be applied to perform clustering of similar documents. However, the embeddings are not used in generating topics [55]. After creating the document embeddings, the algorithm then performs dimensionality reduction using UMAP. This results in a more representative vector space which preserves local and global features. The reduced document embeddings are then clustered into similar groups using HDBSCAN. During the clustering process, HDBSCAN also clusters noise as outliers. Ultimately, this avoids clustering unrelated documents and improves the process of topic assignment. Lastly, topic assignments are done using c-TF-IDF, and are based on the documents in a cluster, where one topic will be a cluster [40, 45, 55, 56].

Optimisation of LDA and BERTopic algorithms. We carried out exploratory topic modelling through an exhaustive search of the number of topics. In order to attain an optimal number of topics k we fine tuned α our hyper-parameter for the LDA algorithm. Additionally, we explored with number of topics from $k = 5$ to $k = 30$, and we looked at values of α that had a uniform distribution with values in the range 0–1, and these where (0.49, 0.504, 0.505, 0.508, 0.50, 0.48). After experimenting with ($k = 12$, $k = 19$, $k = 23$, $k = 29$) topics, we discovered most coherent topics using 19 topics, as most of the words in a topic were associated. It should be

noted that even if we generated high coherence scores for (k = 12, α = 0.54), (k = 23, α = 0.525), and (k = 29, α = 0.53), we settled for (k = 19, α = 0.505), since an increase in topic numbers results in difficulties in terms of human interpretability. Moreover, a large α value leads to similar generation of words in topics. Therefore, we chose lower values that follow a normal distribution and had most diverse representative terms [53]. Fig 2 shows the exhaustive search process that was done to identify the optimal number of topics.

The implementation of the BERTopic algorithm entailed optimisation through the use of UMAP to improve topic assignment. BERTopic model evaluation was done using c_v (it is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity), and c-umass (it takes into consideration the document co-occurrence counts, one-preceding segmentation, and a logarithmic conditional probability as a confirmation measure) [57]. Both measures seek to determine the degree of significance of words in a topic and the level of interpretability, thus assessing topic quality from a human perspective. The c_v measure ranges between 0 and 1. The higher the c_v score, the more understandable and coherent a topic is to a human. On the contrary, the closer the c-umass score to 0 the better [57]. The BERTopic attained a c_v of 0.66 and c-umass of −0.32. The mathematical approach is reported in the study of [40, 48].

Knowledge discovery

Zhu et al. [24] defines knowledge discovery as identifying meaningful information from large volumes of unstructured text. This knowledge can entail implicit or explicit facts, information or descriptions which relates to the contextual understanding of a specific domain. Furthermore, this discovered knowledge can serve as extra information or data that can be further utilised to discover more interesting patterns. In our context we employ knowledge discovery through data mining of the VA narratives in order to find answers to mortality causes and the most prevalent diseases that might have attributed to high mortality numbers within the HDSS. Additionally, further deductions and inferences through integration of our biomedical VA narrative data with other sources of data can give us further insights on our investigations (See Fig 1, Knowledge discovery module).

Buenano-Fernandez et al. [53] points out that at this phase we have to identify relevant topics by adding topic labels through a synthesis of the content generated by the topic assignment algorithms. However, the LDA approach has a challenge of interpretability, as the generated topics are not easily interpretable by humans because the initial assignment of words is random. Therefore, it is of paramount importance to have human interaction from beginning to the end of the process, in order to attain an optimal final model. As such, we engaged with a subject matter expert to assist with the topic labels. The expert synthesised the generated tokens and allocated a probable disease category. Similar steps and consultations were done with experts in order to get the best probable cause of death on the BERTopic algorithm results.

Data analysis

In this study we performed information extraction through word frequencies (n-grams) and keyword extraction using TF-IDF to identify the most frequent terms in the corpus, for our data analysis phase. We removed 2247 cases that had missing textual narratives. We further chucked out terms that had a high and low sparsity respectively, thus they were infrequently occurring in the corpus. We ended up with a 14079 documents for our analysis. Additionally, we also employed inter-topic distance measure through principal component analysis and

applying HBSCAN to understand how our topics are related and clustered in our corpus. Moreover, we also performed structure and semantic analysis of random twenty (20) narratives in order to investigate if the narratives have same or additional variables as compared to the symptoms/variables from the responses to the structured questionnaire.

Statistical packages

All statistical analyses were done using Python 3.11.5 using Jupyter Notebook platform. The following libraries were used; *pandas* and *numpy* for data manipulation and analysis, *nltk* for preprocessing, *matplotlib* for generating visualisation, *sklearn* for feature extraction *bertopic*, *gensim* and *spacy* for topic modelling.

Results

The extracts below describe the results attained from the n-gram natural language processing approach and topic modelling.

Term frequencies using word cloud

Fig 3 depicts the the word cloud of terms that we extracted from our VA corpus. The word cloud captures a visual illustration and representation of word frequency. The more bold and large the text is, the more important it is.

Term frequencies using TF-IDF

Table 1 depicts the top 8 word frequencies using TF-IDF, a process known as keyword extraction. We found out that terms like *swollen, pain, hot, weak, diarrhoea, headache, stomach, vomiting, swollen, stomach, mental confusion, hiv positive, high blood pressure, cough, and chest pain* had high weightings in that listing order signifying their importance in corpus. Additionally, terms like *chest pain, low energy, diarrhoea vomiting* also had high weightings in the corpus. This suggests that these terms had high frequency scores, denoting importance in a document but less common in other collections of documents. Similarly, we can deduce that these term's importance was high in given documents, but very rare in other documents.

Term frequencies using n-grams

We illustrate our word frequencies in Fig 4 using n-grams. Firstly, we depict a visual representation of the most common symptoms from our corpus using uni-grams. In a similar fashion, we managed to discover the same frequent terms as in our TF-IDF approach and the word cloud. We can see that the symptoms such as *pain, diarrhoea, swollen, hot, cough, weak, blood, stomach, chest, headache and vomiting* formed part of the most occurring uni-grams. Moreover, from our bi-grams, we can see that *(chest, pain), (mental, confusion), (swollen, stomach), (diarrhoea, vomiting), (diarrhoea, diarrhoea), (high, blood),(stomach, swollen),(blood, pressure), (stomach, pain) and (low, energy)* were the most frequently occurring symptoms. Tri-grams show that *(chest, pain, cough), (blood, pressure, high), (pain, chest, pain), (cough, chest, pain) and (high, blood, pressure)* were the most occurring symptoms. The quad-grams show that, *(diarrhoea, vomit, diarrhoea, vomit),(vomit, diarrhoea, vomit, diarrhoea), (chest, pain, stomach, pain), (cough, chest, pain, hot), (pain, cough, chest, pain), (chest, pain, low, energy), (blood, result, hiv, positive), (cough, chest, pain, cough), (high, blood, pressure, diabetes), (chest, pain, chest, pain), (high, blood, pressure, high)* as the most occurring symptoms. Based on these word frequencies, one can possibly deduce that, these symptoms were the most common mortality causes of our HDSS population.

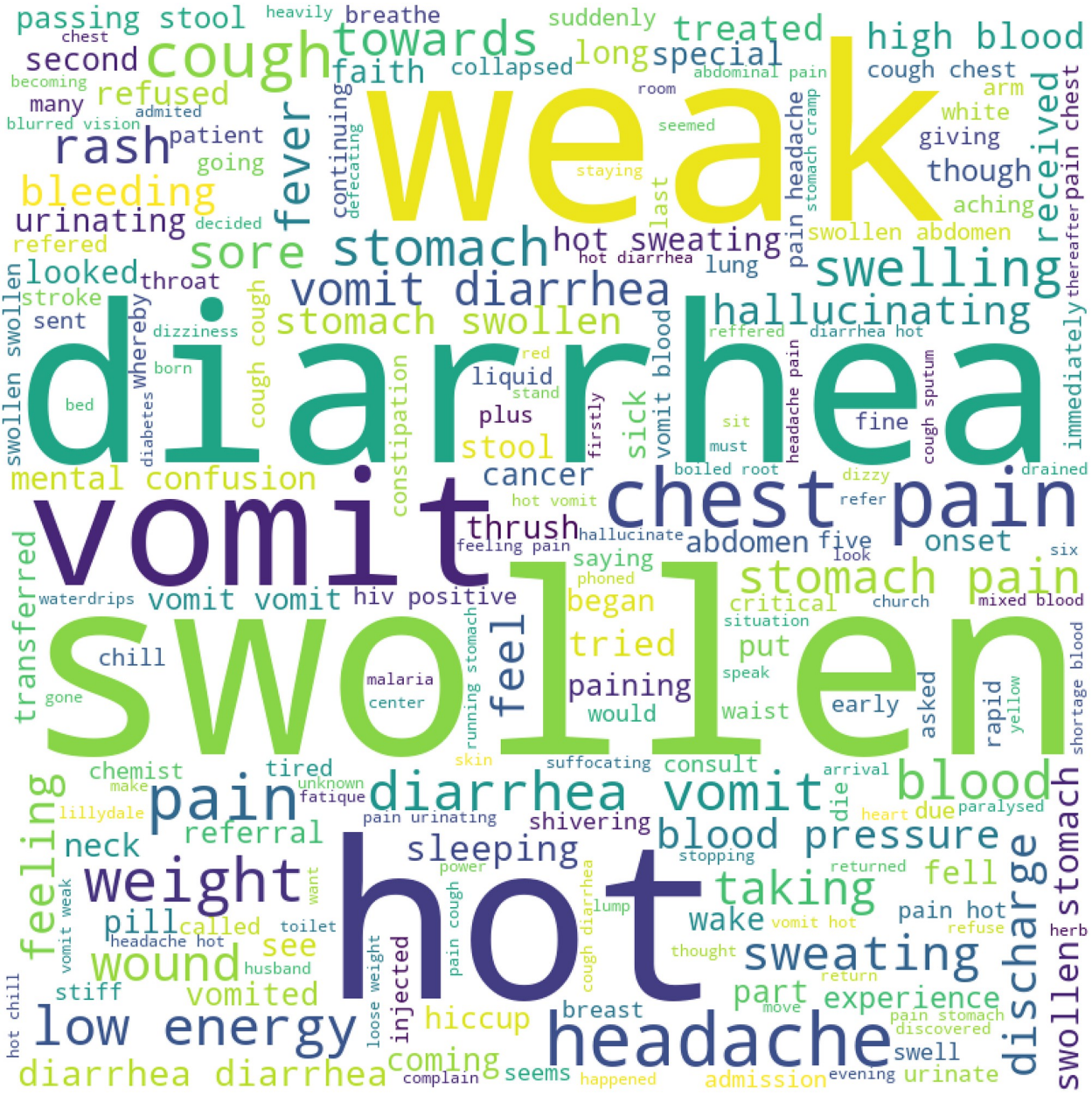


Fig 3. Word frequencies using word cloud.

https://doi.org/10.1371/journal.pone.0308452.g003

Semantic and structure analysis of additional and similar variables

This section discusses results of the structure and semantics analysis of twenty random cases that we extracted to investigate if the narratives had additional information or exhibit the same symptoms as in the responses from the structured questionnaire. Interestingly, we note that the narratives entail the same symptoms as in the responses from the structured questionnaire, and further analyses gave us insights that they also entail additional variables that can be used to help with disease diagnosis especially in cases where the cause of death is unknown or there

Table 1. Most frequently occurring terms.

Term	TF-IDF
pain	0.0926
swollen	0.0925
weak	0.0763
diarrhoea	0.066
vomit	0.058
hot	0.057
cough	0.055
chest pain	0.053

https://doi.org/10.1371/journal.pone.0308452.t001

is a disagreement between physicians. Table 2 below is a comparison of symptoms from the responses from the structured questionnaire and symptoms the extracted textual data. Symptoms highlighted in the colour yellow denote similar symptoms and orange colour additional symptoms. We used anonymous record identifiers. Fig 5 is a VA narrative example that shows how text was preprocessed to generate clean processed symptoms/tokens that were used for the comparative analysis with responses from the structured questionnaire.

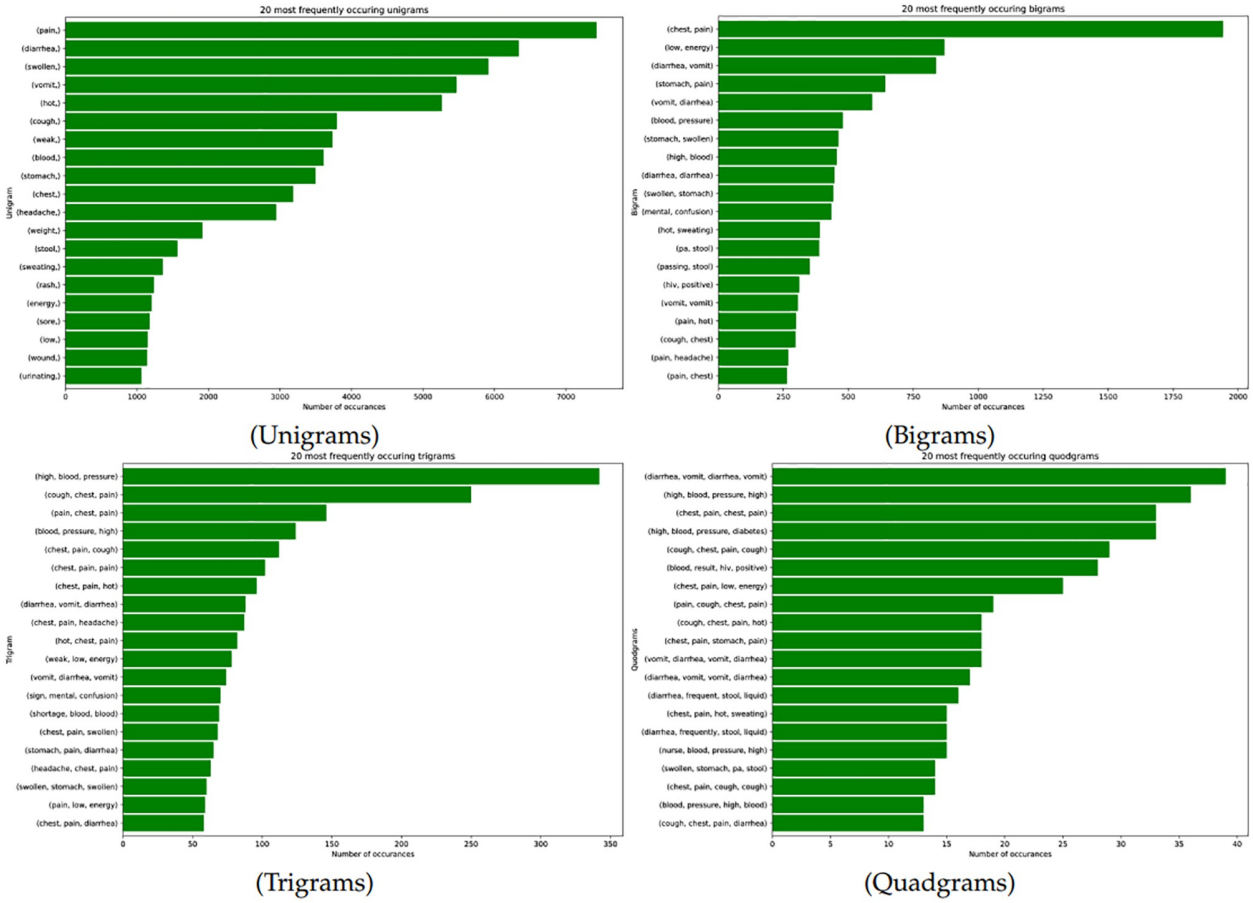


Fig 4. Most prevalent mortality causes using n-grams language processing.

https://doi.org/10.1371/journal.pone.0308452.g004

Table 2. Comparison of symptoms from textual data and responses from structured questionnaire.

RecID: Symptoms from questionnaire	Extracted symptoms from textual data
A1: fever, cough, breath, chest pain, whooping	cough, whooping, cough, screaming, vomited, hot, swollen
A2: cancer, fever, sweating, breath low, yellow, abdomen pain, urine, hair, weight loss	stomach, vaginal bleeding, womb, cancer treatment, painful,stomach bleeding, urinating, swollen breast
A3: hypertension,asthma,low breath, weight loss, swollen legs, alcohol intake	passing stools, swollen
A4: hypertension, skin, urine, weight loss, swollen legs, excessive drinking	develop, sore toe, toes infection, wounds, grow infection, smell, worms, swollen
A5: fever, breath, chest pain, weight loss, excessive drinking, sweating	carrying, heavy box, twisted, pains, toes, affected, toes, swell, pain, blood, clot, removed, swell, huge, lump bursts, holes lump, chest, armpits, groin, throat, lumps, busted, puss coming
A6: TB, epilepsy, fever, cough, breath low, chest pain, diarrhoea, convulsion, vomiting, stiff neck	swollen, diarrhoea, stools, cough, go, toilet
A7: nil	pain, treated, looking better, weak, admitted, drips injection, worse
A8: hiv, fever, cough, low breath, yellow, alcohol, diarrhoea, vomit, weight loss, abdomen, swelling	sweats, lose weight, vomit, weak, hiv, water drips, diarrhoea, hallucinating, swollen stomach
A9: sweating, cough, low breath, urine, weight loss, paralysed, alcohol	cough, heavily, rot, worms, rotting, spread, thigh, buttocks, waist, extent, intestines, visible, abdomen
A10: fever, cough, vomit, weight loss, alcohol	cough, signs, vomits, low, energy, cough critical, tried, natural, fever
A11: hypertension, fever, low breath	fell, forehead, improve, speak, tears, rolled eyes, mixed blood, accident, urinate
A12: fever, cough, low breath, abdomen, swelling, headache, injury	initially, swollen ears, swollen, abdomen, critical, puss, ears, fever, cough
A13: nil	swollen, wounds, top, feet, wounds, burst, puss, swollen, feet, cuts, razor blade, rubbed
A14: fever, low breath, chest pain, diarrhoea, vomit, weight loss, hair,eyes sunken	committing, thinner, diarrhoea, oral, dehydration, solution, vomit
A15: chronic, fever, skin, weight loss, swollen legs, alcohol, smoking	pain, toe, long, kneel, cut, flesh wound, toe, swollen toe, infected
A16: fever, cough, diarrhoea, blood diarrhoea, vomit, weight loss	diarrhoea, vomit, weight loss, oral, dehydration, thrush
A17: hiv, stroke, low breath, weight loss, alcohol, chest pain	fell, weak, touch, hold, feeling, pains, arv, years, clear, pass, stools, giving, arv, sweats, minutes

https://doi.org/10.1371/journal.pone.0308452.t002

We can see that from Table 2, we managed to elicit similar symptoms from the textual data as compared to the responses from the structured questionnaire (see A1, A2, A8, A10, A16, A17). We also note that the extracted textual narratives also exhibit additional symptoms that can be easily interpreted to assist physicians in reaching a diagnosis (A5, A8, A9, A11, A14, A15, A17). On average from our random sample of cases, the retrieval rate of additional is approximately $n > 2$, per every case where n is the number of terms. Moreover, we discovered that the text-extraction algorithm misses a symptom that is in the responses from the questionnaire cases ≈2% of the times. This can be attributed to the fact that some of the narratives entail shallow content and are less informative than others.

UNPROCESSED VA TEXT NARRATIVE
Deceased started to ill while at working place, He came home while experiencing cough with chest pain, difficult in breathing, tiredness and blood vision. The after visited Belfast clinic to get treatment but no improvement. Afterwards deceased complained of stomach pain. Then after experienced diarrhea. He was given traditional medicine but did not change. Afterwards he vomiting worms and diarrhea continued. He continued using traditional medicine and the condition remains the same. Three days before death deceased sneezed a thing like a worm. He died at home and he also experienced hot body. It was examined that his chest and throat developed wounds. Treatment given but no change. His lower lip also had rash that at time chapping and a lot of blood will comes out. After treatment that lip became healed He was taken to traditional healer, but condition unchanged. He was taken Tintswalo hospital, where he was admitted Oxygen supplier was given but he finally passed away on the third day at hospital. A week before death he complained about body pain. At the beginning deceased also had cough and complained of headache during the night only throughout the illness. A month before death he experienced hiccup which continued until death but recurrent, he skips days not defecating When defecate the stool were hard then after yellowish and black few days before death. Deceased also developed ring worms on both checks but healed before death
PROCESSED VA TEXT NARRATIVE
['cough', cough',' chest',' pain',' tiredness',' blood',' vision',' stomach',' pain',' ' vomit',' worms','diarrhea',' sneezed',' worm',' hot',' chest',' throat',' ' lip',' rash',' chapping',' blood',' lip',' pain',' cough',' headache',' hiccup'," defecating',' defecate',' stool',' yellowish',' ring','worms']

Fig 5. An example of a processed VA narrative.

https://doi.org/10.1371/journal.pone.0308452.g005

Topic modelling results using LDA approach

In this section we illustrate through graphs the results of our LDA process. Table 3 shows the results of the common 19 topics that we mined from our VA narratives (the most prevalent diseases). Additional visualisation is also given on the inter topic distance map in Fig 6. Basically, we noted that the most prevalent diseases were mainly communicable and non communicable. The highlights being *diarrhoea, TB, HIV, malaria, chronic ailments, neurological*

Table 3. Most prevalent diseases identified using LDA.

Topic number and name	Tokens
T1: Diarrhoea/respiratory tract infection	cough body diarrhoea weight hot sputum rapid pains vomit headache rash chills
T2: HIV	diarrhoea vomit hiv hot weak positive refuse arv blood cd count stopped
T3: Tuberculosis	cough pain chest sputum hot lungs weight results sharp spit blood lungs
T4: Cervical cancer	bleeding blood cancer urinate swollen urine wound urinating coming vaginal nasal puss
T5:Chronic liver ailment	paining stools pass pains liver weak running constipation stomach admitted shortness tired
T6:Neurological disorders	diarrhoea looked mental confusion hence vomit weight encountered abscess suddenly improve though
T7:Unknown	pains chest headache feeling hot hallucinating abdominal migrant faith breathe energy low
T8: Breast cancer/infection	wound swelling swollen feet sore puss lump burst ointment wounds breast sent hole hip infected
T9: Chronic ailments (liver,lung, kidneys)	swollen abdomen liver discharge diarrhoea vomit yellow drained lungs kidneys lungs swell
T10: Malaria	received malaria breast oral groaning feeding diarrhoea vomit toilet problems dizziness dehydration
T11: Jaundice/malaria/skin infection	rash sores vomit itchy yellow diarrhoea thrush hot eyes sunken weak sweating
T12: Unknown	panado syrup lillydale years liquid problems toilet die go oral whenever going
T13: Diabetes/Hypertension	high blood pressure stroke collapsed diabetes low hot energy shivering paralysed hours
T14: Mental confusion	feels starts refuse injected admission confused weak hiccups care mental makes sick
T15: Accidents	car accident shot road broken truck driving taxi fell collapsed paralyzed hanged
T16: Septic wound/heart ailment	wound received swelling began whilst swollen stomach cramps herbs hot bigger sore
T17: Epilepsy	stiff tongue teeth bent gone onset epilepsy jaws pills uncontrolled neck sudden
T18: Chronic ailments (heart)	heart fast swelling failure beating collapse urinate pass penis catheter bladder paralysis
T19: Maternal and Asthma	crying hours born room pregnant breast delivered delivery asthma breastfeeding cried incubator

https://doi.org/10.1371/journal.pone.0308452.t003

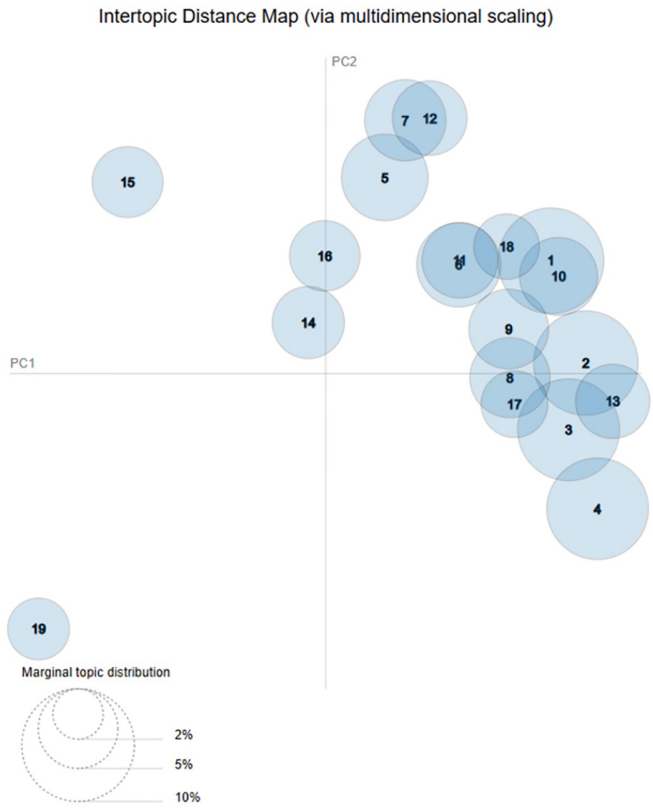


Fig 6. LDA inter-topic distance map using multi-dimensional scaling.

https://doi.org/10.1371/journal.pone.0308452.g006

disorders, asthma, accidents and maternal related deaths. We also had topic 7 and 12 which had fewer symptoms to relate to any disease and we categorised them as unknown. Expectedly, we note that maternal and accident related deaths which are topic 19 and 15 respectively are distinct and far from the other clusters. This is attributed to the fact that they have very dissimilar symptoms from the other topics. We also discovered that topic 14 and 16 which are mental confusion and heart ailment respectively, are on their own cluster. This suggests that the two topics share some common similar symptoms. Another interesting observation is the cluster with 5, 7, and 12 which are topics for chronic liver ailments and unknown topics respectively. This grouping depicts similar symptoms for unknown which are also present in the liver complications. This suggests that cases with such symptoms in the unknown can have symptoms that lead to liver complications.

One interesting discovery is that topic 2 on HIV is central to most of the clusters such as TB, diarrhoea, malaria, cancer, epilepsy, diabetes, hypertension, and chronic ailments. One possible reason for this is that most of these disease are related to HIV and most cases who have these complicated diseases also have HIV, or might be at risk of having HIV, as evidenced by the similar symptoms. Furthermore, we also observe that jaundice/ malaria with skin infection is related to neurological disorders as evidenced by the close cluster embedding of topic 6 in 11. This suggests that some cases had a long untreated jaundice which might have led to brain damage, a condition known as kernicterus. Of particular interest is cluster diarrhoea with respiratory tract infections, malaria and chronic heart ailments (topic 1, 10 and 18). This suggests that these diseases share similar symptoms and if one suffers from the earlier diseases,

it can lead to heart chronic ailments. Another finding is that cluster 8 and 9 on breast cancer and chronic ailments respectively have similar characteristic features. This suggests that if one has cancer they might also have complications of liver, lungs, and kidneys.

Fig 6 depicts the relationship and association between topics using the inter topic distance measure through principal component analysis. This process basically illustrates and models the visual representation of similar topics being in close proximity (one cluster) and dissimilar topics distant to each other. The bigger the topic circle the more tokens it uses from the corpus for topic generation. As such, we can see that topic 1, topic 2, topic 3 and topic 4 use 8.8%, 8.6%, 8.2% and 8.1% of the tokens in the corpus respectively. Topics that are overlapping imply having some similar characteristics symptom wise, hence they fall within the same cluster. We can see that topic *15* and *19* are distinct topics. Noticeably, we can also see clusters *(14 and 16)*, *(5, 7, 12)*, *(3,4)*, *(2, 3, 13)*,*(9,8,17)*, *(11,6)*, *(18,1,10)* having similar characteristics.

Topic modelling results using BERTopic approach

This extract uses figures to depict the results of our BERTopic model in topic assignment. Initially, the model generated 154 topics, identified 6265 documents as outliers. UMAP was applied for dimensionality reduction and we were left with 30 topics. Fig 7 shows the generated thirty topics and corresponding top words with scores for each topic. Topic 0 is about HIV that has top word scores of *hiv, positive, arv, count*. Topic 1 shows skin related diseases. Topic 2 is TB as evidenced by symptoms with high word scores such as *sputum, cough, smoking, chest, lung*. Topic 3 suggests septic wound diseases which might be attributed to other diseases such as diabetes. Topic 4 is unspecified. Topic 5 is a sexually transmitted diseases as evidenced by the genital related symptoms. Topics 6 and 8 are unspecified. Topic 7 and 14 denotes symptoms that might be related to diarrhoea. Topic 9 is about abdominal swelling complications. Topic 10 is about malaria. Topics 11 *shot, gun, bullet, police, shooter*, 17 *stabbed, knife, fight*, 19 *car, accident, road, injured, driving*, and 20 *fire, paraffin, burned, burnt* are accident related causes of death. Topic 12 is maternal related death as highlighted by these high word weightings *infant, feeding, breast, born, milk*. Topics 13 shows symptoms that might be related to other diseases such as HIV. Topic 15 suggests diabetes and hypertension related causes of death as denoted by these terms *diabetes, toe, high, pressure, cut*. Topic 16 suggests ulcers and TB related symptoms. Topic 18 might be related chronic liver ailments and jaundice like disease. This is shown by high word terms like *yellow, eye, liver, urine, yellowish, substance*. Topic 21 is related to stroke and paralysis possibly because of diabetes and hypertension causes, with terms such as *stroke, paralysed, collapsed, paralysis, stiff*. Topic 22 talks to cervical cancer prone in women as shown by these terms *vaginal, bleeding, cancer, cervix, cervical*. Topic 23 shows symptoms that are related to diseases such as malaria and HIV. Topic 24 depicts symptoms that are TB related. Topic 25 is a result of epilepsy evidenced by top word scores such as *epilepsy, seizure, tongue, uncontrolled, stiff, jaw*. Topic 26 exhibits hiccup symptoms coupled with pain and low energy. This suggests possible diseases such as pneumonia, bowel diseases, pancreatitis, pregnancy, bladder irritation, liver cancer or hepatitis. Topic 27 show symptoms related to diabetes such as consciousness, sweating, hot body and sunken eyes. Topic 28 is related to psychiatric conditions because of high weighting on terms like hallucination. Topic 29 is unknown.

Fig 8 shows the inter-topic distance map with 5 dominant clusters. Cluster 1 is made up of topics (5, 18, 22). These are correlated due to genital mortality causes both in men and women. This cluster suggests sexually transmitted diseases and infections that ultimately lead to liver and chronic ailments. Cluster 2 is made up of topics (3, 9, 15). This cluster denotes diabetes and high blood pressure causes of death. These chronic diseases can lead to amputation and

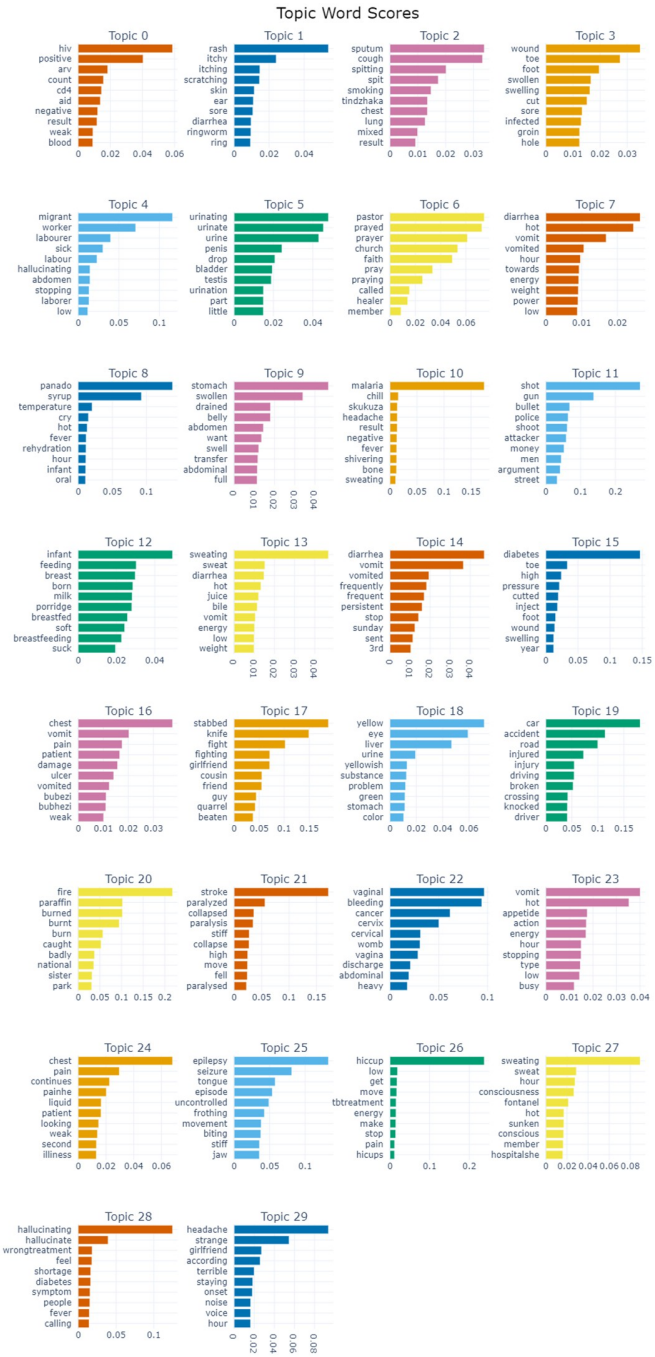


Fig 7. Topics generated using BERTopic model.

https://doi.org/10.1371/journal.pone.0308452.g007

abdominal complications. Cluster 3 shows correlation of topics (11, 19, 20) that are accident and injury related causes of death. Cluster 4 shows the dominant HIV as topic 0 and closely correlated to topics (1, 4, 7, 8, 12, 13, 14, 23, 27, 29). This suggests that most of the HIV diseases are related to diarrhoea, and symptoms such as *low energy, no appetite, headaches, skin*

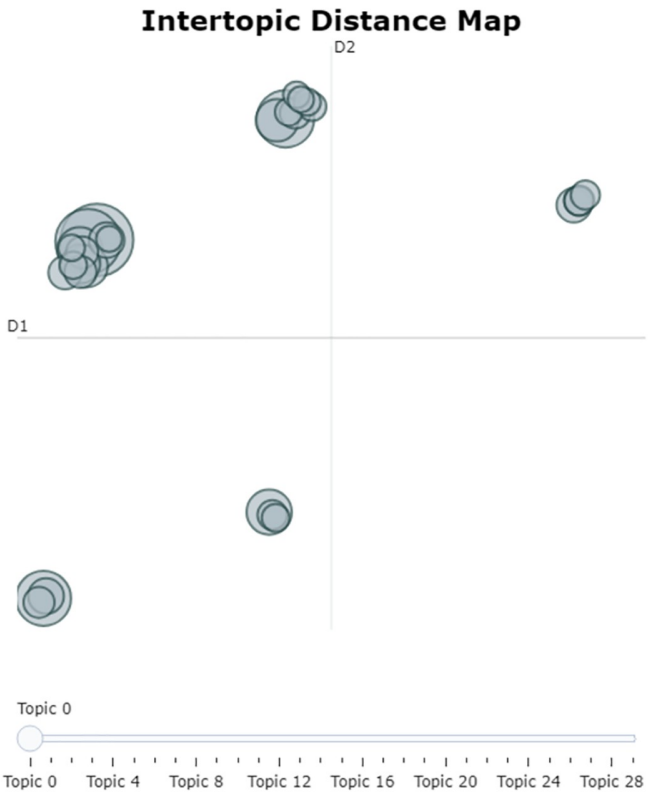


Fig 8. BERTopic inter-topic distance map.

https://doi.org/10.1371/journal.pone.0308452.g008

infections, vomiting, loss of weight, sweating. Interestingly, we observe that similar symptoms are correlated to maternal deaths. Noteworthy, is the striking number of migrant workers who exhibit same symptoms and also succumbed to death from HIV and maternal related diseases. Cluster 5 is made up of topics (2, 6, 16, 21, 24, 26, 28). Topic 2 which is related to TB and lung diseases is closely correlated to topic 16 that talks to chest pains. Noteworthy, we observe epileptic seizures being correlated to loss of consciousness, paralysis and collapsing.

Fig 9 shows a heatmap similarity matrix with distinct topics, generated from BERTopic document embeddings. This was a result of applying the cosine similarity to document embeddings to get the relationship and distinction between topics in a quantifiable way. Noteworthy, the similarity score on the topics on HIV, skin related diseases, migrant workers, malaria, TB, chest pain, hallucination, septic wounds, diabetes and hypertension, and epilepsy are very high amongst others as depicted in Fig 9.

The results illustrated in Fig 8 are also visually depicted in our hierarchical clustering in Fig 10. For example, as observed from the data, the closely clustered linked topics in this group are *11, 17, 19, 20*. These are related to accident and injury mortality causes. This suggests that there were more deaths from road accidents, gender based violence and fire accidents. There is need to raise awareness to avoid the surge of such cases. Interestingly, we also note that HIV, diabetes, hypertension, TB, chronic lung and liver ailments and cervical cancer are closely clustered. This might be attributed to the fact that TB leads to lung and cancer complications. Furthermore, most HIV patients are prone to secondary diseases, that eventually worsen their condition and give rise to more deaths. Noticeably, we observe that patients with liver cirrhosis

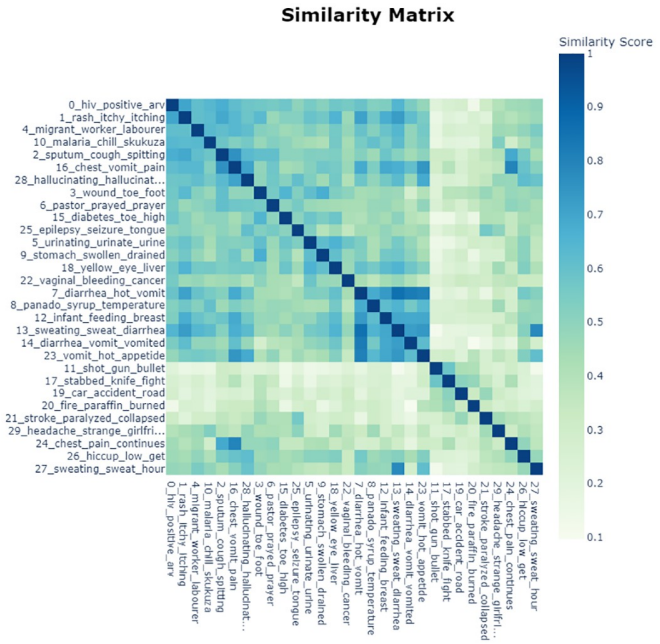


Fig 9. Heatmap similarity matrix.

https://doi.org/10.1371/journal.pone.0308452.g009

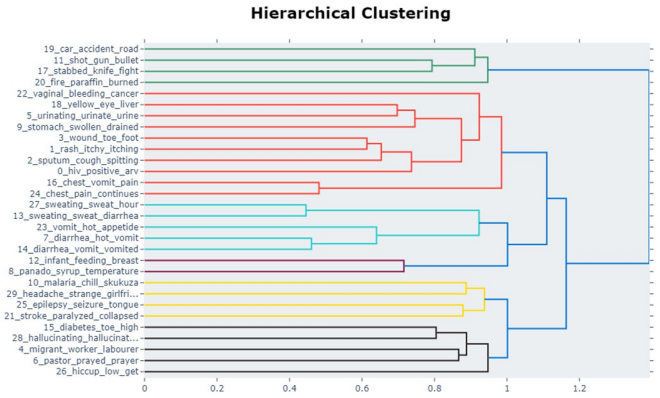


Fig 10. Hierarchical clusters of similar topics.

https://doi.org/10.1371/journal.pone.0308452.g010

often present with urinary complaints as shown in the closely clustered topics 5 and 18. Noteworthy, is the diarrhoea cluster exhibiting typical symptoms such as *sweating, low energy, weight loss, hot body, sunken eyes*. As expected in topic 12, we note an isolated cluster on maternal related death at birth. As observed, we see that epilepsy seizures are closely linked to stroke and paralysis. As such, there is need to invest in more care workers to look after epileptic patients. Interestingly, we note that there is a high correlation on diabetic cases that result in hallucinations. Therefore, there is great need to invest in health awareness programmes, improved primary health care and early screening processes to try and curb the rise of these diseases.

Results from the knowledge discovery process: Identified relevant topics

Our analysis suggests that most of the population in the HDSS could have died from diseases such as diarrhoea, TB, HIV, cancer, maternal, accident and injury and chronic ailments. Our LDA topic clustering shows that accident (T15) and maternal related deaths (T19) were distinct and distant from other clusters. A similar pattern and trend is also noticeable on T4 which is cervical cancer. We notice accident related deaths from the HDSS, but mostly intriguing are suicidal deaths. Of interest is cluster T1, T10 and T18. It shows that cases that succumbed to diarrhoea, respiratory tract infections, malaria and heart ailments related deaths, exhibited mostly similar symptoms. Another paramount discovery is cluster T2, T3 and T13. This cluster depicts that subjects who died from HIV also had other conditions related to TB and diabetes. Furthermore, we can also note that HIV related deaths were also associated with cancers and chronic ailments (T8, T9). T17 shows epilepsy which is also a neurological disorder. This disease is clustered mostly with T8, which is breast cancer. Additionally, we can also see that it shares some similar characteristics with HIV and TB. We can also deduce that cancer of the breast or breast infections have some common symptoms with chronic ailments of the liver, lung and kidneys as shown on the T8 and T9 cluster.

Cluster T5, T7 and T12 show that some deaths from liver ailments can lead to complications from not easy to identify symptoms which are in the unknown topics. One important finding is that HIV related cases share common symptoms with TB related deaths. From the HIV topic, through token analysis, we notice that some of the subjects were defaulting from taking their medication *stopped, refused*. Interestingly, we notice that some cases where affected by hypertension and diabetes, and this led to stroke and paralysis. This finding implies that diabetes and high blood pressure could be indicative of a form of heart metabolic syndrome. Moreover, hallucinations and mental confusion can happen with very high blood sugar levels (T7). We also discovered some cases of jaundice related deaths possibly prone in the new born babies. This calls for better controlled and effective immunisation measures in order to curb such diseases. Similar findings are also exhibited on our BERTopic analysis, as we observed from the discussions in previous paragraphs.

Discussion

The present study's novelty was to apply text mining techniques specifically n-gram natural language processing, LDA and BERT in topic modelling to investigate mortality causes and most prevalent diseases, using VA narrative data. As such, we managed to explore the relevancy of text mining techniques in order to discover mortality causes and the most prevalent diseases. The rich textual narratives availed valuable insights by extracting implicit knowledge on symptoms and cause of death. This is consistent with previous findings by [3]. Moreover, this study, findings are correlated with results from our previous work reported in [18, 58].

We first present the most relevant terms that were the main causes of mortality and identify the main topical diseases that were prevalent in the HDSS population. Additionally, we also managed to perform topic modelling, thus identifying interesting patterns and trends based on disease symptom similarities and differences. The generated associated disease clusters provide a statistical potential on discovering semantic level associations. Of interest, are findings from semantic and structure analysis of random cases, that show that the narratives exhibit similar and in some instances have additional symptoms as compared to those in the responses to the structured questionnaire. These findings can therefore be used to guide the health community and close the civil registration system gap. Our work also provides a methodological framework for text mining and topic clustering. Therefore, robust automated unsupervised

systems that analyse VA narratives and avail interesting insights for public health improvement are vital for the community and world at large.

The findings on the n-gram natural language processing suggest that the HDSS population could have died, mostly from mortality causes such as *diarrhoea, fever, headaches, vomiting, pain, blood pressure, low energy, hot/swollen body*, amongst many. Interestingly, using topic modelling to identify most prevalent diseases, we discovered that communicable diseases such as HIV and tuberculosis were more prevalent in the HDSS. Interestingly, non communicable diseases such as diabetes, chronic ailments (kidney, heart, lung, liver), diarrhoea, cancer, respiratory tract infections, neurological disorders, hypertension and mental health ailments, were also very common in the HDSS. Expectedly, we notice distinct topics for maternal and accident and injury related deaths. This suggests that these topics have least relationships with other topics as they have dissimilar characteristics/ symptoms. A very striking observation is that neurological disorders is embedded in jaundice/ malaria/ wound with skin infection. This can be attributed to the fact that some cases had a long untreated jaundice which might have led to brain damage, a condition known as kernicterus. A similar pattern and trend is exhibited in malaria and diarrhoea and tract respiratory infections. Our topical distance map shows that HIV is central to most clusters (TB, cancer, epilepsy, diabetes, hypertension, chronic ailments), thus implying that most of the diseases exhibit symptoms as in HIV. Consequently, mortality from HIV can also be a result of complications from the earlier diseases. Noteworthy, is the fact that most people who suffer from uncontrolled hypertension and diabetes are prone to stroke and paralysis. The above findings are consistent with our initial work reported in [18, 59].

On the contrary, we note two unknown diseases, that have less informative symptoms to get to a cause of death diagnosis. However, we can see that these topics are also in the same cluster with chronic related liver ailments. Therefore, in such cases the doctors can merge the fewer symptoms in the unknowns and the ones in the known cause of death topic to probably get to a diagnosis as the symptoms are related and associated.

Our structure and semantic analyses can further provide insights that can help the public health community to extract additional implicit knowledge and symptom associations about cause of death from the large VA data. This can prove beneficial especially in cases where the physicians disagree on the cause of death and also where there is limited diagnosis information (symptoms)on the responses from the structured questionnaire. Interestingly, interpretability of CoD is made easier using our word sequence modelling approach that identifies semantics and relations in text. As such, this will save time, is cost effective and improves the diagnosis turnaround time. Of paramount importance is that, this present study methodological approach improves knowledge representation of VA textual data which supports multi disciplinary research across disciplines (See Fig 5).

Our initial study also reports on interesting findings on the most sought after treatment method when the HDSS population is faced with terminal illnesses [18]. Even though this was not the aim of this study, we notice a similar trend and pattern in the current study, as people mostly seek the traditional ways. Noteworthy, we observe diseases such as HIV, that are affecting immigrants in the HDSS. This suggests that that immigrants are also spreading such diseases. Consequently, there is great need for further investigations as this might lead to identifying subjects who default from medication and western ways of getting medical help. As such, this can possibly suggest some of the premature deaths. Informed by the study of Kabudula et al. [60], socio-economic status also plays a big role in morbidity and mortality, hence there is a great need to try address this challenge in order to improve public health.

Literature reports on the threat of non-communicable diseases and chronic diseases. Furthermore, they suggest effective implementation of primary health care systems as well as

scaling up chronic based delivery, in order to curb the high mortality numbers [61]. This present study, also supports the notion raised by these authors in an effort to curb high mortality numbers.

It is imperative to improve population health, through effective intervention and support programmes, and early screening processes that are well monitored and controlled through initiatives by relevant authorities. This can prove beneficial especially for diseases such as cervical cancer.

This present study will lead to improved public health planning through the use of the findings ton inform public health strategies by identifying major health issues contributing to mortality, allowing for targeted interventions and resource allocation. Moreover, the results can be used to develop evidence-based policies addressing the specific health challenges prevalent in a given population, leading to more effective healthcare delivery. The application of topic modelling to VA data represents an innovative approach in health research, showcasing the potential of NLP in extracting valuable insights from unstructured medical narratives.

Limitations and future work

This study was limited in complexity of the VA narratives used. The analysis entailed some terms in topics which were of little significance. This is consistent with the findings of Hacking et al. [34]. Moreover, topic segmentation has a challenge of human interpretation of discovered topics. The narratives were highly unstructured, and entailed high levels of noise, sparsity, had varying technical vocabulary terms, contain misspelled terms, had many grammatical errors, and exhibited issues around word polysemy. Nevertheless, the findings of this study still stand despite all these limitations.

Informed by the findings of this study, specifically the semantic and structure analysis, our future direction will seek to identify under-reported morbidity occurrences at the Agincourt HDSS, by effectively linking HDSS population data and NHLS public health data registries. This can further give us insights on other factors that might be leading to higher mortality numbers.

Conclusion

The application of the unsupervised TM approaches in the current study creates a synergistic framework that leverages the strengths of each method, providing a comprehensive and accurate analysis of mortality causes and prevalent diseases within the VA narratives. The results obtained from this study contribute to the advancement of public health research by offering a sophisticated and interpretable methodology for extracting valuable insights from verbal autopsy data. As such, they can be included in the pipeline for identifying mortality causes and most prevalent diseases alongside human annotation, and interpretation. The increase in textual digital repositories can be fully utilised using automated TM approaches that can extract implicit knowledge on cause of death and avail valuable insights, which could not have been possible through manual processes.

Acknowledgments

We acknowledge the University of the Witwatersrand, Faculty of Health Sciences for providing high computing infrastructure resources.

Author Contributions

Conceptualization: Michael Tonderai Mapundu, Chodziwadziwa Whiteson Kabudula, Eustasius Musenge, Turgay Celik.

Data curation: Michael Tonderai Mapundu.

Formal analysis: Michael Tonderai Mapundu, Victor Olago, Turgay Celik.

Investigation: Michael Tonderai Mapundu.

Methodology: Michael Tonderai Mapundu, Victor Olago.

Project administration: Michael Tonderai Mapundu.

Supervision: Chodziwadziwa Whiteson Kabudula, Eustasius Musenge.

Validation: Michael Tonderai Mapundu.

Visualization: Michael Tonderai Mapundu, Victor Olago.

Writing – original draft: Michael Tonderai Mapundu.

Writing – review & editing: Michael Tonderai Mapundu, Eustasius Musenge, Turgay Celik.

References

1. Nichols EK, Byass P, Chandramohan D, Clark SJ, Flaxman AD, Jakob R, et al. The WHO 2016 verbal autopsy instrument: an international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0. PLoS medicine. 2018; 15(1):e1002486. https://doi.org/10.1371/journal.pmed.1002486 PMID: 29320495
2. Thomas LM, D'Ambruoso L, Balabanova D. Verbal autopsy in health policy and systems: a literature review. BMJ global health. 2018; 3(2):e000639. https://doi.org/10.1136/bmjgh-2017-000639 PMID: 29736271
3. Jeblee S, Gomes M, Jha P, Rudzicz F, Hirst G. Automatically determining cause of death from verbal autopsy narratives. BMC medical informatics and decision making. 2019; 19(1):127. https://doi.org/10.1186/s12911-019-0841-9 PMID: 31288814
4. Soleman N, Chandramohan D, Shibuya K. Verbal autopsy: current practices and challenges. Bulletin of the World Health Organization. 2006; 84:239–245. https://doi.org/10.2471/blt.05.027003 PMID: 16583084
5. Mujtaba G, Shuib L, Idris N, Hoo WL, Raj RG, Khowaja K, et al. Clinical text classification research trends: systematic literature review and open issues. Expert systems with applications. 2019; 116:494–520. https://doi.org/10.1016/j.eswa.2018.09.034
6. Danso S, Atwell E, Johnson O. A comparative study of machine learning methods for verbal autopsy text classification. arXiv preprint arXiv:14024380. 2014;.
7. Ahadh A, Binish GV, Srinivasan R. Text mining of accident reports using semi-supervised keyword extraction and topic modeling. Process Safety and Environmental Protection. 2021; 155:455–465. https://doi.org/10.1016/j.psep.2021.09.022
8. Lozano R, Lopez AD, Atkinson C, Naghavi M, Flaxman AD, Murray CJ. Performance of physician-certified verbal autopsies: multisite validation study using clinical diagnostic gold standards. Population Health Metrics. 2011; 9(1):1–13. https://doi.org/10.1186/1478-7954-9-32
9. Lu HM, Wei CP, Hsiao FY. Modeling healthcare data using multiple-channel latent Dirichlet allocation. Journal of biomedical informatics. 2016; 60:210–223. https://doi.org/10.1016/j.jbi.2016.02.003 PMID: 26898516
10. Kavvadias S, Drosatos G, Kaldoudi E. Supporting topic modeling and trends analysis in biomedical literature. Journal of Biomedical Informatics. 2020; 110:103574. https://doi.org/10.1016/j.jbi.2020.103574 PMID: 32971274
11. Nasser N, Karim L, El Ouadrhiri A, Ali A, Khan N. n-Gram based language processing using Twitter dataset to identify COVID-19 patients. Sustainable Cities and Society. 2021; 72:103048. https://doi.org/10.1016/j.scs.2021.103048 PMID: 34055577
12. Cassim N, Mapundu M, Olago V, Celik T, George JA, Glencross DK. Using text mining techniques to extract prostate cancer predictive information (Gleason score) from semi-structured narrative laboratory

reports in the Gauteng province, South Africa. BMC Medical Informatics and Decision Making. 2021; 21 (1):1–11. https://doi.org/10.1186/s12911-021-01697-2 PMID: 34823522

13. Marrero M, Urbano J, Sánchez-Cuadrado S, Morato J, Gómez-Berbís JM. Named entity recognition: fallacies, challenges and opportunities. Computer Standards & Interfaces. 2013; 35(5):482–489. https://doi.org/10.1016/j.csi.2012.09.004
14. Lucini FR, Fogliatto FS, da Silveira GJ, Neyeloff JL, Anzanello MJ, Kuchenbecker RS, et al. Text mining approach to predict hospital admissions using early medical records from the emergency department. International journal of medical informatics. 2017; 100:1–8. https://doi.org/10.1016/j.ijmedinf.2017.01.001 PMID: 28241931
15. Raja U, Mitchell T, Day T, Hardin JM. Text mining in healthcare. Applications and opportunities. J Healthc Inf Manag. 2008; 22(3):52–6. PMID: 19267032
16. Kim JC, Chung K. Associative feature information extraction using text mining from health big data. Wireless Personal Communications. 2019; 105(2):691–707. https://doi.org/10.1007/s11277-018-5722-5
17. Zaki MJ, Meira W Jr. Data Mining and Machine Learning: Fundamental Concepts and Algorithms. Cambridge University Press; 2019.
18. Mapundu MT, Kabudula CW, Musenge E, Olago V, Celik T. Performance evaluation of machine learning and Computer Coded Verbal Autopsy (CCVA) algorithms for cause of death determination: A comparative analysis of data from rural South Africa. Frontiers in Public Health. 2022;. https://doi.org/10.3389/fpubh.2022.990838 PMID: 36238252
19. Leskovec J, Rajaraman A, Ullman JD. Mining of massive data sets. Cambridge university press; 2020.
20. Danso S, Johnson O, Ten Asbroek A, Soromekun S, Edmond K, Hurt C, et al. A semantically annotated Verbal Autopsy corpus for automatic analysis of cause of death. ICAME Journal. 2013; 37.
21. Danso SO. Text Analytics to Predict Time and Cause of Death from Verbal Autopsies. University of Leeds; 2015.
22. Danso S, Atwell E, Johnson O. Linguistic and statistically derived features for cause of death prediction from verbal autopsy text. In: Language processing and knowledge in the web. Springer; 2013. p. 47–60.
23. Kamath CN, Bukhari SS, Dengel A. Comparative Study between Traditional Machine Learning and Deep Learning Approaches for Text Classification. In: Proceedings of the ACM Symposium on Document Engineering 2018; 2018. p. 1–11.
24. Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, et al. Biomedical text mining and its applications in cancer research. Journal of biomedical informatics. 2013; 46(2):200–211. https://doi.org/10.1016/j.jbi.2012.10.007 PMID: 23159498
25. Senders JT, Cho LD, Calvachi P, McNulty JJ, Ashby JL, Schulte IS, et al. Automating clinical chart review: an open-source natural language processing pipeline developed on free-text radiology reports from patients with glioblastoma. JCO Clinical Cancer Informatics. 2020; 4:25–34. https://doi.org/10.1200/CCI.19.00060 PMID: 31977252
26. Park Y, Park S, Lee M. Analyzing Community Care Research Trends Using Text Mining. Journal of Multidisciplinary Healthcare. 2022; p. 1493–1510. https://doi.org/10.2147/JMDH.S366726 PMID: 35873091
27. Luo Y, Riedlinger G, Szolovits P. Text mining in cancer gene and pathway prioritization. Cancer informatics. 2014; 13:CIN–S13874. https://doi.org/10.4137/CIN.S13874 PMID: 25392685
28. Vyas R, Bapat S, Jain E, Karthikeyan M, Tambe S, Kulkarni BD. Building and analysis of protein-protein interactions related to diabetes mellitus using support vector machine, biomedical text mining and network analysis. Computational biology and chemistry. 2016; 65:37–44. https://doi.org/10.1016/j.compbiolchem.2016.09.011 PMID: 27744173
29. Korhonen A, Ó Séaghdha D, Silins I, Sun L, Högberg J, Stenius U. Text mining for literature review and knowledge discovery in cancer risk assessment and research. PloS one. 2012; 7(4):e33427. https://doi.org/10.1371/journal.pone.0033427 PMID: 22511921
30. Conceição SI, Couto FM. Text Mining for Building Biomedical Networks Using Cancer as a Case Study. Biomolecules. 2021; 11(10):1430. https://doi.org/10.3390/biom11101430 PMID: 34680062
31. Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K. Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study. Journal of forensic and legal medicine. 2018; 57:41–50. https://doi.org/10.1016/j.jflm.2017.07.001 PMID: 29801951
32. Senders JT, Karhade AV, Cote DJ, Mehrtash A, Lamba N, DiRisio A, et al. Natural language processing for automated quantification of brain metastases reported in free-text radiology reports. JCO Clinical Cancer Informatics. 2019; 3:1–9. https://doi.org/10.1200/CCI.18.00138 PMID: 31002562
33. Marir F, Said H, Al-Obeidat F. Mining the web and literature to discover new knowledge about diabetes. Procedia Computer Science. 2016; 83:1256–1261. https://doi.org/10.1016/j.procs.2016.04.261

34. Hacking C, Verbeek H, Hamers JP, Sion K, Aarts S. Text mining in long-term care: Exploring the usefulness of artificial intelligence in a nursing home setting. Plos one. 2022; 17(8):e0268281. https://doi.org/10.1371/journal.pone.0268281 PMID: 36006921

35. DIOUF M, THIAM M, ROCHE M. New approach to discover meaningful terms to specify cause of death from narratives verbal autopsy using TF-IDF and the LDA topic model. In: IEEE EUROCON 2023-20th International Conference on Smart Technologies. IEEE; 2023. p. 502–507.

36. Cheng CH, Hung WL. Tea in benefits of health: A literature analysis using text mining and latent Dirichlet allocation. In: Proceedings of the 2nd International Conference on Medical and Health Informatics; 2018. p. 148–155.

37. Cheng X, Cao Q, Liao SS. An overview of literature on COVID-19, MERS and SARS: Using text mining and latent Dirichlet allocation. Journal of Information Science. 2022; 48(3):304–320. https://doi.org/10.1177/0165551520954674 PMID: 38603038

38. Tran BX, Nghiem S, Sahin O, Vu TM, Ha GH, Vu GT, et al. Modeling research topics for artificial intelligence applications in medicine: latent Dirichlet allocation application study. Journal of medical Internet research. 2019; 21(11):e15511. https://doi.org/10.2196/15511 PMID: 31682577

39. Breuninger TA, Wawro N, Breuninger J, Reitmeier S, Clavel T, Six-Merker J, et al. Associations between habitual diet, metabolic disease, and the gut microbiota using latent Dirichlet allocation. Microbiome. 2021; 9(1):1–18. https://doi.org/10.1186/s40168-020-00969-9 PMID: 33726846

40. Jeon E, Yoon N, Sohn SY. Exploring new digital therapeutics technologies for psychiatric disorders using BERTopic and PatentSBERTa. Technological Forecasting and Social Change. 2023; 186:122130. https://doi.org/10.1016/j.techfore.2022.122130

41. Scarpino I, Zucco C, Vallelunga R, Luzza F, Cannataro M. Investigating topic modeling techniques to extract meaningful insights in Italian long COVID narration. BioTech. 2022; 11(3):41. https://doi.org/10.3390/biotech11030041 PMID: 36134915

42. Silva RPd, Pollettini JT, Pazin Filho A. Unsupervised natural language processing in the identification of patients with suspected COVID-19 infection. Cadernos de Saúde Pública. 2023; 39:e00243722. https://doi.org/10.1590/0102-311XEN243722 PMID: 38055548

43. Williams CY, Li RX, Luo MY, Bance M. Exploring patient experiences and concerns in the online Cochlear implant community: A cross-sectional study and validation of automated topic modelling. Clinical Otolaryngology. 2023; 48(3):442–450. https://doi.org/10.1111/coa.14037 PMID: 36645237

44. Unanue IJ, Borzeshi EZ, Piccardi M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. Journal of biomedical informatics. 2017; 76:102–109. https://doi.org/10.1016/j.jbi.2017.11.007

45. Adewunmi M, Sharma SK, Sharma N, Sushma NS, Mounmo B, et al. Cancer Health Disparities drivers with BERTopic modelling and PyCaret Evaluation. Cancer Health Disparities. 2022; 6.

46. Green ST, Flaxman AD. Machine Learning Methods for Verbal Autopsy in Developing Countries. In: AAAI Spring Symposium: Artificial Intelligence for Development; 2010.

47. Inaam ul haq M, Li Q. Revealing the trends in the academic landscape of the health care system using contextual topic modeling. Data Intelligence. 2023; 5(4):923–946. https://doi.org/10.1162/dint_a_00217

48. Karabacak M, Margetis K. Natural Language Processing Reveals Research Trends and Topics in The Spine Journal Over Two Decades: A Topic Modeling Study. The Spine Journal. 2023;. PMID: 37797843

49. Baird A, Xia Y, Cheng Y. Consumer perceptions of telehealth for mental health or substance abuse: a Twitter-based topic modeling analysis. JAMIA open. 2022; 5(2):ooac028. https://doi.org/10.1093/jamiaopen/ooac028 PMID: 35495736

50. Yang FC, Lee AJ, Kuo SC. Mining health social media with sentiment analysis. Journal of medical systems. 2016; 40(11):1–8. https://doi.org/10.1007/s10916-016-0604-4 PMID: 27663246

51. Kabudula CW, Tollman S, Mee P, Ngobeni S, Silaule B, Gómez-Olivé FX, et al. Two decades of mortality change in rural northeast South Africa. Global health action. 2014; 7(1):25596. https://doi.org/10.3402/gha.v7.25596 PMID: 25377343

52. Watt J, Borhani R, Katsaggelos A. Machine learning refined: foundations, algorithms, and applications. Cambridge University Press; 2020.

53. Buenano-Fernandez D, Gonzalez M, Gil D, Luján-Mora S. Text mining of open-ended questions in self-assessment of university teachers: An LDA topic modeling approach. IEEE Access. 2020; 8:35318–35330. https://doi.org/10.1109/ACCESS.2020.2974983

54. Sievert C, Shirley K. LDAvis: A method for visualizing and interpreting topics. In: Proceedings of the workshop on interactive language learning, visualization, and interfaces; 2014. p. 63–70.

55. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:220305794. 2022;.

56. Grubbs AE, Sinha N, Garg R, Barber EL. Use of topic modeling to assess research trends in the journal Gynecologic Oncology. Gynecologic Oncology. 2023; 172:41–46. https://doi.org/10.1016/j.ygyno.2023.03.001 PMID: 36933402

57. Mifrah S, Benlahmar E. Topic modeling coherence: A comparative study between LDA and NMF models using COVID'19 corpus. International Journal of Advanced Trends in Computer Science and Engineering. 2020; p. 5756–5761. https://doi.org/10.30534/ijatcse/2020/231942020

58. Mapundu MT, Kabudula CW, Musenge E, Olago V, Celik T. Explainable Stacked Ensemble Deep Learning (SEDL) Framework to Determine Cause of Death from Verbal Autopsies. Machine Learning and Knowledge Extraction. 2023; 5(4):1570–1588. https://doi.org/10.3390/make5040079

59. Kabudula CW, Houle B, Collinson MA, Kahn K, Gómez-Olivé FX, Clark SJ, et al. Progression of the epidemiological transition in a rural South African setting: findings from population surveillance in Agincourt, 1993–2013. BMC public health. 2017; 17(1):424. https://doi.org/10.1186/s12889-017-4312-x PMID: 28486934

60. Kabudula CW, Houle B, Collinson MA, Kahn K, Tollman S, Clark S. Assessing changes in household socioeconomic status in rural South Africa, 2001–2013: a distributional analysis using household asset indicators. Social indicators research. 2017; 133:1047–1073. https://doi.org/10.1007/s11205-016-1397-z PMID: 28931968

61. Tollman SM, Kahn K, Sartorius B, Collinson MA, Clark SJ, Garenne ML. Implications of mortality transition for primary health care in rural South Africa: a population-based surveillance study. The Lancet. 2008; 372(9642):893–901. https://doi.org/10.1016/S0140-6736(08)61399-9 PMID: 18790312

D Appendices

machine learning & knowledge extraction

MDPI

Article

Explainable Stacked Ensemble Deep Learning (SEDL) Framework to Determine Cause of Death from Verbal Autopsies

Michael T. Mapundu [1,*,†,‡], Chodziwadziwa W. Kabudula [1,2,‡], Eustasius Musenge [1,‡], Victor Olago [3,‡] and Turgay Celik [4,5]

[1] School of Public Health, Department of Epidemiology and Biostatistics, University of the Witwatersrand, Johannesburg 2193, South Africa; chodziwadziwa.kabudula@wits.ac.za (C.W.K.); eustasius.musenge@wits.ac.za (E.M.)
[2] MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), University of the Witwatersrand, Johannesburg 1360, South Africa
[3] National Health Laboratory Service (NHLS), National Cancer Registry, Johannesburg 2131, South Africa; victoro@nicd.ac.za
[4] Wits Institute of Data Science, University of the Witwatersrand, Johannesburg 2000, South Africa; turgay.celik@wits.ac.za
[5] School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg 2000, South Africa
* Correspondence: michael.mapundu@wits.ac.za; Tel.: +44-74-716-38221
† 27 St Andrews, Parktown, Johannesburg 2193, South Africa.
‡ These authors contributed equally to this work.

Abstract: Verbal autopsies (VA) are commonly used in Low- and Medium-Income Countries (LMIC) to determine cause of death (CoD) where death occurs outside clinical settings, with the most commonly used international gold standard being physician medical certification. Interviewers elicit information from relatives of the deceased, regarding circumstances and events that might have led to death. This information is stored in textual format as VA narratives. The narratives entail detailed information that can be used to determine CoD. However, this approach still remains a manual task that is costly, inconsistent, time-consuming and subjective (prone to errors), amongst many drawbacks. As such, this negatively affects the VA reporting process, despite it being vital for strengthening health priorities and informing civil registration systems. Therefore, this study seeks to close this gap by applying novel deep learning (DL) interpretable approaches for reviewing VA narratives and generate CoD prediction in a timely, easily interpretable, cost-effective and error-free way. We validate our DL models using optimisation and performance accuracy machine learning (ML) curves as a function of training samples. We report on validation with training set accuracy (LSTM = 76.11%, CNN = 76.35%, and SEDL = 82.1%), validation accuracy (LSTM = 67.05%, CNN = 66.16%, and SEDL = 82%) and test set accuracy (LSTM = 67%, CNN = 66.2%, and SEDL = 82%) for our models. Furthermore, we also present Local Interpretable Model-agnostic Explanations (LIME) for ease of interpretability of the results, thereby building trust in the use of machines in healthcare. We presented robust deep learning methods to determine CoD from VAs, with the stacked ensemble deep learning (SEDL) approaches performing optimally and better than Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN). Our empirical results suggest that ensemble DL methods may be integrated in the CoD process to help experts get to a diagnosis. Ultimately, this will reduce the turnaround time needed by physicians to go through the narratives in order to be able to give an appropriate diagnosis, cut costs and minimise errors. This study was limited by the number of samples needed for training our models and the high levels of lexical variability in the words used in our textual information.

Keywords: cause of death; CNN; deep learning; LIME; LSTM; machine learning; NLP; SEDL; verbal autopsy

check for updates

Citation: Mapundu, M.T.; Kabudula, C.W.; Musenge, E.; Olago, V.; Celik, T. Explainable Stacked Ensemble Deep Learning (SEDL) Framework to Determine Cause of Death from Verbal Autopsies. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1570–1588. https://doi.org/10.3390/make5040079

Academic Editor: Luca Longo

Received: 9 August 2023
Revised: 17 October 2023
Accepted: 21 October 2023
Published: 25 October 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

More than 65 percent of the world population lacks high quality information on cause of death (CoD), accounting for two thirds of the sixty million deaths worldwide that are not assigned a medically certified cause [1–3]. This information is vital for informing civil registration systems, policies and strengthening health priorities. Most of these deaths are common in Low- and Medium-Income Countries (LMIC) as they occur outside health facilities [4]. As such, the verbal autopsy (VA) tool is used to determine CoD. A VA is a process that entails non-medical personnel interviewing the next of kin of the deceased using a structured questionnaire where they seek to elicit information on circumstances and events that could have led to CoD. The collected information known as VA narratives are then housed as text and given to two medical doctors who determine the probable CoD, and if they do not agree, a third physician is consulted, a process known as physician-coded verbal autopsy (PCVA) [3]. Detailed information on the VA process is reported in [5]. Even though the PCVA is the only gold standard used in CoD decisions, this process is expensive, time-consuming, inconsistent and subjective (prone to errors in information elicitation), amongst many drawbacks.

This has led to alternative novel approaches that use probabilities, tariff scores and statistical operations to get to the CoD [6,7]. These approaches are known as computer-coded verbal autopsy (CCVA). Recent studies have mainly focussed on statistical approaches that seek to investigate CoD determination using VA narratives by employing CCVA approaches [7–11]. Nevertheless, these approaches perform poorly because of the unstructured narratives that they take as input for CoD classification. Therefore, the CCVA approaches cannot be applied to guide health priorities as they fail to avail enough evidence where there is limited expert diagnosis [1]. On the other hand, there have been efforts to apply machine learning (ML) approaches, which have shown better results than the CCVA approach. Considerable research was conducted by applying various shallow ML algorithms that are trained on very high dimensional (where there are many predictors and outcome variables) and sparse features (predictors and outcome variables are not within the same scale) to obtain meaningful information from the VA narratives [3,12–20]. Most of these ML techniques perform poorly because of high-dimensionality data caused by a lack of feature scaling, missing effective feature engineering techniques, ineffective preprocessing strategies and a lack of data balancing.

Advances in technology, specifically in ML and natural language processing (NLP), have availed new opportunities and proficiencies in the processing and classification of textual narratives. In this study, we present novel robust DL architectures that use NLP, specifically tailored to effectively process VA narratives and predict probable CoD.

Furthermore, we aim to assess the potential impact of the DL methods in healthcare and mortality assessments. In so doing, we will be able to enhance CoD prediction accuracy by leveraging the advanced capabilities of ensemble DL approaches as compared with the existing baseline methods reported in [21]. Our DL methods will further give us insights into how it arrived at a CoD prediction, a step crucial for enforcing transparency and trustworthiness in the use of AI in disease diagnosis in the public health space. As such, this will allow us to explore the potential application of the SEDL in the real world, opening avenues for assessing its applicability beyond this study's context, such as different populations, regions and datasets, and consequently discovering new knowledge and presenting valuable insights and semantic relationships between the VA narratives and their corresponding probable CoD.

Even though DL techniques are more accurate in terms of prediction, they have a limitation in terms of complexity and model interpretability (understanding how the model performed in the way it did). Despite these challenges, DL has shown promising results because of its efficiency, superior performance, integrated feature learning and effective capabilities of attaining end-to-end learning from complex and multi-modality data [22]. Conversely, conventional statistical and ML techniques require firstly to perform feature engineering to attain effective and robust features representative of the data and then

build predictive models. There is scant literature, especially on applying advanced ML techniques known as DL models as approaches to solving VA problems.

This study enhances the transparency, trustworthiness and accountability in the use of AI-driven healthcare applications. Integrating such an explainable system in CoD diagnosis will minimise diagnosis errors, reduce bias and improve diagnosis turnaround time and cut costs, amongst many benefits. As such, it will enforce effective collaborations on the use of AI systems in aiding human interpretation and annotation of disease diagnosis, thus improving the quality of VA reporting and decision making, which is key in informing civil registration systems and strengthening health priorities.

2. Deep Learning in VA

DL functionalities, such as automated feature extraction and engineering, have triggered the application of automated diagnosis systems, something that was initially impossible with conventional ML approaches [23]. DL has produced good results and has already been applied in the health domain as in the work of [24–27]. Other studies by [27–29] also show the positive results of DL approaches. Ref. [30] pointed out that the DL approaches can improve ML results in various fields, such as speech recognition, drug discovery and object detection, amongst many. Moreover, DL models with optimal hidden layers have been developed to reveal information not easily detectable with traditional statistical and ML models. These DL architectures can achieve learning of data representation with varying levels of data abstraction when even computational techniques use few processing layers [31]. Feature engineering in DL to extract, represent and select features is performed using an end-to-end system that learns in an automated fashion from training data. Therefore, these advanced approaches can be relevant in VA data with high-dimensional sparse data.

Initially, the Convolutional Neural Network (CNN) architecture was used for image recognition using multiple data arrays. Nevertheless, further inquiries with this architecture showed better results when used for text classification using one-dimensional data structures and character levels [29,32]. Various researchers applied the basic Recurrent Neural Network (RNN), with some investigators modifying the approach for sentiment analysis and producing good results [33–38]. Other work by [30,39] used a fusion of a CNN and RNN and attained good results using a hierarchical implementation. Ref. [40] investigated a multitask learning model for CoD classification. They used a CNN and Linear Discriminant Analysis (LDA) for topic segmentation. The predicted key phrase clusters outperformed the LDA and CNN models in extracting keywords from the VA data. They employed a CNN feed-forward network that took the input of word embeddings, using 10-fold cross validation for optimisation, to generate distinct LDA topics. They attained a precision of 0.779, a recall of 0.778 and an F1-score of 0.774. However, they used a small dataset and had word clusters that were more frequent and longer than others, creating non-representative features. Ref. [41] also explored character embeddings on VA data to try and establish if their approach could improve CoD classification. They used four datasets with varying disease categories and deduced that character information improves accuracy when used with smaller datasets as the models can handle unknown words and different forms of spelling. They reported a significant improvement using models, such as a CNN. However, their study only used small datasets of less than 1000 cases, which varied in size. Ref. [42] used DL interpretable methods to extract the CoD from VAs by applying logistic regression, Bayes classifier, random forest ensemble and two variants of RNN DL architectures. Their experiments used three datasets, but high scores were attained only for the DL architecture on the neonatal dataset with a precision of 63.2%, an accuracy of 63.0%, a recall of 63.0% and an F-score of 61.3% .

Ref. [43] experimented on the prediction of death status on treatment course in SARS-COV-2 patients using DL and ML methods and reported an accuracy of 97.15% on the DL approaches and 92.15% for random forest, 93.4% for k-nearest neighbour and 99.7% for the ensemble classifier XGBoost. Ref. [44] performed an analysis of railway accident narratives using DL distributed vector representations and a RNN and CNN. They achieved

a high accuracy of 75% and an F-score of 0.65 using word2vec as the distributed vector representation of words and a CNN.

2.1. Ensemble DL

Apart from the conventional DL approaches discussed above, there have been studies that applied what are known as ensemble DL approaches (combination of several individual models) in a quest to attain a better generalisation performance. Ensemble DL makes use of an averaging or voting process of the combined models to get to a final model prediction. Ref. [23] conducted a study where they applied SEDL using the CNN for paediatric pneumonia diagnosis by using chest X-ray images. They combined conventional traditional ML approaches with DL approaches. They reported an accuracy of 98.3%, a precision of 99.29%, a recall of 98.36%, an F1-score of 98.83% and an AUC of 98.24%. They concluded that their stacked ensemble models attained optimal model performance and the findings can be used to assist clinicians in the diagnosis process. Similar work was reported in [45], which pointed out the robustness of ensemble classifiers on attaining high model performance. However, they highlighted the importance of addressing class imbalance for improved results. Ref. [46] conducted a study on the application of ensemble DL for COVID-19 case detection using chest X-ray images. They attained an accuracy of 95% and reported that ensemble classifiers can yield better performance. Nevertheless, they pointed out the need to have more data points for model improvement. Ref. [47] conducted a review on the various types of ensemble DL approaches and reported that ensemble approaches are better in terms of model performance. However, they argued that these DL models are difficult to train with smaller training samples, they are complex in nature and selecting models to include in the ensemble architecture is also a challenge. The authors further pointed out that there is a need for researchers to investigate how best to find a strategy to define the number of base learners to be included in the architecture, find a criterion for model selection and combine different fusion strategies. Ref. [23] argued that stacking is one effective ensemble DL approach, which entails creating a stack of predictions from base classifiers, also known as individual classifiers. These predictions on base classifiers are used as features, which are taken as input for training the meta learner, which is the final classifier. This results in model improvement as stacking combines the strengths of the base classifiers. Ref. [48] reported on the application of ensemble DL models for heart disease classification using data from Mexico. They reported on the accuracy and F1-scores within the range of 91% and 96%. They concluded that ensemble DL can attain high model performance, which can be used for real-time reporting and diagnosis in the health space.

2.2. Explainable AI

Although DL methods have shown promising results in various domains, they are difficult to interpret. In other words, it is difficult to tell how a model got to a final prediction due to the complexity. This is the reason why DL models are known as a black box [49–51]. As such, this has given rise to the field of Explainable Artificial Intelligence (XAI), which strives to ease DL model interpretability. Ref. [50] pointed out that XAI approaches can be categorised into intrinsic and post hoc. The authors elaborated that intrinsic approaches are easy to understand as to how they got to the decision-making process. On the contrary, post hoc approaches help us understand the context of the input data that leads to a classification decision in any classifier. They usually use visual explanations, local explanations and explanation by simplification, amongst many functionalities. Ref. [52] argued that XAI can be applied in the medical field to aid clinicians in getting to a diagnosis in a way that is transparent, understandable and explainable. This can be effectively implemented only if ML models are able to justify how they arrived at a decision. Additionally, the authors further elaborated that if XAI is employed properly, it will increase the chances of trusting and implementing such strategies within the medical domain. The authors applied XAI to aid in human decision support within the medical space. They applied three different strategies and concluded that if properly implemented and improved, these approaches

may prove key in the decision-making process for clinicians. Ref. [49] also supported the above described notion and argued that most of the DL model predictions are black boxes and they lack interpretability on how the model got to a decision. As such, this makes it difficult for the health experts to be able to understand and trust the results generated by ML models, hence the limited use within the public health space. In order to build trust in model predictions, there is a need to incorporate the explainability and interpretation of the results generated by the models. This may position DL models as an encouraging choice in its application and usage in healthcare to improve healthcare outcomes.

In this study, we used Local Interpretable Model-agnostic Explanations (LIME), because it eases interpretability and provides key meaningful information for decision making [51]. The XAI approach gets to a decision based on validating model classifier behaviour in close proximity with the cases to be explained based on the local models, such as logistic regression or decision trees [52]. Ref. [49] pointed out that XAI can be thought of as features that describe how a particular complex model derived its predictions. Therefore, it makes it easier for a user to understand how it got to a decision. XAI highlights the most important and relevant features that it used to get to a prediction. It takes data input from the model whilst observing any prediction changes in its response. Moreover, it generates a prediction explanation of a single case and not the entire dataset using unseen or test data. It uses individual feature prediction probabilities to explain how it arrived at a prediction [49]. As such, the results generated by XAI approaches should be visualised, thus making generalisations of findings easier, specifically in clinical settings. Therefore, this will make the DL approaches more reliable and will thus improve such application acceptance levels in healthcare.

This present study investigated the application of SEDL architectures to determine CoDs from VA narratives. There is limited research in the application of XAI on SEDL architectures, specifically in the VA domain. One notable study [42] used XAI methods; however, they did not employ stacked ensemble approaches. The contribution of the current study is to extend the current XAI research and its application in healthcare in order to build trustworthy and easy-to-understand DL models that are cost-effective, timely, error-free, consistent and accurate in order to improve VA reporting.

In this study, we apply LSTM, a CNN and SEDL to determine easily interpretable CoDs from VA narratives, thus making it easier to understand how the model arrived at its predicted outcome. The XAI DL approaches can therefore be effectively integrated into the process of identifying mortality causes, alongside human annotation and interpretation. This can prove very beneficial for clinicians and other interested researchers.

3. Materials and Methods

This section will describe the methods used in this study.

3.1. Study Design

This study is a retrospective cross-sectional study that uses secondary data analysis. We standardised and normalised our dataset within the Python Spyder environment and created data frames for cleaned VA narratives, model performance and classified VA categories, which are stored within a PostgreSQL version 4.2 Relational Database Management System.

3.2. Study Population

In this study, we used VA data from the Agincourt Health and Demographic Surveillance System (HDSS). The Agincourt HDSS study area came into existence in 1992 and is located in rural northeastern South Africa. Specifically, it is situated in the rural subdistrict of Bushbuckridge under the Ehlanzeni District, in the Mpumalanga Province. The Agincourt study area covers approximately 420 km^2. According to the Agincourt fact sheet of 2019, the population comprised 116,247 individuals residing in 28 villages with

22,716 households, with 55,961 males, 60,280 females, 11,724 children under 5 years and 928 school-going children aged from 5 to 19 [53].

3.3. Data Source and Description

This study used the Agincourt HDSS dataset for the period of 1993 to 2015. Our DL models specifically used input from VA narratives to predict CoD in twelve disease categories, which were described in the literature [3,9,54,55]. The dataset has 287 columns/features and 16,338 records/observations. We, therefore, used all the symptoms and the VA narrative column as our predictors X and the corresponding doctors diagnosis International Classification of Diseases 10th Revision (ICD-10) code as our target variable Y with 16,338 records.

Table 1 depicts the data labelling as in this study for the twelve disease categories.

Table 1. Labelling of the twelve disease classes.

Disease Category	Class Label	Number of Samples
HIV/TB	0	3388
Other Infectious	1	964
Metabolic	2	242
Cardiovascular	3	140
Indeterminate	4	1468
Maternal and Neonatal	5	121
Abdominal	6	117
Neoplasms	7	93
Neurological	9	57
Respiratory	10	46
Other NCD	11	21

Legend: HIV—human immunodeficiency virus, TB—tuberculosis, NCD—non-communicable diseases.

3.4. Preprocessing

Preprocessing addressed the data that were incomplete, noisy and inconsistent, through data cleaning, munging, transformation and reduction [56]. Data cleaning was also performed to simplify text by removing meaningless text that was deemed irrelevant for the models. This was achieved by converting text to lower case and removing all punctuation marks, spaces, numbers and special characters. Spelling correction was applied using the TextBlob Python library. Tokenisation was employed to create string tokens by splitting a document. Stopword removal was performed to remove irrelevant words using the NLTK library of English stopwords. Lemmatisation was applied to convert all possible word variations into the root form, known as the lemma. All nulls were dropped to remove bias in the modelling. Feature scaling was performed by using the Python StandardScaler to set the magnitude of our predictors X and response variables Y to be within the same range. To remove bias after exploratory data analysis, we applied data balancing using the Synthetic Minority Oversampling Technique (SMOTE) for our training dataset, as we had imbalanced classes.

3.5. Word Embedding and Representation

In this study, we used the global vector (GloVe) distributed vector word embedding and representation because we aimed to address the issue of complex and high data dimensionality, resulting in a distributed representation of words in low-dimensional space. Distributed vectors mark the beginning of a data processing layer in a DL model and imply that sentences with similar meanings tend to occur in similar contexts. The novel Glove was applied instead of term frequency (TF) and term frequency with inverse document frequency (TF-IDF), because the Glove is robust in discovering relationships as compared with the traditional TF which has common words dominating the vector space and TF-IDF fails to find similarity between words. The Glove is described in detail in [57,58].

3.6. Text Classification with Deep Learning Architectures

In this study, we employed the Long-Term Short Memory (LSTM), CNN and SEDL architectures as our text classification techniques. These approaches are able to deal with the challenge of large vocabularies that are made up of unknown words or out of vocabulary words (words that appear a few times in the test set but are not in the training corpus of words), unlike distributed vector representations. In the literature, they are known as character embeddings because each word is considered as no more than a composition of individual letters [59].

3.6.1. Long-Term Short Memory

LSTM is a variant of the RNN which uses a feed-back loop between layers and thus is dynamic [60]. RNNs are used to process information that is sequential and apply the same operation on each sequential instance. RNNs use input that relies on output of prior steps or previous computational output [61]. They make use of memory from previous computations to effectively process the current tasks. RNNs are made up of three layers: input, hidden layers and output [60]. However, RNNs have issues of learning long distance associations known as the vanishing gradient (where weighted activation function inputs of a neural network increase or decrease whilst the derivative function approaches zero). Consequently, this ultimately affects accuracy and makes it difficult for the architecture to learn and hypertune parameters for optimum results. RNNs are described elsewhere in [59].

One variant of a RNN that overcomes the vanishing gradient issue is LSTM. LSTM has three gates, namely, forget, input and output. It is distinctly efficient as compared with other RNN models in that it can remove information from memory, can save selected information to memory and can focus only on instant important aspects that are relevant [61]. The forget gate differentiates itself from other basic RNN architectures and makes it possible to backpropagate an infinite number of times. Furthermore, it calculates the hidden layer by averaging the three layers [59]. It calculates the hidden state by combining the three gates [59]. Figure 1 depicts the LSTM architecture and the conventional RNN.

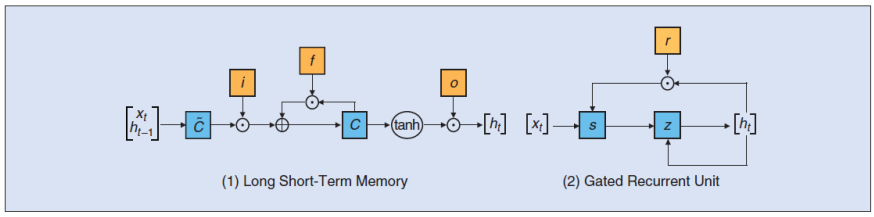


Figure 1. LSTM deep learning framework. Source: [59].

3.6.2. Convolutional Neural Network

The Convolutional Neural Network (CNN) was developed as a feature extraction function to extract high-level features from n-grams, and thus, these features are used for various NLP tasks [59,62,63]. The CNN uses the weight constraint, meaning that a single convolutional layer must have the same weights on the inputs for all the nodes in that layer. As such, this makes the CNN more efficient than training a basic neural network [61]. The CNN is made up of an input layer, convolutional layers, pooling layers, fully connected layers and the softmax function for classification. The functionality of the CNN is described in [59]. Figure 2 illustrates the pictorial representation of the CNN architecture.

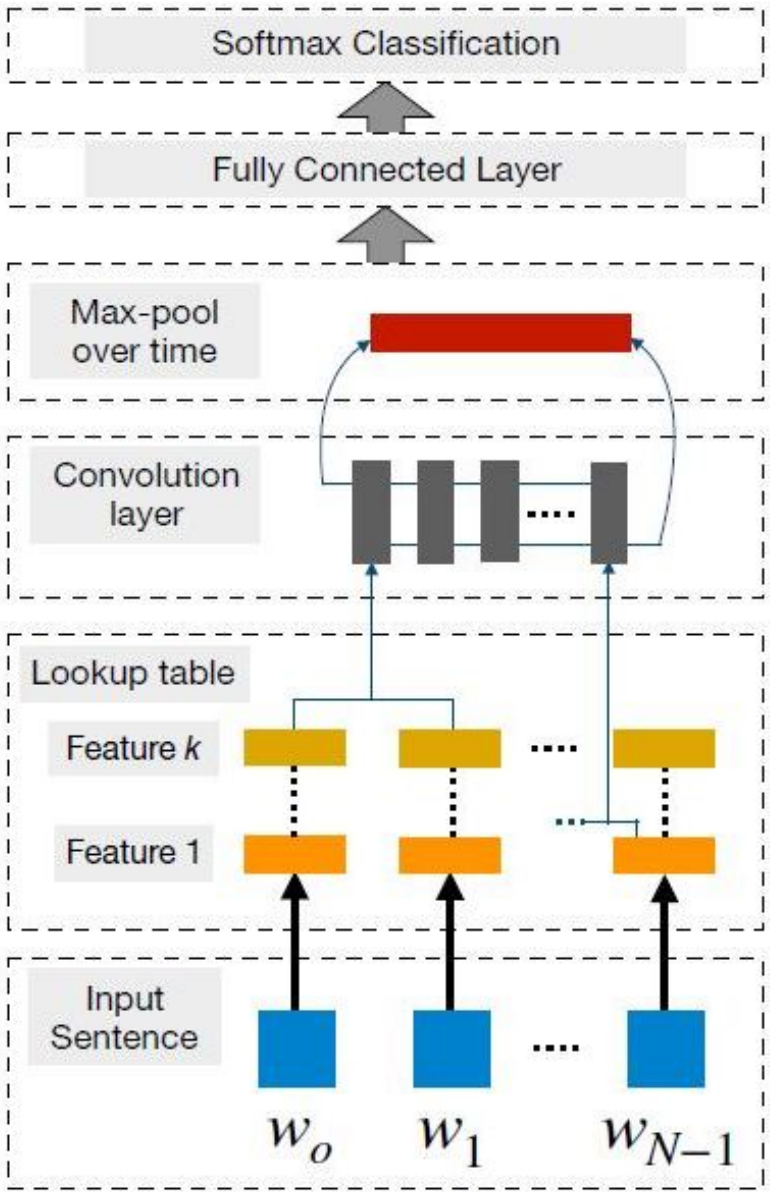


Figure 2. CNN deep learning framework. Source: [59].

3.7. Stacked Ensemble DL Models

Figure 3 depicts the ensemble DL methods using stacking. We employed various base learners ($B_1 - B_n$) that go through several iterations ($Level_1 - Level_n$) and feed into the meta learner (M_{output}) for the final model prediction. This study employed the stacking ideas formulated in the work of [64]. The dataset was first split into n equal parts, and each n^{th}-fold cross-validation set was taken for testing and the remainder for training the model. We created pairs of train–test sub-datasets to attain various model predictions that were used as input in our meta model for the final prediction. As such, this approach improves model performance and reduces room for bias in results [47].

3.8. Explainable AI Using LIME

Figure 4 is a visual representation of a general schematic diagram of the implementation of XAI using post hoc explanations by a simplification approach, LIME. We first fed the VA narratives as the input into the black box, SEDL models. This was followed by explaining the model output using the LIME component to provide the CoD diagnosis

and explanations to the experts. In this study, we followed the XAI mathematical approach (LIME) reported on in [50,52].

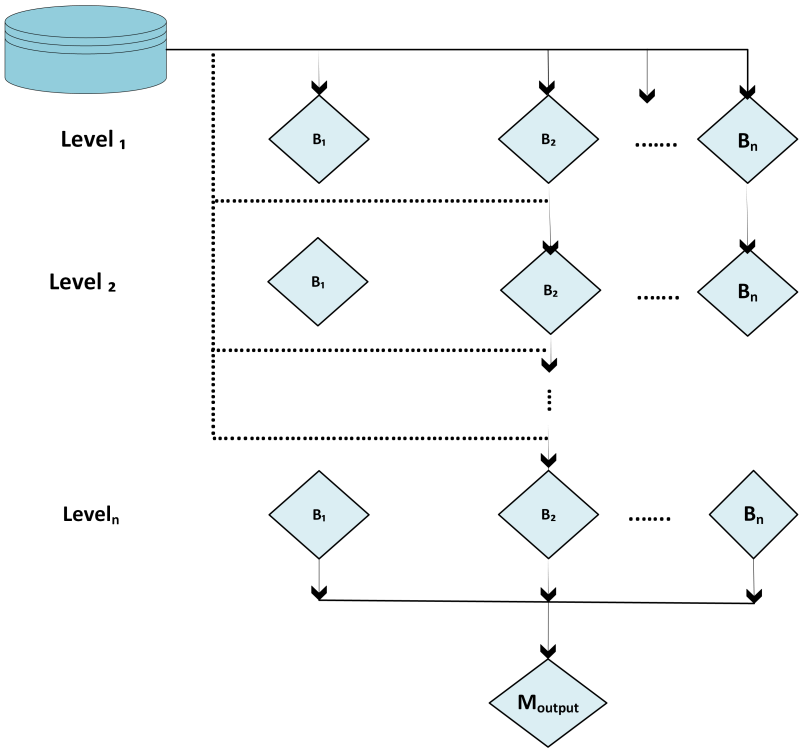


Figure 3. Stacking of our models. Source: [47].

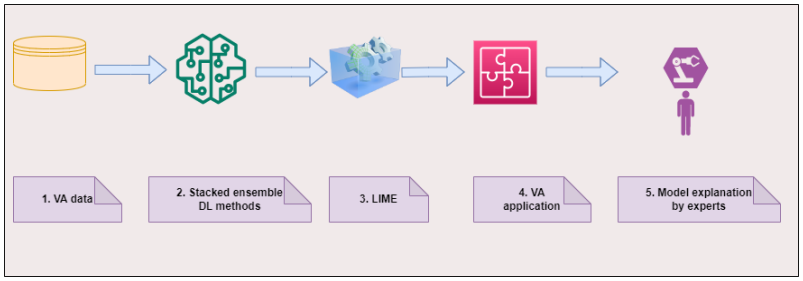


Figure 4. General pipeline of XAI implementation using VA data.

4. Experiments

Deep Learning Models

This section details the distributed word vector representations, DL architectures and Python libraries used in this study. The global vectors (Gloves) were used in order to attain a vectorised representation of words using the Keras Python library, and we re-trained it with our own corpus. We generated a 128-dimension vector for each word using a window size of 5. The input VA narratives were padded to be of the same size of 250 words for all the narratives. As we experimented, it was discovered that changing the dimensions to higher values would not have any significant effect. The CNN architecture was implemented which consisted of four 1D convolutional layers (128, 256, 256 and 512) with corresponding maximum pooling and dropout layers. The kernel size of the convolutional and max pooling layers was set at 3. We had a flattening layer for generating all input of the convolutional layers into a single long feature vector that was connected to the fully connected layer. The fully connected layer was set to 128 units and also used a dropout layer. All the layers used the rectified linear unit (ReLU) activation function except the classification layer, which used softmax. For optimisation, cross entropy for loss was

used and adam was used as the optimiser. The model was set to early stopping if there was no further improvement in training.

The RNN implementation using the LSTM variant included two Gated Recurrent layers with 256 and 512 nodes, respectively. The fully connected layer had 12 nodes at the end and also had a dropout layer with the rate set at 0.002. The classification layer used softmax as the activation function and we optimised using cross entropy for loss and set our optimiser as adam. We also set our model to early stopping if there was no effective further training.

We fine-tuned the batch size and number of epochs as part of batch normalisation where we sought to normalise the distribution of each input dimension of our dataset [65]. Furthermore, we tuned the optimization algorithm used to train the network, each with default parameters. Additionally, we optimised the learning rate which controls how much to update the weight at the end of each batch and the momentum controls how much to let the previous update influence the current weight update. Optimisation was also applied to the selection of network weight initialization by evaluating all of the available techniques. Moreover, we fine-tuned the activation function, which controls the non-linearity of individual neurons and when to fire. The dropout rate fused with the weight constraint were optimised by tuning the dropout rate for regularisation in an effort to limit overfitting and improve the model's ability to generalise. We further tuned the number of neurons in the hidden layer as a larger network requires more training and at least the batch size and number of epochs should ideally be optimised with the number of neurons.

Our conventional DL models failed to converge, as they failed to attain our optimal model performance set on 80% in our initial work reported in [21]. Therefore, we explored with ensemble DL methods, specifically employing SEDL architectures. A sequential model from the Keras ML library was used. It had one global average pooling layer and four dense layers with 1024, 512, 128, and 64 units, respectively. Additionally, to address model complexity and avoid overfitting, we added dropout layers and employed regularisation using Lasso regularisation. We used adam as our optimiser, set the patience to 3 and used 500 epochs and a batch size of 128. This formed the basis of our base learners, which were made up of 100 members.

5. DL Model Evaluation

In this study, we applied optimisation and performance learning curves to evaluate our DL models, specifically using training and validation datasets to visualise the cross entropy loss and model performance. This was performed in various iterations. We used the optimisation learning curves that depict a training loss and validation loss graph over time. The training loss depicts the extent to which a model is fitting the training data and, on the other hand, the validation loss indicates how well the model fits new unseen data. Performance learning curves which use accuracy as a metric to evaluate models were employed. Training accuracy indicates how well the model fits the training data, whilst validation accuracy denotes the extent to which a model fits new unseen data. Learning curves are described elsewhere in the studies of [61,66,67].

We also report on the accuracy of our models. Accuracy denotes all classes with classified results that have been predicted correctly in fraction terms [18,20]. True Positives (TPs) and True Negatives (TNs) represent the number of outcomes in which our prediction model correctly classifies positive and negative cases, respectively. In our study, TP denotes predicted positive VA narratives with a particular disease category from the twelve classes and are actually positive, while TN denotes predicted negative VA narratives with a particular disease category from the twelve classes and are actually negative. Conversely, False Positives (FPs) and False Negatives (FNs) denote the number of outcomes where our models incorrectly predicted the positive and negative classes. In our case, the FPs imply predicted positive VA narratives with a particular disease category from the twelve classes but are actually negative, and FN depicts the predicted negative VA narratives with a par-

ticular disease category from the twelve classes but are actually positive. The mathematical approach of accuracy is given by

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

6. Results

In this section, we discuss the results attained from our DL models in the classification of VA narratives into twelve disease categories.

We removed 2247 VA narratives that had null values and were left with 14,091 cases of textual information that were fed into our DL models. We also dropped 5170 target outcome values that were also missing. The longest narrative had 715 words, and thus 3949 characters. On the contrary, the shortest narrative had seven words, and thus 55 characters. Generally our narratives had a mean sentence length of 99.76 words.

Twelve disease classes with 3388 samples each for our training dataset were created after data balancing using SMOTE. This implied that the majority of classes had more data points compared with other minority classes that were less represented in terms of the data samples. Models that lack proper implementation of these data handling strategies usually suffer from model bias and can cause huge misclassification errors and misinterpretations.

Each VA narrative was taken as a single short document that is made up of a sequence of words. These sequences of words are unigrams (n = 1) and bigrams (n = 2) that were taken as input by our DL architectures for predicting the probable CoD into any of the twelve disease categories (target class). We sought to predict the values of our target features for the test set as unseen data. These text sequences were converted into numeric vector representations of word embeddings using global vectors (Gloves) where our models applied mathematical operations for prediction.

6.1. Model Performance

Table 2 below presents our performance evaluation metrics for the three DL architectures (LSTM, CNN and SEDL) attained by employing optimisation and performance ML curves. SEDL outperformed LSTM and the CNN in disease classification. Our ML curve depicts SEDL achieving a training accuracy of 82.1%, as compared with LSTM at 76.11% and the CNN at 76.35%. In terms of validation accuracy, SEDL achieved 82%, compared with LSTM and the CNN that attained 67.05% and 66.16%, respectively. A test accuracy of 82%, 67% and 66.2% for SEDL, LSTM and the CNN was achieved, respectively. In terms of test loss, SEDL attained 1.15%, LSTM 11.95% and the CNN 12.64%. SEDL, LSTM and the CNN were, respectively, approximately 82%, 67% and 66.2% accurate. This implies that all three models predicted the class correctly for 82%, 67% and 66.2% of the samples in the test dataset.

Table 2. Performance evaluation of models.

Model	Training Accuracy (%)	Validation Accuracy (%)	Test Accuracy (%)	Test Loss (%)
LSTM	76.11	67.05	67	11.95
CNN	76.35	66.16	66.2	12.64
SEDL	82.1	82	82	1.15

Figures 5 and 6 depict the DL model performances using learning curves. Our DL models were able to generalise well on new unseen data. This implies that our models are able to make accurate predictions on new data, illuminating the same characteristics of the training data. In order to be able to generalise with our models and avoid overfitting and underfitting, we reduced the model complexity by using dropout (20%). We also used early stopping (monitor = validation loss), patience (3) and the minimum delta (0.001%) to

force the model to stop when there was no further model improvement. The small variance between our validation and test accuracies denotes a good fit of our models.

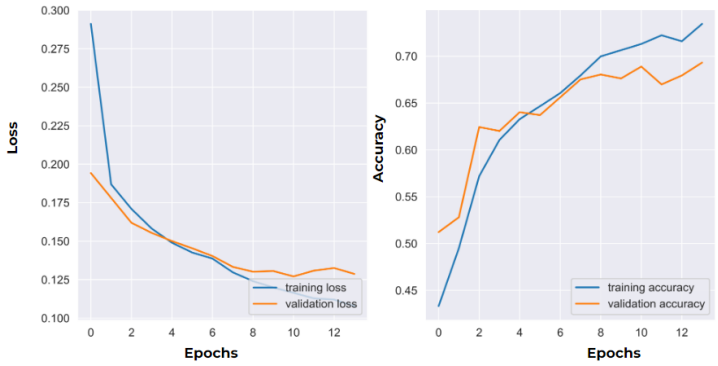


Figure 5. Learning curve depicting performance evaluation of CNN.

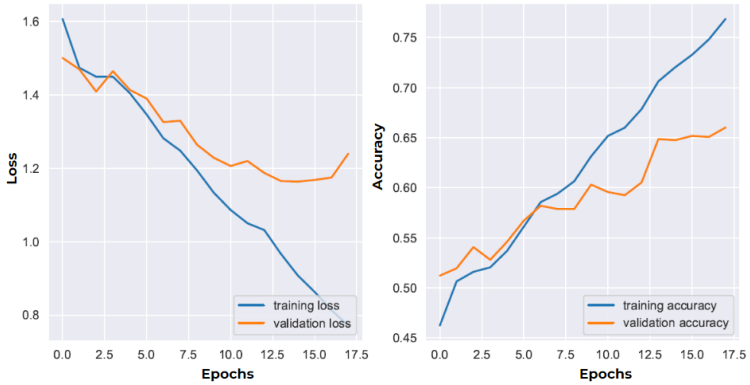


Figure 6. Learning curve depicting performance evaluation of LSTM.

Figures 7 and 8 show a learning curve performance evaluation of the stacked ensemble model. The variance between the train and test curve is minimal and thus shows that we did not have any underfitting or overfitting in our ensemble models.

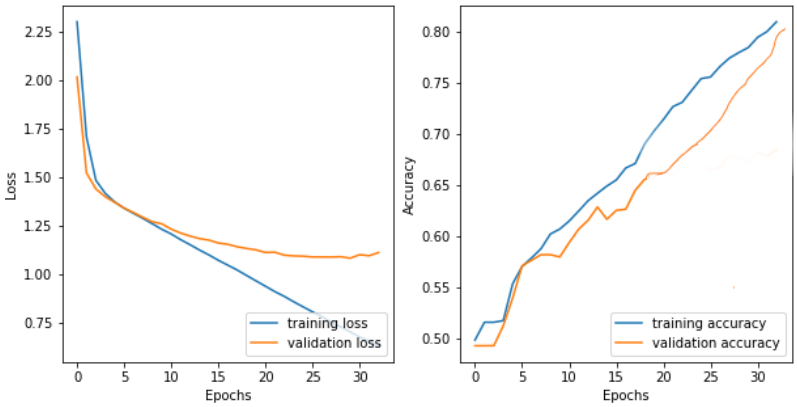


Figure 7. Learning curve depicting SEDL performance evaluation during model training.

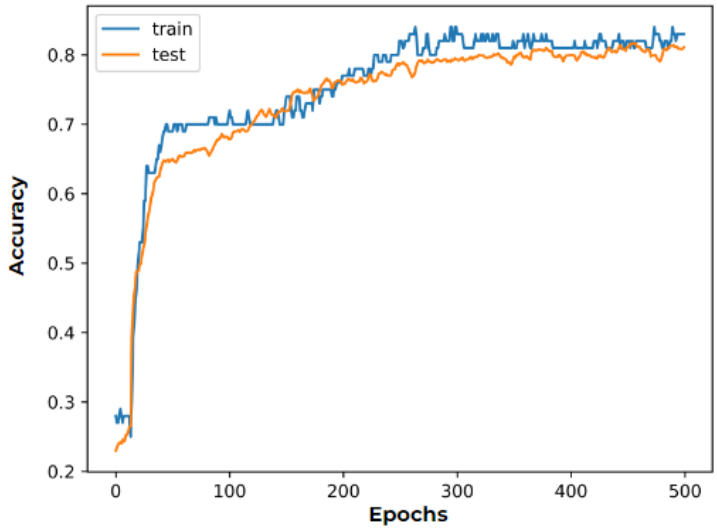


Figure 8. Learning curve depicting SEDL performance evaluation.

6.2. Model Explainability

This extract discusses the model interpretability of our SEDL architectures using LIME. Generally, we attained different CoD predictions and interpretability from our models. However, further insight was derived from the results attained by using LIME to ease explainability.

Figures 9 and 10 show the text narratives of cases with metabolic, maternal and neonatal diagnoses, respectively, that were correctly predicted by our SEDL model. The case of a patient who succumbed from metabolic diabetes disease shows that the respondent raised symptoms, such as (*"high sugar" and "sugar diabetic"*). These terms also constituted high probabilities in our XAI LIME model. Figure 9 depicts another correct prediction by our model. Interestingly, LIME managed to arrive at a prediction based on the patient's death description based on the keywords (*"baby", "incubator" , "mother", "delivery" and "birth"*), which all had high probabilities. It is noteworthy that in certain instances, our models got to some CoDs by inferring the symptoms, treatment and laboratory tests performed. Moreover, such cases had more correct predictions given by our models.

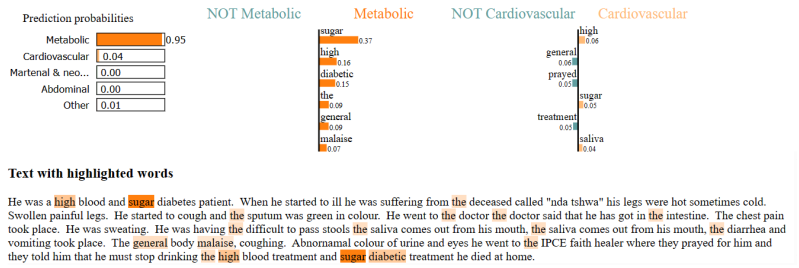


Figure 9. Example of LIME correct prediction.

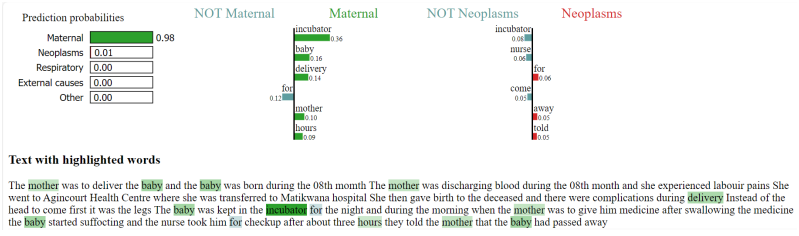


Figure 10. Example of LIME correct prediction.

Figures 11 and 12 show cases that were incorrectly predicted by our models as HIV and TB being the CoDs, whilst the gold standard was non-communicable diseases and respiratory causes, respectively. On the contrary, we found that in certain cases the models had challenges in NLP proficiencies, because some extracts of the narratives had irrelevant historical information of the patients that was insignificant to achieve a proper CoD. As such, the correct predictions were achieved by paying attention to the relevant descriptions that led to death, rather than historical patient information. This is evident in the textual extract given in Figure 12.

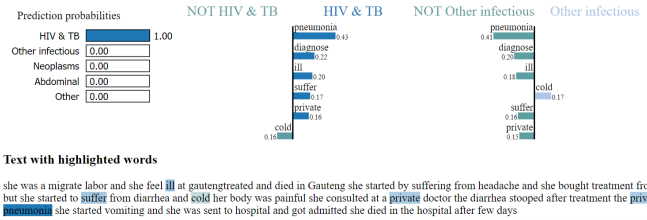


Figure 11. Example of LIME incorrect prediction.

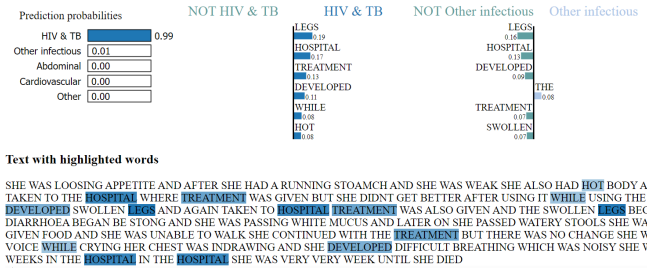


Figure 12. Example of LIME incorrect prediction.

We designed an experiment that sought to elicit mean opinion scores (MOSs) from doctors, based on the evaluation of the LIME results and model predictions. The goal of this was to assess if the LIME explanations made sense or not. We randomly selected 20 LIME predictions and gave them to 10 doctors for review and assessment. The MOSs were denoted by the following terms and corresponding weights:

1. Bad.
2. Poor.
3. Fair.
4. Good.
5. Excellent.

This calculated MOS value represents the average quality perception of the LIME predictions. We attained a high MOS of 4.21, which implies that the LIME predictions made sense.

7. Discussion

Determining CoD from VAs using narratives only largely remains a manual task that is tedious, time-consuming, prone to errors and costly. Specifically, the turnaround time to obtain a proper diagnosis on CoD from physicians after reviewing the VA narratives is still a challenge. Despite efforts to improve the VA processes, the VA elicitation process suffers from many drawbacks and still lags in the determination of CoD from VAs. This ultimately affects VA reporting as it does not happen in real time, even though it is key in informing civil registration systems and strengthening health priorities.

Generally, we employed novel DL architectures specifically using only VA narratives, with the aim of reducing the turnaround time needed by physicians to give an appropriate diagnosis after reviewing the narratives. Whilst the literature mainly uses the structured responses to the VA questionnaire for information extraction, we found the VA narratives

have rich and valuable information [3]. In comparison with the related work that usually uses the closed-ended responses to the structured questionnaires to build predictive models, we have shown that using the VA narratives only produces comparative results that are in close agreement with a physician-determined CoD. Additionally, to obtain improved information from the VA narratives, we need to consider using new novel technological ways, such as using tablets for recording the VA process with relatives of the deceased and transcribing it in an automated fashion. Consequently, doing away with summarising the interview may produce much more reliable results with an error percentage below 6% [42].

In this study, we used LSTM and a CNN to determine CoD using VA narratives only. Consistent with previous studies, we discovered that the DL approaches do not always produce the best performance as expected, as they require vast amounts of data and computing resources. Therefore, we faced challenges of few training data samples, as well as a high prevalence of out of vocabulary words with high lexical variability in the terms (words with broad meaning in the medical domain) [42]. Even though our DL models underperformed, LSTM produced better results than the CNN. Nevertheless, LSTM's underperformance signifies that our model was not effective for the sequential modelling of narratives and could not address the vanishing gradient issue. This implies that during the training of our model, using gradient methods, our gradient was getting smaller, thus preventing our models from training and learning effectively as the weights remained unchanged or not updated. Consequently, this issue might have negatively affected LSTM's capability of using its forget gate functionality (to forget previous computations by removing it from memory, the ability to save only useful information to memory and focus only on components of memory that are recently relevant), possibly suggesting such a performance.

It follows that the CNN also performed moderately. The performance of this model suggests that, during training, it did not effectively constraint input weights to be the same for all nodes in each convolutional layer. Furthermore, we noted that we did not have a good number of parameters at each layer in order to attain balanced, efficient training of the data points in this model. Moreover, this model did not fully utilise its ability to take the combined results from previous layers as input to other layers to attain a cutting-edge experience to actually learn more abstract and complex relationships.

As stated by [31], DL approaches require vast amounts of training data in order to perform optimally in classification; thus, in cases where there are limited data points, it does not achieve the best results as it needs to learn various feature weights and select the discriminative ones for model prediction. Conversely, our initial work reported in [21] proved that conventional approaches are better in CoD determination from VAs as compared with basic DL techniques. In a quest to prove the capabilities of DL approaches, we explored SEDL models.

Our SEDL architectures attained optimal results as compared with conventional DL approaches, as they surpassed the 80% model performance threshold set in our initial work reported in [21]. This can be attributed to the fact that the stacked ensemble classifier managed to effectively improve the performance during the iterative model training, by combining all the capabilities of all the base classifiers. Our empirical findings suggest that these SEDL architectures can be effectively used for CoD determination using narratives only.

DL models are a black box and difficult to implement as an alternative for the physician CoD gold standard. Additionally, the public health domain still lacks trust in implementing machines for disease diagnosis and prognosis, because this sensitive field requires more accurate models that are easily interpretable. However, the findings of this study position our SEDL framework as an encouraging alternative to physician diagnosis. This is consistent with the findings reported in the study of [42]. This is due to automated techniques that enforce a reduction in human error and inconsistencies, thus enforcing standardisation. Additionally, ML models are scalable and can be used to analyse and interpret large datasets, something not possible through human processes. Moreover, automated models are time-saving (reduce time needed to review the narratives), cost-effective (less costly

than the physician), free from human error and enforce consistency in CoD determination (the machine can improve CoD diagnosis and prognosis through model training).

The process of eliciting information from relatives of the deceased remains a huge challenge of the VA process, in terms of interpretability, because of the subjective nature of recall bias. Our findings can result in improved efficiency by automating the VA analysis process, thus reducing the time and effort required for human analysis. Therefore, this improves VA reporting and decision making.

Interestingly, it should be noted that clinical data are not always found in large volumes; thus, in cases where the datasets are limited or small in terms of training samples, it is better to use conventional ML approaches. From a comparative analysis of the results attained in this study and our previous study where we used conventional ML approaches [21], we found that traditional ML approaches are better in terms of performance when using VA narratives for CoD classification. Nevertheless, the novelty of this study is in the use of SEDL frameworks that are easily interpretable using LIME. Moreover, the evaluation conducted by the physicians presents this study as a promising accurate and cost-effective alternative. This further builds trust in the application of machines in the public health space.

This study was limited in the number of training samples that are needed by DL models to improve performance. Moreover, model performance was also affected by high lexical variability in the terms in the medical domain. The study was also limited by the length of the recall period, which could have created a bias in the collected VA data. In addition, some of the data collectors might have had inadequate training or inconsistencies in interviewing techniques, which could have introduced bias into the data collection process. The VA data that we used did not include native terms, which thus may have impacted the generalisability of the findings and limited a chance for model improvement.

8. Conclusions

This study showed that our SEDL framework may be included in the process for determining mortality causes, alongside human annotation and interpretation. As such, the findings of this study may be used as a baseline of building trust in the implementation of machines in disease prognosis and diagnosis in the public health field. Ultimately, this will reduce the turnaround time needed for CoD determination, cut costs, strengthen civil registration systems and health priorities as well as inform policy and practice. Future work will entail the application of record linkage to identify under-reported morbidity occurrences by effectively linking HDSS population data and NHLS public health data registries.

Author Contributions: Conceptualization, M.T.M., C.W.K., T.C. and E.M.; methodology, M.T.M. and V.O.; software, M.T.M. and V.O.; validation, M.T.M., T.C. and V.O.; formal analysis, M.T.M. and V.O.; investigation, M.T.M.; data curation, M.T.M. and C.W.K.; writing—original draft preparation, M.T.M.; writing—review and editing, M.T.M.; visualization, M.T.M. and V.O.; supervision, C.W.K., T.C. and E.M.; project administration, M.T.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of the University of the Witwatersrand Faculty of Health Sciences, Human Research Ethics Committee (Medical), (protocol code M1911132 and approved 23 September 2020).

Informed Consent Statement: Patient consent was waived. This study is a retrospective study that used secondary data analysis. We obtained permission from the Agincourt Health and Demographic Surveillance Site to use their data for research purposes and maintained confidentiality through anonymity.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy and ethical restrictions.

Acknowledgments: This work was supported by the Developing Excellence in Leadership, Training and Science (DELTAS) Africa Initiative Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) (Grant No. DEL-15-005). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS) Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (Grant No. 107754/Z/15/Z) and the United Kingdom government. Furthermore, we received support from the University of the Witwatersrand Faculty of Health Sciences Seed funding.

Conflicts of Interest: The authors declare that they have no competing interests.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CCVA	Computer-Coded Verbal Autopsy
CoD	Cause of Death
CNN	Convolutional Neural Network
DL	Deep Learning
XAI	Explainable Artificial Intelligence
LIME	Local Interpretable Model-agnostic Explanations
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Natural Language Processing
PCVA	Physician-Coded Verbal Autopsy
RNN	Recurrent Neural Network
SEDL	Stacked Ensemble Deep Learning
VA	Verbal Autopsy

References

1. Nichols, E.K.; Byass, P.; Chandramohan, D.; Clark, S.J.; Flaxman, A.D.; Jakob, R.; Leitao, J.; Maire, N.; Rao, C.; Riley, I.; et al. The WHO 2016 verbal autopsy instrument: An international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0. *PLoS Med.* **2018**, *15*, e1002486. [CrossRef]
2. Thomas, L.M.; D'Ambruoso, L.; Balabanova, D. Verbal autopsy in health policy and systems: A literature review. *BMJ Glob. Health* **2018**, *3*, e000639. [CrossRef]
3. Jeblee, S.; Gomes, M.; Jha, P.; Rudzicz, F.; Hirst, G. Automatically determining cause of death from verbal autopsy narratives. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 127. [CrossRef]
4. Soleman, N.; Chandramohan, D.; Shibuya, K. Verbal autopsy: Current practices and challenges. *Bull. World Health Organ.* **2006**, *84*, 239–245. [CrossRef] [PubMed]
5. Bailo, P.; Gibelli, F.; Ricci, G.; Sirignano, A. Verbal autopsy as a tool for defining causes of death in specific healthcare contexts: Study of applicability through a traditional literature review. *Int. J. Environ. Res. Public Health* **2022**, *19*, 11749. [CrossRef] [PubMed]
6. Clark, S.J. A Guide to Comparing the Performance of VA Algorithms. *arXiv* **2018**, arXiv:1802.07807.
7. Desai, N.; Aleksandrowicz, L.; Miasnikof, P.; Lu, Y.; Leitao, J.; Byass, P.; Tollman, S.; Mee, P.; Alam, D.; Rathi, S.K.; et al. Performance of four computer-coded verbal autopsy methods for cause of death assignment compared with physician coding on 24,000 deaths in low-and middle-income countries. *BMC Med.* **2014**, *12*, 20. [CrossRef]
8. James, S.L.; Flaxman, A.D.; Murray, C.J. Performance of the Tariff Method: Validation of a simple additive algorithm for analysis of verbal autopsies. *Popul. Health Metrics* **2011**, *9*, 31. [CrossRef]
9. Byass, P.; Herbst, K.; Fottrell, E.; Ali, M.M.; Odhiambo, F.; Amek, N.; Hamel, M.J.; Laserson, K.F.; Kahn, K.; Kabudula, C.; et al. Comparing verbal autopsy cause of death findings as determined by physician coding and probabilistic modelling: A public health analysis of 54 000 deaths in Africa and Asia. *J. Glob. Health* **2015**, *5*.
10. McCormick, T.H.; Li, Z.R.; Calvert, C.; Crampin, A.C.; Kahn, K.; Clark, S.J. Probabilistic cause-of-death assignment using verbal autopsies. *J. Am. Stat. Assoc.* **2016**, *111*, 1036–1049. [CrossRef]
11. Miasnikof, P.; Giannakeas, V.; Gomes, M.; Aleksandrowicz, L.; Shestopaloff, A.Y.; Alam, D.; Tollman, S.; Samarikhalaj, A.; Jha, P. Naive Bayes classifiers for verbal autopsies: Comparison to physician-based classification for 21,000 child and adult deaths. *BMC Med.* **2015**, *13*, 286. [CrossRef] [PubMed]
12. Boulle, A.; Chandramohan, D.; Weller, P. A case study of using artificial neural networks for classifying cause of death from verbal autopsy. *Int. J. Epidemiol.* **2001**, *30*, 515–520. [CrossRef] [PubMed]

13. Flaxman, A.D.; Vahdatpour, A.; Green, S.; James, S.L.; Murray, C.J. Random forests for verbal autopsy analysis: Multisite validation study using clinical diagnostic gold standards. *Popul. Health Metrics* **2011**, *9*, 29. [CrossRef]
14. Quigley, M.A.; Chandramohan, D.; Setel, P.; Binka, F.; Rodrigues, L.C. Validity of data-derived algorithms for ascertaining causes of adult death in two African sites using verbal autopsy. *Trop. Med. Int. Health* **2000**, *5*, 33–39. [CrossRef] [PubMed]
15. Mwanyangala, M.A.; Urassa, H.M.; Rutashobya, J.C.; Mahutanga, C.C.; Lutambi, A.M.; Maliti, D.V.; Masanja, H.M.; Abdulla, S.K.; Lema, R.N. Verbal autopsy completion rate and factors associated with undetermined cause of death in a rural resource-poor setting of Tanzania. *Popul. Health Metrics* **2011**, *9*, 41. [CrossRef] [PubMed]
16. Koopman, B.; Karimi, S.; Nguyen, A.; McGuire, R.; Muscatello, D.; Kemp, M.; Truran, D.; Zhang, M.; Thackway, S. Automatic classification of diseases from free-text death certificates for real-time surveillance. *BMC Med Informatics Decis. Mak.* **2015**, *15*, 53. [CrossRef]
17. Koopman, B.; Zuccon, G.; Nguyen, A.; Bergheim, A.; Grayson, N. Extracting cancer mortality statistics from death certificates: A hybrid machine learning and rule-based approach for common and rare cancers. *Artif. Intell. Med.* **2018**, *89*, 1–9. [CrossRef]
18. Mujtaba, G.; Shuib, L.; Raj, R.G.; Rajandram, R.; Shaikh, K.; Al-Garadi, M.A. Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. *PLoS ONE* **2017**, *12*, e0170242. [CrossRef]
19. Mujtaba, G.; Shuib, L.; Raj, R.G.; Rajandram, R.; Shaikh, K.; Al-Garadi, M.A. Classification of forensic autopsy reports through conceptual graph-based document representation model. *J. Biomed. Inform.* **2018**, *82*, 88–105. [CrossRef]
20. Mujtaba, G.; Shuib, L.; Raj, R.G.; Rajandram, R.; Shaikh, K. Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study. *J. Forensic Leg. Med.* **2018**, *57*, 41–50. [CrossRef]
21. Mapundu, M.T.; Kabudula, C.W.; Musenge, E.; Olago, V.; Celik, T. Performance evaluation of machine learning and Computer Coded Verbal Autopsy (CCVA) algorithms for cause of death determination: A comparative analysis of data from rural South Africa. *Front. Public Health* **2022**, *10*, 990838. [CrossRef]
22. Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; Dudley, J.T. Deep learning for healthcare: Review, opportunities and challenges. *Briefings Bioinform.* **2017**, *19*, 1236–1246. [CrossRef] [PubMed]
23. Prakash, J.A.; Ravi, V.; Sowmya, V.; Soman, K. Stacked ensemble learning based on deep convolutional neural networks for pediatric pneumonia diagnosis using chest X-ray images. *Neural Comput. Appl.* **2022**, *35*, 8259–8279. [CrossRef]
24. Ravì, D.; Wong, C.; Deligianni, F.; Berthelot, M.; Andreu-Perez, J.; Lo, B.; Yang, G.Z. Deep learning for health informatics. *IEEE J. Biomed. Health Inform.* **2016**, *21*, 4–21. [CrossRef] [PubMed]
25. Kwak, G.H.J.; Hui, P. Deephealth: Deep learning for health informatics. *ACM Trans. Comput. Healthc.* **2019**.
26. Srivastava, S.; Soman, S.; Rai, A.; Srivastava, P.K. Deep learning for health informatics: Recent trends and future directions. In Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, 13–16 September 2017; pp. 1665–1670.
27. Zhao, R.; Yan, R.; Chen, Z.; Mao, K.; Wang, P.; Gao, R.X. Deep learning and its applications to machine health monitoring. *Mech. Syst. Signal Process.* **2019**, *115*, 213–237. [CrossRef]
28. Zhang, T.; Oles, F.J. Text categorization based on regularized linear classification methods. *Inf. Retr.* **2001**, *4*, 5–31. [CrossRef]
29. Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. *arXiv* **2015**, arXiv:1509.01626.
30. Kowsari, K.; Brown, D.E.; Heidarysafa, M.; Meimandi, K.J.; Gerber, M.S.; Barnes, L.E. Hdltex: Hierarchical deep learning for text classification. In Proceedings of the 2017 16th IEEE international conference on machine learning and applications (ICMLA), Cancun, Mexico, 18–21 December 2017; pp. 364–371.
31. Mujtaba, G.; Shuib, L.; Idris, N.; Hoo, W.L.; Raj, R.G.; Khowaja, K.; Shaikh, K.; Nweke, H.F. Clinical text classification research trends: Systematic literature review and open issues. *Expert Syst. Appl.* **2019**, *116*, 494–520. [CrossRef]
32. Johnson, R.; Zhang, T. Effective use of word order for text categorization with convolutional neural networks. *arXiv* **2014**, arXiv:1412.1058.
33. Irsoy, O.; Cardie, C. Opinion mining with deep recurrent neural networks. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 720–728.
34. Tang, D.; Qin, B.; Liu, T. Document modeling with gated recurrent neural network for sentiment classification. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1422–1432.
35. Liu, F.; Zheng, J.; Zheng, L.; Chen, C. Combining attention-based bidirectional gated recurrent neural network and two-dimensional convolutional neural network for document-level sentiment classification. *Neurocomputing* **2020**, *371*, 39–50. [CrossRef]
36. Ghosh, M.; Sanyal, G. Document modeling with hierarchical deep learning approach for sentiment classification. In Proceedings of the 2nd International Conference on Digital Signal Processing, Tokyo, Japan, 25–27 February 2018; pp. 181–185.
37. Xu, J.; Chen, D.; Qiu, X.; Huang, X. Cached long short-term memory neural networks for document-level sentiment classification. *arXiv* **2016**, arXiv:1610.04989.
38. Jelodar, H.; Wang, Y.; Orji, R.; Huang, S. Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 2733–2742. [CrossRef]
39. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489.

40. Jeblee, S.; Gomes, M.; Hirst, G. Multi-task learning for interpretable cause of death classification using key phrase prediction. In Proceedings of the BioNLP 2018 Workshop, Melbourne, Australia, 19 July 2018; pp. 12–17.
41. Yan, Z.; Jeblee, S.; Hirst, G. Can Character Embeddings Improve Cause-of-Death Classification for Verbal Autopsy Narratives? In Proceedings of the 18th BioNLP Workshop and Shared Task, Florence, Italy, 1 August 2019; pp. 234–239.
42. Blanco, A.; Perez, A.; Casillas, A.; Cobos, D. Extracting Cause of Death from Verbal Autopsy with Deep Learning interpretable methods. *IEEE J. Biomed. Health Inform.* **2020**, *25*, 1315–1325. [CrossRef]
43. Kivrak, M.; Guldogan, E.; Colak, C. Prediction of death status on the course of treatment in SARS-COV-2 patients with deep learning and machine learning methods. *Comput. Methods Programs Biomed.* **2021**, *201*, 105951. [CrossRef]
44. Heidarysafa, M.; Kowsari, K.; Barnes, L.; Brown, D. Analysis of Railway Accidents' Narratives Using Deep Learning. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018; pp. 1446–1453.
45. El Asnaoui, K. Design ensemble deep learning model for pneumonia disease classification. *Int. J. Multimed. Inf. Retr.* **2021**, *10*, 55–68. [CrossRef]
46. Tang, S.; Wang, C.; Nie, J.; Kumar, N.; Zhang, Y.; Xiong, Z.; Barnawi, A. EDL-COVID: Ensemble deep learning for COVID-19 case detection from chest X-ray images. *IEEE Trans. Ind. Inform.* **2021**, *17*, 6539–6549. [CrossRef]
47. Ganaie, M.A.; Hu, M.; Malik, A.; Tanveer, M.; Suganthan, P. Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105151. [CrossRef]
48. Baccouche, A.; Garcia-Zapirain, B.; Castillo Olea, C.; Elmaghraby, A. Ensemble deep learning models for heart disease classification: A case study from Mexico. *Information* **2020**, *11*, 207. [CrossRef]
49. Loh, H.W.; Ooi, C.P.; Seoni, S.; Barua, P.D.; Molinari, F.; Acharya, U.R. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Comput. Methods Programs Biomed.* **2022**, *226*, 107161. [CrossRef] [PubMed]
50. Zhang, Y.; Weng, Y.; Lund, J. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics* **2022**, *12*, 237. [CrossRef] [PubMed]
51. Javed, A.R.; Khan, H.U.; Alomari, M.K.B.; Sarwar, M.U.; Asim, M.; Almadhor, A.S.; Khan, M.Z. Toward explainable AI-empowered cognitive health assessment. *Front. Public Health* **2023**, *11*, 1024195. [CrossRef]
52. Knapič, S.; Malhi, A.; Saluja, R.; Främling, K. Explainable artificial intelligence for human decision support system in the medical domain. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 740–770. [CrossRef]
53. Kabudula, C.W.; Tollman, S.; Mee, P.; Ngobeni, S.; Silaule, B.; Gómez-Olivé, F.X.; Collinson, M.; Kahn, K.; Byass, P. Two decades of mortality change in rural northeast South Africa. *Glob. Health Action* **2014**, *7*, 25596. [CrossRef] [PubMed]
54. Danso, S.; Atwell, E.; Johnson, O. A comparative study of machine learning methods for verbal autopsy text classification. *arXiv* **2014**, arXiv:1402.4380.
55. King, G.; Lu, Y. Verbal autopsy methods with multiple causes of death. *Stat. Sci.* **2008**, *23*, 78–91. [CrossRef]
56. Shah, C. *A Hands-On Introduction to Data Science*; Cambridge University Press: Cambridge, UK, 2020.
57. Naili, M.; Chaibi, A.H.; Ghezala, H.H.B. Comparative study of word embedding methods in topic segmentation. *Procedia Comput. Sci.* **2017**, *112*, 340–349. [CrossRef]
58. Alami, N.; Meknassi, M.; En-nahnahi, N. Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning. *Expert Syst. Appl.* **2019**, *123*, 195–211. [CrossRef]
59. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *arXiv* **2017**, arXiv:1708.02709.
60. Zaki, M.J.; Meir, W., Jr. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*; Cambridge University Press: Cambridge, UK, 2019.
61. Leskovec, J.; Rajaraman, A.; Ullman, J.D. *Mining of Massive Data Sets*; Cambridge University Press: Cambridge, UK, 2020.
62. Kirillov, A.; Schlesinger, D.; Forkel, W.; Zelenin, A.; Zheng, S.; Torr, P.; Rother, C. Efficient likelihood learning of a generic CNN-CRF model for semantic segmentation. *arXiv* **2015**, arXiv:1511.05067.
63. Malinowski, M.; Rohrbach, M.; Fritz, M. Ask your neurons: A neural-based approach to answering questions about images. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1–9.
64. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
65. Watt, J.; Borhani, R.; Katsaggelos, A. *Machine Learning Refined: Foundations, Algorithms, and Applications*; Cambridge University Press: Cambridge, UK, 2020.
66. Anzanello, M.J.; Fogliatto, F.S. Learning curve models and applications: Literature review and research directions. *Int. J. Ind. Ergon.* **2011**, *41*, 573–583. [CrossRef]
67. Hoiem, D.; Gupta, T.; Li, Z.; Shlapentokh-Rothman, M.M. Learning Curves for Analysis of Deep Networks. *arXiv* **2020**, arXiv:2010.11029.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

E Appendices

Overview of Statistical and Machine Learning techniques for Determining Causes of Death from Verbal Autopsies: A Systematic Literature Review

Michael Mapundu[a], *Chodziwadziwa Kabudula*[a,b], *Eustasius Musenge*[a], *Turgay Celik*[c,d]

[a] *School of Public Health Faculty of Health Sciences , University of the Witwatersrand South Africa*
[b] *MRC / Wits Rural Public Health and Health Transitions Research Unit (Agincourt)*
[c] *Wits Institute of Data Science , University of the Witwatersrand South Africa*
[d] *School of Computer Science and Applied Mathematics , University of the Witwatersrand South Africa*

Abstract

Background: The process of determining causes of death in areas where there is limited clinical services using verbal autopsies has become a key issue in terms of accuracy on cause of death (prone to errors and subjective), quality of data among many drawbacks. This is mainly because there is no proper standard available in performing verbal autopsy, even though it is important for civil registration systems and strengthening of health priorities. Physician diagnosis is the only gold standard in reviewing verbal autopsy narratives. In practice, conventional statistical methods are used to perform verbal autopsies due to their simplicity and transparency. However, in literature complex machine learning models can be found that can replace the traditional statistical methods. There has not been much application of machine learning techniques in verbal autopsy to determine cause of death, despite the advances in technology. As such, there is a need for a thorough survey of recent literature on statistical and machine learning approaches applied in verbal autopsy to determine cause of death.
Methods: A systematic review was conducted and included a search from six databases. Our study only included scientific articles published in last decade that reported on verbal autopsy and: (1) algorithms; (2) statistical techniques; (3) machine learning and (4) deep learning. The search yielded 110 articles, after methodological systematic analysis, we identified 85 articles as being relevant and discarded the other 25. We investigated and compared the most commonly used statistical and machine learning techniques in VAs, identified limitations of each of these techniques, proposed a guiding machine learning framework and pointed to future directions.
Results: Eighty five studies met the inclusion criteria. Apart from physician diagnosis, statistical methods are the most currently applied tools to determine cause of death from verbal autopsies. However, there has been little application of traditional machine learning and emerging techniques, even though they have shown promising results in other domains.
Conclusions: Technological application of machine learning to determine cause of death, should focus on effective ideal strategies of pre-processing, transparency, robust feature engineering techniques and data balancing in order to attain optimal model performance.

Keywords:

Verbal Autopsy, Machine learning, Algorithms, Natural Language Processing, Deep learning, Artificial intelligence.

Corresponding author: michael.mapundu@wits.ac.za

I. Background

Half of the countries of the world do not meet the United Nations 90 percent death registration coverage requirement because deaths in many developing countries are not captured in civil registration systems [1; 2]. In addition, 65 percent of the population in the world lack high quality information on cause of death since every year about sixty million deaths worldwide are not assigned a medically certified cause [3]. Since cause of death information is required to inform critical health policies and priorities, in the absence of the more traditional, clinically oriented sources, cause of death information should be derived from alternative sources. One widely adopted alternative source of cause of death information is Verbal autopsy (VA). As stated by [2], the VA concept emanated in Africa and Asia in the 1950's and involves having dialogue with the family member of the deceased who provides information about any possible preceding conditions that may have led to the cause of death in their opinion. Literature to date suggests that the VA approach was first adopted in India and subsequently research studies using VA surfaced in the late 1970s and early 1980s in Niakhar in Senegal and Matlab in Bangladesh respectively [2]. The use of VA as a source of cause of death information is now common in many developing countries where access to clinical services, death registration and consequently medical certification of causes of death is limited such as Bangladesh, India, Malawi, Niger, South Africa and Nigeria [4].

The process of determining causes of death from VAs is traditionally done through the Physician Certified Verbal Autopsy (PCVA) approach. This approach involves two physicians reviewing the responses to questions from VA interviews to reach an agreement on cause of death, where

they do not agree a third physician decides on cause of death [5]. The PCVA approach is regarded as the gold standard for determining causes of death from VA and is used for training, testing and evaluation of automated approaches for determining causes of death from VA. However, the PCVA has a number of shortcomings as discussed in section IV. In order to improve the determination of causes of death from VA and address some of the weaknesses of the PCVA approach, automated approaches also known as Computer Coded Verbal Autopsies (CCVA) are used as alternative approaches. CCVA methods are data driven and employ automated processing and reasoning. They take as input, the responses to the questions of the VA interviews and maps them to corresponding target ICD-10 cause of death codes or other recognized standard categorization [4; 6; 7]. As argued by [8], one problem with data driven methods of determining causes of death from VA is reliance on clinically confirmed causes. Handling of missing data also poses a further challenge as the norm is to treat them as unknown, thus creating a bias in the interpretations.

Driven by recent advancement in technology, there has been a surge in application of computational tools that accurately analyse data for decision making. However, the quality of data is compromised by incompleteness and inaccuracies, amongst other drawbacks [9]. Awareness of these problems is required in research communities in developing and underdeveloped countries [10; 11]. Computational tools use machine learning (ML) techniques that learn from experience relative to some tasks within a class and make use of a performance measure or weighting to improve the performance. Machine learning techniques employ some statistical, probabilistic and optimisation technique in order to discover patterns and trends in complex data through some analysis to get to a decision [12]. There are various statistical and ML techniques that have been adopted in health care to provide required services and address specific problems, such as various classification and regression models that exist to date [13–16]. [3] argues that to date, ML approaches have been predominantly employed to data generated from only the structured questions of VA interviews, attaining true positive prediction scores of 0.60 for individual cause of death categorization. However, they included several numbers of cause of death classes. Available evidence indicates that ML can generate real time cause of death information that is similar to that of physicians/experts [17]. Moreover, there has not been full utilisation of deep learning and its capabilities in the health domain, even though they demonstrated great performance and potential in other artificial intelligence (AI) areas and NLP problems. Further insights and critique of existing literature on CCVA is discussed in later sections of this methodological systematic analysis.

In existing research, there is evidence of application of complex machine learning (ML) algorithms that can substitute conventional PCVA and some existing CCVA algorithms as approaches for determining cause of death from VAs [18–22]. Even though these ML approaches yield high accuracies in model performance in terms of classification, they lack clear explanations on model interpretability (lack of justification by the researcher on model performance) [17]. There is need to have ML models that are fair, accountable and transparent. Using only hospital based deaths [5] conducted a study where they compared PCVA and automated methods and concluded that out of the two methods, no particular method is considered best in terms of performance. There is need for further inquiries with computer coded techniques in order to improve performance, thus there is need for publicly available datasets such that comparisons, inferences and deductions can be done.

Subsequently, to the best of our knowledge as of the time of this write-up, there is little evidence that suggest existence of studies that have done a methodological systematic analysis of literature on the use of statistical and ML approaches (including deep learning) in VA. Therefore, there is a need to have a systematic and comprehensive literature review in order to have an objective understanding of these statistical and ML applications in VA. The methodological systematic analysis process highlights various studies in which statistical and ML were applied in VA, thus giving researchers and any other interested parties an idea of the problems and issues encountered. By so doing this improves the application of statistical and ML in VA and strengthen health priorities. Although ML models often achieve high accuracy scores, they generally struggle to explain their predictions. This is due to factors such as human errors, subjectivity, and the challenges posed by fragmented and imbalanced datasets, which are used as inputs for VA narratives and mapped to corresponding cause of death target values according to ICD-10 codes. In this systematic review, our goal is to do a methodological systematic analysis of existing literature on statistical and machine learning models which were applied in VA between 2010 and 2020 and propose ML framework guidelines for VA. This paper is organized as follows. Section I is the background and has subsections of; feature engineering and feature selection techniques, statistical and machine learning models, evaluation metrics, ML optimisation techniques, imbalanced and fragmented datasets, limitations and assumptions of models, emerging trends and current technological dynamics in VA and limitations in literature. Section II briefly highlights our methodology. In Section III we present the results and proposed guiding framework for machine learning in VA. Section IV provides a critique of the discussion. Section V is the conclusion. Section VI is a list of abbreviations. All declarations are in section VII. Figure legends are presented in section VIII.

A. Feature selection and feature engineering

This extract highlights familiar and mostly applied feature selection (FS) and feature engineering (FE) approaches in VA. The aim of FS and FE is to come up with a vector space that is efficient, scalable and accurate. FS is a process of generating new relevant input features with highest score or

weighting from existing ones and FE is generating a subset of features from current features [23; 24]. FE has three distinct sub-processes and commences with feature extraction, then feature value representation and lastly feature selection [25].

*1) **Feature extraction**:* Feature extraction is a process that seeks to pull out relevant features from a set of reports/narratives or any source of information. Thus feature extraction allocates a document from a pre-determined set of categories [19]. Various fully automated feature extraction approaches exist in literature [25]. The following extracts discuss various automated feature extraction approaches.

Bag of words (BoW) extracts words that are unique from all narratives existent in the file regardless of their categories and stores in a list (BoW). The BoW technique splits text into words and only includes useful words in the feature vector as dimensions (words) extracted. Each word represents an autonomous and particular feature [25; 26]. However, the BoW ignores the order of words and outputs a generated reduced form of occurrences of each word in text narratives using word frequency [22].

n-grams are a set of words that are sequential as they make use of continuous number of items such as characters or words from a given sequence of narratives. n-gram models can be of the form; a) n = 1 (unigram), b) when n = 2 (bigram), c) n = 3 (trigram) and d) hybrid-grams (mixture of unigram, bigram, and trigram) [25].

Word2vec, exists in two forms namely; skip-gram and continuous bag of words (CBoW) models. The skip-gram model does it's learning repeatedly using existing words presented in a sentence to predict the next word. On the contrary, the CBoW model predicts the prevailing word using nearby words. Both approaches determine a limit on number of words by applying the parameter window size technique [25].

Concept-based feature extraction is mainly used to extract medical concepts using medically based ontologies where the relationship between terms is difficult to establish. This is common when there are various terms that mean the same concept. Concept-based feature extraction can be done through a bag of concepts or bag of phrases [25].

Graph-based or graph of word (GoW) features are a graphical representation of features using content-based features. The aim of GoW is to represent the order of words, where each given narrative is depicted in a graph structure. The vertex of each graph contains a particular selective feature of the given narrative. Edges are used to connect two co-existing features, where each edge has the weight calculated using the frequency of nearby vertices in the graph [25].

Doc2vec is an approach that applies an unsupervised learning algorithm to come up with vector representations of paragraphs or documents. The algorithm is a modification and application of the concepts of word2vec, and aims at finding similarities in paragraphs or documents [22].

Rule-based feature extraction is used in case-based reasoning which employs transfer learning (mainly applied when input has few labelled instances that can be used to train an accurate model through transferring capabilities from existing systems to untrained ones). The rule-based approach extracts features that are taken as input by the system's reasoning inference engine. Manually crafted rules are fed into the information extraction module, and searching can be done to attain features and mapped to corresponding values. Conversion can now be done to transform them into a machine readable format for processing. The whole process is automated and makes use of predefined rules [27].

Of all the above discussed feature extraction techniques, [25] suggest the BoW as the commonly used approach for feature extraction as it maintains word frequency. However, this approach does not take into account the order of words and syntax. As an alternative to the BoW and to address some of the above drawbacks, the n-gram technique for feature extraction was proposed. Nevertheless, the n-gram approach fails to acquire the meaning of opposite words and does not consider similar words (synonyms) when applied to clinical narratives. This has a negative effect of increasing the feature space with redundant features, thus the number of features increases enormously with increasing n. In trying to address these issues, the literature discusses other feature extraction approaches such as bag of phrases and bag of contents. These approaches have shown improved accuracy and attained good results in identifying words with similar meaning as matching and also capturing vagueness in words. Although these concept-based techniques are useful, nevertheless they still ignore grammar and word order. This has seen the rise of GoW feature extraction techniques which are better in terms of performance to traditional BoW and n-gram models. However, the GoW approaches are computationally expensive [25].

*2) **Feature value Representation Techniques**:* It is imperative for one to transform extracted features into numeric vectors after the feature extraction phase, a process known as feature value representation [25]. The aim is to represent extracted features in a master feature vector made up of rows (documents) and columns (features) [19]. Feature representation is done through indexing (allocation of indexing to documents) and weighting (assigning a weight to each term in the document that defines its importance) [23]. There are various weighting schemes and all can take any of the forms below.

Binary representation (BR) defines feature presence as "1" and not present as "0" [19; 28].

Term frequency (TF) defines the count or number of terms found in a given document. TF allocates a weight to a feature using the number of times a word repeats in a given document. This implies that mostly frequently occurring items are a true portrayal of the content and get high scores [26].

Document frequency (DF) is basically the commonness of documents that comprise a given term or the count of documents in which that feature occurs [19; 28].

Concept-Based Representation is term based frequency

that classifies conceptual features as graph-based document representation [22].

Term Frequency-Inverse Document Frequency Representation (TF-IDF) is defined as term frequency with inverse class frequency. TF-IDF identifies a term that is a distinctive feature and frequently occurs in a particular category but is less frequent in other categories, hence has little weight or magnitude [26].

Normalised Term Frequency-Inverse Document Frequency Representation (NTF-IDF) is a technique that uses a normalised factor in relation to TF-IDF as a document length. Terms of same frequency in various documents that apply the normalised factor ensure equal importance and relevance of features derived from short and long documents [28].

*3) **Feature selection**:* Feature selection in VA can be either expert driven or fully automated. The extracts below highlight the various feature selection schemes that exist in VA literature.

Expert driven feature selection also known as PCVA is when a team of experts is specifically responsible for identifying and discovering relevant particular selective features from the VA narratives. The most frequently counted features are extracted and ranked on the basis of their weighting (using discriminative weights) and stored in a vocabulary list for classification [25; 29]. The PCVA process has many drawbacks despite being the gold-standard in VA cause of death determination. The use of scarce various experts to reach a consensus on cause of death, makes the process expensive. Moreover, it lacks consistency as it relies on assessment and review of VA forms [30]. One problem with analysis of VA forms is the variation of performance in the population being studied relative to the true cause of death [30; 31]. Consequently, PCVA is compromised due to poor repeatability of results amongst many drawbacks [4; 32; 33]. PCVA tends to focus on one cause of death even though there might be multiple causes [4]. PCVAs follow own methods when they capture data and they do not have standard questions as features extracted are inconsistent.

[5] report on a study where they compared PCVA and CCVA, and deduced that the Random Forest approach performed better compared to PCVA methods. Moreover, they conclude that there is no optimum approach for VA coding and suggested further inquiries with large dataset and training of models especially where death is undocumented medically. [34] investigated completion rates and factors associated with undetermined cause of death and concluded that, completion rates are high in the early phases of VA, but in later phases of the VA process a significant number of deaths are still lost. The reason for this being after getting VA forms, physicians do not allocate a definitive cause of death. [35] investigated use of PCVA and InterVA and concluded that, the two approaches comparative analysis was slight at individual level and poor at population level, hence they recommended more inquiries and research on VA methods. The wide spectrum of drawbacks above, gave rise to alternative automated feature selection approaches.

*4) **Automated feature selection**:* Automated feature selection entails the use of computer programs applying various statistical approaches to automatically extract features from text narratives. These approaches do not need expert intervention. Previous research provides strong evidence that most investigators have applied automated feature extraction techniques to extract content-based features, concept-based features, structural features and linguistic features [25]. The next step after feature representation is to create a vector that comprises only selected relevant words useful for categorization, a process known as dimensionality reduction. Various forms of dimensionality reduction are discussed below.

Pearson Correlation (PC), which measures the correlation between two variables thus independent X and dependent variables Y [19]. The correlation coefficient r is between -1 and +1 where -1 denotes a negative correlation and +1 denotes a positive correlation between two instances [25; 29].

Chi-squared test (CS), which determines the dependency of two variables by measuring the statistical independence [19; 25]. This statistical test measures the significance level of the observed frequencies of tokens O from the expected frequencies E. Furthermore, this approach is a feature reduction technique where O denotes observed or collected data and E_i denotes expected values.

Information Gain (IG), which measures the importance of a particular feature in the feature vector space. This metric computes the reduced degree in uncertainty given the value of the uncertainty [19; 29].

Principal component analysis (PCA), generates principal components through statistical means by applying orthogonal transformation to transform correlated features [25]. The aim of PCA is to come up with a correlation matrix made up of fewer principal components that are less than or equal to initial number of observations. PCA makes it possible to reduce large datasets into smaller ones but avoiding information loss. The benefit of PCA is that it seeks to identify meaningful patterns in a dataset enabling easier exploration and visualization of data [36].

Local Semi-Supervised Feature Selection (LSFS), which is a technique that applies a boundary for every data sample in the dataset. Thus it further applies vectors using feature weights whilst incrementing margins to get the most relevant and useful features [25].

Fisher Markov Selector (FMS), which is an automated feature selection scheme can effectively and efficiently handle high dimensionality of the feature space as it can choose optimally the most useful important subset of features from various classes [29].

Improved Global Feature Selection (IGFS), which is a hybrid approach employs the ensemble method where it merges the strengths of global feature selection method and a one-sided local feature selection [29].

B. *Semi Supervised learning*

VA is a semi supervised learning problem. Specifically it is a multi classification problem, where the aim is to classify VA narratives to various disease categories as per ICD-10 codes. The following sections discuss the semi supervised learning problem and then highlight the frequently used statistical and ML methods in VA.

*1) **Semi supervised learning problem**:* Semi supervised ML approaches seek to optimise classification accuracy with few labelled data points. This approach is relevant in cases where there is a large volume of unlabelled data. However, process of labelling is costly, difficult and labour intensive [25]. On applying semi supervised machine learning approach, the aim is to maximise the training data set using input X and output y features. Learning takes place using the mapping function $[y = f(X)]$, where training data is provided as an input to the learning algorithm. As such, semi supervised learning aims to predict efficiently and optimally the mapping function, using unseen data X [25].

Classification in VA

A classification method is defined as a description of a specific analytical technique that employs a classifier with predefined rules derived from empirical data applying any classification approach [37]. Classes are derived from classification methods and have attributes that describe them. Classification approaches can fall in any of the following: i) Linear and discriminant techniques which use least squares, hyperplane, quadratic and logistic regression function to separate the dimensional feature space by fitting a line that maximises categories of data points for each class., ii) Probability density function which utilizes the k-nearest neighbour approach and applies the concept that a new data point has a high chance of being positioned close to other data points of the same class in the feature space. Therefore, new data points in the feature space are categorized to clusters using class probabilities of their positions and iii) Decision trees and rule-based methods which iteratively divides the feature space into smaller regions until branches are pure, thus there are only branches belonging to a specific single class [37].

Classification requires the creation of a classifier from training data to predict the VA classification/categories [38]. In simple terms in the VA domain, classes can be thought of as validated cause of death and attributes are symptoms/signs that are collected through interviews with the relatives of the deceased. The feature space consists of observations that have certain conditional probabilities with corresponding scores [37]. Therefore the aim of classification methods is to divide the attribute space into predominant clusters of a single class of observation [37; 39].

*2) **Statistical models**:* Various statistical learning models have been applied in VA domain to date. The following extracts discuss statistical approaches.

Logistic regression

The study by [14] elaborate on logistic regression models as being widely used in biomedicine due to their user friendliness, high level of interpretability and their optimum use in solving problems. In this statistical model we assume $w \in \{0, 1\}$, where 0 represents a bad class and 1 a good class. Therefore, the logistic regression model defines X as the data set and θ as a vector of coefficients derived from maximum likelihood estimation [40].

Naive Bayes

The Naive Bayes classifier (NBC) is based on the Bayes rule and makes the assumption that features in the dataset are mutually exclusive or independent from each other [24]. This implies that the probability each feature contributes, is independent to the final class probability. This approach can be trained efficiently as it assumes value independence of a specific feature from the value of other features, when given a class variable [41]. NBC is a statistical approach that derives conditional probability between each class and corresponding attribute values by analysing the relationship between each feature of a class and the class. This computes the probability of instance x given class y assuming the features are independent [21].

[42] argue that the NBC has shown better results in terms of VA classification. [21] implemented the NBC in VA as; each VA record applies NBC to assign a possible cause of death using probabilistic means. Each possible cause of death had a label that corresponded to independent conditional probabilities given specific symptoms and the absolute probability of that same label. The label with the maximum probability is then allocated as the appropriate cause of death to the VA record. In mathematical terms, each record is given cause of death C_{j*} with the maximum probability, for any set of n recorded features in the verbal autopsy, represented by $F1, ..., Fn$.

*3) **Machine learning techniques**:* ML approaches specifically can be powerful, effective and efficient tools in developing better semantic structures to solve issues around automated VA classification [43]. This is evidenced through their superior capabilities such as; use independent parameters which are non parametric, fast in testing, produce results in complex domains, reduce problem complexity, easy to implement and employ more resembling characteristics of the model [24]. [32] points out that machine learning models can understand better the input-output relationship. Moreover, ML techniques are extensive, comprehensive, highly reliable and consistent relative to changes [4]. The study of [44] concurs with the above assertions and points out that, ML approaches have been employed due to their ability to handle high dimensional data and producing good results with limited training samples.

[32] applied Artificial Neural Networks (ANN) to classify cause of death from VA narratives. However, they reported issues of over fitting and suggested increasing the number of training examples could improve ANN. [3] did a study on

VA classification and they reported that there is no technique of automating VA classification that can be better than the PCVA. Albeit, their study only used conventional ML approaches. [45] applied Natural Language Processing (NLP) for classification as a deep learning model (advanced ML approach). Nevertheless, they used a small sample size and their results cannot be generalised across studies. [46] did a study on "multi-task learning for interpretable cause-of-death classification using key phrase prediction" using clustering techniques. In spite of that, they reported a low score in terms of accuracy. [41] used unstructured clinical datasets applying text mining and ML for automatic identification of protected health information and got promising results. [20] did an investigation on application of character embeddings (representing each character in a narrative as a numeric in the vector space) in VA and they reported a low score in terms of accuracy. Regardless, they concluded that application of character embeddings can be useful in VA.

Machine learning approaches have been used effectively for epidemic forecasting including surveillance for health systems. Nonetheless, there is need to focus on real-time monitoring and forecasting [47]. Literature reports on application of conventional shallow ML models in VA domains which perform poorly, suggesting that as they learn they suffer from dimensionality and sparsity of the feature space [6; 18; 18; 19; 21–24; 26; 28; 29; 31; 45; 46; 48]. We discuss various ML approaches in the following extracts.

k-nearest neighbour

The k- nearest neighbour model derives it's concept from the distance or similarity function for given pairs of data points [24]. The k-nearest neighbour determines similarity function by calculating the Euclidean distance or the cosine similarity measures. However, the best choice of k depends on data. Thus, a good choice can reduce the noise effect [24]. Furthermore, it uses all available instances to categorise a new instance based on the distance function [19]. K-nearest neighbour can also learn by analogy through n dimensional numeric attributes [49]. Therefore, data points which are close to each other are called nearest neighbours. The final class of a specific data point is known as the majority class.

k-means clustering

k-means clustering is a technique that applies unsupervised learning and aims at grouping similar objects with similar characteristics together [50]. In performing this approach one has to first define required number of clusters k for n data points, allocate each data point to a cluster randomly, work-out cluster centroids, re-allocate each data point to the nearest cluster centroid, re-calculate cluster centroids and iterate on last two steps until exhaustive to get to global optima [36]. There are a few studies that have applied this approach to date. One study by [46] used clustering, however they concluded that this approach can improve cause of death classification and can be interpreted easily even though in future they have to improve on accuracy.

Decision trees

According to [19], decision trees identify feature conditions as nodes, and the outcome of a condition as a branch and class labels as leaf nodes. The start of the tree node is known as the root node and is identified as the most important feature in the feature set. The root node can be recognized as an attribute that can split the feature space in a best optimal way. Branches can also be labelled as terms with a corresponding weight. Leaf nodes represent class labels and new data point can be generated and assigned a class based on the majority vote of the leaf nodes.

Random Forests

Random Forests has been used in the VA domain because of easier implementation. This approach uses many individual learner trees to perform classification based on votes on an overall category from a given set of inputs [51]. A majority vote is applied to determine a class of a new data point through combining random forests and decision trees. Combined decision trees k are built on randomised algorithm from the bootstrap samples with n observations [31].

Support Vector Machine

The Support Vector Machine (SVM) approach has been applied in the VA field because of the ability to handle high dimensional data, non-linearity and producing good results with limited training samples. This technique employs the idea of a decision boundary known as a hyperplane that distinguishes between classes in a high dimensional space [52]. The documents that are close to the hyperplane are called the support vectors [24]. The SVM approach aims at maximising the margin. SVM computes the maximum linear distance between separated classes in feature space. In non-linear cases the SVM has the ability to use kernels which can easily handle non-linearity between classes and the feature space.

Artificial Neural Networks

Artificial Neural Networks (ANNs) which form the basis of deep learning has been applied in VA domain, and uses a neural classifier that can be thought of being a network of units where each unit can be a term and the output unit represents a category [24]. The ANN consists of the input, hidden and output layers. In document text classification, input units are allocated weights. The activation of these units is transimitted via a network to get an output value that determines the categorisation [24]. The layering of the nodes provides a map of the decision space also known as the neural network where a program can learn rules from massive data amounts being processed [13; 53]. The ANN applies an error function at the output layer to generate updated weights using back propagation (process of fine tuning weights based on error rate computed from previous iteration). Effective and efficient back propagation makes models genaralizable and reliable. ANNs have been applied in VA and have shown promising results even though there is need to have more training examples for the model to be effective [32]. [3] applied the various ML approaches and got best results from the ANN using a feed-forward network (no back propagation) with one hidden layer and achieved sensitivity

of 0.695. Figure 1 below depicts a typical ANN model of a single neuron in a feed-forward ANN. Input information x_i is received from other neurons, and takes the product of input information and each data point corresponding weight denoted by w_{ij}, and produces a weighted output applying an activation function $f(x_j)$. The process iterates taking the weighted output as input to another next layer neuron in the next layer. The same process is continues up to the output layer.

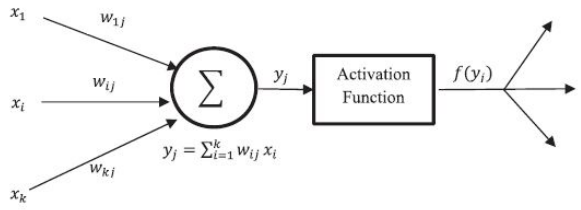


Fig. 1. ANN structure of single neuron.

Decision rule

Decision rule is a classification approach that uses an inference rule-based engine to classify documents to their respective categories. Furthermore this approach can entail building a knowledge base with a set of terms or keywords derived from consultations with domain experts [54].

Case-based reasoning Case-based reasoning (CBR) has been applied in literature by [27] and employs reasoning through a rule-based approach to construct meaning through synthesis of information to derive a new case in an autopsy report. This new meaning constructed through extraction, is converted into features and values. The new features and values are taken as input by the system for similarity analysis. In the process these new generated feature values are then matched with the previous cases by calculating individual feature similarity scores using the k-nearest neighbour approach. It employs the Naive Bayes learner to assign weights to features based on significance and is used as a computation to the final similarity analysis. Only top scores are selected using similarity values and identified by user as the desired outcome, and thus stored in the case base [27].

4) Ensemble classifiers: The following extract discusses ensemble approaches, which are hybrid schemes of various ML approaches and use a voting scheme or weighting in order to do the classification. These approaches mainly use several prediction tools and aggregate their prediction, such as such as voting, associative classifiers, centroid-based classifiers discussed in the study of [24]. Implementing ensemble methods improves prediction quality [6; 42]. However, these approaches suffer from very low interpretability and high computational time [14].

Boosting

Boosting also known as an iterative ensemble classifier is a ML approach that improves a weak algorithm into a stronger one through aggregation of multiple classifiers [55]. It employs a weighted vote when labelling unlabelled instances that ultimately becomes a composite classifier [39]. The most common boosting algorithm is The ADABOOST which is short for Adaptive Boosting. The approach has been used in many studies such as [6; 55].

Bagging

Bagging or bootstrap aggregation is an approach that builds a classifier for every bootstrap based on n bootstraps from given datasets [39]. Given each trial denoted by $(t = 1, 2, 3...T)$, where the dataset is of size N. The learning environment generates a classifier C^t from the dataset and a final classifier C^* is created from aggregation of the T classifiers that are classified from instances x through voting for class k where it is recorded for every classifier $C^x = k$ and C^* which denotes the class with the most votes.

5) VA algorithms: VA algorithms have been the alternative traditional method of assigning cause of death apart from the PCVA technique. VA algorithms require VA data derived from real deaths and Symptom Cause Information (SCI) (which is a repository of information about symptoms that are related to each probable cause of death). The SCI may be derived from the physician's reports and logic (that entails a logical algorithm) that combines the two to identify cause-specific mortality fractions (CSMF) to assign specific causes of death. The literature to date identifies five various free online VA computational algorithms that have been widely employed and most of them apply statistical approaches [56].

The InterVA algorithm uses the Bayes rule (formula that computes conditional probabilities of X when given evidence) and Symptom Cause Information (SCI) as conditional probabilities of symptoms given a specific cause of death using Cause Specific Mortality Fractions (CSMFs). These conditional probabilities are sourced from physicians and the Population Health Metrics Research Consortium (PHMRC) gold standard dataset [57]. The PHMRC is responsible for supporting survey design, primary data collection, analysis, development, and to provide tools for quantifying population health. VA datasets cannot distinguish between the confirmed absence of a symptom and missing data. Therefore, to address this limitation of the data, InterVA-4 only utilizes the confirmed presence of a symptom. Moreover, this approach uses propensities (which are specific to a particular cause of death) and not probabilities. [21] applied the Inter VA and reported a sensitivity of 0.43 and CSMF accuracy of 0.71.

PHMRC 'gold-standard' data have both VA and medically certified cause of death. This contains information about the relationship between VA symptoms and medically certified causes. The Tariff algorithm uses fewer symptoms and it employs the SCI as a tariff score to rank causes of death, thus determining the association between specific symptoms and causes in the PHMRC gold standard dataset. CSMFs are arrived at by identifying a single cause with the highest rank for each death in the VA dataset and sum them up [56]. The standardized Tariff Score Q_{ki} for death k and cause i is the

quantile value of S_{ki} in the distribution of S_{*i} formed by a set of deaths sampled with replacement from the PHMRC data set so that they have a uniform cause distribution. Values for Q_{ki} are bounded between 0 and 1 so they have the same scale and can be compared. The cause chosen for each death k is the one with the largest Q_{ki}. [58] applied the Tariff approach and reported 0.505 Chance-corrected concordance (CCC). CCC computes the degree to which the predicted cause of death classes relate to the actual cause of death classes. They further reported a CSFM accuracy of 0.770.

The InSilicoVA algorithm uses a statistical approach employing joint probabilities to identify the most likely significant cause of death in relation to CSMFs for all deaths in a VA data set [59]. This implies that the algorithm uses SCI and derives this based on the Naive Bayes model based on conditional probabilities (symptoms given cause of death and cause of death given symptoms). The probabilities can be sourced from physicians or use a gold standard dataset that can be a de-biased physician-coded VAs. The study by [7] applied the Insilico approach and reported true positive predictions of 0.341 on all 34 cause of death classes and 0.85 CSMF accuracy.

The Naïve Bayes Classifier (NBC) algorithm applies the bayes rule to categorise cause of death. The study by [21] applied the NBC and achieved better results as compared to InterVA and the Tariff approach with a sensitivity of 0.57 and CSMF accuracy of 0.88. Nevertheless, their algorithms made use of structured questionnaire data only.
The King-Lu algorithm uses symptom conditional probabilities to approximate cause of death of a dataset over 13 classes. This approach does not avail a cause of death for individual records [3]. This algorithm relies on SCI training data which defines clusters of symptoms rather than a single symptom. Moreover, it is recommended to use gold standard deaths and if possible they should be from the same population as the VA deaths to get credible and better results [56]. [60] used the King-Lu and reported a CSMF accuracy of 0.96

[1] argue that these VA algorithms can not avail enough evidence where there is limited expert diagnosis, hence they cannot be used to guide health priorities. These VA algorithms mainly employ statistical approaches and tariff scores to rank causes of death [56; 60]. Various VA algorithms are dependent on sample size, age group, causes of death, data set size or characteristics of the sample in order to produce best results [4; 61]. There has been challenges and issues in terms of various VA techniques and there has been proposals to improve VA approaches through minimising the number of features under study (dimensionality reduction) and also combining various algorithms in order to improve on performance, accuracy and efficiency [21]. Moreover, the validity of the VA approaches performance in terms of sensitivity, specificity and predictive values vary with regard to causes of death across populations [62]. It is difficult to generalize and standardise VA classification practices, since there is no gold standard and patient's records vary in terms of socio-economic status [21; 56]. Another issue is associated with standardising VA questionnaires so that they have the same structure and content. The format and standard of VA questionnaires differs considerably, thus their administration requires appropriate training so as to elicit relevant and appropriate symptoms and causes. There are also often language barriers and the interviewer and interviewee need to speak the same language so as to derive best results. [4] recommended incorporating fully trained multiple translators. A standardised and uniformly reliable method is required to determine cause of death and also a standardised questionnaire which can be applied globally [4].

C. Deep learning

Deep learning approaches have been implemented successfully in literature for the past decades, with the aim of of improving classification. Moreover, we have witnessed the implementation of emerging deep learning frameworks that have successfully managed to extract and reveal critical information, a task that conventional statistical and machine learning models could not. Additionally, deep learning achieves learning of data representation with varying levels of data abstraction when even computational techniques use few processing layers [25]. Therefore, these advanced approaches can be relevant in VA data with high-dimensional sparse data.

Literature reports a wide gap, especially in applying advanced machine learning techniques known as deep learning models as approaches to solving VA problems. [20; 25; 46] are some of few studies to have used deep learning approaches to date. [20] investigated "the application of character embeddings to improve cause of death classification using VA narratives." In their study they applied distributed word vectors (Sentence2vec, GLoVE and Word2vec) and combined two character embeddings techniques (CNN and GRN) for classification and they concluded an improvement in cause of death classification. Application of Natural Language Processing (NLP) as a deep learning approach to solve VA problems has shown promising results [3; 28; 63–65]. NLP is an approach that entails converting human language into formal easy to understand language for computers. It performs automated analysis and human language representation through use of a variety of automated techniques [66; 67]. NLP has efficient, superior performance, integrated feature learning and effective capabilities of attaining end to end learning from complex and multi modality data [43]. NLP word embeddings (representation of text in numeric form stored as vectors) on dense vector representations, enabling multi-level automated feature representation learning, has to date produced excellent results [66]. [68] used a method based on NLP to extract relevant clinical information and discovered that the approach was better than PCVA diagnosis. [69] investigated use of NLP on suicide note classification and discovered that NLP can assist in the classification process. [70] did a comparative study of existing ML and deep learning approaches and found out that deep learning was better in terms of performance as

compared to machine learning. Recent deep learning methods such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM) and the Transformer architecture have managed to explicitly solve most NLP tasks [43]. The following sections discuss the deep learning frameworks that are used in VA. Figure 2 below compares the conventional ANN with one hidden layer and the deep learning architectures with multiple hidden layers.

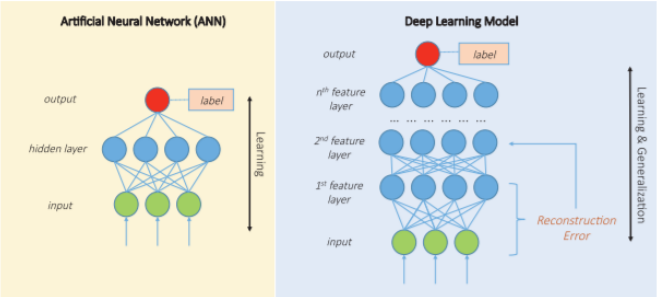


Fig. 2. Comparison of ANN and deep learning architecture.

1) Distributed vector representations: Distributed representations emanated from the issue of complex and high data dimensionality for traditional statistical NLP. Their end result is the distributed representation of words in low dimensional space. The start of the data processing layer in deep learning models is represented by distributed vectors. In distributed vector representation, the logic is that analogous words usually come, about the same context. In order to compute similarity mathematically, the cosine similarity coefficient is used [66]. The study of [46] successfully applied distributed vector representations.

Word2vec

Word2vec is an approach that applies the concept that words within a vector space are of fixed length [71]. Word2vec builds vocabulary from a training text corpus and creates a vector space for each word through learning. It further groups alike words using the cosine distance function [72]. It consists two models to develop and construct distributed vector presentations which are; Continuous Bag-of -Words (CBOW) and Skip-gram models. CBOW predicts a word based on its context whilst the Skip-gram model predicts context based on the neighbouring words, thus it discovers semantic relations amongst terms in the corpus [73; 74]. The Word2vec uses the input, projection and output layers [73]. A one hot word vector of V neurons is taken by the input layer and corresponds to hidden layer N neurons. The softmax is output that contains all words in the vocabulary. Each word in the dictionary is denoted by learned vectors v_c and v_w where the connected layers use a weight matrix [66]

Sentence2vec

[75] discusses the Sentence2vec which applies the same concept of Word2vec. They argue that a sentence can be seen as text unit and the summary of the document being made up of sentences. They point out that each word in the corpus has an *id* that denotes its position in the dictionary. Sentence2vec represents each sentence in a document as a vector, unlike Word2vec which represents each word as a vector [75].

Global vectors The Glove technique is count-based and makes the supposition that the meaning of a word is implied by it's own contextualisation [73; 75]. It is mainly based on word occurrences in a textual knowledge base. This implies that words with similar contexts should also have similar embeddings.

Doc2vec

This technique uses an unsupervised learning algorithm to come up with vector representations of paragraphs or documents. The algorithm is a modification and application of the concepts of Word2vec and it aims at finding similarities in paragraphs or documents [22]. The following extracts present deep learning architectures that focus on character embeddings.

2) Character embeddings: Character embeddings assume each word to be composed of not more that one character. They can handle the issue of out of vocabulary words (words that appear a few times in the test set but are not in the training corpus of words), unlike distributed vector representations [66].

Convolutional Neural Networks

Convolutional Neural Networks (CNN) is an example of a character embedding technique. CNN was developed as a feature extraction function to extract high level features from n-grams, thus the features would be used for various NLP tasks [66; 76–78]. The CNN consists of input, convolutional, pooling and fully connected layers. It also has a lookup table and softmax function for classification. After taking input, a look-up table is used for transforming each word into a vector of user-defined dimensions. This implies that given an input sequence of $(s_1, s_2, ...s_n)$ of n words is changed into a series of vectors $(w_0, w_1, ...w_n)$. Learning in this network process is done using weights. Data representations and extractions are done on the convolutional and pooling layers. The convolutional layer makes use of a various numbers of filters (kernels), where one kernel extracts a particular pattern of n-gram. The pooling layer employs a maximum pooling strategy which performs sub-sampling on each input specifically by applying a maximum operation on each filter. Maximum pooling therefore, maps the input to a fixed dimension of outputs regardless of the size of filters and thus a reduction in the output's dimensionality. Lastly, a fully connected layer employs a feed forward ANN. The fully connected layer consists of an input, hidden and an output layer(s). At the output layer, a loss/cost function is applied in this context known as Softmax classification [66]. The study of [46] successfully applied the CNN and Gated Recurrent Neural network discussed below. [25] argues that the CNN has the ability to effectively learn multifaceted features from the medical data sets as compared with traditional statistical and machine learning approaches. Figure 3 shows the visual representations of the CNN architecture.

Gated Recurrent Neural network

Recurrent Neural Networks (RNNs) is a form of character

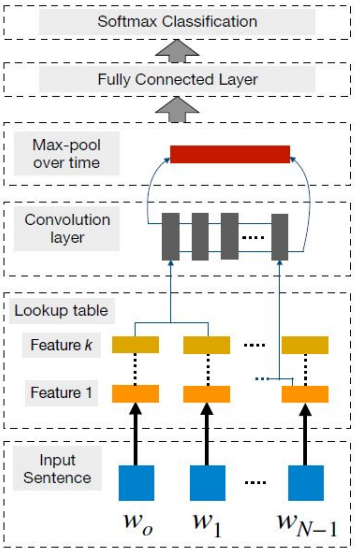


Fig. 3. Convolutional neural network framework.

embedding architecture. RNNs are employed in NLP to process sequential information. RNNs produces output that is dependent on previous tasks and results by performing same task on each instance. They make use of memory to pull previous computations for processing with current operations [66]. However, RNNs have issues of "vanishing gradient" (where weighted activation function inputs of a neural network increase or decrease whilst the derivative function approaches zero) which ultimately affects accuracy and makes it difficult for the architecture to learn and hyper tune parameters for optimum results. The Gated Recurrent Neural network (GRNN) has the rest and update gate that control flow of information. This approach further has no control of the hidden layer and has no memory unit. The GRNN has been applied in the study of [20; 46] and showed promising results.

D. ML Performance Metrics in VA

This section gives an overview of the ML evaluation metrics that have been applied in VA to date. More approaches to evaluation, accuracy and validation of algorithmic performance have already been reported on. Nonetheless, in medical data focus is given to Recall than Accuracy [9]. [57] points out that performance evaluation to determine the predictive accuracy of VA algoriths can be done at individual level or population level. At individual level we seek to determine specific cause of death categorisation, whilst at population level we seek to estimate Cause Specific Mortality Fractions (CSMFs). These processes are key to achieve comparability. Assuming a supervised learning problem with a labelled deaths dataset, there is a training set $T = (x1, y1...; x_n...; x_n)$ where each x represents a symptom and each known cause of death is indicated by y . As such, we have a model of a VA algorithm f to derive the predicted label for a new dataset (X, Y) is denoted by function f. Given the above proposition we can therefore attain the performance of the algorithm that depends on the training data.

A confusion matrix is the benchmark of our evaluation matrices, and consists of True Positives (TP), True Negatives (TN) that denote our correct predictions, False Positives (FP) and False Negatives (FN) that signify incorrect predictions. Below we discuss various performance evaluation measures in VA.

Actual labels	Predicted labels		
		Positives	Negatives
	Positives	TP	FP
	Negatives	FN	TN

Fig. 4. ML performance evaluation confusion matrix

Macro precision or Precision/ Positive Predictive Value (PPV) refers to the proportion of correctly predicted positive VA reports to the total positively predicted VA reports in fraction terms. This denotes how good a classifier identifies positive observations without much on the negative ones.

Macro recall is also known as Sensitivity or True Positive Rate (TPR) defines the ratio of correctly predicted positive VA reports to the all VA reports in actual positive category. This implies the rate of correctly predicted positive observations or diagnoses (cause of death) [29].

F-measure is a test score to measure accuracy of a particular model. It is a weighted combination of Precision and Recall scores and is computed as the average or mean of the Precision and Recall [29].

Overall accuracy denotes all classes with classified results that have been predicted correctly in fraction terms, thus it is the ratio of correctly predicted VA to the total VA reports [29].

The Area Under the Curve (AUC) is used to measure and validate the performance of machine learning (ML) algorithms. It is a graphical plot that depicts the True Positive Rate (TPR) against the False Positive Rate (FPR). The AUC evaluates the classifier's performance, with good performance indicated by an AUC value close to 1, meaning the curve is closer to the upper left corner. Poor performance is indicated by an AUC value of 0.5 or lower. [9; 29].

Specificity is a function that computes the percentage of actual negative medical reports that are correctly predicted as

a negative [25].

Overall chance-corrected concordance (CCC) defines CCC for cause j as the overall weighted sum of cause-specific CCC [57]. [79] points out that given a proposition of measuring agreement on a cause of death, sensitivity and concordance never considers the agreement expected by chance alone. As such, given a VA algorithm that assigns randomly cause of deaths by chance, we would expect concordance of $\frac{1}{n}$.

Cause Specific Mortality Fractions (CSMF) accuracy defines accuracy as having a value between 0 and 1 and it assumes the worst possible case for predicting CSFM and assigns a weight on the least possible CSFM value that matches the total absolute error [57; 79].

Top cause accuracy is the number of correct cause of death of being first assignment divided by number of causes.

Top 3 cause accuracy is the number of correct cause of death within the first three cause assignment divided by number of causes [57].

Kappa metric applies the true test set elements under the assumption of chance prediction and is dependent on the structure of the test dataset. Thus it computes the relationship of all causes except cause specific of concordance [79].

Case similarity measure is used as performance metric in Case-Based Reasoning (CBR). CBR evaluates case similarity by using the nearest neighbour approach because of its simplicity and closeness to human judgement. Features are compared to determine the matching of all previous cases in a knowledge base of cases and all cases and feature values should exist in the system dictionary. Once the matching and comparison is done, the ecosystem performs matching cases using a ranking system based on likeliness scores applying the k-nearest neighbour. Only cases with highest weighting are considered as results [27].

E. ML Optimisation techniques in VA

ML learning entails applying a learning algorithm where a classifier is induced from training data, hence every classifier has a measure of performance (through bias and variance) known as the prediction error which is unknown and is estimated from data [38; 80].

1) k-fold cross validation: Validation analysis is crucial in ML as it finds optimum models to fit with new observations,thus improving accuracy and model prediction. One approach that has evidently been applied in the VA field as an optimisation technique is the k-fold cross validation. k-fold cross validation is a technique that splits data into k folds, and learning of a classifier is done through $k-1$ folds where the test error is calculated using the balance folds and gives the mean error for each fold. [81] points out that k-fold is the widely used procedure to test the effectiveness of a classification algorithm by comparing the performance between two algorithms. Moreover, performance can be evaluated through by average k-accuracies from k-fold cross validation. [3] used 10-fold cross validation and deduced the ANN as the best classifier.

F. Imbalanced and fragmented datasets

There is little literature in VA that discusses the issue of fragmented or imbalanced datasets, also known as skewed datasets, even though it is a common situation in VA domain to have these natural datasets due to the rare situation of the events under consideration. VA data has suffered from high text scarcity, imbalanced text and also in the cause of death categories which is an issue for NLP [82]. [25] points out that imbalanced class distribution (majority class out numbers minority class in terms of records) is when one class can be of more interest and is insufficiently represented compared to other class. In such instances, when categorisation is performed there can be misinterpretations and misclassification. Henceforth, imbalanced datasets are a drawback in algorithmic predictions as there is a always a bias of rules that concentrate mainly within the majority class and tend to frequent less the minority class. This leads to misinterpretation of the results in terms of accuracy in the minority class even though overall performance maybe high. As such, there is need for special techniques to address this thereof.

One approach that has been proposed is over-sampling. The effect of over-sampling in trying to create a balanced dataset without adding new data points, is that it grows the minority class size by duplicating of some positive cases in a random fashion. However, through the replication of data points over sampling can suffer from issues of model over-fitting. As a substitute to this approach and in seeking to address over-fitting is the Synthetic Minority Over-sampling Technique (SMOTE).

SMOTE uses data generation and interpolation to create synthetic minority examples of data points that are near to each other using already prevailing positive cases. On the contrary is an approach known as under-sampling which takes out negative cases from the dataset haphazardly to come up with a balanced dataset [25]. Nevertheless, under-sampling can result in information loss from the majority class. To curb information loss, one can still use random over-sampling as it can replace lost data through minority class replication of data points. However, this can result in over-fitting. However, despite a plethora of benefits such as ease of implementation and improved accuracy levels, these approaches are worse in terms of performance as compared to other techniques [14]

[14] discusses other approaches of overcoming this data imbalance, which are random sampling, applying the k-nn and selecting an appropriate way for categorization quality

using techniques such as Area Under Curve or ROC can prove beneficial. Furthermore, they argue that to address issues of imbalanced data, sampling techniques are the most common solutions. Other alternative approaches of addressing imbalanced data issues such as nearest neighbours and points on the line segments use random techniques. Other sampling techniques which integrate machine learning approaches such as k-nearest neighbour is the Balance Cascade ensemble algorithm. This approach takes out past sample cases that where appropriately categorized from the main class and uses the balance of the data points to come up with a new classifier. Another approach is NearMiss algorithm that uses the k-nearest neighbour approach to choose data points that are nearest to each other for each minority case, such that every minority case is in close proximity with some majority ones. The following extract discusses limitations in statistical and ML models.

G. Model limitations and imposed assumptions

The VA community has witnessed the application of statistical approaches such as the NBC and logistic regression. The NBC assumes that features are conditionally independent and normally distributed numeric values. However, the proposition of normal distribution in the NBC does not always hold in that there might be need to approximate other continuous distributions. The NBC model produces poor results in terms of performance if there is a high degree of feature correlation. In most cases if one does not do a rigorous feature selection process, then this can also compromise the model [24]. NBC assumes conditional independence of extracted features, however this proposition becomes more complex with the increasing number of features, which ultimately reduces the performance of NBC [18]. Moreover, if there is an imbalanced dataset in terms of training and test data, the NBC can lead to erroneous results and misinterpretations [21]. Consequently, the NBC performance is affected negatively in instances where, NBC assumes conditional independence among features which might not be the case for that specific dataset. As a result of an increase in features, it becomes difficult to manage conditional dependence of features [19].

LR is another statistical technique in VA, and is of the notion that the association between inputs and log odds is linear. On the contrary, it is not always the case in terms of linearity, for this reason we can have a non-linearity existence in the LR. To overcome the non-linearity problem in LR we recommend the SVM, which uses the kernel trick and is efficient in non-linear classification techniques [14]. As such, SVM is costly in terms of computation specifically when applying the kernel technique during learning. Moreover, achieving real space division is difficult or rather impossible, thus the SVM applies a margin that allows misclassification of some examples whilst increasing the overall performance. In such cases in biomedicine we use the soft margin SVM through introducing penalties to the objective function [14].

In addition, some studies provide evidence that one can employ successfully the k-nearest neighbour as a non-parametric technique in VA scoring. Yet, the k-nearest neighbour is costly when it comes to computation of weighting distances for each case warehoused during categorization. [19] argues that there is little of learning done by this technique during training, rather it mirrors the training set to do categorisation. Because of this, the model performs poorly and over-fitted. As a consequence, the model fails to handle noisy data and fails to identify patterns for prediction when using unseen or test data, thus it emulates the training set.

The application of the RF in VA has witnessed issues in terms of performance, specifically when the dataset contains large volume of features and very few discriminative features. This therefore, leads to trees in a forest populated by less relevant features that lead to misinterpretations and incorrect predictions [19]. [19] explored the J48 decision tree to classify cause of death from forensic autopsy reports and pointed out that, the model achieved the lowest performance. This was due to issues of having continuous data representing all features in the master feature vector, thus it was difficult to find the global minima needed to construct the J48 decision tree. They concluded that the J48 decision tree is inappropriate for categorization of VA reports. On contrary the ANNs exhibit better performance because of their ability to handle complex data, when equalled to other techniques in terms of accuracy but difficult to interpret. [32] argues that the ANN has limitation of input selection not being straight forward, time consuming in terms of designing optimal network for each cause of death, generalisation of predictions has low sensitivity and specificity and comparing and prioritising CSFM over sensitivity and specificity is a manual process.

There is little evidence that exists in terms of application of emerging deep learning frameworks in VA. Even though deep learning models suffer from interpretability, they show promising results. [25] argues that, for the most part the CNN architecture has resulted in improved accuracy, nonetheless for the model to be more effective in terms of performance it requires lots of training data to learn efficiently. As a result, with little training cases the model is bound not to perform optimally when it comes to accuracy. In other words, the model requires a large number of features in order to iteratively learn the most useful and relevant ones to use for classification. In general within the VA domain it is difficult to find data in large volumes. In such cases, it is recommended to employ other choices of deep learning frameworks (I-H) to optimise VA report categorization besides CNN. The following section discusses emerging trends in deep learning.

H. Emerging Trends: Deep Learning in VA

Machine learning techniques suffer from various drawbacks, hence the need to use other emerging technologies of deep

learning. [43] argues that there are various challenges with traditional ML techniques, such as supervised ML requiring a labelled dataset by experts, going through data cleaning, performing feature engineering among many. This therefore, leads to poor scaling of the feature space and consequently hides possible opportunities to discover new trends. [32] argues that conventional ML approaches lack the capability to process natural data in raw form. This is as a result of the structure, linearity and challenging transformation processes of the input feature. As an option, deep learning architectures provide representation techniques of the target value from the raw text, thus discovering new patterns and trends.

Deep learning entails various layers made up of neural networks that process data to learn hierarchical abstract representations of data. Deep learning involves NLP, which uses a variety of computational methods for automated analysis and illustration of the human language [66]. NLP tasks done by computers include among many; "parsing, tagging part-of-speech, machine translation, computer vision, pattern recognition, semantic role labelling, named-entity recognition, deep learning methods and dialogue systems [67]." The following extracts discuss other forms of character embeddings that can prove essential in the VA domain.

1) Other Recurrent Neural Networks approaches: Recurrent Neural Networks (RNNs) is a form of character embedding architecture that is employed in NLP to process sequential information. Therefore apart from the GRNN applied in [20], there is need to explore with other RNNs techniques in VA research such as the Long Short Term Memory (LSTM) architecture. The LSTM technique has the capability to learn and hyper tune parameters for optimum results, thus overcoming the issue of the vanishing gradient. Figure 5 is the RNN model.

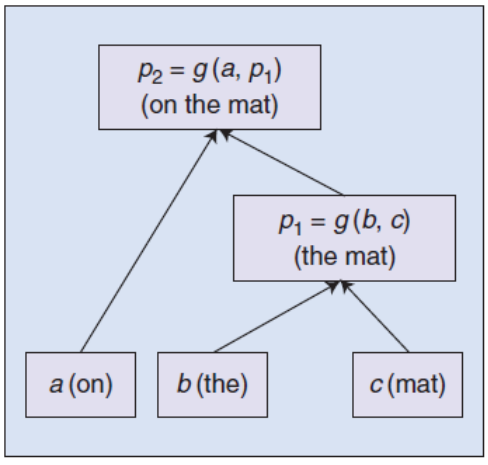


Fig. 5. Recurrent neural network framework.

Long term short memory

Long Short Term Memory (LSTM) character embedding architecture that has the forget, input and output gate. The forget gate makes it possible for it to error back propagate a number of times and calculates the hidden layer by averaging the three layers [66]. It calculates the hidden state by combining the three gates. Figure 6 is the LSTM model.

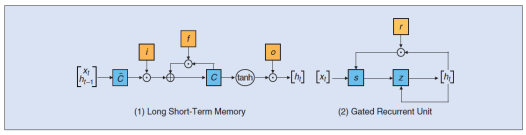


Fig. 6. Long term short memory framework.

Matrix Vector Recursive Neural Network

The Matrix Vector Recursive Neural Network is a character embedding architecture, that seeks to represent every word and phrase as both matrix and vector [83]. It creates a combination of one matrix multiplied by the vector of the other. Recursive Neural Tensor Network (RNTN) applies less parameters and promotes more interaction between input vectors [66].

Transformer

Developments in deep learning have led to new architectures such as the transformer, which has an encoder that maps input sequences of symbols to a sequence of continuous representations, such that the decoder generates an output sequence of symbols one element at a time. The transformer takes input of previous instances to generate next feature representations [84]. [84] points out that the transformer architecture uses a decoder and encoder that are connected layers with structured attention. The encoder takes symbol representations in a sequence $(x_1, ..., x_n)$ to continuous sequenced representations $(z_1, ..., z_n)$, thus decoder uses z sequenced outputs $(y_1, ..., y_m)$ features one at a time. At each stage, the model takes previous input to generate the next symbols. The encoder has 6 stacks of similar layers which have 2 sub layers apiece for attention and feed forward functionality. The decoder consists of 6 stacked similar layers and 2 sub layers apiece. It then has a third sub layer which functions as multi head attention over output from the encoder. The attention has mappings to requests and uses keys with values that correspond to output. It creates vectors for requests/ queries, keys, values and output. The end result is a calculated weight of all values using a compatibility function of the request and the related key. It applies the scaled dot product that takes input as queries of dimension d_k and values d_v. It applies the softmax function to obtain weight by taking the dot product of the query and keys and divide each by $\sqrt{d_k}$. The multi-head attention attends to varying representations at differing positions in space by permitting the model in a combined way.

I. Limitations in VA literature

Below we discuss and analyse limitations that we identified from the VA literature;

- Missing data and varying VA input
- No model transparency performed,

- Data balancing techniques not applied,
- Absence of standardised tools and techniques that are used in VA studies,
- Scarcity and cost of experts,
- Various medical ontologies resulting in heterogeneous datasets,
- Relying on expert's diagnosis as gold standard,
- Scarce availability of publicly available biomedical datasets with standard training and testing data,
- High dimensional low sample size datasets and noisy data points,
- No exploratory data analysis is done to explore variable distribution and check for any outliers in datasets and
- Model Complexity is not taken into account.

This section synthesises the main limitations of VA literature. In general all the discussed methods take input of VA narratives, however the end result of categorisation is affected by main limitations of VA studies. We have witnessed use of smaller data cases in VA studies, and usually gives a biased outcome in terms of the precision of estimates. Apart from this, there is bias in interviewees remembering true cause of deaths, which can result in misclassification. Furthermore, because of lack of resources to determine cause of deaths, in most cases there are cases of undetermined or unknown deaths and in fact as a result these are not included in studies because of missing information [85].

Most statistical and machine models in VA do not perform model transparency and this reduces the confidence in model interpretability, hence there is need to incorporate model transparency in order to be able to justify and explain model prediction. The VA community can apply any choice of transparency models such as Local Interpretable Model-Agnostic Explanations (LIME) amongst many to cater for this limitation. Few studies report on data balancing to improve classification accuracy, even though there is scarce literature on the actual application of data balancing techniques in VA studies. Depending on the dataset type and characteristics, one can incorporate any choice of an optimal data balancing approach.

There is also a need to have standardization of tools, processes and methods used in VA studies, one example is the diverse options in data acquisition and preprocessing. Moreover, one key challenge and limitation is the scarcity of physicians that perform VAs. Thus, there is need to prioritise resources in order to address the imbalance. This synthesis of the literature revealed the one main limitation to be heterogeneous datasets. Hence, there is a great need to have standard ontologies that house VA data.

Further inquiries from the literature revealed that there is no gold standard that can be used to test the performance of VA. This lack of a true gold standard limits the explainability and interpretation of validation study findings. To add more to these limitations is the issue of scarce availability of biomedical datasets that can be used to train and test models in VA in order to improve performance. It is key going forward to have more publicly available datasets that can be used by the VA community to train and test models effectively, as this improves model performance and avail better results.

There is need to have more robust dimensionality reduction schemes in place to improve model performance. Clearly, we need to apply special preprocessing techniques and algorithms that can handle noisy data. Although previous research provides insights on how feature engineering improves model accuracy, we need to put in place other forms of feature engineering techniques in order to generate new relevant features from existing ones and improve model output. We propose to apply and explore with genetically inspired feature selection or engineering approaches. Model complexity is a result of models that are highly parameterised, which is the norm in biomedical data. For this reason it is imperative to reduce the feature space and only retain relevant and useful features during training of the algorithms. The next extracts discuss our methodology.

II. Methods

In this paper we apply a methodological systematic analysis to;

1) systematically review the most commonly used statistical and ML techniques in determining causes of deaths from VAs;
2) identify limitations of each of these approaches;
3) propose machine learning framework guidelines for determining causes of death from VAs;
4) give an insight to future emerging directions;

This methodological systematic analysis takes into account statistical methods such as, LR and NB, and the ML approaches include k-nn, DT, SVM, ANN, RF, CBR, Ensemble classifiers (Boosting and Bagging), k-means clustering, distributed word vectors (Word2vec, Sentence2vec, Doc2vec, Glove), Convolutional Neural Network (CNN) and Gated Recurrent Neural Network (GRNN). The statistical and ML methods discussed in this methodological systematic analysis is not complete considering various approaches that have been applied in VAs. This literature review only highlights most of the regularly used methods. In this paper we present a flowchart that details our methodological systematic analysis in Figure 8. This study considers the procedure of newly published articles on VA. The study used the following search terms: "Algorithms in Verbal Autopsy", "Statistical techniques in Verbal Autopsy", "Machine Learning in Verbal Autopsy" and "Deep Learning in Verbal Autopsy" to retrieve relevant articles. The inclusion criterion included published articles on VA over the period 2010 to 2020. In this study we chose to include high quality papers from reputable journals such as; peer-reviewed journals and conference articles. As part of the filtering process we selected papers through analysis by evaluating the abstract, conclusion and reading the whole paper. Only articles that were written

in English were considered. The repositories used in the search for papers were Google Scholar, PlosOne, Science Direct, IEEEXplore, BioMed Central and Springer-Link. We selected these databases as they are reputable and consists of research work that covers the whole globe. Our final list of papers did not comprise any unpublished work and dissertations. The search yielded 110 articles, of which 76 were peer-reviewed journal articles and 34 were conference papers. Furthermore, we had a database that served as a point of reference for further statistical analysis. The main intention was to quantitatively establish the research status of statistical and machine learning application in VA. The methodological systematic analysis identified 85 articles as being relevant for our study and the remaining 25 were discarded. As a result of the methodological systematic analysis of the literature we generate and present tables, pie charts and histograms. Figure 7 depicts a histogram of the initial peer reviewed journal articles and conference papers.

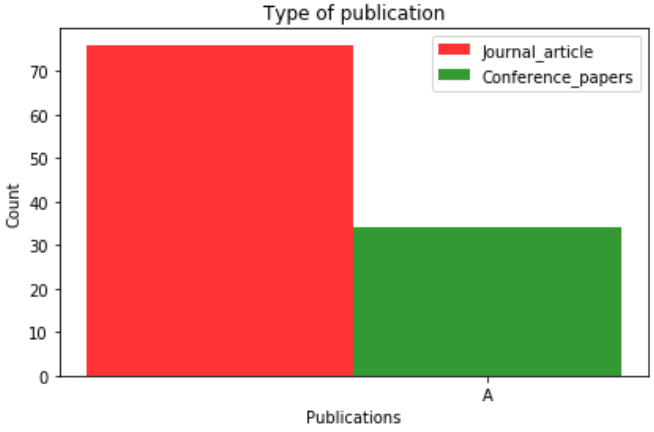


Fig. 7. Distribution of initial searched journal articles vs conference papers

A. *Existing literature survey on VA*

There are few literature surveys on VA [5; 8; 25; 28]. In this study we give an insight into the key aims of each literature study and perform a comparative analysis to distinctively highlight key differences with our study. The identified studies mainly air their views in terms of model comparison in the determination of causes of deaths from VAs based on their performances. However, much of the studies do not consider critical issues in the determination of causes of deaths from VAs, such as pre-processing, nature of datasets, data exploration and model explainability (transparency). Furthermore, deep learning models are not considered in the listed table below of literature surveys. We have also included a study done by [25], which provides the most recent and the most comprehensive comparative analysis of existing methods in automated text classification in VA. The included studies cover mainly performance evaluation of traditional techniques and propose new methods for VA classification. The literature discussed in [25] covers most automated text classification techniques in VA, model performance, propose new methods for VA, feature selection, feature engineering and other issues. However, in contrast to this study, at the time of this write-up, there is little evidence to suggest any methodological systematic analysis in VA that has been published which discusses, conventional statistical and ML approaches and emerging advanced deep learning techniques. Table I provides a comparison of existing literature survey papers in VA against this study.

III. Results

This section gives an account of results from the primary studies, based on the meta-analysis. The frequency distributions of, feature extraction techniques, feature value representation schemes, feature selection approaches, algorithms and performance validation metrics from all principal studies are synthesized and analysed. The pie-charts for pre-processing techniques, studies that report on data transparency and balancing techniques, the distribution of selected VA articles compared to other domains are also shown. We present a tag cloud of all articles used in this methodological systematic analysis.

A. *Preprocessing*

Pre-processing is a key stage in cleaning data. Our methodological systematic analysis suggests that only 38% of the primary studies applied pre-processing to normalise their data. There is 62% of primary studies that did not apply pre-processing to their data. As such, this adds to noise and high dimensional data and consequently low accuracy prediction models. Figure 9 below denotes the distribution of studies that applied pre-processing.

B. *Transparency*

Transparency is an approach that makes machine learning models more explainable and justified. As such, there is not a single study in VA that applies transparency models and we only present 5% of the studies that report on transparency. There is great need to introduce transparency in VA models. Figure 10 below shows the studies that discuss transparency models even though they were not specifically applied.

C. *Data balancing*

VA data generally entails data imbalance, where there is more bias in terms of the majority class. Approaches such as oversampling and undersampling have been applied to address this imbalance. However, few notable studies apply such approaches. Figure 11 shows only 6% of the primary studies applied data balancing.

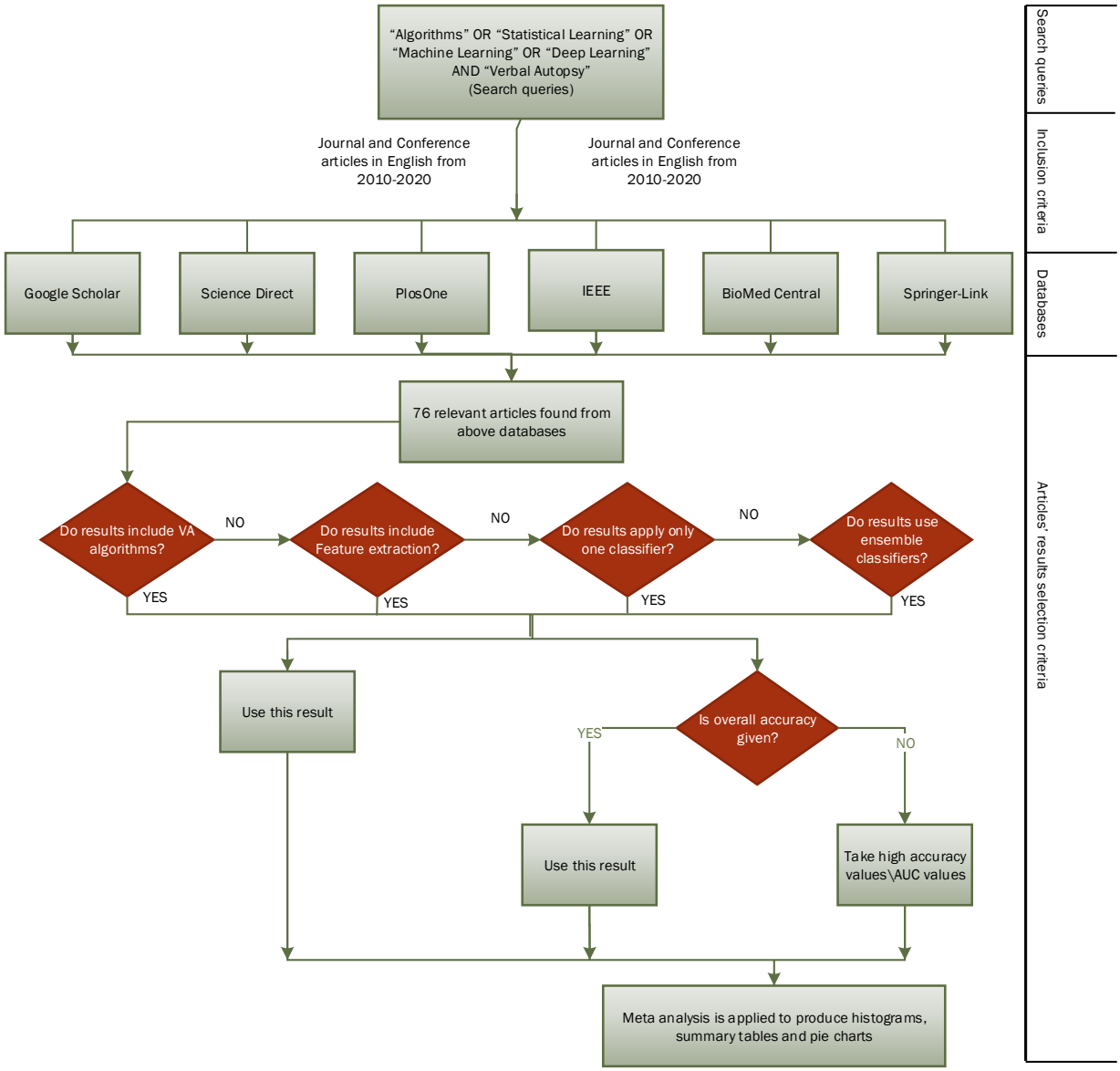


Fig. 8. Flowchart that details our methodological systematic analysis.

D. Feature extraction

Table II and Figure 12 details feature extraction techniques from primary studies. Our analysis shows that out of 38 studies, most of the VA models use expert driven feature selection, followed by the bag of words and n-gram models respectively.

E. Feature value representation schemes

Table III and Figure 13 shows the distribution of feature value representation schemes. This investigation suggests that out of the 11 studies analysed, term frequency value representation and binary representation are the most widely used approaches.

TABLE I
EXISTING LITERATURE SURVEYS ON VA AND THEIR DIFFERENCES FROM THIS SURVEY PAPER: LEGEND:FEATURE ENGINEERING (FE), FEATURE SELECTION (FS).

Comparative analysis of existing literature papers				
Survey paper	Articles searched	Years	Objectives	Differences from this study
[28]	Unspecified	1997-2009	Model performance evaluation	Deep learning models not considered
			Investigation of model types	Data balancing ignored
			Analysis of FE approaches	Not all ML approaches are discussed
				Model transparency not covered
[25]	1497	2013–2018	Model performance evaluation	Deep learning not covered
			Investigation of model types	
			Analysis of pre-processing approaches	Not all ML approaches are discussed
			Analysis of FE approaches	
			Analysis of performance metrics	
[5]	19	1997–2014	Model performance evaluation	Deep learning not covered
			Investigation of model types	
				Not all ML approaches are discussed
			Analysis of FE approaches	
			Analysis of performance metrics	
				Model transparency not covered
[8]	392	1978–2008	Model performance evaluation	Deep learning not covered
			Comparative analysis of VA methods and concepts	Nature of datasets not reported
				Not all ML approaches are discussed
				Model transparency not covered

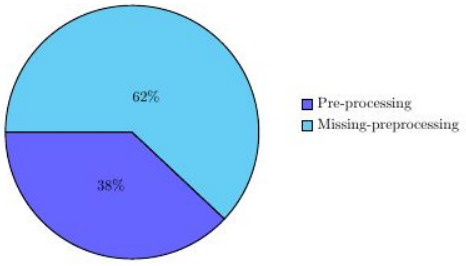


Fig. 9. Distribution of VA primary studies that include pre-processing.

F. Feature selection

Table IV and Figure 14 shows the distribution of feature selection techniques applied in primary studies. Out of 18 studies, NLP is the most common approach possibly depicting deep learning as an emerging solution, Pearson Correlation and Information Gain coming after.

G. Comparative analysis of studies reviewed in this paper

Figure 15 shows the VA study distribution compared to other fields. The distribution shows that we used 66% of the primary studies which are in the VA domain and the remainder were from other fields. Figure 16 below depicts the journal articles and conference papers used in this methodological systematic analysis. From the primary studies synthesised, we used 66% from reputable and peer reviewed journal articles and 34% from conference proceedings.

H. Evaluation metrics

The section below discusses various performance evaluation metrics used in this systematic methodological systematic

TABLE II
FEATURE EXTRACTION METHODS USED IN PRIMARY STUDIES. LEGEND: BAG OF WORDS (BOW), N-GRAMS (NG), HYBRID (HB), WORD2VEC (WV), CONCEPT-BASED (CB), GRAPH-BASED (GB), RULE BASE (RB), DOC2VEC (DV), EXPERT DRIVEN (ED), SENTENCE2VEC (SV), GLOVE (GV).

	Feature extraction methods from primary studies										
Study	BoW	HB	NG	WV	CB	GB	RB	DV	SV	GV	ED
[3]	✓										
[4]											✓
[5]											✓
[56]											✓
[6]											✓
[18]		✓	✓								
[28]	✓										
[19]		✓	✓								
[31]											✓
[21]											✓
[22]	✓	✓	✓	✓	✓	✓	✓	✓			
[7]											✓
[29]											✓
[8]											✓
[57]											✓
[58]											✓
[59]											✓
[60]											✓
[61]											✓
[62]											✓
[27]							✓				
[25]	✓	✓	✓	✓	✓	✓	✓				
[18]	✓								✓		
[30]											✓
[32]											✓
[34]											✓
[35]											✓
[42]											✓
[45]	✓										
[46]				✓							
[20]				✓						✓	
[48]			✓		✓						
[54]					✓						
[63]	✓										
[64]	✓										
[69]											✓
[79]											✓
[82]											✓

analysis. Table V and Figure 17 show the distribution of performance evaluation approaches applied in this methodological systematic analysis. Out of the thirty four (34) studies that were reported in literature to have applied evaluation metrics, the most used evaluation metrics are Precision and Recall followed by CSFM accuracy, CCC, Specificity and F- score.

I. Tag cloud of common and frequent terms of reviewed studies

Figure 18 is a visual model that illustrates the highest frequency terms appearing in the title and abstract of the peer reviewed literature, where higher occurring terms appear in larger font size. From our cloud tag we can deduce that ML task of classification appears more, also we can can report that verbal autopsy is common within the health field through text or data that forms part of our narratives.

J. Models

The section below discusses various models in this methodological systematic analysis. Table VI and Figure 19 show the distribution of statistical and machine learning models applied in this study. In the table we have the following abbreviations; VA algorithms (VA), Logistic Regression (LR), Random Forests (RF), Decision Tree (DT),

TABLE III
FEATURE VALUE REPRESENTATION SCHEMES FROM PRIMARY STUDIES. LEGEND: BINARY REPRESENTATION (BR), TERM FREQUENCY (TF), DOCUMENT FREQUENCY (DF), TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF), NORMALISED TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (NTF-IDF).

Feature value representation from primary studies				
Study	BR	TF	TF-IDF	NTF-IDF
[3]		✓		
[18]	✓	✓	✓	✓
[28]	✓	✓	✓	✓
[22]			✓	
[25]	✓	✓	✓	✓
[45]		✓		
[48]	✓	✓		
[29]		✓		
[54]	✓	✓		
[63]				✓
[82]		✓		

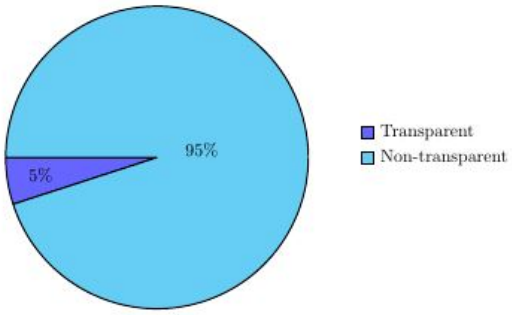


Fig. 10. Pictorial distribution of VA primary studies that include transparency.

TABLE IV
FEATURE SELECTION TECHNIQUES FROM PRIMARY STUDIES. LEGEND: PEARSON CORRELATION (PC), CHI-SQUARED (CS), INFORMATION GAIN (IG), PRINCIPAL COMPONENT ANALYSIS (PCA),FISHER MARKOV SELECTOR (FMS), IMPROVED GLOBAL FEATURE SELECTION (IGFS), LOCAL SEMI SUPERVISED (LSS), NATURAL LANGUAGE PROCESSING (NLP)

Feature selection techniques from primary studies								
Study	PC	CS	IG	PCA	FMS	IGFS	LSS	NLP
[3]	✓							
[28]							✓	
[57]	✓							
[62]		✓						
[19]	✓	✓	✓					
[62]			✓					
[29]	✓	✓	✓		✓	✓		
[25]	✓	✓	✓	✓				
[18]	✓	✓	✓		✓	✓		
[45]			✓					
[46]								✓
[20]								✓
[54]								✓
[63]								✓
[64]								✓
[82]								✓
[85]								✓
[40]		✓						

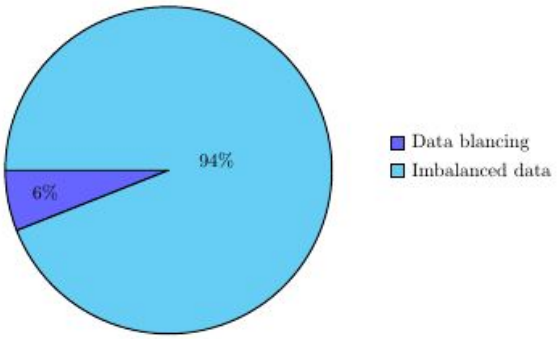


Fig. 11. Pictorial distribution of VA related studies that include data balancing.

K- nearest neighbor (k-nn), Support Vector Machine (SVM), Artificial Neural Network (ANN), Naive Bayes Classifier (NBC), K-means clustering (k-means), Rule-based (RB), Ensemble Classifier (EC), Convolutional neural network (CNN), Gated Recurrent neural network (GRN). This methodological systematic analysis suggests that out of 36 studies, conventional VA algorithms are the most widely used technique, followed by Naive Bayes Classifier, RF and SVM and k-nn.

TABLE V
EVALUATION METRICS TECHNIQUES APPLIED IN PRIMARY STUDIES. LEGEND: SPECIFICITY (SP), PRECISION (PR), RECALL (RE), F-SCORE (FS), ACCURACY (AC), CSFM (CS), CCC, PCCC, KAPPA (KA), CASE SIMILARITY MEASURE (CSM), TOP CAUSE (TC), TOP3COD (T3), AREA UNDER CURVE (AUC)

	Evaluation metrics of primary studies												
Study	Sp	Pr	Re	Ac	CSFM	CCC	PCCC	Ka	CSM	Fs	TC	T3	AUC
[3]		✓	✓		✓		✓						
[4]	✓	✓	✓		✓								
[1]					✓								
[5]	✓	✓			✓	✓							
[6]		✓											
[18]		✓	✓										
[28]		✓	✓							✓			
[19]		✓	✓	✓						✓			
[31]					✓	✓							
[21]	✓	✓			✓		✓						
[22]		✓	✓	✓						✓			
[7]					✓						✓		
[8]	✓	✓	✓		✓			✓					
[57]					✓	✓					✓	✓	
[58]					✓	✓							
[59]					✓	✓							
[60]		✓			✓		✓						
[61]					✓	✓		✓					
[62]	✓			✓		✓							
[25]		✓	✓	✓					✓			✓	
[29]		✓	✓	✓						✓			✓
[30]	✓	✓	✓		✓	✓		✓					
[32]	✓		✓	✓		✓							
[35]						✓		✓					
[42]			✓				✓		✓				
[45]				✓									
[46]		✓	✓							✓			
[20]		✓	✓		✓					✓			
[48]		✓	✓							✓			
[54]			✓	✓							✓		
[63]		✓	✓							✓			
[69]			✓										
[79]	✓	✓	✓		✓	✓		✓					
[85]		✓	✓										
[40]													✓

K. Machine learning framework guidelines for VA

Our proposed guiding framework for VA researchers and practitioners is based on the results of this methodological systematic analysis. First and foremost, due to the nature of VA datasets, which may be balanced or imbalanced, we recommend performing exploratory data analysis and data preprocessing as the initial steps in data cleaning. After successful data cleaning, we suggest conducting well-structured and logical feature selection and engineering to identify the most discriminative and representative features for our vector space. To address class imbalance in the VA domain, and depending on the dataset's type and characteristics, we propose applying techniques such as under-sampling, over-sampling, resampling or hybrid methods to balance the class distribution. Moreover, it is crucial to ensure that the applied techniques do not compromise the integrity of the data, particularly in cases of rare causes of death. It is also imperative to evaluate different data balancing methods and their impact on model performance. This evaluation allows for the selection of the optimal approach while maintaining the clinical relevance of predictions. Notably, caution should be exercised to avoid eliminating or overly simplifying minority classes, as this

TABLE VI
MODELS APPLIED IN PRIMARY STUDIES.

	Models applied in of primary studies												
Study	VA	LR	RF	DT	k-nn	SVM	ANN	NBC	k-means	RB	EC	CNN	GRN
[3]		✓	✓	✓		✓	✓	✓					
[1]	✓												
[5]	✓		✓										
[56]	✓												
[6]						✓					✓		
[18]			✓			✓		✓					
[28]			✓			✓		✓					
[19]			✓	✓	✓	✓		✓		✓			
[31]	✓	✓	✓										
[21]	✓							✓					
[22]			✓	✓	✓	✓		✓			✓		
[7]	✓												
[8]	✓							✓					
[57]	✓												
[58]	✓												
[59]	✓												
[60]	✓		✓										
[27]					✓			✓		✓			
[61]	✓												
[62]	✓												
[25]		✓	✓	✓	✓	✓	✓	✓				✓	
[29]			✓	✓	✓	✓		✓					
[30]	✓												
[32]		✓					✓						
[34]		✓											
[35]	✓												
[42]								✓			✓		
[45]		✓		✓		✓		✓					
[46]									✓			✓	✓
[20]													✓
[48]					✓				✓	✓			
[54]					✓				✓	✓			
[63]						✓							
[69]	✓												
[79]	✓												
[82]	✓												
[40]		✓											

could result in the loss of critical information related to less common causes of death during class balancing.

The next vital step involves establishing baseline models. Start with simple models such as logistic regression and decision trees as benchmarks to evaluate the effectiveness of more complex models. This approach helps in understanding the basic patterns in the data before moving on to advanced techniques. Additionally, compare these baseline models against more sophisticated machine learning models, such as Random Forests, Gradient Boosting Machines, SVM and Neural Networks. During this process, ensure that each model is properly tuned and cross-validated to prevent overfitting and provide a fair comparison. It is also important to compare ensemble model performance with individual models to determine the best approach for VA data. Periodically re-evaluate the benchmark models with updated datasets to ensure they remain relevant and effective as new data becomes available. However, this systematic methodological analysis shows that SVM and ensemble classifiers generally perform better than other models. After this process, we recommend

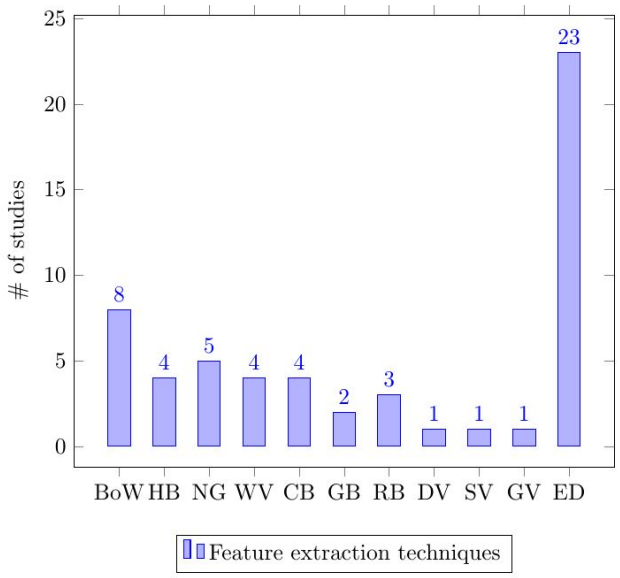


Fig. 12. Distribution of feature extraction techniques reviewed.

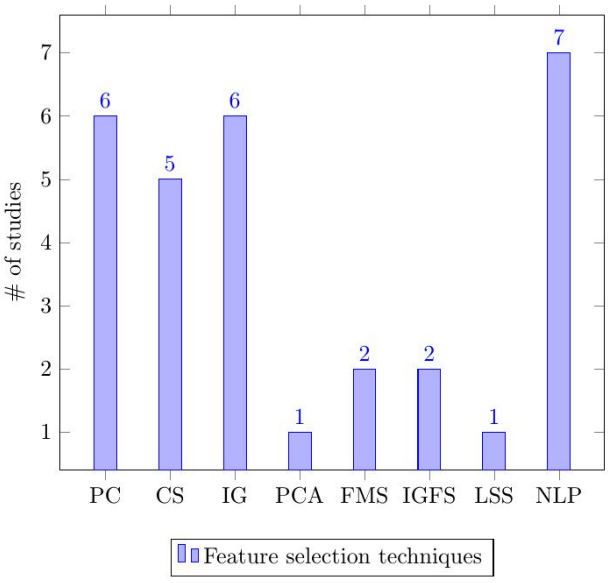


Fig. 14. Distribution of feature selection techniques.

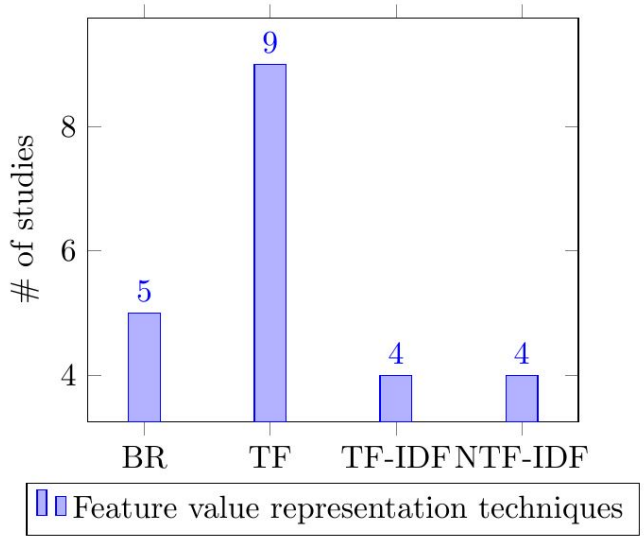


Fig. 13. Distribution of feature value representation schemes.

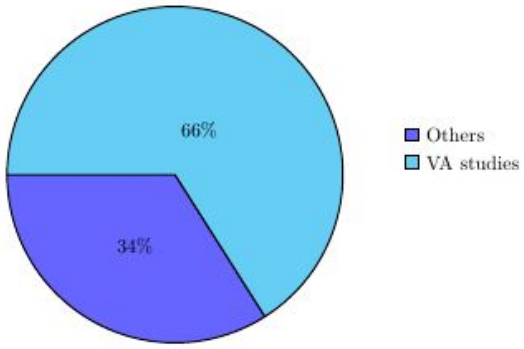


Fig. 15. Distribution of VA related studies in comparison to other fields used in this methodological systematic analysis.

applying evaluation metrics to assess the performance of the approaches and choose the optimal one for classification.

This analysis also showed that the most common feature extraction techniques are bag of words and n-grams, while the most frequently used feature value representation schemes are Binary Representation (BR) and Term Frequency (TF). Furthermore, our investigation revealed that the most common feature selection or dimensionality reduction techniques include expert-driven feature selection by physicians, information gain, and Pearson correlation.

Moreover, the few studies that have used the emerging trend of advanced machine learning known as deep learning have proved that distributed vector representations (Word2vec, Sentence2vec, Doc2vec and Glove) show promising results. CNN improves VA performance through use of character embeddings. New models exist in deep learning such as the Recursive Neural Network and Transformer, which have shown promising results in other domains and can prove useful to explore with, in our VA field. Performance evaluation metrics such as Precision, Recall, Accuracy, AUC and F-score can be applied to evaluate models.

Below Figure 20 is the proposed visual representation of the guiding framework.

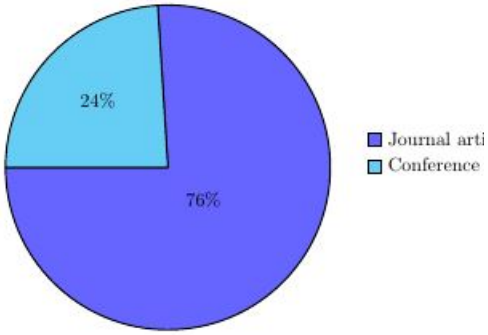


Fig. 16. Distribution of journal articles and conference articles used in this methodological systematic analysis.

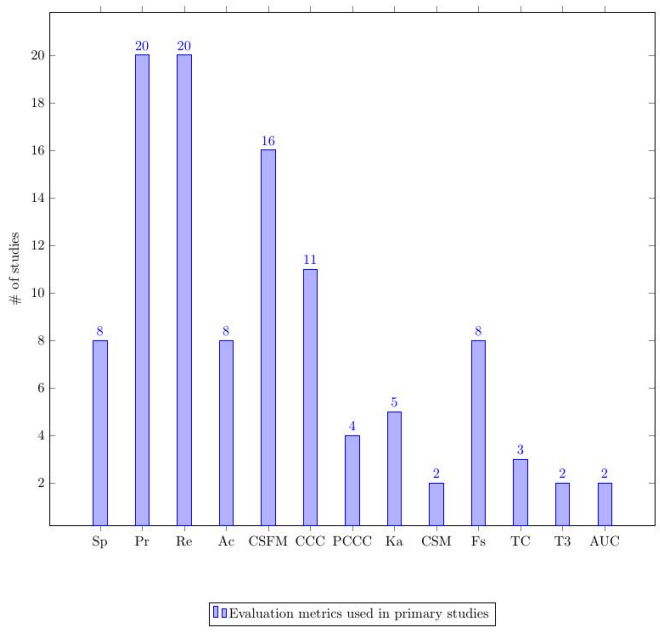


Fig. 17. Distribution of evaluation metrics from primary studies reviewed.

IV. Discussion

Our study conducted a comprehensive methodological analysis of publications that are in the VA field in terms of statistical and machine learning application to determine cause of death. Subsequently, several main sub fields related to ML in VA were presented. We further gave an insight on the application of statistical, ML and advanced deep learning approaches in VA, which has not been done at this stage. Therefore from this detailed methodological systematic analysis, one can be able to identify various statistical and ML applications in VA and thus also identify opportunities

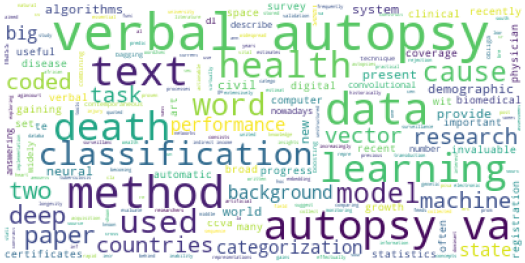


Fig. 18. Tag cloud of this methodological systematic analysis

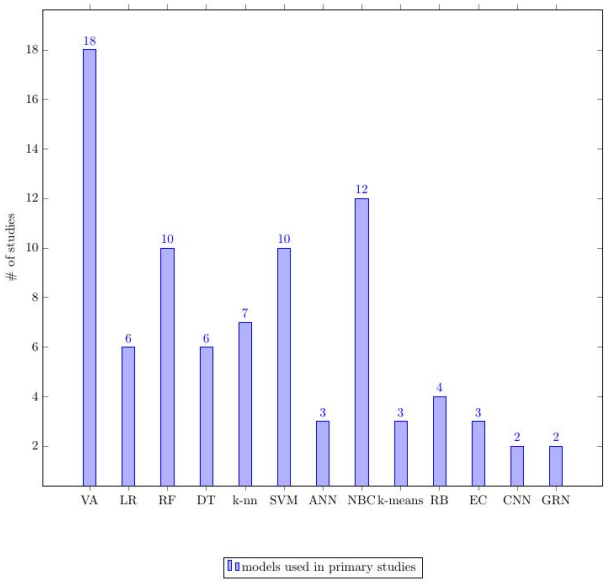


Fig. 19. Distribution of statistical and ML models used in primary studies.

and challenges in the VA domain. Most of the studies carried out in VA have been using mostly supervised learning techniques with a few reports on unsupervised learning approaches. Moreover, there is little or few studies that apply deep learning as a ML approach in VA. Deep learning model application can improve results within the field of VA, such as better knowledge representation systems for VA, thus enforcing standardisation and a shared data repository for further research.

Generally all ML studies in VA have been trained on sparse high dimensional data and high training cost of samples, thus applying alternative approaches can effectively handle increased training data samples and improve efficiency. There is also a limitation in terms of having publicly available training datasets, which can be used by researchers to train and improve models. This should be one key area to be looked into within the VA community.

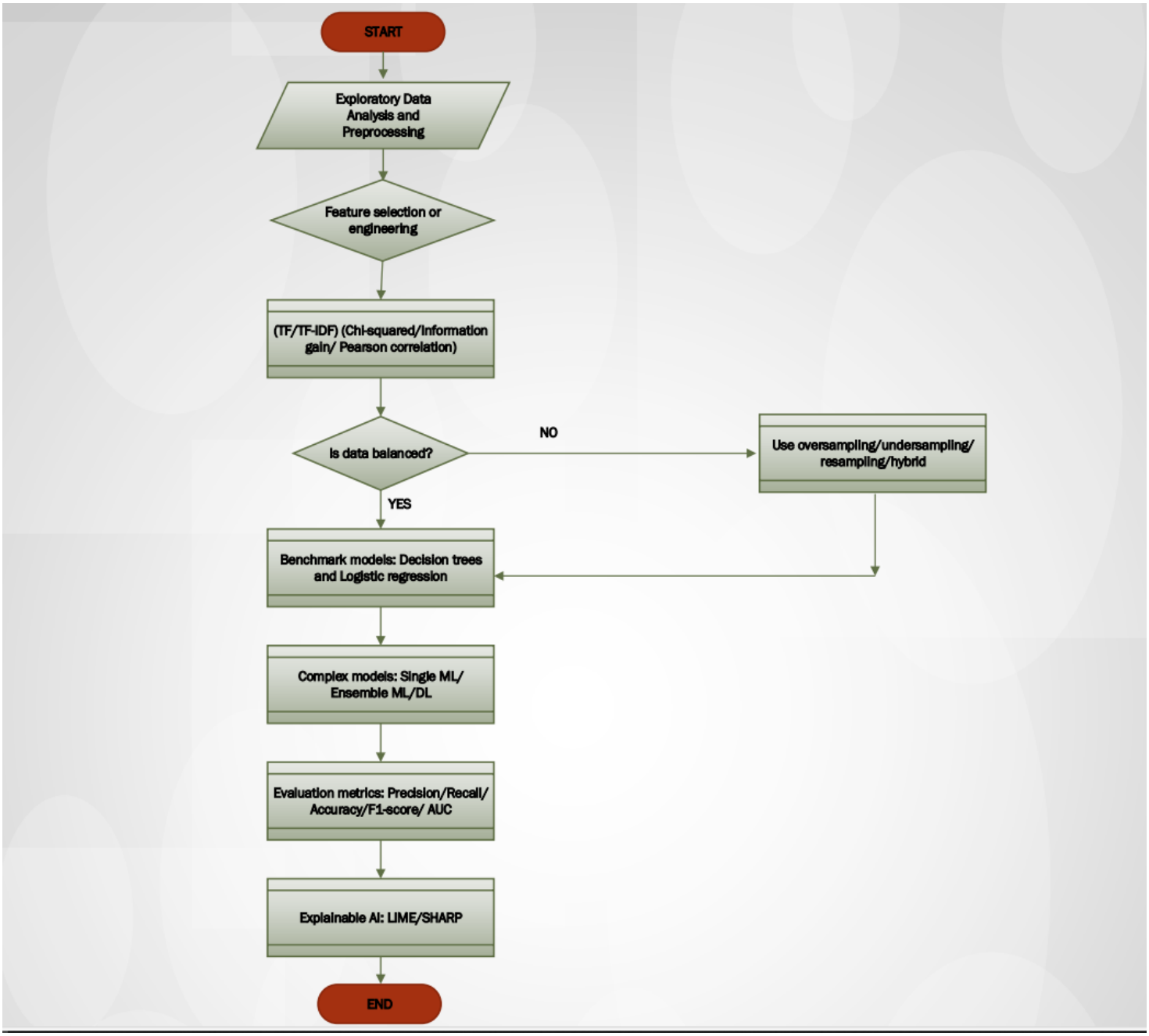


Fig. 20. Proposed guidelines for the VA machine learning framework

Future VA studies should incorporate and utilise well structured and logical feature selection and feature engineering techniques such as filter, wrapper and embedded methods, as they have shown promising and excellent results in other fields. These methods make it possible to attain relevant features that are discriminative, thereby addressing issues of noise and high hyper parameter space amongst many drawbacks [63].

Literature to date has presented various studies that have applied model evaluation in terms of performance for various statistical and ML frameworks in VA. Yet little evidence exists that shows a consensus on optimal best performing algorithm. Our methodological systematic analysis investigated commonly applied statistical and ML approaches including deep learning approaches and we found out that the SVM as the most optimum classifier and we deduced that hybrid models are better than single models classifiers. However, there is less application of deep learning models in VA, even though the application of the CNN from literature showed improvements to classification accuracy as well as better knowledge representation using character embeddings in the VA domain. This was better in comparison to conventional statistical and classical machine learning

models.

Our study further flags limitations such as lack of standardised VA questionnaires, scarcity and cost of experts to perform VAs, various medical ontologies resulting in heterogeneous datasets, relying on expert's diagnosis as the gold standard, scarce availability of publicly available biomedical datasets, various ways of acquiring data, lack of effective preprocessing procedures, few data instances, high and complex feature space that id highly dimensional amongst many. The proposed guiding framework for VA practitioners and researchers in this study suggests performing robust and logical exploratory data analysis and pre-processing for data cleaning, applying feature selection and feature engineering in the most effective way, performing data balancing, employing benchmark models and applying evaluation metrics for model performance. This survey pointed to emerging directions in VA such as the applications of deep learning models in VA and explainable AI.

Prospectively the VA community should come up with strategies of effective data balancing of datasets. They should seek to apply emerging techniques apart from common over-sampling and under-sampling. Newer novice approaches like linear dependence approach and wavelet data transformation should be employed in VA. Upcoming work should consider doing more effective exploratory data analysis before model development. This gives the researcher the opportunity to fully understand the feature space in terms of complexity, feature distribution, association and correlation. As a result of this one can be able to identify key important and relevant features that are the best predictors of our outcome variable. In other words we can esily narrow down our feature space without losing any key information and this improves model performance. Future research should introduce and apply standard VA questionnaires which are key input to the VA process. Moreover, there should be standardisation of medical ontologies and incorporation of homogeneous datasets. Data acquisition and pre-processing techniques should be standardised and be performed in a robust and logical manner.

There is scanty evidence from past studies in employing effective exploratory data analysis. There is great need to incorporate such in order to to better understand distributions of features and address model complexity to allow efficiency in model development in VAs. As a result of exploratory analysis, one can easily acquire knowledge prior to model development such as; identifying the correlation between the predictors and outcome feature and predictors that are more important in predicting the target feature. Apart from this it helps in the reduction of the feature space, hence addressing complexity and improve on performance of algorithms. Furthermore, there is need to make available public datasets in order to train models for improved performance. If resources permit, there should be a team of experts that specifically deal with VAs.

The future of the VA community should focus on transparency, to justify and make non-transparent models explainable. One example of a transparent model is Local Interpretable Model-Agnostic Explanations (LIME). LIME can explain any classifier prediction through learning an interpretable model around a prediction. Transparency is key in that it builds trust in humans in terms of predictions, such that if there is need for interventions or any form of action then it can be taken. Transparency gives the leverage for an expert to accept or reject a model.

Most VA studies have employed supervised machine learning approaches. There is need for the VA research community to consider other forms of machine learning such as; unsupervised learning (which do not need a labelled dataset, hence performs classification or clustering based on unlabelled points) , semi-supervised machine learning (makes use of few labelled data points to make predictions), transfer learning (relevant where the input data set has few labelled features to effectively train a model) and reinforcement learning (uses environment interaction to learn through past exposures to accurately classify VA narratives using a a reward system).
To get better improved results going forward it is vital for all researchers and stakeholders that deal with VA studies to collaborate effectively and build strategic partnership that seek to address the confines within the VA domain with particular emphasis on tools, approaches, data sources among many and put in place effective strategies that can address the problems at hand.

V. Conclusion

In this study, we reviewed in detail traditional statistical and ML approaches to determine cause of deaths from VAs, identified limitations of each of these techniques, proposed machine learning framework guidelines for determining causes of death from VAs and pointed to emerging future directions. Most VA models use the PCVA approach as a gold standard, even though it is expensive, inconsistent and inaccurate (subjective and prone to errors because of the VA narratives that are taken as input). In practice, as an alternative to PCVA, mostly statistical approaches such as LR and NB are employed because of simplicity and transparency. Nevertheless, in literature complex machine learning and emerging deep learning frameworks exist, that can be applied to improve cause of death classification. Most of the machine learning models reviewed apply the RF and SVM. However, incorporating ensemble classifiers can yield better results. To achieve optimum predictions in determining cause of deaths from VAs, models need to employ effective data exploration (fully understand datasets before modelling), data balancing, pre-processing (improve knowledge representation of medical datasets) and robust emerging feature engineering techniques (extract relevant features and narrow down the feature space, thus improving model performance). Additionally, other types of machine learning like transfer learning and reinforcement learning need to be explored further as they have shown

TABLE VII
LIST OF ABBREVIATIONS USED IN THIS STUDY

List of abbreviations	
Abbreviation	Full term
AI	Artificial intelligence
ANN	Artificial neural network
AUC	Area under curve
BoW	Bag of words
BR	Binary representation
CBR	Case-based reasoning
CCC	Chance corrected concordance
CCVA	Computer coded verbal autopsy
CNN	Convolutional neural network
CSFM	Cause specific mortality fractions
DT	Decision tree
FE	Feature engineering
FMS	Fisher markov selector
FN	False negative
FP	False positive
FS	Feature selection
GRNN	Gated recurrent neural network
GoW	Graph of words
ICD-10	International classification of diseases, tenth revision codes
IG	Information gain
IGFS	Improved global feature selection
k-nn	k-nearest neighbour
LIME	Local interpretable model-agnostic explanations
LSFS	Local semi-supervised feature selection
LR	Logistic regression
LSTM	Long term short memory
ML	Machine learning
NB	Naive bayes
NLP	Natural language processing
NTF-IDF	Normalised term frequency-inverse document frequency
PCA	Principal component analysis
PC	Pearson correlation
PCVA	Physician certified verbal autopsy
RF	Random forest
RNN	Recurrent neural network
SMOTE	Synthetic Minority over-sampling technique
SVM	Support vector machine
TF	Term frequency
TF-IDF	Term frequency-inverse document frequency
TN	True negative
TP	True positive
VA	Verbal autopsy

promising results in other domains. Since VAs are key inputs to the civil registration, there is need to build transparent models for cause of death classification.

The significant contribution of VA data in informing interventions, health priorities, monitoring and evaluation programmes, is of importance. Therefore, it is imperative for the VA community to continue exploring and developing new strategies in line with the long term efforts of improving the civil registration systems. Advances in VA methods should seek to achieve consistency, comparability, and generalisation to other studies.

VI. ABBREVIATIONS

The list of abbreviations is given in Table VII below.

VII. DECLARATIONS

A. Ethics approval and consent to participate

This is not applicable to this study.

B. Consent for publication

This is not applicable to this study.

C. Availability of data and materials

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

D. Competing interests

The authors declare that they have no financial or personal relationship(s) that may have inappropriately influenced them in writing this article.

E. Declaration of interest

The authors declare that they do not have any financial and personal relationships that could inappropriately influence this study.

F. Funding

This research was not grant funded.

G. Authors' contributions

All authors listed below have read and approved the manuscript.

MM is the main author and conducted the whole review process including the write up.

EM reviewed and contributed on statistical models.

CK reviewed and contributed on verbal autopsy theory.

TC reviewed and contributed on the machine learning concepts.

H. Corresponding authors

Michael T. Mapundu

University of the Witwatersrand

School of Public Health

Department of Biostatistics and Epidemiology

7 York Road, Parktown, Johannesburg, South Africa

Tel: +27 11 71-72627

Email: michael.mapundu@wits.ac.za

I. Acknowledgements

This work was supported by the Developing Excellence in Leadership, Training and Science (DELTAS) Africa Initiative Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) [Grant No.DEL-15-005]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS) Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [Grant No. 107754/Z/15/Z] and the United Kingdom government.

REFERENCES

[1] Erin K Nichols, Peter Byass, Daniel Chandramohan, Samuel J Clark, Abraham D Flaxman, Robert Jakob, Jordana Leitao, Nicolas Maire, Chalapati Rao, Ian Riley, et al. The who 2016 verbal autopsy instrument: An international standard suitable for automated analysis by interva, insilicova, and tariff 2.0. *PLoS medicine*, 15(1):e1002486, 2018.

[2] Lisa-Marie Thomas, Lucia D'Ambruoso, and Dina Balabanova. Verbal autopsy in health policy and systems: a literature review. *BMJ global health*, 3(2): e000639, 2018.

[3] Serena Jeblee, Mireille Gomes, Prabhat Jha, Frank Rudzicz, and Graeme Hirst. Automatically determining cause of death from verbal autopsy narratives. *BMC medical informatics and decision making*, 19(1):127, 2019.

[4] Nadia Soleman, Daniel Chandramohan, and Kenji Shibuya. Verbal autopsy: current practices and challenges. *Bulletin of the World Health Organization*, 84:239–245, 2006.

[5] Jordana Leitao, Nikita Desai, Lukasz Aleksandrowicz, Peter Byass, Pierre Miasnikof, Stephen Tollman, Dewan Alam, Ying Lu, Suresh Kumar Rathi, Abhishek Singh, et al. Comparison of physician-certified verbal autopsy with computer-coded verbal autopsy for cause of death assignment in hospitalized patients in low-and middle-income countries: systematic review. *BMC medicine*, 12(1):22, 2014.

[6] Sean T Green and Abraham D Flaxman. Machine learning methods for verbal autopsy in developing countries. In *AAAI Spring Symposium: Artificial Intelligence for Development*, 2010.

[7] Tyler H McCormick, Zehang Richard Li, Clara Calvert, Amelia C Crampin, Kathleen Kahn, and Samuel J Clark. Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association*, 111(515):1036–1049, 2016.

[8] Edward Fottrell and Peter Byass. Verbal autopsy: methods in transition. *Epidemiologic reviews*, 32(1): 38–55, 2010.

[9] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*, 5: 8869–8879, 2017.

[10] Sharathkumar Anbu and Bhaskarjit Sarmah. Machine learning approach for predicting womens health risk. In *Advanced Computing and Communication Systems (ICACCS), 2017 4th International Conference on*, pages 1–4. IEEE, 2017.

[11] Victor Mudenda, Sebastian Lucas, Aaron Shibemba, Justin O'grady, Matthew Bates, Nathan Kapata, Samana Schwank, Peter Mwaba, Rifat Atun, Michael Hoelscher, et al. Tuberculosis and tuberculosis/hiv/aids–associated mortality in africa: The urgent need to expand and invest in routine and research autopsies. *Journal of Infectious Diseases*, 205(suppl_2):S340–S346, 2012.

[12] B Nithya and V Ilango. Predictive analytics in health care using machine learning tools and techniques. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 492–499. IEEE, 2017.

[13] Zahid Iqbal, Rafia Ilyas, Waseem Shahzad, and Irum Inayat. A comparative study of machine learning techniques used in non-clinical systems for continuous healthcare of independent livings. In *2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*. IEEE, 2018.

[14] Anton Kocheturov, Panos M Pardalos, and Athanasia Karakitsiou. Massive datasets and machine learning for computational biomedicine: trends and challenges. *Annals of Operations Research*, pages 1–30, 2018.

[15] Lei Liu. Research on logistic regression algorithm of breast cancer diagnose data by machine learning. In *2018 International Conference on Robots & Intelligent System (ICRIS)*. IEEE, 2018.

[16] Ivo D Dinov, Ben Heavner, Ming Tang, Gustavo Glusman, Kyle Chard, Mike Darcy, Ravi Madduri, Judy Pa, Cathie Spino, Carl Kesselman, et al. Predictive big data analytics: a study of parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. *PloS one*, 11 (8):e0157077, 2016.

[17] Abraham D Flaxman and Theo Vos. Machine learning in population health: Opportunities and threats. *PLoS medicine*, 15(11), 2018.

[18] Ghulam Mujtaba, Liyana Shuib, Ram Gopal Raj, Retnagowri Rajandram, and Khairunisa Shaikh. Automatic text classification of icd-10 related cod from complex and free text forensic autopsy reports. In *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*, pages 1055–1058. IEEE, 2016.

[19] Ghulam Mujtaba, Liyana Shuib, Ram Gopal Raj, Retnagowri Rajandram, and Khairunisa Shaikh. Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study. *Journal of forensic and legal medicine*, 57:41–50, 2018.

[20] Zhaodong Yan, Serena Jeblee, and Graeme Hirst. Can character embeddings improve cause-of-death classification for verbal autopsy narratives? In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 234–239, 2019.

[21] Pierre Miasnikof, Vasily Giannakeas, Mireille Gomes, Lukasz Aleksandrowicz, Alexander Y Shestopaloff, Dewan Alam, Stephen Tollman, Akram Samarikhalaj, and Prabhat Jha. Naive bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths. *BMC medicine*, 13 (1):286, 2015.

[22] Ghulam Mujtaba, Liyana Shuib, Ram Gopal Raj, Retnagowri Rajandram, Khairunisa Shaikh, and Mohammed Ali Al-Garadi. Classification of forensic autopsy reports through conceptual graph-based document representation model. *Journal of biomedical informatics*, 82:88–105, 2018.

[23] Suresh Kumar and Shivani Goel. Enhancing text classification by stochastic optimization method and support vector machine. *International Journal of Computer Science and Information Technologies*, 6(4), 2015.

[24] Vandana Korde and C Namrata Mahender. Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications*, 3(2): 85, 2012.

[25] Ghulam Mujtaba, Liyana Shuib, Norisma Idris, Wai Lam Hoo, Ram Gopal Raj, Kamran Khowaja, Khairunisa Shaikh, and Henry Friday Nweke. Clinical text classification research trends: systematic literature review and open issues. *Expert Systems with Applications*, 116:494–520, 2019.

[26] Vianney Jouhet, Georges Defossez, Anita Burgun, Pierre Le Beux, P Levillain, Pierre Ingrand, Vincent Claveau, et al. Automated classification of free-text pathology reports for registration of incident cases of cancer. *Methods of information in medicine*, 51(3):242, 2012.

[27] Wei Liang Yeow, Rohana Mahmud, and Ram Gopal Raj. An application of case-based reasoning with machine learning for forensic autopsy. *Expert Systems with Applications*, 41(7):3497–3505, 2014.

[28] Samuel Danso, Eric Atwell, and Owen Johnson. A comparative study of machine learning methods for verbal autopsy text classification. *arXiv preprint arXiv:1402.4380*, 2014.

[29] Ghulam Mujtaba, Liyana Shuib, Ram Gopal Raj, Retnagowri Rajandram, Khairunisa Shaikh, and Mohammed Ali Al-Garadi. Automatic icd-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. *PloS one*, 12(2):e0170242, 2017.

[30] Christopher JL Murray, Rafael Lozano, Abraham D Flaxman, Peter Serina, David Phillips, Andrea Stewart, Spencer L James, Alireza Vahdatpour, Charles Atkinson, Michael K Freeman, et al. Using verbal autopsy to measure causes of death: the comparative performance of existing methods. *BMC medicine*, 12 (1):5, 2014.

[31] Abraham D Flaxman, Alireza Vahdatpour, Sean Green, Spencer L James, and Christopher JL Murray. Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Population health metrics*, 9(1):29, 2011.

[32] Andrew Boulle, Daniel Chandramohan, and Peter Weller. A case study of using artificial neural networks for classifying cause of death from verbal autopsy. *International journal of epidemiology*, 30(3):515–520, 2001.

[33] Maria A Quigley, Daniel Chandramohan, and Laura C Rodrigues. Diagnostic accuracy of physician review, expert algorithms and data-derived algorithms in adult verbal autopsies. *International Journal of*

Epidemiology, 28(6):1081–1087, 1999.

[34] Mathew A Mwanyangala, Honorathy M Urassa, Jensen C Rutashobya, Chrisostom C Mahutanga, Angelina M Lutambi, Deodatus V Maliti, Honorati M Masanja, Salim K Abdulla, and Rose N Lema. Verbal autopsy completion rate and factors associated with undetermined cause of death in a rural resource-poor setting of tanzania. *Population health metrics*, 9(1):41, 2011.

[35] Noriah Maraba, Aaron S Karat, Kerrigan McCarthy, Gavin J Churchyard, Salome Charalambous, Kathleen Kahn, Alison D Grant, and Violet Chihota. Verbal autopsy-assigned causes of death among adults being investigated for tb in south africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 110 (9):510–516, 2016.

[36] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[37] Barnaby C Reeves and Maria Quigley. A review of data-derived methods for assigning causes of death from verbal autopsy data. *International journal of epidemiology*, 26(5):1080–1089, 1997.

[38] Juan D Rodriguez, Aritz Perez, and Jose A Lozano. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, 32(3): 569–575, 2010.

[39] Khaled Fawagreh, Mohamed Medhat Gaber, and Eyad Elyan. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering: An Open Access Journal*, 2(1):602–609, 2014.

[40] Nuntaporn Klinjun, Apiradee Lim, and Kanitta Bundhamcharoen. A logistic regression model for estimating transport accident deaths using verbal autopsy data. *Asia Pacific Journal of Public Health*, 27 (3):286–292, 2015.

[41] Kunal Rajput, Girija Chetty, and Rachel Davey. Phis (protected health information) identification from free text clinical records based on machine learning. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–9. IEEE, 2017.

[42] Syed Shariyar Murtaza, Patrycja Kolpak, Ayse Bener, and Prabhat Jha. Automated verbal autopsy classification: using one-against-all ensemble method and naïve bayes classifier. *Gates open research*, 2, 2018.

[43] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2017.

[44] William H Crown. Potential application of machine learning in health outcomes research and some statistical cautions. *Value in health*, 18(2):137–140, 2015.

[45] John Pestian, Henry Nasrallah, Pawel Matykiewicz, Aurora Bennett, and Antoon Leenaars. Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights*, 3: BII–S4706, 2010.

[46] Serena Jeblee, Mireille Gomes, and Graeme Hirst. Multi-task learning for interpretable cause of death classification using key phrase prediction. In *Proceedings of the BioNLP 2018 workshop*, pages 12–17, 2018.

[47] Ekkarat Boonchieng and Khanita Duangchaemkarn. Digital disease detection: Application of machine learning in community health informatics. In *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–5. IEEE, 2016.

[48] Bevan Koopman, Guido Zuccon, Anthony Nguyen, Anton Bergheim, and Narelle Grayson. Extracting cancer mortality statistics from death certificates: A hybrid machine learning and rule-based approach for common and rare cancers. *Artificial intelligence in medicine*, 2018.

[49] Indrajani Sutedja, Hendro Nindito, Azani Cempaka Sari, et al. Building a mobile health application and heart diagnose to assist patients: Analysis and design. In *Information Management and Technology (ICIMTech), 2017 International Conference on*, pages 19–24. IEEE, 2017.

[50] Khaled Alsabti, Sanjay Ranka, and Vineet Singh. An efficient k-means clustering algorithm. 1997.

[51] Frederick Livingston. Implementation of breiman's random forest machine learning algorithm. *ECE591Q Machine Learning Journal Paper*, 2005.

[52] Jiafan Peng, Shunong Zhang, Dongmu Peng, and Kan Liang. Application of machine learning method in bridge health monitoring. In *Reliability Systems Engineering (ICRSE), 2017 Second International Conference on*, pages 1–7. IEEE, 2017.

[53] Matthew D Byrne. Machine learning in health care. *Journal of PeriAnesthesia Nursing*, 32(5):494–496, 2017.

[54] Bevan Koopman, Sarvnaz Karimi, Anthony Nguyen, Rhydwyn McGuire, David Muscatello, Madonna Kemp, Donna Truran, Ming Zhang, and Sarah Thackway. Automatic classification of diseases from free-text death certificates for real-time surveillance. *BMC medical informatics and decision making*, 15(1):53, 2015.

[55] Robert E Schapire. Using output codes to boost multiclass learning problems. In *ICML*, volume 97, pages 313–321. Citeseer, 1997.

[56] Samuel J Clark. A guide to comparing the performance of va algorithms. *arXiv preprint arXiv:1802.07807*, 2018.

[57] Samuel J Clark, Zehang Li, and Tyler H McCormick. Quantifying the contributions of training data and algorithm logic to the performance of automated cause-assignment algorithms for verbal autopsy. *arXiv preprint arXiv:1803.07141*, 2018.

[58] Spencer L James, Abraham D Flaxman, and Christopher JL Murray. Performance of the tariff

method: validation of a simple additive algorithm for analysis of verbal autopsies. *Population Health Metrics*, 9(1):31, 2011.

[59] Zehang Richard Li, Tyler H McCormick, and Samuel J Clark. Non-confirming replication of "performance of insilicova for assigning causes of death to verbal autopsies: multisite validation study using clinical diagnostic gold standards," by flaxman et al. *BMC medicine*, 18(1):1–5, 2020.

[60] Nikita Desai, Lukasz Aleksandrowicz, Pierre Miasnikof, Ying Lu, Jordana Leitao, Peter Byass, Stephen Tollman, Paul Mee, Dewan Alam, Suresh Kumar Rathi, et al. Performance of four computer-coded verbal autopsy methods for cause of death assignment compared with physician coding on 24,000 deaths in low-and middle-income countries. *BMC medicine*, 12(1):20, 2014.

[61] Henry D Kalter, Jamie Perin, and Robert E Black. Validating hierarchical verbal autopsy expert algorithms in a large data set with known causes of death. *Journal of global health*, 6(1), 2016.

[62] Daniel Chandramohan, Philip Setel, and Maria Quigley. Effect of misclassification of causes of death in verbal autopsy: can it be adjusted? *International Journal of Epidemiology*, 30(3):509–514, 2001.

[63] Samuel Danso, Eric Atwell, and Owen Johnson. Linguistic and statistically derived features for cause of death prediction from verbal autopsy text. In *Language processing and knowledge in the web*, pages 47–60. Springer, 2013.

[64] SA Danso, Owen Johnson, A Ten Asbroek, S Soromekun, K Edmond, C Hurt, L Hurt, C Zandoh, C Tawiah, J Fenty, et al. A semantically annotated verbal autopsy corpus for automatic analysis of cause of death. *ICAME Journal*, 37, 2013.

[65] Samuel Odei Danso. *Text Analytics to Predict Time and Cause of Death from Verbal Autopsies*. PhD thesis, University of Leeds, 2015.

[66] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *arXiv preprint arXiv:1708.02709*, 2017.

[67] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[68] Carol Friedman, Lyudmila Shagina, Yves Lussier, and George Hripcsak. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, 11(5):392–402, 2004.

[69] Vendhan Gajalakshmi and Richard Peto. Suicide rates in rural tamil nadu, south india: verbal autopsy of 39 000 deaths in 1997–98. *International journal of epidemiology*, 36(1):203–207, 2007.

[70] Cannannore Nidhi Kamath, Syed Saqib Bukhari, and Andreas Dengel. Comparative study between traditional machine learning and deep learning approaches for text classification. In *Proceedings of the ACM Symposium on Document Engineering 2018*, pages 1–11, 2018.

[71] Wenfeng Tian, Jun Li, and Hongguang Li. A method of feature selection based on word2vec in text categorization. In *2018 37th Chinese Control Conference (CCC)*, pages 9452–9455. IEEE, 2018.

[72] Long Ma and Yanqing Zhang. Using word2vec to process big text data. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2895–2897. IEEE, 2015.

[73] Marwa Naili, Anja Habacha Chaibi, and Henda Hajjami Ben Ghezala. Comparative study of word embedding methods in topic segmentation. *Procedia Computer Science*, 112:340–349, 2017.

[74] Eissa M Alshari, Azreen Azman, Shyamala Doraisamy, Norwati Mustapha, and Mustafa Alkeshr. Improvement of sentiment analysis based on clustering of word2vec features. In *2017 28th International Workshop on Database and Expert Systems Applications (DEXA)*, pages 123–126. IEEE, 2017.

[75] Nabil Alami, Mohammed Meknassi, and Noureddine En-nahnahi. Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning. *Expert Systems with Applications*, 2019.

[76] Alexander Kirillov, Dmitrij Schlesinger, Walter Forkel, Anatoly Zelenin, Shuai Zheng, Philip Torr, and Carsten Rother. Efficient likelihood learning of a generic cnn-crf model for semantic segmentation. *arXiv preprint arXiv:1511.05067*, 2015.

[77] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2015.

[78] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[79] Christopher JL Murray, Rafael Lozano, Abraham D Flaxman, Alireza Vahdatpour, and Alan D Lopez. Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. *Population health metrics*, 9(1):28, 2011.

[80] Gaoxia Jiang and Wenjian Wang. Error estimation based on variance analysis of k-fold cross-validation. *Pattern Recognition*, 69:94–106, 2017.

[81] Tzu-Tsung Wong. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9):2839–2846, 2015.

[82] Samuel Danso, Eric Atwell, Owen Johnson, Guus ten Asbroek, Karen Edmond, C Hurt, L Hurt, C Zandoh, C Tawiah, Z Hill, et al. A verbal autopsy corpus for machine learning of cause of death. In *Proceedings of the Corpus Linguistics Conference*, 2011.

[83] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and

Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[84] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[85] Samantha Herrera, Yeetey Enuameh, George Adjei, Kenneth Ayuurebobi, Kwaku Poku Asante, Osman Sankoh, Seth Owusu-Agyei, Yazoume Yé, et al. A systematic review and synthesis of the strengths and limitations of measuring malaria mortality through verbal autopsy. *Malaria journal*, 16(1):421, 2017.

VIII. Figure Legends

This section presents a list of figures in this paper.

- Fig. 1 ANN structure of single neuron.
- Fig. 2 Comparison of ANN and deep learning architecture (Young 2017).
- Fig. 3 Convolutional neural network framework (Young 2017).
- Fig. 4 ML performance evaluation confusion matrix.
- Fig. 5 Recurrent neural network network framework (Young 2017).
- Fig. 6 Long term short memory framework (Young 2017).
- Fig. 7 Distribution of initial searched journal articles vs conference papers.
- Fig. 8 Flowchart that details our methodical meta-analysis.
- Fig. 9 Distribution of VA primary studies that include pre-processing.
- Fig. 10 Pictorial distribution of VA primary studies that include transparency.
- Fig. 11 Pictorial distribution of VA related studies that include data balancing.
- Fig. 12 Distribution of feature extraction techniques reviewed. LEGEND: BAG OF WORDS (BOW), N-GRAMS (NG), HYBRID (HB), WORD2VEC (WV), CONCEPT-BASED (CB), GRAPH-BASED (GB), RULE BASE (RB), DOC2VEC (DV), EXPERT DRIVEN (ED), SENTENCE2VEC (SV), GLOVE (GV).
- Fig. 13 Distribution of feature value representation schemes. LEGEND: BINARY REPRESENTATION (BR), TERM FREQUENCY (TF), DOCUMENT FREQUENCY (DF), TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF), NORMALISED TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (NTF-IDF).
- Fig. 14 Distribution of feature selection techniques. LEGEND: PEARSON CORRELATION (PC), CHI-SQUARED (CS), INFORMATION GAIN (IG), PRINCIPAL COMPONENT ANALYSIS (PCA),FISHER MARKOV SELECTOR (FMS), IMPROVED GLOBAL FEATURE SELECTION (IGFS), LOCAL SEMI SUPERVISED (LSS), NATURAL LANGUAGE PROCESSING (NLP).
- Fig. 15 Distribution of VA related studies in comparison to other fields used in this review.
- Fig. 16 Distribution of journal articles and conference articles used in this review.
- Fig. 17 Distribution of evaluation metrics from primary studies reviewed. LEGEND: SPECIFICITY (SP), PRECISION (PR), RECALL (RE), F-SCORE (FS), ACCURACY (AC), CSFM (CS), CCC, PCCC, KAPPA (KA), CASE SIMILARITY MEASURE (CSM), TOP CAUSE (TC), TOP3COD (T3), AREA UNDER CURVE (AUC).
- Fig. 18 Tag cloud of this systematic review.
- Fig. 19 Distribution of statistical and ML models used in primary studies. LEGEND: VA ALGORITHMS (VA), LOGISTIC REGRESSION (LR), RANDOM FORESTS (RF), DECISION TREE (DT), K- NEAREST NEIGHBOR (K-NN), SUPPORT VECTOR MACHINE (SVM), ARTIFICIAL NEURAL NETWORK (ANN), NAIVE BAYES CLASSIFIER (NBC), K-MEANS CLUSTERING (K-MEANS), RULE-BASED (RB), ENSEMBLE CLASSIFIER (EC), CONVOLUTIONAL NEURAL NETWORK (CNN), GATED RECURRENT NEURAL NETWORK (GRN).
- Fig. 20 The guiding machine learning VA that is proposed in this literature survey.

F Appendices

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Mr Michael Mapundu

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M1911132

NAME: (Principal Investigator)	Mr Michael Mapundu
DEPARTMENT:	Epidemiology and Biostatistics Agincourt Health and Demographic Surveillance Site National Health Laboratory Service
PROJECT TITLE:	Computational approaches to characterizing morbidity and mortality patterns in rural South Africa
DATE CONSIDERED:	29/11/2019
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Prof T Celik, Prof E Musenge and Dr C Kabudula
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	23/09/2020

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions.Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. I **agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in **November** and will therefore be due in the month of **November** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

25/09/2020
Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

G Appendices

ML Script Python

Combined dataset

```
# -*- coding: utf-8 -*-
"""
Created on Wed Jan  6 13:06:50 2021


"""

print('importing required packages...')
import pandas as pd, numpy as np, os, re, time, matplotlib.pyplot
as plt, seaborn as sns, warnings, psycopg2
from sqlalchemy import create_engine
from sklearn.linear_model import LinearRegression,
LogisticRegression, Ridge, Lasso, ElasticNet
from sklearn.ensemble import RandomForestClassifier,
AdaBoostClassifier, RandomForestRegressor
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
from imblearn.over_sampling import SMOTE
#from imblearn.over_sampling import ROSE
from imblearn.combine import SMOTETomek
import statsmodels.api as sm
import xgboost as xgb
from scipy.sparse import hstack
from scipy import interp
from itertools import cycle
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
from imblearn.over_sampling import RandomOverSampler

#from sklearn.preprocessing import LabelEncoder, label_binarize
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import TruncatedSVD

from sklearn.dummy import DummyClassifier
from sklearn.linear_model import SGDClassifier
from sklearn.naive_bayes import MultinomialNB, GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier,
AdaBoostClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.multiclass import OneVsRestClassifier
from sklearn.svm import SVC
from sklearn.neural_network import MLPClassifier
from xgboost import XGBRegressor
from xgboost import XGBClassifier

#from sklearn.cross_validation import train_test_split
from sklearn import metrics
from sklearn.preprocessing import LabelEncoder, label_binarize
'''Metrics/Evaluation'''
from sklearn.model_selection import train_test_split,
GridSearchCV
from sklearn.metrics import accuracy_score, precision_score,
recall_score, f1_score, roc_curve, auc, confusion_matrix
```

```
'''Display'''
from IPython.core.display import display, HTML
import nltk
print('required packages loaded and view set successfully...')

#seting a few parameters
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
#pd.set_option('display.max_colwidth', -13)
pd.options.display.float_format = '{:,.2f}'.format
display(HTML("<style>.container { width:95% !important;
}</style>"))
sns.set_style('darkgrid')
warnings.filterwarnings('ignore')

start = time.time()
startptd = time.strftime('%X %x %Z')
print('\n','The program start time and Date','\n',startptd)

#Setting the working directory
os.chdir('C:\\Users\\a0056407\\Desktop\\Michael_Mapundu__Docs\\Ph
DPythonScript\\FinalMLCombinedVAS\\')
#df =
pd.read_csv('C:\\Users\\a0056407\\Desktop\\Michael_Mapundu__Docs\
\PhDPythonScript\\FinalMLCombinedVAS\\MichaelData.csv',
encoding='latin1')
print('reading the dataset, please wait...')

#Converting the columns headers to lower case
#df.columns = df.columns.str.lower()

conn = psycopg2.connect(user="postgres", password="postgres",
                                    host="127.0.0.1", port="5432",
database="mike")
print('deleting table from database if exist')
with conn:
    cursor = conn.cursor()
    cursor.execute('drop table if exists encryted_ncr')
    conn.commit()

print('exporting the dataframe to postgres table...')
engine =
create_engine('postgresql://postgres:postgres@localhost:5432/mike
')
#df.to_sql('verbal_autopsy', engine, chunksize=10000,
index=False)
print('data loaded to postgres successfully...')
cursor = conn.cursor()
sql_cmd1 = "select * from verbal_autopsy"
cursor.execute(sql_cmd1)
#cursor.execute(sql_cmd2)
dfA = pd.read_sql(sql_cmd1, conn)
#df2 = pd.read_sql(sql_cmd2, conn)
#closing the connection
cursor.close()
```

```
conn.close()
print('data read successfully and connection closed...')

for col in dfA.columns:
    print(col)

#drop less important rows
del dfA['dob'], dfA['dod'], dfA['elder'], dfA['midage'],
dfA['adult'], dfA['child'], dfA['province'], dfA['womandeath']
del dfA['under5'], dfA['infant'], dfA['neonate'],
dfA['deathlocal'], dfA['diedat'], dfA['diedatplace'],
dfA['hospital']
del dfA['deathregistration'], dfA['respondentpresent'],
dfA['whynorespondent'], dfA['respondentdeceasedrelation']
del dfA['otherrelation'], dfA['famcause1'], dfA['famcause2'],
dfA['receiv_biotreatment']
del dfA['biomedi_received'], dfA['receiv_tradi_treatment'],
dfA['tradimed_received'], dfA['treatm_sougth_first'],
dfA['other_remarks']
del dfA['icdimed_consensus'], dfA['icdcont_consensus'],
dfA['maincod_assess1'], dfA['icdmain_assess1'], dfA['lik1'],
dfA['lik3']
del dfA['immcod_assess1'], dfA['contrcod_assess1'],
dfA['doctor_name_assess1'], dfA['maincod_assess2'], dfA['lik2'],
dfA['cause3']
del dfA['icdmain_assess2'], dfA['immcod_assess2'],
dfA['contrcod_assess2'], dfA['doctor_name_assess2'],
dfA['cause2']
del dfA['malprev'], dfA['hivprev'], dfA['pregstat'],
dfA['preglik'], dfA['prmat'], dfA['indet'], dfA['cause1']

#recoding data
dfA['death_age'] = dfA['death_age'].fillna(0)
dfA['death_age'] = pd.to_numeric(dfA['death_age'],
errors='coerce')
dfA['male'] = np.where(dfA['male']=='y', '1', '0')
dfA['male'] = pd.to_numeric(dfA['male'], errors='coerce')
dfA['female'] = np.where(dfA['female']=='y', '1', '0')
dfA['female'] = pd.to_numeric(dfA['female'], errors='coerce')
dfA['acute'] = np.where(dfA['acute']=='y', '1', '0')
dfA['acute'] = pd.to_numeric(dfA['acute'], errors='coerce')
dfA['chronic'] = np.where(dfA['chronic']=='y', '1', '0')
dfA['chronic'] = pd.to_numeric(dfA['chronic'], errors='coerce')
dfA['sudden'] = np.where(dfA['female']=='y', '1', '0')
dfA['sudden'] = pd.to_numeric(dfA['sudden'], errors='coerce')
dfA['wet_seas'] = np.where(dfA['wet_seas']=='y', '1', '0')
dfA['wet_seas'] = pd.to_numeric(dfA['wet_seas'], errors='coerce')
dfA['dry_seas'] = np.where(dfA['dry_seas']=='y', '1', '0')
dfA['dry_seas'] = pd.to_numeric(dfA['dry_seas'], errors='coerce')
dfA['heart_dis'] = np.where(dfA['heart_dis']=='y', '1', '0')
dfA['heart_dis'] = pd.to_numeric(dfA['heart_dis'],
errors='coerce')
dfA['tuber'] = np.where(dfA['tuber']=='y', '1', '0')
dfA['tuber'] = pd.to_numeric(dfA['tuber'], errors='coerce')
dfA['hiv_aids'] = np.where(dfA['hiv_aids']=='y', '1', '0')
dfA['hiv_aids'] = pd.to_numeric(dfA['hiv_aids'], errors='coerce')
```

```
dfA['hypert'] = np.where(dfA['hypert']=='y', '1', '0')
dfA['hypert'] = pd.to_numeric(dfA['hypert'], errors='coerce')
dfA['diabetes'] = np.where(dfA['diabetes']=='y', '1', '0')
dfA['diabetes'] = pd.to_numeric(dfA['diabetes'], errors='coerce')
dfA['asthma'] = np.where(dfA['asthma']=='y', '1', '0')
dfA['asthma'] = pd.to_numeric(dfA['asthma'], errors='coerce')
dfA['epilepsy'] = np.where(dfA['epilepsy']=='y', '1', '0')
dfA['epilepsy'] = pd.to_numeric(dfA['epilepsy'], errors='coerce')
dfA['cancer'] = np.where(dfA['cancer']=='y', '1', '0')
dfA['cancer'] = pd.to_numeric(dfA['cancer'], errors='coerce')
dfA['copd'] = np.where(dfA['copd']=='y', '1', '0')
dfA['copd'] = pd.to_numeric(dfA['copd'], errors='coerce')
dfA['dement'] = np.where(dfA['dement']=='y', '1', '0')
dfA['dement'] = pd.to_numeric(dfA['dement'], errors='coerce')
dfA['depress'] = np.where(dfA['depress']=='y', '1', '0')
dfA['depress'] = pd.to_numeric(dfA['depress'], errors='coerce')
dfA['stroke'] = np.where(dfA['stroke']=='y', '1', '0')
dfA['stroke'] = pd.to_numeric(dfA['stroke'], errors='coerce')
dfA['sickle'] = np.where(dfA['sickle']=='y', '1', '0')
dfA['sickle'] = pd.to_numeric(dfA['sickle'], errors='coerce')
dfA['kidney_dis'] = np.where(dfA['kidney_dis']=='y', '1', '0')
dfA['kidney_dis'] = pd.to_numeric(dfA['kidney_dis'],
errors='coerce')
dfA['liver_dis'] = np.where(dfA['liver_dis']=='y', '1', '0')
dfA['liver_dis'] = pd.to_numeric(dfA['liver_dis'],
errors='coerce')
dfA['measles'] = np.where(dfA['measles']=='y', '1', '0')
dfA['measles'] = pd.to_numeric(dfA['measles'], errors='coerce')
dfA['men_con'] = np.where(dfA['men_con']=='y', '1', '0')
dfA['men_con'] = pd.to_numeric(dfA['men_con'], errors='coerce')
dfA['mencon3'] = np.where(dfA['mencon3']=='y', '1', '0')
dfA['mencon3'] = pd.to_numeric(dfA['mencon3'], errors='coerce')
dfA['malaria'] = np.where(dfA['malaria']=='y', '1', '0')
dfA['malaria'] = pd.to_numeric(dfA['malaria'], errors='coerce')
dfA['malarneg'] = np.where(dfA['malarneg']=='y', '1', '0')
dfA['malarneg'] = pd.to_numeric(dfA['malarneg'], errors='coerce')
dfA['fever'] = np.where(dfA['fever']=='y', '1', '0')
dfA['fever'] = pd.to_numeric(dfA['fever'], errors='coerce')
dfA['ac_fever'] = np.where(dfA['ac_fever']=='y', '1', '0')
dfA['ac_fever'] = pd.to_numeric(dfA['ac_fever'], errors='coerce')
dfA['ch_fever'] = np.where(dfA['ch_fever']=='y', '1', '0')
dfA['ch_fever'] = pd.to_numeric(dfA['ch_fever'], errors='coerce')
dfA['night_sw'] = np.where(dfA['night_sw']=='y', '1', '0')
dfA['night_sw'] = pd.to_numeric(dfA['night_sw'], errors='coerce')
dfA['cough'] = np.where(dfA['cough']=='y', '1', '0')
dfA['cough'] = pd.to_numeric(dfA['cough'], errors='coerce')
dfA['ac_cough'] = np.where(dfA['ac_cough']=='y', '1', '0')
dfA['ac_cough'] = pd.to_numeric(dfA['ac_cough'], errors='coerce')
dfA['ch_cough'] = np.where(dfA['ch_cough']=='y', '1', '0')
dfA['ch_cough'] = pd.to_numeric(dfA['ch_cough'], errors='coerce')
dfA['pr_cough'] = np.where(dfA['pr_cough']=='y', '1', '0')
dfA['pr_cough'] = pd.to_numeric(dfA['pr_cough'], errors='coerce')
dfA['bl_cough'] = np.where(dfA['bl_cough']=='y', '1', '0')
dfA['bl_cough'] = pd.to_numeric(dfA['bl_cough'], errors='coerce')
dfA['whoop'] = np.where(dfA['whoop']=='y', '1', '0')
dfA['whoop'] = pd.to_numeric(dfA['whoop'], errors='coerce')
```

```
dfA['breath'] = np.where(dfA['breath']=='y', '1', '0')
dfA['breath'] = pd.to_numeric(dfA['breath'], errors='coerce')
dfA['rapid_br'] = np.where(dfA['rapid_br']=='y', '1', '0')
dfA['rapid_br'] = pd.to_numeric(dfA['rapid_br'], errors='coerce')
dfA['ac_rpbr'] = np.where(dfA['ac_rpbr']=='y', '1', '0')
dfA['ac_rpbr'] = pd.to_numeric(dfA['ac_rpbr'], errors='coerce')
dfA['ch_rpbr'] = np.where(dfA['ch_rpbr']=='y', '1', '0')
dfA['ch_rpbr'] = pd.to_numeric(dfA['ch_rpbr'], errors='coerce')
dfA['br_less'] = np.where(dfA['br_less']=='y', '1', '0')
dfA['br_less'] = pd.to_numeric(dfA['br_less'], errors='coerce')
dfA['ac_brl'] = np.where(dfA['ac_brl']=='y', '1', '0')
dfA['ac_brl'] = pd.to_numeric(dfA['ac_brl'], errors='coerce')
dfA['ch_brl'] = np.where(dfA['ch_brl']=='y', '1', '0')
dfA['ch_brl'] = pd.to_numeric(dfA['ch_brl'], errors='coerce')
dfA['exert_br'] = np.where(dfA['exert_br']=='y', '1', '0')
dfA['exert_br'] = pd.to_numeric(dfA['exert_br'], errors='coerce')
dfA['lying_br'] = np.where(dfA['lying_br']=='y', '1', '0')
dfA['lying_br'] = pd.to_numeric(dfA['lying_br'], errors='coerce')
dfA['chest_in'] = np.where(dfA['chest_in']=='y', '1', '0')
dfA['chest_in'] = pd.to_numeric(dfA['chest_in'], errors='coerce')
dfA['wheeze'] = np.where(dfA['wheeze']=='y', '1', '0')
dfA['wheeze'] = pd.to_numeric(dfA['wheeze'], errors='coerce')
dfA['ch_pain'] = np.where(dfA['ch_pain']=='y', '1', '0')
dfA['ch_pain'] = pd.to_numeric(dfA['ch_pain'], errors='coerce')
dfA['yellow'] = np.where(dfA['yellow']=='y', '1', '0')
dfA['yellow'] = pd.to_numeric(dfA['yellow'], errors='coerce')
dfA['diarr'] = np.where(dfA['diarr']=='y', '1', '0')
dfA['diarr'] = pd.to_numeric(dfA['diarr'], errors='coerce')
dfA['ac_diarr'] = np.where(dfA['ac_diarr']=='y', '1', '0')
dfA['ac_diarr'] = pd.to_numeric(dfA['ac_diarr'], errors='coerce')
dfA['pe_diarr'] = np.where(dfA['pe_diarr']=='y', '1', '0')
dfA['pe_diarr'] = pd.to_numeric(dfA['pe_diarr'], errors='coerce')
dfA['ch_diarr'] = np.where(dfA['ch_diarr']=='y', '1', '0')
dfA['ch_diarr'] = pd.to_numeric(dfA['ch_diarr'], errors='coerce')
dfA['bl_diarr'] = np.where(dfA['bl_diarr']=='y', '1', '0')
dfA['bl_diarr'] = pd.to_numeric(dfA['bl_diarr'], errors='coerce')
dfA['vomiting'] = np.where(dfA['vomiting']=='y', '1', '0')
dfA['vomiting'] = pd.to_numeric(dfA['vomiting'], errors='coerce')
dfA['bl_vomit'] = np.where(dfA['bl_vomit']=='y', '1', '0')
dfA['bl_vomit'] = pd.to_numeric(dfA['bl_vomit'], errors='coerce')
dfA['abdom'] = np.where(dfA['abdom']=='y', '1', '0')
dfA['abdom'] = pd.to_numeric(dfA['abdom'], errors='coerce')
dfA['abd_pain'] = np.where(dfA['abd_pain']=='y', '1', '0')
dfA['abd_pain'] = pd.to_numeric(dfA['abd_pain'], errors='coerce')
dfA['ac_abdp'] = np.where(dfA['ac_abdp']=='y', '1', '0')
dfA['ac_abdp'] = pd.to_numeric(dfA['ac_abdp'], errors='coerce')
dfA['ch_abdp'] = np.where(dfA['ch_abdp']=='y', '1', '0')
dfA['ch_abdp'] = pd.to_numeric(dfA['ch_abdp'], errors='coerce')
dfA['swe_abd'] = np.where(dfA['swe_abd']=='y', '1', '0')
dfA['swe_abd'] = pd.to_numeric(dfA['swe_abd'], errors='coerce')
dfA['ac_swab'] = np.where(dfA['ac_swab']=='y', '1', '0')
dfA['ac_swab'] = pd.to_numeric(dfA['ac_swab'], errors='coerce')
dfA['ch_swab'] = np.where(dfA['ch_swab']=='y', '1', '0')
dfA['ch_swab'] = pd.to_numeric(dfA['ch_swab'], errors='coerce')
dfA['abd_mass'] = np.where(dfA['abd_mass']=='y', '1', '0')
dfA['abd_mass'] = pd.to_numeric(dfA['abd_mass'], errors='coerce')
```

```
dfA['ac_abdm'] = np.where(dfA['ac_abdm']=='y', '1', '0')
dfA['ac_abdm'] = pd.to_numeric(dfA['ac_abdm'], errors='coerce')
dfA['ch_abdm'] = np.where(dfA['ch_abdm']=='y', '1', '0')
dfA['ch_abdm'] = pd.to_numeric(dfA['ch_abdm'], errors='coerce')
dfA['headache'] = np.where(dfA['headache']=='y', '1', '0')
dfA['headache'] = pd.to_numeric(dfA['headache'], errors='coerce')
dfA['skin'] = np.where(dfA['skin']=='y', '1', '0')
dfA['skin'] = pd.to_numeric(dfA['skin'], errors='coerce')
dfA['skin_les'] = np.where(dfA['skin_les']=='y', '1', '0')
dfA['skin_les'] = pd.to_numeric(dfA['skin_les'], errors='coerce')
dfA['b_norm'] = np.where(dfA['b_norm']=='y', '1', '0')
dfA['b_norm'] = pd.to_numeric(dfA['b_norm'], errors='coerce')
dfA['b_assist'] = np.where(dfA['b_assist']=='y', '1', '0')
dfA['b_assist'] = pd.to_numeric(dfA['b_assist'], errors='coerce')
dfA['b_caes'] = np.where(dfA['b_caes']=='y', '1', '0')
dfA['b_caes'] = pd.to_numeric(dfA['b_caes'], errors='coerce')
dfA['b_first'] = np.where(dfA['b_first']=='y', '1', '0')
dfA['b_first'] = pd.to_numeric(dfA['b_first'], errors='coerce')
dfA['b_more4'] = np.where(dfA['b_more4']=='y', '1', '0')
dfA['b_more4'] = pd.to_numeric(dfA['b_more4'], errors='coerce')
dfA['b_mbpr'] = np.where(dfA['b_mbpr']=='y', '1', '0')
dfA['b_mbpr'] = pd.to_numeric(dfA['b_mbpr'], errors='coerce')
dfA['b_msmds'] = np.where(dfA['b_msmds']=='y', '1', '0')
dfA['b_msmds'] = pd.to_numeric(dfA['b_msmds'], errors='coerce')
dfA['b_mcon'] = np.where(dfA['b_mcon']=='y', '1', '0')
dfA['b_mcon'] = pd.to_numeric(dfA['b_mcon'], errors='coerce')
dfA['b_mbvi'] = np.where(dfA['b_mbvi']=='y', '1', '0')
dfA['b_mbvi'] = pd.to_numeric(dfA['b_mbvi'], errors='coerce')
dfA['b_mvbl'] = np.where(dfA['b_mvbl']=='y', '1', '0')
dfA['b_mvbl'] = pd.to_numeric(dfA['b_mvbl'], errors='coerce')
dfA['b_bfac'] = np.where(dfA['b_bfac']=='y', '1', '0')
dfA['b_bfac'] = pd.to_numeric(dfA['b_bfac'], errors='coerce')
dfA['b_bhome'] = np.where(dfA['b_bhome']=='y', '1', '0')
dfA['b_bhome'] = pd.to_numeric(dfA['b_bhome'], errors='coerce')
dfA['b_bway'] = np.where(dfA['b_bway']=='y', '1', '0')
dfA['b_bway'] = pd.to_numeric(dfA['b_bway'], errors='coerce')
dfA['b_bprof'] = np.where(dfA['b_bprof']=='y', '1', '0')
dfA['b_bprof'] = pd.to_numeric(dfA['b_bprof'], errors='coerce')
dfA['injury'] = np.where(dfA['injury']=='y', '1', '0')
dfA['injury'] = pd.to_numeric(dfA['injury'], errors='coerce')
dfA['traffic'] = np.where(dfA['traffic']=='y', '1', '0')
dfA['traffic'] = pd.to_numeric(dfA['traffic'], errors='coerce')
dfA['o_trans'] = np.where(dfA['o_trans']=='y', '1', '0')
dfA['o_trans'] = pd.to_numeric(dfA['o_trans'], errors='coerce')
dfA['fall'] = np.where(dfA['fall']=='y', '1', '0')
dfA['fall'] = pd.to_numeric(dfA['fall'], errors='coerce')
dfA['fire'] = np.where(dfA['fire']=='y', '1', '0')
dfA['fire'] = pd.to_numeric(dfA['fire'], errors='coerce')
dfA['drown'] = np.where(dfA['drown']=='y', '1', '0')
dfA['drown'] = pd.to_numeric(dfA['drown'], errors='coerce')
dfA['assault'] = np.where(dfA['assault']=='y', '1', '0')
dfA['assault'] = pd.to_numeric(dfA['assault'], errors='coerce')
dfA['vemon'] = np.where(dfA['vemon']=='y', '1', '0')
dfA['vemon'] = pd.to_numeric(dfA['vemon'], errors='coerce')
dfA['force'] = np.where(dfA['force']=='y', '1', '0')
dfA['force'] = pd.to_numeric(dfA['force'], errors='coerce')
```

```
dfA['smoking'] = np.where(dfA['smoking']=='y', '1', '0')
dfA['smoking'] = pd.to_numeric(dfA['smoking'], errors='coerce')
dfA['alcohol'] = np.where(dfA['alcohol']=='y', '1', '0')
dfA['alcohol'] = pd.to_numeric(dfA['alcohol'], errors='coerce')
dfA['poison'] = np.where(dfA['poison']=='y', '1', '0')
dfA['poison'] = pd.to_numeric(dfA['poison'], errors='coerce')
dfA['inflict'] = np.where(dfA['inflict']=='y', '1', '0')
dfA['inflict'] = pd.to_numeric(dfA['inflict'], errors='coerce')
dfA['suicide'] = np.where(dfA['suicide']=='y', '1', '0')
dfA['suicide'] = pd.to_numeric(dfA['suicide'], errors='coerce')
dfA['married'] = np.where(dfA['married']=='y', '1', '0')
dfA['married'] = pd.to_numeric(dfA['married'], errors='coerce')
dfA['vaccin'] = np.where(dfA['vaccin']=='y', '1', '0')
dfA['vaccin'] = pd.to_numeric(dfA['vaccin'], errors='coerce')
dfA['disch'] = np.where(dfA['disch']=='y', '1', '0')
dfA['disch'] = pd.to_numeric(dfA['disch'], errors='coerce')
dfA['smore2'] = np.where(dfA['smore2']=='y', '1', '0')
dfA['smore2'] = pd.to_numeric(dfA['smore2'], errors='coerce')
dfA['scosts'] = np.where(dfA['scosts']=='y', '1', '0')
dfA['scosts'] = pd.to_numeric(dfA['scosts'], errors='coerce')
dfA['smobph'] = np.where(dfA['smobph']=='y', '1', '0')
dfA['smobph'] = pd.to_numeric(dfA['smobph'], errors='coerce')
dfA['stradm'] = np.where(dfA['stradm']=='y', '1', '0')
dfA['stradm'] = pd.to_numeric(dfA['stradm'], errors='coerce')
dfA['sdoubt'] = np.where(dfA['sdoubt']=='y', '1', '0')
dfA['sdoubt'] = pd.to_numeric(dfA['sdoubt'], errors='coerce')
dfA['sur_1m'] = np.where(dfA['sur_1m']=='y', '1', '0')
dfA['sur_1m'] = pd.to_numeric(dfA['sur_1m'], errors='coerce')
dfA['treat'] = np.where(dfA['treat']=='y', '1', '0')
dfA['treat'] = pd.to_numeric(dfA['treat'], errors='coerce')
dfA['streat'] = np.where(dfA['streat']=='y', '1', '0')
dfA['streat'] = pd.to_numeric(dfA['streat'], errors='coerce')
dfA['smedic'] = np.where(dfA['smedic']=='y', '1', '0')
dfA['smedic'] = pd.to_numeric(dfA['smedic'], errors='coerce')
dfA['strans'] = np.where(dfA['strans']=='y', '1', '0')
dfA['strans'] = pd.to_numeric(dfA['strans'], errors='coerce')
dfA['sadmit'] = np.where(dfA['sadmit']=='y', '1', '0')
dfA['sadmit'] = pd.to_numeric(dfA['sadmit'], errors='coerce')
dfA['shospf'] = np.where(dfA['shospf']=='y', '1', '0')
dfA['shospf'] = pd.to_numeric(dfA['shospf'], errors='coerce')
dfA['t_ort'] = np.where(dfA['t_ort']=='y', '1', '0')
dfA['t_ort'] = pd.to_numeric(dfA['t_ort'], errors='coerce')
dfA['antib_i'] = np.where(dfA['antib_i']=='y', '1', '0')
dfA['antib_i'] = pd.to_numeric(dfA['antib_i'], errors='coerce')
dfA['surgery'] = np.where(dfA['surgery']=='y', '1', '0')
dfA['surgery'] = pd.to_numeric(dfA['surgery'], errors='coerce')
dfA['blood_tr'] = np.where(dfA['blood_tr']=='y', '1', '0')
dfA['blood_tr'] = pd.to_numeric(dfA['blood_tr'], errors='coerce')
dfA['t_iv'] = np.where(dfA['t_iv']=='y', '1', '0')
dfA['t_iv'] = pd.to_numeric(dfA['t_iv'], errors='coerce')
dfA['t_ngt'] = np.where(dfA['t_ngt']=='y', '1', '0')
dfA['t_ngt'] = pd.to_numeric(dfA['t_ngt'], errors='coerce')
dfA['sk_feet'] = np.where(dfA['sk_feet']=='y', '1', '0')
dfA['sk_feet'] = pd.to_numeric(dfA['sk_feet'], errors='coerce')
dfA['rash'] = np.where(dfA['rash']=='y', '1', '0')
dfA['rash'] = pd.to_numeric(dfA['rash'], errors='coerce')
```

```
dfA['ac_rash'] = np.where(dfA['ac_rash']=='y', '1', '0')
dfA['ac_rash'] = pd.to_numeric(dfA['ac_rash'], errors='coerce')
dfA['ch_rash'] = np.where(dfA['ch_rash']=='y', '1', '0')
dfA['ch_rash'] = pd.to_numeric(dfA['ch_rash'], errors='coerce')
dfA['measrash'] = np.where(dfA['measrash']=='y', '1', '0')
dfA['measrash'] = pd.to_numeric(dfA['measrash'], errors='coerce')
dfA['mttv'] = np.where(dfA['mttv']=='y', '1', '0')
dfA['mttv'] = pd.to_numeric(dfA['mttv'], errors='coerce')
dfA['mlf_sh'] = np.where(dfA['mlf_sh']=='y', '1', '0')
dfA['mlf_sh'] = pd.to_numeric(dfA['mlf_sh'], errors='coerce')
dfA['mlf_lh'] = np.where(dfA['mlf_lh']=='y', '1', '0')
dfA['mlf_lh'] = pd.to_numeric(dfA['mlf_lh'], errors='coerce')
dfA['mlf_bk'] = np.where(dfA['mlf_bk']=='y', '1', '0')
dfA['mlf_bk'] = pd.to_numeric(dfA['mlf_bk'], errors='coerce')
dfA['born_malf'] = np.where(dfA['born_malf']=='y', '1', '0')
dfA['born_malf'] = pd.to_numeric(dfA['born_malf'],
errors='coerce')
dfA['devel'] = np.where(dfA['devel']=='y', '1', '0')
dfA['devel'] = pd.to_numeric(dfA['devel'], errors='coerce')
dfA['b_yellow'] = np.where(dfA['b_yellow']=='y', '1', '0')
dfA['b_yellow'] = pd.to_numeric(dfA['b_yellow'], errors='coerce')
dfA['umbinf'] = np.where(dfA['umbinf']=='y', '1', '0')
dfA['umbinf'] = pd.to_numeric(dfA['umbinf'], errors='coerce')
dfA['cold'] = np.where(dfA['cold']=='y', '1', '0')
dfA['cold'] = pd.to_numeric(dfA['cold'], errors='coerce')
dfA['unw_d2'] = np.where(dfA['unw_d2']=='y', '1', '0')
dfA['unw_d2'] = pd.to_numeric(dfA['unw_d2'], errors='coerce')
dfA['unw_d1'] = np.where(dfA['unw_d1']=='y', '1', '0')
dfA['unw_d1'] = pd.to_numeric(dfA['unw_d1'], errors='coerce')
dfA['font_lo'] = np.where(dfA['font_lo']=='y', '1', '0')
dfA['font_lo'] = pd.to_numeric(dfA['font_lo'], errors='coerce')
dfA['font_hi'] = np.where(dfA['font_hi']=='y', '1', '0')
dfA['font_hi'] = pd.to_numeric(dfA['font_hi'], errors='coerce')
dfA['uri_haem'] = np.where(dfA['uri_haem']=='y', '1', '0')
dfA['uri_haem'] = pd.to_numeric(dfA['uri_haem'], errors='coerce')
dfA['ac_stnk'] = np.where(dfA['ac_stnk']=='y', '1', '0')
dfA['ac_stnk'] = pd.to_numeric(dfA['ac_stnk'], errors='coerce')
dfA['convul'] = np.where(dfA['convul']=='y', '1', '0')
dfA['convul'] = pd.to_numeric(dfA['convul'], errors='coerce')
dfA['herpes'] = np.where(dfA['herpes']=='y', '1', '0')
dfA['herpes'] = pd.to_numeric(dfA['herpes'], errors='coerce')
dfA['stiff_neck'] = np.where(dfA['stiff_neck']=='y', '1', '0')
dfA['stiff_neck'] = pd.to_numeric(dfA['stiff_neck'],
errors='coerce')
dfA['ch_stnk'] = np.where(dfA['ch_stnk']=='y', '1', '0')
dfA['ch_stnk'] = pd.to_numeric(dfA['ch_stnk'], errors='coerce')
dfA['coma'] = np.where(dfA['coma']=='y', '1', '0')
dfA['coma'] = pd.to_numeric(dfA['coma'], errors='coerce')
dfA['co_ons'] = np.where(dfA['co_ons']=='y', '1', '0')
dfA['co_ons'] = pd.to_numeric(dfA['co_ons'], errors='coerce')
dfA['exc_urine'] = np.where(dfA['exc_urine']=='y', '1', '0')
dfA['exc_urine'] = pd.to_numeric(dfA['exc_urine'],
errors='coerce')
dfA['uri_ret'] = np.where(dfA['uri_ret']=='y', '1', '0')
dfA['uri_ret'] = pd.to_numeric(dfA['uri_ret'], errors='coerce')
dfA['urine'] = np.where(dfA['urine']=='y', '1', '0')
```

```
dfA['urine'] = pd.to_numeric(dfA['urine'], errors='coerce')
dfA['ac_conv'] = np.where(dfA['ac_conv']=='y', '1', '0')
dfA['ac_conv'] = pd.to_numeric(dfA['ac_conv'], errors='coerce')
dfA['ch_conv'] = np.where(dfA['ch_conv']=='y', '1', '0')
dfA['ch_conv'] = pd.to_numeric(dfA['ch_conv'], errors='coerce')
dfA['unc_con'] = np.where(dfA['unc_con']=='y', '1', '0')
dfA['unc_con'] = pd.to_numeric(dfA['unc_con'], errors='coerce')
dfA['hyster'] = np.where(dfA['hyster']=='y', '1', '0')
dfA['hyster'] = pd.to_numeric(dfA['hyster'], errors='coerce')
dfA['bpr_preg'] = np.where(dfA['bpr_preg']=='y', '1', '0')
dfA['bpr_preg'] = pd.to_numeric(dfA['bpr_preg'], errors='coerce')
dfA['fit_preg'] = np.where(dfA['fit_preg']=='y', '1', '0')
dfA['fit_preg'] = pd.to_numeric(dfA['fit_preg'], errors='coerce')
dfA['vis_bl'] = np.where(dfA['vis_bl']=='y', '1', '0')
dfA['vis_bl'] = pd.to_numeric(dfA['vis_bl'], errors='coerce')
dfA['bleed'] = np.where(dfA['bleed']=='y', '1', '0')
dfA['bleed'] = pd.to_numeric(dfA['bleed'], errors='coerce')
dfA['e_bleed'] = np.where(dfA['e_bleed']=='y', '1', '0')
dfA['e_bleed'] = pd.to_numeric(dfA['e_bleed'], errors='coerce')
dfA['s_bleed'] = np.where(dfA['s_bleed']=='y', '1', '0')
dfA['s_bleed'] = pd.to_numeric(dfA['s_bleed'], errors='coerce')
dfA['d_bleed'] = np.where(dfA['d_bleed']=='y', '1', '0')
dfA['d_bleed'] = pd.to_numeric(dfA['d_bleed'], errors='coerce')
dfA['p_bleed'] = np.where(dfA['p_bleed']=='y', '1', '0')
dfA['p_bleed'] = pd.to_numeric(dfA['p_bleed'], errors='coerce')
dfA['placent_r'] = np.where(dfA['placent_r']=='y', '1', '0')
dfA['placent_r'] = pd.to_numeric(dfA['placent_r'],
errors='coerce')
dfA['disch_sm'] = np.where(dfA['disch_sm']=='y', '1', '0')
dfA['disch_sm'] = pd.to_numeric(dfA['disch_sm'], errors='coerce')
dfA['term_att'] = np.where(dfA['term_att']=='y', '1', '0')
dfA['term_att'] = pd.to_numeric(dfA['term_att'], errors='coerce')
dfA['abort'] = np.where(dfA['abort']=='y', '1', '0')
dfA['abort'] = pd.to_numeric(dfA['abort'], errors='coerce')
dfA['arch_b'] = np.where(dfA['arch_b']=='y', '1', '0')
dfA['arch_b'] = pd.to_numeric(dfA['arch_b'], errors='coerce')
dfA['conv_d2'] = np.where(dfA['conv_d2']=='y', '1', '0')
dfA['conv_d2'] = pd.to_numeric(dfA['conv_d2'], errors='coerce')
dfA['conv_d1'] = np.where(dfA['conv_d1']=='y', '1', '0')
dfA['conv_d1'] = pd.to_numeric(dfA['conv_d1'], errors='coerce')
dfA['ab_posit'] = np.where(dfA['ab_posit']=='y', '1', '0')
dfA['ab_posit'] = pd.to_numeric(dfA['ab_posit'], errors='coerce')
dfA['st_suck'] = np.where(dfA['st_suck']=='y', '1', '0')
dfA['st_suck'] = pd.to_numeric(dfA['st_suck'], errors='coerce')
dfA['cord'] = np.where(dfA['cord']=='y', '1', '0')
dfA['cord'] = pd.to_numeric(dfA['cord'], errors='coerce')
dfA['born_early'] = np.where(dfA['born_early']=='y', '1', '0')
dfA['born_early'] = pd.to_numeric(dfA['born_early'],
errors='coerce')
dfA['born_3437'] = np.where(dfA['born_3437']=='y', '1', '0')
dfA['born_3437'] = pd.to_numeric(dfA['born_3437'],
errors='coerce')
dfA['born_38'] = np.where(dfA['born_38']=='y', '1', '0')
dfA['born_38'] = pd.to_numeric(dfA['born_38'], errors='coerce')
dfA['born_nobr'] = np.where(dfA['born_nobr']=='y', '1', '0')
```

```
dfA['born_nobr'] = pd.to_numeric(dfA['born_nobr'],
errors='coerce')
dfA['waters'] = np.where(dfA['waters']=='y', '1', '0')
dfA['waters'] = pd.to_numeric(dfA['waters'], errors='coerce')
dfA['ab_size'] = np.where(dfA['ab_size']=='y', '1', '0')
dfA['ab_size'] = pd.to_numeric(dfA['ab_size'], errors='coerce')
dfA['baby_br'] = np.where(dfA['baby_br']=='y', '1', '0')
dfA['baby_br'] = pd.to_numeric(dfA['baby_br'], errors='coerce')
dfA['born_big'] = np.where(dfA['born_big']=='y', '1', '0')
dfA['born_big'] = pd.to_numeric(dfA['born_big'], errors='coerce')
dfA['born_small'] = np.where(dfA['born_small']=='y', '1', '0')
dfA['born_small'] = pd.to_numeric(dfA['born_small'],
errors='coerce')
dfA['fed_d1'] = np.where(dfA['fed_d1']=='y', '1', '0')
dfA['fed_d1'] = pd.to_numeric(dfA['fed_d1'], errors='coerce')
dfA['mushy'] = np.where(dfA['mushy']=='y', '1', '0')
dfA['mushy'] = pd.to_numeric(dfA['mushy'], errors='coerce')
dfA['twin'] = np.where(dfA['twin']=='y', '1', '0')
dfA['twin'] = pd.to_numeric(dfA['twin'], errors='coerce')
dfA['comdel'] = np.where(dfA['comdel']=='y', '1', '0')
dfA['comdel'] = pd.to_numeric(dfA['comdel'], errors='coerce')
dfA['no_life'] = np.where(dfA['no_life']=='y', '1', '0')
dfA['no_life'] = pd.to_numeric(dfA['no_life'], errors='coerce')
dfA['cried'] = np.where(dfA['cried']=='y', '1', '0')
dfA['cried'] = pd.to_numeric(dfA['cried'], errors='coerce')
dfA['cyanosis'] = np.where(dfA['cyanosis']=='y', '1', '0')
dfA['cyanosis'] = pd.to_numeric(dfA['cyanosis'], errors='coerce')
dfA['move_lb'] = np.where(dfA['move_lb']=='y', '1', '0')
dfA['move_lb'] = pd.to_numeric(dfA['move_lb'], errors='coerce')
dfA['wt_loss'] = np.where(dfA['wt_loss']=='y', '1', '0')
dfA['wt_loss'] = pd.to_numeric(dfA['wt_loss'], errors='coerce')
dfA['wasting'] = np.where(dfA['wasting']=='y', '1', '0')
dfA['wasting'] = pd.to_numeric(dfA['wasting'], errors='coerce')
dfA['or_cand'] = np.where(dfA['or_cand']=='y', '1', '0')
dfA['or_cand'] = pd.to_numeric(dfA['or_cand'], errors='coerce')
dfA['rigidity'] = np.where(dfA['rigidity']=='y', '1', '0')
dfA['rigidity'] = pd.to_numeric(dfA['rigidity'], errors='coerce')
dfA['anaemia'] = np.where(dfA['anaemia']=='y', '1', '0')
dfA['anaemia'] = pd.to_numeric(dfA['anaemia'], errors='coerce')
dfA['swe_legs'] = np.where(dfA['swe_legs']=='y', '1', '0')
dfA['swe_legs'] = pd.to_numeric(dfA['swe_legs'], errors='coerce')
dfA['swell'] = np.where(dfA['swell']=='y', '1', '0')
dfA['swell'] = pd.to_numeric(dfA['swell'], errors='coerce')
dfA['swe_oral'] = np.where(dfA['swe_oral']=='y', '1', '0')
dfA['swe_oral'] = pd.to_numeric(dfA['swe_oral'], errors='coerce')
dfA['swe_oth'] = np.where(dfA['swe_oth']=='y', '1', '0')
dfA['swe_oth'] = pd.to_numeric(dfA['swe_oth'], errors='coerce')
dfA['swe_gen'] = np.where(dfA['swe_gen']=='y', '1', '0')
dfA['swe_gen'] = pd.to_numeric(dfA['swe_gen'], errors='coerce')
dfA['swe_breast'] = np.where(dfA['swe_breast']=='y', '1', '0')
dfA['swe_breast'] = pd.to_numeric(dfA['swe_breast'],
errors='coerce')
dfA['swe_armp'] = np.where(dfA['swe_armp']=='y', '1', '0')
dfA['swe_armp'] = pd.to_numeric(dfA['swe_armp'], errors='coerce')
dfA['swe_neck'] = np.where(dfA['swe_neck']=='y', '1', '0')
dfA['swe_neck'] = pd.to_numeric(dfA['swe_neck'], errors='coerce')
```

```
dfA['lab_24'] = np.where(dfA['lab_24']=='y', '1', '0')
dfA['lab_24'] = pd.to_numeric(dfA['lab_24'], errors='coerce')
dfA['died_lab'] = np.where(dfA['died_lab']=='y', '1', '0')
dfA['died_lab'] = pd.to_numeric(dfA['died_lab'], errors='coerce')
dfA['death_24'] = np.where(dfA['death_24']=='y', '1', '0')
dfA['death_24'] = pd.to_numeric(dfA['death_24'], errors='coerce')
dfA['baby_al'] = np.where(dfA['baby_al']=='y', '1', '0')
dfA['baby_al'] = pd.to_numeric(dfA['baby_al'], errors='coerce')
dfA['baby_pos'] = np.where(dfA['baby_pos']=='y', '1', '0')
dfA['baby_pos'] = pd.to_numeric(dfA['baby_pos'], errors='coerce')
dfA['mon_early'] = np.where(dfA['mon_early']=='y', '1', '0')
dfA['mon_early'] = pd.to_numeric(dfA['mon_early'],
errors='coerce')
dfA['breast_fd'] = np.where(dfA['breast_fd']=='y', '1', '0')
dfA['breast_fd'] = pd.to_numeric(dfA['breast_fd'],
errors='coerce')
dfA['del_cs'] = np.where(dfA['del_cs']=='y', '1', '0')
dfA['del_cs'] = pd.to_numeric(dfA['del_cs'], errors='coerce')
dfA['prof_ass'] = np.where(dfA['prof_ass']=='y', '1', '0')
dfA['prof_ass'] = pd.to_numeric(dfA['prof_ass'], errors='coerce')
dfA['del_ass'] = np.where(dfA['del_ass']=='y', '1', '0')
dfA['del_ass'] = pd.to_numeric(dfA['del_ass'], errors='coerce')
dfA['del_norm'] = np.where(dfA['del_norm']=='y', '1', '0')
dfA['del_norm'] = pd.to_numeric(dfA['del_norm'], errors='coerce')
dfA['del_fac'] = np.where(dfA['del_fac']=='y', '1', '0')
dfA['del_fac'] = pd.to_numeric(dfA['del_fac'], errors='coerce')
dfA['del_home'] = np.where(dfA['del_home']=='y', '1', '0')
dfA['del_home'] = pd.to_numeric(dfA['del_home'], errors='coerce')
dfA['del_else'] = np.where(dfA['del_else']=='y', '1', '0')
dfA['del_else'] = pd.to_numeric(dfA['del_else'], errors='coerce')
dfA['vb_bet'] = np.where(dfA['vb_bet']=='y', '1', '0')
dfA['vb_bet'] = pd.to_numeric(dfA['vb_bet'], errors='coerce')
dfA['vb_men'] = np.where(dfA['vb_men']=='y', '1', '0')
dfA['vb_men'] = pd.to_numeric(dfA['vb_men'], errors='coerce')
dfA['vb_after'] = np.where(dfA['vb_after']=='y', '1', '0')
dfA['vb_after'] = pd.to_numeric(dfA['vb_after'], errors='coerce')
dfA['hair'] = np.where(dfA['hair']=='y', '1', '0')
dfA['hair'] = pd.to_numeric(dfA['hair'], errors='coerce')
dfA['exc_drink'] = np.where(dfA['exc_drink']=='y', '1', '0')
dfA['exc_drink'] = pd.to_numeric(dfA['exc_drink'],
errors='coerce')
dfA['paral_one'] = np.where(dfA['paral_one']=='y', '1', '0')
dfA['paral_one'] = pd.to_numeric(dfA['paral_one'],
errors='coerce')
dfA['eye_sunk'] = np.where(dfA['eye_sunk']=='y', '1', '0')
dfA['eye_sunk'] = pd.to_numeric(dfA['eye_sunk'], errors='coerce')
dfA['bl_orif'] = np.where(dfA['bl_orif']=='y', '1', '0')
dfA['bl_orif'] = pd.to_numeric(dfA['bl_orif'], errors='coerce')
dfA['diff_sw'] = np.where(dfA['diff_sw']=='y', '1', '0')
dfA['diff_sw'] = pd.to_numeric(dfA['diff_sw'], errors='coerce')
dfA['not_preg'] = np.where(dfA['not_preg']=='y', '1', '0')
dfA['not_preg'] = pd.to_numeric(dfA['not_preg'], errors='coerce')
dfA['pregnant'] = np.where(dfA['pregnant']=='y', '1', '0')
dfA['pregnant'] = pd.to_numeric(dfA['pregnant'], errors='coerce')
dfA['more4'] = np.where(dfA['more4']=='y', '1', '0')
dfA['more4'] = pd.to_numeric(dfA['more4'], errors='coerce')
```

```
dfA['pend_6w'] = np.where(dfA['pend_6w']=='y', '1', '0')
dfA['pend_6w'] = pd.to_numeric(dfA['pend_6w'], errors='coerce')
dfA['multip'] = np.where(dfA['multip']=='y', '1', '0')
dfA['multip'] = pd.to_numeric(dfA['multip'], errors='coerce')
dfA['cs_prev'] = np.where(dfA['cs_prev']=='y', '1', '0')
dfA['cs_prev'] = pd.to_numeric(dfA['cs_prev'], errors='coerce')
dfA['del_6wks'] = np.where(dfA['del_6wks']=='y', '1', '0')
dfA['del_6wks'] = pd.to_numeric(dfA['del_6wks'], errors='coerce')
dfA['first_p'] = np.where(dfA['first_p']=='y', '1', '0')
dfA['first_p'] = pd.to_numeric(dfA['first_p'], errors='coerce')

print(len(dfA))
print(dfA.head())

print('drop nulls...')
dfA_subA = dfA.dropna(subset=['icdmain_consensus'])
dfA_subA["icdmain_consensus"] =
dfA_subA["icdmain_consensus"].str.upper()
dfA_subA['length'] = dfA_subA['icdmain_consensus'].str.len()
dfA_subA = dfA_subA[dfA_subA.length > 2]
print(len(dfA_subA))
print(dfA_subA.head())
del dfA_subA['length']

print('recod the cause of death categories...')
dfA_subA['broad_cause_cat'] = ''
dfA_subA.loc[(dfA_subA['icdmain_consensus'] == 'B20') |
(dfA_subA['icdmain_consensus'] == 'B24')|
            (dfA_subA['icdmain_consensus'] == 'A16') |
(dfA_subA['icdmain_consensus'] == 'A15'),
            'broad_cause_cat'] = 'HIV/AIDS & TB'

dfA_subA.loc[(dfA_subA['icdmain_consensus'] == 'A40') |
(dfA_subA['icdmain_consensus'] == 'A41')|
            (dfA_subA['icdmain_consensus'] == 'J00') |
(dfA_subA['icdmain_consensus'] == 'J22')|
            (dfA_subA['icdmain_consensus'] == 'A00') |
(dfA_subA['icdmain_consensus'] == 'A09')|
            (dfA_subA['icdmain_consensus'] == 'B50') |
(dfA_subA['icdmain_consensus'] == 'B54')|
            (dfA_subA['icdmain_consensus'] == 'B05') |
(dfA_subA['icdmain_consensus'] == 'A39')|
            (dfA_subA['icdmain_consensus'] == 'G00') |
(dfA_subA['icdmain_consensus'] == 'G05')|
            (dfA_subA['icdmain_consensus'] == 'A33') |
(dfA_subA['icdmain_consensus'] == 'A35')|
            (dfA_subA['icdmain_consensus'] == 'A37') |
(dfA_subA['icdmain_consensus'] == 'A92')|
            (dfA_subA['icdmain_consensus'] == 'A99') |
(dfA_subA['icdmain_consensus'] == 'A90')|
            (dfA_subA['icdmain_consensus'] == 'A91') |
(dfA_subA['icdmain_consensus'] == 'A17')|
            (dfA_subA['icdmain_consensus'] == 'A19') |
(dfA_subA['icdmain_consensus'] == 'A20')|
            (dfA_subA['icdmain_consensus'] == 'A38') |
(dfA_subA['icdmain_consensus'] == 'A42')|
```

```
            (dfA_subA['icdmain_consensus'] == 'A89') |
(dfA_subA['icdmain_consensus'] == 'B00')|
            (dfA_subA['icdmain_consensus'] == 'B19') |
(dfA_subA['icdmain_consensus'] == 'B25')|
            (dfA_subA['icdmain_consensus'] == 'B49') |
(dfA_subA['icdmain_consensus'] == 'B55')|
            (dfA_subA['icdmain_consensus'] ==
'B99'),'broad_cause_cat'] = 'Other infectious'

dfA_subA.loc[(dfA_subA['icdmain_consensus'] == 'O00') |
(dfA_subA['icdmain_consensus'] == 'O03')|
            (dfA_subA['icdmain_consensus'] == 'O08') |
(dfA_subA['icdmain_consensus'] == 'O10')|
            (dfA_subA['icdmain_consensus'] == 'O16') |
(dfA_subA['icdmain_consensus'] == 'O46')|
            (dfA_subA['icdmain_consensus'] == 'O67') |
(dfA_subA['icdmain_consensus'] == 'O72')|
            (dfA_subA['icdmain_consensus'] == 'O63') |
(dfA_subA['icdmain_consensus'] == 'O66')|
            (dfA_subA['icdmain_consensus'] == 'O75.3') |
(dfA_subA['icdmain_consensus'] == 'O85')|
            (dfA_subA['icdmain_consensus'] == 'O99.0') |
(dfA_subA['icdmain_consensus'] == 'O71')|
            (dfA_subA['icdmain_consensus'] == 'P95') |
(dfA_subA['icdmain_consensus'] == 'P36')|
            (dfA_subA['icdmain_consensus'] == 'A33') |
(dfA_subA['icdmain_consensus'] == 'Q00')|
            (dfA_subA['icdmain_consensus'] == 'Q99') |
(dfA_subA['icdmain_consensus'] == 'P00')|
            (dfA_subA['icdmain_consensus'] == 'P04') |
(dfA_subA['icdmain_consensus'] == 'P08')|
            (dfA_subA['icdmain_consensus'] == 'P15') |
(dfA_subA['icdmain_consensus'] == 'P26')|
            (dfA_subA['icdmain_consensus'] == 'P35') |
(dfA_subA['icdmain_consensus'] == 'P37')|
            (dfA_subA['icdmain_consensus'] == 'P94') |
(dfA_subA['icdmain_consensus'] == 'P96')|
            (dfA_subA['icdmain_consensus'] == 'P05') |
(dfA_subA['icdmain_consensus'] == 'P07')|
            (dfA_subA['icdmain_consensus'] == 'P20') |
(dfA_subA['icdmain_consensus'] == 'P22')|
            (dfA_subA['icdmain_consensus'] == 'P23') |
(dfA_subA['icdmain_consensus'] == 'P25')|
            (dfA_subA['icdmain_consensus'] == 'O01') |
(dfA_subA['icdmain_consensus'] == 'O02')|
            (dfA_subA['icdmain_consensus'] == 'O20') |
(dfA_subA['icdmain_consensus'] == 'O45')|
            (dfA_subA['icdmain_consensus'] == 'O47') |
(dfA_subA['icdmain_consensus'] == 'O62')|
            (dfA_subA['icdmain_consensus'] == 'O68') |
(dfA_subA['icdmain_consensus'] == 'O70')|
            (dfA_subA['icdmain_consensus'] == 'O73') |
(dfA_subA['icdmain_consensus'] == 'O84')|
            (dfA_subA['icdmain_consensus'] == 'O86') |
(dfA_subA['icdmain_consensus'] == 'O99'),
            'broad_cause_cat'] = 'Martenal & neonatal'
```

```
dfA_subA.loc[(dfA_subA['icdmain_consensus'] == 'V01') |
(dfA_subA['icdmain_consensus'] == 'V89')|
             (dfA_subA['icdmain_consensus'] == 'V99') |
(dfA_subA['icdmain_consensus'] == 'V90')|
             (dfA_subA['icdmain_consensus'] == 'W00') |
(dfA_subA['icdmain_consensus'] == 'W19')|
             (dfA_subA['icdmain_consensus'] == 'W65') |
(dfA_subA['icdmain_consensus'] == 'W74')|
             (dfA_subA['icdmain_consensus'] == 'X00') |
(dfA_subA['icdmain_consensus'] == 'X19')|
             (dfA_subA['icdmain_consensus'] == 'X20') |
(dfA_subA['icdmain_consensus'] == 'X29')|
             (dfA_subA['icdmain_consensus'] == 'X40') |
(dfA_subA['icdmain_consensus'] == 'X49')|
             (dfA_subA['icdmain_consensus'] == 'X60') |
(dfA_subA['icdmain_consensus'] == 'X84')|
             (dfA_subA['icdmain_consensus'] == 'X85') |
(dfA_subA['icdmain_consensus'] == 'Y09')|
             (dfA_subA['icdmain_consensus'] == 'X30') |
(dfA_subA['icdmain_consensus'] == 'X39')|
             (dfA_subA['icdmain_consensus'] == 'S00') |
(dfA_subA['icdmain_consensus'] == 'T99')|
             (dfA_subA['icdmain_consensus'] == 'W20') |
(dfA_subA['icdmain_consensus'] == 'W64')|
             (dfA_subA['icdmain_consensus'] == 'W75') |
(dfA_subA['icdmain_consensus'] == 'W99')|
             (dfA_subA['icdmain_consensus'] == 'X50') |
(dfA_subA['icdmain_consensus'] == 'X59')|
             (dfA_subA['icdmain_consensus'] == 'Y10') |
(dfA_subA['icdmain_consensus'] == 'Y98'),
             'broad_cause_cat'] = 'External causes'

dfA_subA.loc[(dfA_subA['icdmain_consensus'] == 'J40') |
(dfA_subA['icdmain_consensus'] == 'J44')|
             (dfA_subA['icdmain_consensus'] == 'J45') |
(dfA_subA['icdmain_consensus'] == 'J46'),
             'broad_cause_cat'] = 'Respiratory'

dfA_subA.loc[(dfA_subA['icdmain_consensus'] == 'R10') |
(dfA_subA['icdmain_consensus'] == 'K70')|
             (dfA_subA['icdmain_consensus'] == 'K76') |
(dfA_subA['icdmain_consensus'] == 'N17')|
             (dfA_subA['icdmain_consensus'] == 'N19'),
'broad_cause_cat'] = 'Abdominal'

dfA_subA.loc[(dfA_subA['icdmain_consensus'] == 'G40') |
(dfA_subA['icdmain_consensus'] == 'G41'),
             'broad_cause_cat'] = 'Neurological'

dfA_subA.loc[(dfA_subA['icdmain_consensus'] == 'R95') |
(dfA_subA['icdmain_consensus'] == 'R99'),
             'broad_cause_cat'] = 'Indeterminate'

dfA_subA.loc[(dfA_subA['icdmain_consensus'] == 'D50') |
(dfA_subA['icdmain_consensus'] == 'D64')|
```

```
            (dfA_subA['icdmain_consensus'] == 'E40') |
(dfA_subA['icdmain_consensus'] == 'E46')|
            (dfA_subA['icdmain_consensus'] == 'E10') |
(dfA_subA['icdmain_consensus'] == 'E14')
            ,'broad_cause_cat'] = 'Metabolic'

dfA_subA.loc[(dfA_subA['icdmain_consensus'] == 'I20') |
(dfA_subA['icdmain_consensus'] == 'I25')|
            (dfA_subA['icdmain_consensus'] == 'I60') |
(dfA_subA['icdmain_consensus'] == 'I69')|
            (dfA_subA['icdmain_consensus'] == 'D57') |
(dfA_subA['icdmain_consensus'] == 'I00')|
            (dfA_subA['icdmain_consensus'] == 'I09') |
(dfA_subA['icdmain_consensus'] == 'I10')|
            (dfA_subA['icdmain_consensus'] == 'I15') |
(dfA_subA['icdmain_consensus'] == 'I26')|
            (dfA_subA['icdmain_consensus'] == 'I52') |
(dfA_subA['icdmain_consensus'] == 'I70')|
            (dfA_subA['icdmain_consensus'] ==
'I99'),'broad_cause_cat'] = 'Cardiovascular'

dfA_subA.loc[(dfA_subA['icdmain_consensus'] == 'C00') |
(dfA_subA['icdmain_consensus'] == 'C06')|
            (dfA_subA['icdmain_consensus'] == 'C15') |
(dfA_subA['icdmain_consensus'] == 'C26')|
            (dfA_subA['icdmain_consensus'] == 'C30') |
(dfA_subA['icdmain_consensus'] == 'C39')|
            (dfA_subA['icdmain_consensus'] == 'C50') |
(dfA_subA['icdmain_consensus'] == 'C51')|
            (dfA_subA['icdmain_consensus'] == 'C58') |
(dfA_subA['icdmain_consensus'] == 'C60')|
            (dfA_subA['icdmain_consensus'] == 'C63') |
(dfA_subA['icdmain_consensus'] == 'C07')|
            (dfA_subA['icdmain_consensus'] == 'C14') |
(dfA_subA['icdmain_consensus'] == 'C40')|
            (dfA_subA['icdmain_consensus'] == 'C49') |
(dfA_subA['icdmain_consensus'] == 'C48'),
            'broad_cause_cat'] = 'Neoplasms'

dfA_subA.loc[(dfA_subA['icdmain_consensus'] == 'M99') |
(dfA_subA['icdmain_consensus'] == 'N00')|
            (dfA_subA['icdmain_consensus'] == 'N16') |
(dfA_subA['icdmain_consensus'] == 'N20')|
            (dfA_subA['icdmain_consensus'] == 'N99') |
(dfA_subA['icdmain_consensus'] == 'R00')|
            (dfA_subA['icdmain_consensus'] == 'R09') |
(dfA_subA['icdmain_consensus'] == 'R11')|
            (dfA_subA['icdmain_consensus'] == 'R94') |
(dfA_subA['icdmain_consensus'] == 'D55')|
            (dfA_subA['icdmain_consensus'] == 'D89') |
(dfA_subA['icdmain_consensus'] == 'E00')|
            (dfA_subA['icdmain_consensus'] == 'E07') |
(dfA_subA['icdmain_consensus'] == 'E15')|
            (dfA_subA['icdmain_consensus'] == 'E35') |
(dfA_subA['icdmain_consensus'] == 'E50')|
```

```
            (dfA_subA['icdmain_consensus'] == 'E90') |
(dfA_subA['icdmain_consensus'] == 'F00')|
            (dfA_subA['icdmain_consensus'] == 'F99') |
(dfA_subA['icdmain_consensus'] == 'G06')|
            (dfA_subA['icdmain_consensus'] == 'G09') |
(dfA_subA['icdmain_consensus'] == 'G10')|
            (dfA_subA['icdmain_consensus'] == 'G37') |
(dfA_subA['icdmain_consensus'] == 'G50')|
            (dfA_subA['icdmain_consensus'] == 'G99') |
(dfA_subA['icdmain_consensus'] == 'H00')|
            (dfA_subA['icdmain_consensus'] == 'H95') |
(dfA_subA['icdmain_consensus'] == 'J30')|
            (dfA_subA['icdmain_consensus'] == 'J99') |
(dfA_subA['icdmain_consensus'] == 'K00')|
            (dfA_subA['icdmain_consensus'] == 'K31') |
(dfA_subA['icdmain_consensus'] == 'K35')|
            (dfA_subA['icdmain_consensus'] == 'K38') |
(dfA_subA['icdmain_consensus'] == 'K40')|
            (dfA_subA['icdmain_consensus'] == 'K93') |
(dfA_subA['icdmain_consensus'] == 'L00')|
            (dfA_subA['icdmain_consensus'] == 'L99') |
(dfA_subA['icdmain_consensus'] == 'M00')|
            (dfA_subA['icdmain_consensus'] == 'J39') |
(dfA_subA['icdmain_consensus'] == 'J47'),
            'broad_cause_cat'] = 'Other NCD'

print(dfA_subA['broad_cause_cat'].value_counts())

#keep the two categories
#dfA_subB = dfA_subA[(dfA_subA['broad_cause_cat'] == 'HIV/AIDS &
TB') |
#                    (dfA_subA['broad_cause_cat'] == 'Other
infectious')]
dfA_subB = dfA_subA[(dfA_subA['broad_cause_cat'] != '')]

print('encoding the labels in the dataset, please wait...')
#Turning labels into numbers
LE = LabelEncoder()
dfA_subB['label_num'] =
LE.fit_transform(dfA_subB['broad_cause_cat'])
print(dfA_subB.head())

# saving the dataframe
dfA_subB.to_csv('StatsTests.csv')

# saving the dataframe
dfA_subB.to_csv(r'C:\Users\a0056407\Desktop\StatsTests\file3.csv'
, index=False)

print('generating features, please wait...')
#splitting the data into features and targets
X1 =
dfA_subB[dfA_subB.columns[~dfA_subB.columns.isin(['icdmain_consen
sus','broad_cause_cat',

'label_num','id', 'disease_description'])]]
```

```
print (X1.shape)
#tfidf vectorizer with 1 and 2 grams
tfidf_vectorizer = TfidfVectorizer(ngram_range=(1, 2), min_df =
2, max_df = .95)
print('generating features using tfidf vectorizor, please
wait...')
#splitting the data into features and targets
X2 =
tfidf_vectorizer.fit_transform(dfA_subB['disease_description'].va
lues.astype(str)) #features
print (X2.shape)

X = hstack([X1, X2])

y = dfA_subB['label_num'].values #target
print (X.shape)
print(y.shape)

# from imblearn.over_sampling import RandomOverSampler
# ros = RandomOverSampler(random_state=0)
# X_resampled, y_resampled = ros.fit_resample(X, y)

print('balancing the data')
# # transform the dataset
strategy = {0:3388, 1:3388, 2:3388, 3:3388, 4:3388, 5:3388,
6:3388, 7:3388, 8:3388, 9:3388, 10:3388, 11:3388}
ros = RandomOverSampler(sampling_strategy=strategy)
X, y = ros.fit_resample(X, y)

print('reducing the features to best 100, please wait...')
#Dimenionality reduction. Only using the 100 best features
lsa = TruncatedSVD(n_components=100, n_iter=10, random_state=3)
X = lsa.fit_transform(X)
print(X.shape)



#Preliminary model evaluation using default parameters
print('generating algorithms performance before optimization,
please wait...')
#Creating a dict of the models
model_dict = {'Random Forest':
RandomForestClassifier(random_state=3),
              'NBC': MultinomialNB(),
              'Logistic Regression':LogisticRegression(),
              'K Nearest Neighbor': KNeighborsClassifier(),
              'ANN': MLPClassifier(),
              'XGBoost': XGBClassifier(),
              'DT':DecisionTreeClassifier(),
             'Support Vector Machine': SVC()}

#Train test split with stratified sampling for evaluation
X_train, X_test, y_train, y_test = train_test_split(X,
                                                    y,
```

```
                                                        test_size =
.3,
                                                        shuffle =
True,
                                                        stratify = y,
                                                        random_state
= 3)

from sklearn.preprocessing import MinMaxScaler #fixed import

scaler = MinMaxScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

#Function to get the scores for each model in a df
def model_score_df(model_dict):
    model_name, ac_score_list, p_score_list, r_score_list,
f1_score_list = [], [], [], [], []
    for k,v in model_dict.items():
        model_name.append(k)
        v.fit(X_train, y_train)
        y_pred = v.predict(X_test)
        ac_score_list.append(accuracy_score(y_test, y_pred))
        p_score_list.append(precision_score(y_test, y_pred,
average='macro'))
        r_score_list.append(recall_score(y_test, y_pred,
average='macro'))
        f1_score_list.append(f1_score(y_test, y_pred,
average='macro'))
        model_comparison_df = pd.DataFrame([model_name,
ac_score_list, p_score_list, r_score_list, f1_score_list]).T
        model_comparison_df.columns = ['model_name',
'accuracy_score', 'precision_score', 'recall_score', 'f1_score']
        model_comparison_df =
model_comparison_df.sort_values(by='f1_score', ascending=False)
    return model_comparison_df

#print(model_score_df(model_dict))
models = pd.DataFrame(model_score_df(model_dict))
print(models)

print('saving models performance, please wait...')
conn = psycopg2.connect(user="postgres", password="postgres",
                                    host="127.0.0.1", port="5432",
database="mike")

print('deleting table from database if exist')
with conn:
    cursor = conn.cursor()
    cursor.execute('drop table if exists models_performance')
    conn.commit()

print('exporting the dataframe to postgres table...')
engine =
create_engine('postgresql://postgres:postgres@localhost:5432/mike
')
```

```
models.to_sql('models_performance', engine, chunksize=1000)
print('data loaded to postgres successfully...')

#Define the best models with the selected params from the
grdsearch
lr_best_model = LogisticRegression(penalty='l2', dual=False,
tol=0.0001,
                                   C=1.0, fit_intercept=True,
intercept_scaling=1,
                                   class_weight=None,
random_state=None,
                                   solver='lbfgs',
max_iter=1000, multi_class='auto',
                                   verbose=0, warm_start=False,
n_jobs=None,
                                   l1_ratio=None)

ann_best_model = MLPClassifier(activation='relu', alpha=0.0001,
batch_size='auto', beta_1=0.9,
                               beta_2=0.999,
early_stopping=False, epsilon=1e-08,
                               hidden_layer_sizes=(50, 50, 50),
learning_rate='constant',
                               learning_rate_init=0.001,
max_iter=200, momentum=0.9,
                               nesterovs_momentum=True,
power_t=0.5, random_state=None,
                               shuffle=True, solver='adam',
tol=0.0001, validation_fraction=0.1,
                               verbose=False, warm_start=False)
dt_best_model = DecisionTreeClassifier(class_weight=None,
criterion='gini', max_depth=None,
                                       max_features=None,
max_leaf_nodes=None,
                                          min_samples_leaf=1,

min_samples_split=2, min_weight_fraction_leaf=0.0,
                                          random_state=None,
splitter='best')

xgboost_best_model =  XGBClassifier(objective="multi:softproba",
alpha=0, n_estimators=20, random_state=42, num_class = 11 ,
learning_rate=0.1,
                                 max_depth=10,
eval_metric='mlogloss')

nbc_best_model = GaussianNB(var_smoothing=2e-9)

rf_best_model =RandomForestClassifier(bootstrap=True,
ccp_alpha=0.0,
                                      class_weight=None,
                                      criterion='gini',
max_depth=None,

max_features='auto',
```

```
max_leaf_nodes=None,
                                                max_samples=None,
min_impurity_decrease=0.0,
                                                min_samples_leaf=1,
min_samples_split=2,
min_weight_fraction_leaf=0.0,
                                                n_estimators=100,
n_jobs=None,
                                                oob_score=False,
                                                random_state=None,
verbose=0,
                                                warm_start=False)

svm_best_model = SVC(C=1.0, break_ties=False, cache_size=200,
                           class_weight=None, coef0=0.0,
                           decision_function_shape='ovr',
degree=3,
                           gamma='scale', kernel='rbf',
max_iter=-1,
                           probability=False, random_state=None,
shrinking=True,
                           tol=0.001, verbose=False)

knn_best_model = KNeighborsClassifier(algorithm='auto',
leaf_size=30,
                                            metric='minkowski',
                                            metric_params=None,
n_jobs=None,
                                            n_neighbors=5, p=2,
                                            weights='uniform')
print('printing confusion matrix for LR...')
#Confusion Matrix - SGD
#Train test split with stratified sampling. Using non-binarized
labels
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size = .3, shuffle = True, stratify = y, random_state = 3)
#Fit the training data
lr_best_model.fit(X_train, y_train)
#Predict the testing data
y_pred = lr_best_model.predict(X_test)
#Get the confusion matrix and put it into a df
cm = confusion_matrix(y_test, y_pred)
cm_df = pd.DataFrame(cm, index = ['HIV/AIDS & TB','Other
infectious', 'Metabolic',
                                  'Cardiovascular',
'Indeterminate', 'Martenal & neonatal', 'Abdominal',
                                  'Neoplasms', 'External causes',
'Neurological',
                                  'Respiratory', 'Other NCD'],
                     columns = ['HIV/AIDS &
TB','Indeterminate','Other infectious', 'Metabolic',
```

```
                                    'Cardiovascular', 'Martenal &
neonatal', 'Abdominal',
                                    'Neoplasms', 'External causes',
'Neurological',
                                    'Respiratory', 'Other NCD'])
#Plot the heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(cm_df, center=0, cmap=sns.diverging_palette(220, 15,
as_cmap=True), annot=True, fmt='g')
plt.title('LR (loss = log) \nF1 Score (avg = macro) :
{0:.2f}'.format(f1_score(y_test, y_pred, average='macro')),
fontsize = 13)
plt.ylabel('True label', fontsize = 13)
plt.xlabel('Predicted label', fontsize = 13)
plt.savefig('cm_lr.png', dpi = 900)
plt.savefig('cm_lr.pdf', dpi = 900)
plt.show()

print('printing roc curve for LR...')
#Plot AUC - SGD
#Binarize the labels
y_b = label_binarize(y, classes = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9,
10, 11])
n_classes = y_b.shape[1]
#Shuffle and split training and test sets with stratified
sampling and binarized labels
X_train_b, X_test_b, y_train_b, y_test_b = train_test_split(X,
                                                            y_b,

test_size = .3,

shuffle = True,

stratify = y,

random_state = 3)

lr_classifier = OneVsRestClassifier(lr_best_model)
y_score = lr_classifier.fit(X_train_b,
y_train_b).predict(X_test_b)
# Compute ROC curve and ROC area for each class
fpr = dict()
tpr = dict()
roc_auc = dict()
for i in range(n_classes):
    fpr[i], tpr[i], _ = roc_curve(y_test_b[:, i], y_score[:, i])
    roc_auc[i] = auc(fpr[i], tpr[i])
#Compute micro-average ROC curve and ROC area
fpr['micro'], tpr['micro'], _ = roc_curve(y_test_b.ravel(),
y_score.ravel())
roc_auc['micro'] = auc(fpr['micro'], tpr['micro'])
#First aggregate all false positive rates
all_fpr = np.unique(np.concatenate([fpr[i] for i in
range(n_classes)]))
#Then interpolate all ROC curves at this points
mean_tpr = np.zeros_like(all_fpr)
```

```
for i in range(n_classes):
    mean_tpr += interp(all_fpr, fpr[i], tpr[i])
#Finally average it and compute AUC
mean_tpr /= n_classes
fpr['macro'] = all_fpr
tpr['macro'] = mean_tpr
roc_auc['macro'] = auc(fpr['macro'], tpr['macro'])
#Plot all ROC curves
plt.figure(figsize=(13,10))
sns.set_style('darkgrid')
lw=2
plt.plot(fpr['micro'],
         tpr['micro'],
         label='micro-average ROC curve (area =
{0:0.3f})'''.format(roc_auc['micro']),
         color='deeppink',
         linestyle=':',
         linewidth=4)
plt.plot(fpr['macro'],
         tpr['macro'],
         label='macro-average ROC curve (area =
{0:0.3f})'''.format(roc_auc['macro']),
         color='navy',
         linestyle=':',
         linewidth=4)
colors = cycle(['#41924F', '#FFC300', '#a98ff3', '#59C7EA',
'#9467bd', '#e377c2',
                '#1f77b4', '#ff7f0e', '#2ca02c', '#d62728',
'#8c564b'])
for i, color in zip(range(n_classes), colors):
    plt.plot(fpr[i], tpr[i],
             color=color,
             lw=lw,
             label='ROC curve of class {0} (area =
{1:0.3f})'''.format(i, roc_auc[i]))

plt.plot([0, 1], [0, 1], 'k--', lw=lw)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('1-Specificity (False Positive Rate)', fontsize = 14)
plt.ylabel('Sensitivity (True Positive Rate)', fontsize = 14)
plt.title('Receiver Operating Characteristic - LR', fontsize =
16)
plt.legend(loc="lower right", fontsize = 13)
plt.savefig('roc_lr.png', dpi = 900)
plt.savefig('roc_lr.pdf', dpi = 900)
plt.show()

#####################################################################
##################################

print('printing confusion matrix for ANN...')
#Confusion Matrix - SGD
#Train test split with stratified sampling. Using non-binarized
labels
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size = .3, shuffle = True, stratify = y, random_state = 3)
#Fit the training data
ann_best_model.fit(X_train, y_train)
#Predict the testing data
y_pred = ann_best_model.predict(X_test)
#Get the confusion matrix and put it into a df
cm = confusion_matrix(y_test, y_pred)
cm_df = pd.DataFrame(cm, index = ['HIV/AIDS & TB','Other
infectious', 'Metabolic',
                                  'Cardiovascular',
'Indeterminate', 'Martenal & neonatal', 'Abdominal',
                                  'Neoplasms', 'External causes',
'Neurological',
                                  'Respiratory', 'Other NCD'],
                     columns = ['HIV/AIDS &
TB','Indeterminate','Other infectious', 'Metabolic',
                                  'Cardiovascular', 'Martenal &
neonatal', 'Abdominal',
                                  'Neoplasms', 'External causes',
'Neurological',
                                  'Respiratory', 'Other NCD'])
#Plot the heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(cm_df, center=0, cmap=sns.diverging_palette(220, 15,
as_cmap=True), annot=True, fmt='g')
plt.title('ANN (loss = log) \nF1 Score (avg = macro) :
{0:.2f}'.format(f1_score(y_test, y_pred, average='macro')),
fontsize = 13)
plt.ylabel('True label', fontsize = 13)
plt.xlabel('Predicted label', fontsize = 13)
plt.savefig('cm_ann.png', dpi = 900)
plt.savefig('cm_ann.pdf', dpi = 900)
plt.show()

print('printing roc curve for ANN...')
#Plot AUC - RF
#Learn to predict each class against the other
ann_classifier = OneVsRestClassifier(ann_best_model)
y_score = ann_classifier.fit(X_train_b,
y_train_b).predict(X_test_b)
# Compute ROC curve and ROC area for each class
fpr = dict()
tpr = dict()
roc_auc = dict()
for i in range(n_classes):
    fpr[i], tpr[i], _ = roc_curve(y_test_b[:, i], y_score[:, i])
    roc_auc[i] = auc(fpr[i], tpr[i])
#Compute micro-average ROC curve and ROC area
fpr['micro'], tpr['micro'], _ = roc_curve(y_test_b.ravel(),
y_score.ravel())
roc_auc['micro'] = auc(fpr['micro'], tpr['micro'])
#First aggregate all false positive rates
all_fpr = np.unique(np.concatenate([fpr[i] for i in
range(n_classes)]))
#Then interpolate all ROC curves at this points
```

```
mean_tpr = np.zeros_like(all_fpr)
for i in range(n_classes):
    mean_tpr += interp(all_fpr, fpr[i], tpr[i])
#Finally average it and compute AUC
mean_tpr /= n_classes
fpr['macro'] = all_fpr
tpr['macro'] = mean_tpr
roc_auc['macro'] = auc(fpr['macro'], tpr['macro'])
#Plot all ROC curves
plt.figure(figsize=(13,10))
sns.set_style('darkgrid')
lw=2
plt.plot(fpr['micro'],
         tpr['micro'],
         label='micro-average ROC curve (area =
{0:0.3f})'''.format(roc_auc['micro']),
         color='deeppink',
         linestyle=':',
         linewidth=4)
plt.plot(fpr['macro'],
         tpr['macro'],
         label='macro-average ROC curve (area =
{0:0.3f})'''.format(roc_auc['macro']),
         color='navy',
         linestyle=':',
         linewidth=4)
colors = cycle(['#41924F', '#FFC300', '#a98ff3', '#59C7EA',
'#9467bd', '#e377c2',
                '#1f77b4', '#ff7f0e', '#2ca02c', '#d62728',
'#8c564b'])
for i, color in zip(range(n_classes), colors):
    plt.plot(fpr[i], tpr[i],
             color=color,
             lw=lw,
             label='ROC curve of class {0} (area =
{1:0.3f})'''.format(i, roc_auc[i]))

plt.plot([0, 1], [0, 1], 'k--', lw=lw)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('1-Specificity (False Positive Rate)', fontsize = 14)
plt.ylabel('Sensitivity (True Positive Rate)', fontsize = 14)
plt.title('Receiver Operating Characteristic - ANN', fontsize =
16)
plt.legend(loc="lower right", fontsize = 13)
plt.savefig('roc_ann.png', dpi = 900)
plt.savefig('roc_ann.pdf', dpi = 900)
plt.show()

#################################################################
##############################
print('printing confusion matrix for DT...')
#Confusion Matrix - SGD
#Train test split with stratified sampling. Using non-binarized
labels
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size = .3, shuffle = True, stratify = y, random_state = 3)
#Fit the training data
dt_best_model.fit(X_train, y_train)
#Predict the testing data
y_pred = dt_best_model.predict(X_test)
#Get the confusion matrix and put it into a df
cm = confusion_matrix(y_test, y_pred)
cm_df = pd.DataFrame(cm, index = ['HIV/AIDS & TB','Other
infectious', 'Metabolic',
                                  'Cardiovascular',
'Indeterminate', 'Martenal & neonatal', 'Abdominal',
                                  'Neoplasms', 'External causes',
'Neurological',
                                  'Respiratory', 'Other NCD'],
                     columns = ['HIV/AIDS &
TB','Indeterminate','Other infectious', 'Metabolic',
                                  'Cardiovascular', 'Martenal &
neonatal', 'Abdominal',
                                  'Neoplasms', 'External causes',
'Neurological',
                                  'Respiratory', 'Other NCD'])
#Plot the heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(cm_df, center=0, cmap=sns.diverging_palette(220, 15,
as_cmap=True), annot=True, fmt='g')
plt.title('DT (loss = log) \nF1 Score (avg = macro) :
{0:.2f}'.format(f1_score(y_test, y_pred, average='macro')),
fontsize = 13)
plt.ylabel('True label', fontsize = 13)
plt.xlabel('Predicted label', fontsize = 13)
plt.savefig('cm_dt.png', dpi = 900)
plt.savefig('cm_dt.pdf', dpi = 900)
plt.show()

print('printing roc curve for DT...')
#Plot AUC - RF
#Learn to predict each class against the other
dt_classifier = OneVsRestClassifier(dt_best_model)
y_score = dt_classifier.fit(X_train_b,
y_train_b).predict(X_test_b)
# Compute ROC curve and ROC area for each class
fpr = dict()
tpr = dict()
roc_auc = dict()
for i in range(n_classes):
    fpr[i], tpr[i], _ = roc_curve(y_test_b[:, i], y_score[:, i])
    roc_auc[i] = auc(fpr[i], tpr[i])
#Compute micro-average ROC curve and ROC area
fpr['micro'], tpr['micro'], _ = roc_curve(y_test_b.ravel(),
y_score.ravel())
roc_auc['micro'] = auc(fpr['micro'], tpr['micro'])
#First aggregate all false positive rates
all_fpr = np.unique(np.concatenate([fpr[i] for i in
range(n_classes)]))
#Then interpolate all ROC curves at this points
```

```
mean_tpr = np.zeros_like(all_fpr)
for i in range(n_classes):
    mean_tpr += interp(all_fpr, fpr[i], tpr[i])
#Finally average it and compute AUC
mean_tpr /= n_classes
fpr['macro'] = all_fpr
tpr['macro'] = mean_tpr
roc_auc['macro'] = auc(fpr['macro'], tpr['macro'])
#Plot all ROC curves
plt.figure(figsize=(13,10))
sns.set_style('darkgrid')
lw=2
plt.plot(fpr['micro'],
         tpr['micro'],
         label='micro-average ROC curve (area =
{0:0.3f})'''.format(roc_auc['micro']),
         color='deeppink',
         linestyle=':',
         linewidth=4)
plt.plot(fpr['macro'],
         tpr['macro'],
         label='macro-average ROC curve (area =
{0:0.3f})'''.format(roc_auc['macro']),
         color='navy',
         linestyle=':',
         linewidth=4)
colors = cycle(['#41924F', '#FFC300', '#a98ff3', '#59C7EA',
'#9467bd', '#e377c2',
                '#1f77b4', '#ff7f0e', '#2ca02c', '#d62728',
'#8c564b'])
for i, color in zip(range(n_classes), colors):
    plt.plot(fpr[i], tpr[i],
             color=color,
             lw=lw,
             label='ROC curve of class {0} (area =
{1:0.3f})'''.format(i, roc_auc[i]))

plt.plot([0, 1], [0, 1], 'k--', lw=lw)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('1-Specificity (False Positive Rate)', fontsize = 14)
plt.ylabel('Sensitivity (True Positive Rate)', fontsize = 14)
plt.title('Receiver Operating Characteristic - DT', fontsize =
16)
plt.legend(loc="lower right", fontsize = 13)
plt.savefig('roc_dt.png', dpi = 900)
plt.savefig('roc_dt.pdf', dpi = 900)
plt.show()

######################################################################
##################################
print('printing confusion matrix for XGBoost...')
#Confusion Matrix - SGD
#Train test split with stratified sampling. Using non-binarized
labels
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size = .3, shuffle = True, stratify = y, random_state = 3)
#Fit the training data
xgboost_best_model.fit(X_train, y_train)
#Predict the testing data
y_pred = xgboost_best_model.predict(X_test)
#Get the confusion matrix and put it into a df
cm = confusion_matrix(y_test, y_pred)
cm_df = pd.DataFrame(cm, index = ['HIV/AIDS & TB','Other
infectious', 'Metabolic',
                                  'Cardiovascular',
'Indeterminate', 'Martenal & neonatal', 'Abdominal',
                                  'Neoplasms', 'External causes',
'Neurological',
                                  'Respiratory', 'Other NCD'],
                     columns = ['HIV/AIDS &
TB','Indeterminate','Other infectious', 'Metabolic',
                                  'Cardiovascular', 'Martenal &
neonatal', 'Abdominal',
                                  'Neoplasms', 'External causes',
'Neurological',
                                  'Respiratory', 'Other NCD'])
#Plot the heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(cm_df, center=0, cmap=sns.diverging_palette(220, 15,
as_cmap=True), annot=True, fmt='g')
plt.title('XGBoost (loss = log) \nF1 Score (avg = macro) :
{0:.2f}'.format(f1_score(y_test, y_pred, average='macro')),
fontsize = 13)
plt.ylabel('True label', fontsize = 13)
plt.xlabel('Predicted label', fontsize = 13)
plt.savefig('cm_xgb.png', dpi = 900)
plt.savefig('cm_xgb.pdf', dpi = 900)
plt.show()

print('printing roc curve for XGBoost...')
#Plot AUC - RF
#Learn to predict each class against the other
xgb_classifier = OneVsRestClassifier(xgboost_best_model)
y_score = xgb_classifier.fit(X_train_b,
y_train_b).predict(X_test_b)
# Compute ROC curve and ROC area for each class
fpr = dict()
tpr = dict()
roc_auc = dict()
for i in range(n_classes):
    fpr[i], tpr[i], _ = roc_curve(y_test_b[:, i], y_score[:, i])
    roc_auc[i] = auc(fpr[i], tpr[i])
#Compute micro-average ROC curve and ROC area
fpr['micro'], tpr['micro'], _ = roc_curve(y_test_b.ravel(),
y_score.ravel())
roc_auc['micro'] = auc(fpr['micro'], tpr['micro'])
#First aggregate all false positive rates
all_fpr = np.unique(np.concatenate([fpr[i] for i in
range(n_classes)]))
#Then interpolate all ROC curves at this points
```

```
mean_tpr = np.zeros_like(all_fpr)
for i in range(n_classes):
    mean_tpr += interp(all_fpr, fpr[i], tpr[i])
#Finally average it and compute AUC
mean_tpr /= n_classes
fpr['macro'] = all_fpr
tpr['macro'] = mean_tpr
roc_auc['macro'] = auc(fpr['macro'], tpr['macro'])
#Plot all ROC curves
plt.figure(figsize=(13,10))
sns.set_style('darkgrid')
lw=2
plt.plot(fpr['micro'],
         tpr['micro'],
         label='micro-average ROC curve (area =
{0:0.3f})'''.format(roc_auc['micro']),
         color='deeppink',
         linestyle=':',
         linewidth=4)
plt.plot(fpr['macro'],
         tpr['macro'],
         label='macro-average ROC curve (area =
{0:0.3f})'''.format(roc_auc['macro']),
         color='navy',
         linestyle=':',
         linewidth=4)
colors = cycle(['#41924F', '#FFC300', '#a98ff3', '#59C7EA',
'#9467bd', '#e377c2',
                '#1f77b4', '#ff7f0e', '#2ca02c', '#d62728',
'#8c564b'])
for i, color in zip(range(n_classes), colors):
    plt.plot(fpr[i], tpr[i],
             color=color,
             lw=lw,
             label='ROC curve of class {0} (area =
{1:0.3f})'''.format(i, roc_auc[i]))

plt.plot([0, 1], [0, 1], 'k--', lw=lw)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('1-Specificity (False Positive Rate)', fontsize = 14)
plt.ylabel('Sensitivity (True Positive Rate)', fontsize = 14)
plt.title('Receiver Operating Characteristic - XGBoost', fontsize
= 16)
plt.legend(loc="lower right", fontsize = 13)
plt.savefig('roc_xgb.png', dpi = 900)
plt.savefig('roc_xgb.pdf', dpi = 900)
plt.show()

##################################################################
###################################

print('printing confusion matrix for NBC...')
#Confusion Matrix - SGD
#Train test split with stratified sampling. Using non-binarized
labels
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size = .3, shuffle = True, stratify = y, random_state = 3)
#Fit the training data
nbc_best_model.fit(X_train, y_train)
#Predict the testing data
y_pred = nbc_best_model.predict(X_test)
#Get the confusion matrix and put it into a df
cm = confusion_matrix(y_test, y_pred)
cm_df = pd.DataFrame(cm, index = ['HIV/AIDS & TB','Other
infectious', 'Metabolic',
                                  'Cardiovascular',
'Indeterminate', 'Martenal & neonatal', 'Abdominal',
                                  'Neoplasms', 'External causes',
'Neurological',
                                  'Respiratory', 'Other NCD'],
                     columns = ['HIV/AIDS &
TB','Indeterminate','Other infectious', 'Metabolic',
                                  'Cardiovascular', 'Martenal &
neonatal', 'Abdominal',
                                  'Neoplasms', 'External causes',
'Neurological',
                                  'Respiratory', 'Other NCD'])
#Plot the heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(cm_df, center=0, cmap=sns.diverging_palette(220, 15,
as_cmap=True), annot=True, fmt='g')
plt.title('NBC (loss = log) \nF1 Score (avg = macro) :
{0:.2f}'.format(f1_score(y_test, y_pred, average='macro')),
fontsize = 13)
plt.ylabel('True label', fontsize = 13)
plt.xlabel('Predicted label', fontsize = 13)
plt.savefig('cm_nbc.png', dpi = 900)
plt.savefig('cm_nbc.pdf', dpi = 900)
plt.show()

print('printing roc curve for NBC...')
#Plot AUC - RF
#Learn to predict each class against the other
nbc_classifier = OneVsRestClassifier(nbc_best_model)
y_score = nbc_classifier.fit(X_train_b,
y_train_b).predict(X_test_b)
# Compute ROC curve and ROC area for each class
fpr = dict()
tpr = dict()
roc_auc = dict()
for i in range(n_classes):
    fpr[i], tpr[i], _ = roc_curve(y_test_b[:, i], y_score[:, i])
    roc_auc[i] = auc(fpr[i], tpr[i])
#Compute micro-average ROC curve and ROC area
fpr['micro'], tpr['micro'], _ = roc_curve(y_test_b.ravel(),
y_score.ravel())
roc_auc['micro'] = auc(fpr['micro'], tpr['micro'])
#First aggregate all false positive rates
all_fpr = np.unique(np.concatenate([fpr[i] for i in
range(n_classes)]))
#Then interpolate all ROC curves at this points
```

```
mean_tpr = np.zeros_like(all_fpr)
for i in range(n_classes):
    mean_tpr += interp(all_fpr, fpr[i], tpr[i])
#Finally average it and compute AUC
mean_tpr /= n_classes
fpr['macro'] = all_fpr
tpr['macro'] = mean_tpr
roc_auc['macro'] = auc(fpr['macro'], tpr['macro'])
#Plot all ROC curves
plt.figure(figsize=(13,10))
sns.set_style('darkgrid')
lw=2
plt.plot(fpr['micro'],
         tpr['micro'],
         label='micro-average ROC curve (area =
{0:0.3f})'''.format(roc_auc['micro']),
         color='deeppink',
         linestyle=':',
         linewidth=4)
plt.plot(fpr['macro'],
         tpr['macro'],
         label='macro-average ROC curve (area =
{0:0.3f})'''.format(roc_auc['macro']),
         color='navy',
         linestyle=':',
         linewidth=4)
colors = cycle(['#41924F', '#FFC300', '#a98ff3', '#59C7EA',
'#9467bd', '#e377c2',
                '#1f77b4', '#ff7f0e', '#2ca02c', '#d62728',
'#8c564b'])
for i, color in zip(range(n_classes), colors):
    plt.plot(fpr[i], tpr[i],
             color=color,
             lw=lw,
             label='ROC curve of class {0} (area =
{1:0.3f})'''.format(i, roc_auc[i]))

plt.plot([0, 1], [0, 1], 'k--', lw=lw)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('1-Specificity (False Positive Rate)', fontsize = 14)
plt.ylabel('Sensitivity (True Positive Rate)', fontsize = 14)
plt.title('Receiver Operating Characteristic - NBC', fontsize =
16)
plt.legend(loc="lower right", fontsize = 13)
plt.savefig('roc_nbc.png', dpi = 900)
plt.savefig('roc_nbc.pdf', dpi = 900)
plt.show()

#####################################################################
################################
print('printing confusion matrix for KNN...')
#Confusion Matrix - KNN
#Train test split with stratified sampling. Using non-binarized
labels
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size = .3, shuffle = True, stratify = y, random_state = 3)
#Fit the training data
knn_best_model.fit(X_train, y_train)
#Predict the testing data
y_pred = knn_best_model.predict(X_test)
#Get the confusion matrix and put it into a df
cm = confusion_matrix(y_test, y_pred)
cm_df = pd.DataFrame(cm, index = ['HIV/AIDS & TB','Other
infectious','Indeterminate', 'Metabolic',
                                  'Cardiovascular', 'Martenal &
neonatal', 'Abdominal',
                                  'Neoplasms', 'External causes',
'Neurological',
                                  'Respiratory', 'Other NCD'],
                     columns = ['HIV/AIDS & TB','Other
infectious','Indeterminate', 'Metabolic',
                                  'Cardiovascular', 'Martenal &
neonatal', 'Abdominal',
                                  'Neoplasms', 'External causes',
'Neurological',
                                  'Respiratory', 'Other NCD'])
#Plot the heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(cm_df, center=0, cmap=sns.diverging_palette(220, 15,
as_cmap=True), annot=True, fmt='g')
plt.title('KNN \nF1 Score (avg = macro) :
{0:.2f}'.format(f1_score(y_test, y_pred, average='macro')),
fontsize = 20)
plt.ylabel('True label', fontsize = 18)
plt.xlabel('Predicted label', fontsize = 18)
plt.savefig('cm_knn.png', dpi = 900)
plt.savefig('cm_knn.pdf', dpi = 900)
plt.show()

print('printing roc curve for KNN...')
#Plot AUC - RF
#Learn to predict each class against the other
knn_classifier = OneVsRestClassifier(knn_best_model)
y_score = knn_classifier.fit(X_train_b,
y_train_b).predict(X_test_b)
# Compute ROC curve and ROC area for each class
fpr = dict()
tpr = dict()
roc_auc = dict()
for i in range(n_classes):
    fpr[i], tpr[i], _ = roc_curve(y_test_b[:, i], y_score[:, i])
    roc_auc[i] = auc(fpr[i], tpr[i])
#Compute micro-average ROC curve and ROC area
fpr['micro'], tpr['micro'], _ = roc_curve(y_test_b.ravel(),
y_score.ravel())
roc_auc['micro'] = auc(fpr['micro'], tpr['micro'])
#First aggregate all false positive rates
all_fpr = np.unique(np.concatenate([fpr[i] for i in
range(n_classes)]))
#Then interpolate all ROC curves at this points
```

```
mean_tpr = np.zeros_like(all_fpr)
for i in range(n_classes):
    mean_tpr += interp(all_fpr, fpr[i], tpr[i])
#Finally average it and compute AUC
mean_tpr /= n_classes
fpr['macro'] = all_fpr
tpr['macro'] = mean_tpr
roc_auc['macro'] = auc(fpr['macro'], tpr['macro'])
#Plot all ROC curves
plt.figure(figsize=(13,10))
sns.set_style('darkgrid')
lw=2
plt.plot(fpr['micro'],
         tpr['micro'],
         label='micro-average ROC curve (area =
{0:0.3f})'''.format(roc_auc['micro']),
         color='deeppink',
         linestyle=':',
         linewidth=4)
plt.plot(fpr['macro'],
         tpr['macro'],
         label='macro-average ROC curve (area =
{0:0.3f})'''.format(roc_auc['macro']),
         color='navy',
         linestyle=':',
         linewidth=4)
colors = cycle(['#41924F', '#FFC300', '#a98ff3', '#59C7EA',
'#9467bd', '#e377c2',
                '#1f77b4', '#ff7f0e', '#2ca02c', '#d62728',
'#8c564b'])
for i, color in zip(range(n_classes), colors):
    plt.plot(fpr[i], tpr[i],
             color=color,
             lw=lw,
             label='ROC curve of class {0} (area =
{1:0.3f})'''.format(i, roc_auc[i]))

plt.plot([0, 1], [0, 1], 'k--', lw=lw)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('1-Specificity (False Positive Rate)', fontsize = 14)
plt.ylabel('Sensitivity (True Positive Rate)', fontsize = 14)
plt.title('Receiver Operating Characteristic - KNN', fontsize =
16)
plt.legend(loc="lower right", fontsize = 13)
plt.savefig('roc_knn.png', dpi = 900)
plt.savefig('roc_knn.pdf', dpi = 900)
plt.show()

#####################################################################
################################


print('printing confusion matrix for SVM...')
#Confusion Matrix - SVM
```

```
#Train test split with stratified sampling. Using non-binarized
labels
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size = .3, shuffle = True, stratify = y, random_state = 3)
#Fit the training data
svm_best_model.fit(X_train, y_train)
#Predict the testing data
y_pred = svm_best_model.predict(X_test)
#Get the confusion matrix and put it into a df
cm = confusion_matrix(y_test, y_pred)
cm_df = pd.DataFrame(cm, index = ['HIV/AIDS & TB','Other
infectious', 'Metabolic',
                                  'Cardiovascular', 'Martenal &
neonatal','Indeterminate', 'Abdominal',
                                  'Neoplasms', 'External causes',
'Neurological',
                                  'Respiratory', 'Other NCD'],
                     columns = ['HIV/AIDS & TB','Other
infectious', 'Metabolic','Indeterminate',
                                  'Cardiovascular', 'Martenal &
neonatal', 'Abdominal',
                                  'Neoplasms', 'External causes',
'Neurological',
                                  'Respiratory', 'Other NCD'])
#Plot the heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(cm_df, center=0, cmap=sns.diverging_palette(220, 15,
as_cmap=True), annot=True, fmt='g')
plt.title('SVM  \nF1 Score (avg = macro) :
{0:.2f}'.format(f1_score(y_test, y_pred, average='macro')),
fontsize = 20)
plt.ylabel('True label', fontsize = 18)
plt.xlabel('Predicted label', fontsize = 18)
plt.savefig('cm_svm.png', dpi = 900)
plt.savefig('cm_svm.pdf', dpi = 900)
plt.show()

print('printing roc curve for SVM...')
#Plot AUC - RF
#Learn to predict each class against the other
svm_classifier = OneVsRestClassifier(svm_best_model)
y_score = svm_classifier.fit(X_train_b,
y_train_b).predict(X_test_b)
# Compute ROC curve and ROC area for each class
fpr = dict()
tpr = dict()
roc_auc = dict()
for i in range(n_classes):
    fpr[i], tpr[i], _ = roc_curve(y_test_b[:, i], y_score[:, i])
    roc_auc[i] = auc(fpr[i], tpr[i])
#Compute micro-average ROC curve and ROC area
fpr['micro'], tpr['micro'], _ = roc_curve(y_test_b.ravel(),
y_score.ravel())
roc_auc['micro'] = auc(fpr['micro'], tpr['micro'])
#First aggregate all false positive rates
```

```
all_fpr = np.unique(np.concatenate([fpr[i] for i in
range(n_classes)]))
#Then interpolate all ROC curves at this points
mean_tpr = np.zeros_like(all_fpr)
for i in range(n_classes):
    mean_tpr += interp(all_fpr, fpr[i], tpr[i])
#Finally average it and compute AUC
mean_tpr /= n_classes
fpr['macro'] = all_fpr
tpr['macro'] = mean_tpr
roc_auc['macro'] = auc(fpr['macro'], tpr['macro'])
#Plot all ROC curves
plt.figure(figsize=(13,10))
sns.set_style('darkgrid')
lw=2
plt.plot(fpr['micro'],
         tpr['micro'],
         label='micro-average ROC curve (area =
{0:0.3f})'''.format(roc_auc['micro']),
         color='deeppink',
         linestyle=':',
         linewidth=4)
plt.plot(fpr['macro'],
         tpr['macro'],
         label='macro-average ROC curve (area =
{0:0.3f})'''.format(roc_auc['macro']),
         color='navy',
         linestyle=':',
         linewidth=4)
colors = cycle(['#41924F', '#FFC300', '#a98ff3', '#59C7EA',
'#9467bd', '#e377c2',
                '#1f77b4', '#ff7f0e', '#2ca02c', '#d62728',
'#8c564b'])
for i, color in zip(range(n_classes), colors):
    plt.plot(fpr[i], tpr[i],
             color=color,
             lw=lw,
             label='ROC curve of class {0} (area =
{1:0.3f})'''.format(i, roc_auc[i]))

plt.plot([0, 1], [0, 1], 'k--', lw=lw)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('1-Specificity (False Positive Rate)', fontsize = 14)
plt.ylabel('Sensitivity (True Positive Rate)', fontsize = 14)
plt.title('Receiver Operating Characteristic - SVM', fontsize =
16)
plt.legend(loc="lower right", fontsize = 13)
plt.savefig('roc_svm.png', dpi = 900)
plt.savefig('roc_svm.pdf', dpi = 900)
plt.show()



##################################################################
############################
```

```
print('printing confusion matrix for RF...')
#Confusion Matrix - RF
#Fit the training data
rf_best_model.fit(X_train, y_train)
#Predict the testing data
y_pred = rf_best_model.predict(X_test)
#Get the confusion matrix and put it into a df
cm = confusion_matrix(y_test, y_pred)
cm_df = pd.DataFrame(cm, index = ['HIV/AIDS & TB','Other
infectious', 'Indeterminate','Metabolic',
                                  'Cardiovascular', 'Martenal &
neonatal', 'Abdominal',
                                  'Neoplasms', 'External causes',
'Neurological',
                                  'Respiratory', 'Other NCD'],
                     columns = ['HIV/AIDS & TB','Other
infectious', 'Indeterminate','Metabolic',
                                  'Cardiovascular', 'Martenal &
neonatal', 'Abdominal',
                                  'Neoplasms', 'External causes',
'Neurological',
                                  'Respiratory', 'Other NCD'])
#Plot the heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(cm_df, center=0, cmap=sns.diverging_palette(220, 15,
as_cmap=True), annot=True, fmt='g')
plt.title('Random Forest \nF1 Score (avg = macro) :
{0:.2f}'.format(f1_score(y_test, y_pred, average='macro')),
fontsize = 20)
plt.ylabel('True label', fontsize = 18)
plt.xlabel('Predicted label', fontsize = 18)
plt.tight_layout()
plt.savefig('cm_rf.png', dpi = 900)
plt.savefig('cm_rf.pdf', dpi = 900)
plt.show()

print('printing roc curve for RF...')
#Plot AUC - RF
#Learn to predict each class against the other
rf_classifier = OneVsRestClassifier(rf_best_model)
y_score = rf_classifier.fit(X_train_b,
y_train_b).predict(X_test_b)
# Compute ROC curve and ROC area for each class
fpr = dict()
tpr = dict()
roc_auc = dict()
for i in range(n_classes):
    fpr[i], tpr[i], _ = roc_curve(y_test_b[:, i], y_score[:, i])
    roc_auc[i] = auc(fpr[i], tpr[i])
#Compute micro-average ROC curve and ROC area
fpr['micro'], tpr['micro'], _ = roc_curve(y_test_b.ravel(),
y_score.ravel())
roc_auc['micro'] = auc(fpr['micro'], tpr['micro'])
```

```
#First aggregate all false positive rates
all_fpr = np.unique(np.concatenate([fpr[i] for i in
range(n_classes)]))
#Then interpolate all ROC curves at this points
mean_tpr = np.zeros_like(all_fpr)
for i in range(n_classes):
    mean_tpr += interp(all_fpr, fpr[i], tpr[i])
#Finally average it and compute AUC
mean_tpr /= n_classes
fpr['macro'] = all_fpr
tpr['macro'] = mean_tpr
roc_auc['macro'] = auc(fpr['macro'], tpr['macro'])
#Plot all ROC curves
plt.figure(figsize=(13,10))
sns.set_style('darkgrid')
lw=2
plt.plot(fpr['micro'],
         tpr['micro'],
         label='micro-average ROC curve (area =
{0:0.3f})'''.format(roc_auc['micro']),
         color='deeppink',
         linestyle=':',
         linewidth=4)
plt.plot(fpr['macro'],
         tpr['macro'],
         label='macro-average ROC curve (area =
{0:0.3f})'''.format(roc_auc['macro']),
         color='navy',
         linestyle=':',
         linewidth=4)
colors = cycle(['#41924F', '#FFC300', '#a98ff3', '#59C7EA',
'#9467bd', '#e377c2',
                '#1f77b4', '#ff7f0e', '#2ca02c', '#d62728',
'#8c564b'])
for i, color in zip(range(n_classes), colors):
    plt.plot(fpr[i], tpr[i],
             color=color,
             lw=lw,
             label='ROC curve of class {0} (area =
{1:0.3f})'''.format(i, roc_auc[i]))

plt.plot([0, 1], [0, 1], 'k--', lw=lw)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('1-Specificity (False Positive Rate)', fontsize = 14)
plt.ylabel('Sensitivity (True Positive Rate)', fontsize = 14)
plt.title('Receiver Operating Characteristic - RF', fontsize =
16)
plt.legend(loc="lower right", fontsize = 13)
plt.savefig('roc_rf.png', dpi = 900)
plt.savefig('roc_rf.pdf', dpi = 900)
plt.show()

##################################################################
############################
```

```
print('performing actual classification, please wait')
#Use the saved models to transform the unseen text with tf-idf
and lsa
X_unseen_x1 =
dfA[dfA.columns[~dfA.columns.isin(['icdmain_consensus','broad_cau
se_cat',

'label_num','id', 'disease_description'])]]
X_unseen_x2 =
tfidf_vectorizer.transform(dfA['disease_description'].values.asty
pe(str))

X_unseen_x = hstack([X_unseen_x1, X_unseen_x2])

X_unseen = lsa.transform(X_unseen_x)


#Fit the models with the best params on the full data
lr_best_model.fit(X, y)
rf_best_model.fit(X, y)
svm_best_model.fit(X, y)
knn_best_model.fit(X, y)
ann_best_model.fit(X, y)
dt_best_model.fit(X, y)
xgboost_best_model.fit(X, y)
nbc_best_model.fit(X, y)


#Make the prediction on the unseen articles with the fitted best
models and put it into a df alongside the correct labels
dfA['pred_lr'] = lr_best_model.predict(X_unseen)
dfA['pred_rf'] = rf_best_model.predict(X_unseen)
dfA['pred_svm'] = svm_best_model.predict(X_unseen)
dfA['pred_knn'] = knn_best_model.predict(X_unseen)
dfA['pred_xgboost'] = xgboost_best_model.predict(X_unseen)
dfA['pred_ann'] = ann_best_model.predict(X_unseen)
dfA['pred_dt'] = dt_best_model.predict(X_unseen)
dfA['pred_nbc'] = nbc_best_model.predict(X_unseen)


print('reset the index...','\n')
#dfA.set_index(['id'], inplace=True)
print(len(dfA))
print(dfA.head())

conn = psycopg2.connect(user="postgres", password="postgres",
                                    host="127.0.0.1", port="5432",
database="mike")

print('deleting table from database if exist')
with conn:
    cursor = conn.cursor()
    cursor.execute('drop table if exists ml_classified_cases2')
    conn.commit()
```

```
print('exporting the dataframe to postgres table...')
engine =
create_engine('postgresql://postgres:postgres@localhost:5432/mike
')
dfA.to_sql('ml_classified_case2', engine, chunksize=1000,
index=False)
print('data loaded to postgres successfully...')

print('halting...')
stoptd = time.strftime('%X %x %Z')
print('\n','The program stop time and Date','\n',stoptd)
print('It took', (time.time()-start)/60, 'minutes to run the
script.')


print('self clean up...')
del dfA, dfA_subA, conn, cursor, sql_cmd1, start, startptd,
stoptd
del X, X_test, X_train, X_unseen, X_unseen_tfidf, engine,
knn_best_model, y_pred
del lsa, model_dict, models, rf_best_model, nbc_best_model,
lr_best_model, xbg_best_model, ann_best_model, dt_best_model,
svm_best_model
del y_test, y_train, strategy, df, sql_cmd2, df1, df2, y,
tfidf_vectorizer
del dfA_subB, cm, cm_df, oversample
print('complete...')
```

ML Script CCVA R

```
# Chunk 1: setup
# load necessary R packages
list.of.packages <- c("knitr", "xtable", "gtools", "gridExtra",
"car" )
# identify required packages that are not already installed
new.packages <- list.of.packages[!(list.of.packages %in%
installed.packages()[,"Package"])]
# install required packages that are not already installed
if(length(new.packages)) install.packages(new.packages)
#load the required packages
library(knitr)          # knitting support
library(lubridate)      # working with dates a little easier
library(tidyverse)      # data management collection (dplyr, plurrr,
ggplot2, tidyr...)
library(readxl)         # importing excel data. Functionality absent
in tidyverse
library(caret)          # for computing confusion matrix
library(openVA)         # Composed of InSilicoVA, InterVA4,
InterVA5, Tariff packages
#library(InSilicoVA)     # InsilicoVA implementation
#library(InterVA4)       # InterVA implementation
# library and load the anaconda environment to use interface to
python
library(reticulate)
use_condaenv(condaenv = NULL,
conda = "C:/Users/dell/anaconda3/python.exe",
required = FALSE
)
# set global chunk options
knitr::opts_chunk$set(
error = FALSE,
fig.align = "center",
fig.path = "figure/",
fig.show = "hold",
include = FALSE,
message = FALSE,
echo = FALSE,
cache = TRUE,
size = "small",
tidy = TRUE
)
opts_chunk$set(warning=FALSE, message=FALSE, error=FALSE,
cache=FALSE)
opts_chunk$set(tidy=TRUE)
opts_chunk$set(cache=TRUE)
options(formatR.arrow=TRUE, width=60)
opts_chunk$set(size="small")
# read_chunk('code/VA-Algorithms.R') # put R code here with
# opts_knit$set(root.dir = "code/")
# working directory
setwd("C:/Users/a0056407/Desktop/CCVAFINAL2022/VAMachine - 22-05-
2022")
rm(list=ls())
# Chunk 2: data_import
```

```
#py_run_file('VAScript_WHO_approach.py') NOT IMPLEMENTED ANYMORE
rawdata <-
read_excel("C:/Users/a0056407/Desktop/CCVAFINAL2022/VAMachine - 22-
05-2022/MichaelData.xlsx",
progress = readxl_progress()
)
# Chunk 3
dim(rawdata)
names(rawdata) <- unlist(lapply(names(rawdata), tolower))
# Chunk 4
rawdata$id <- as.numeric(rawdata$id)
nrow(rawdata[is.na(rawdata$id),])
# Chunk 5
# filter peculiar ids to create a staging file
stagedata <- rawdata %>%
# remove missing values from id
filter(!is.na(id)) %>%
# drop duplicated id. 7 observations will be dropped from the
dataset
filter(!duplicated(id))
print(nrow(duplicated(stagedata$id)))
# Chunk 6
# introduce a new variable to unify sex
stagedata$sex <- ""
stagedata[ (stagedata$male=="y" & !is.na(stagedata$male)), c("sex")]
<- "men"
stagedata[ (stagedata$female=="y" & !is.na(stagedata$female)),
c("sex")] <- "women"
# drop missing sex value
stagedata <- stagedata[stagedata$sex=="men" |
stagedata$sex=="women", ]
# Distribution of stratifying variable `sex`
table(stagedata$sex)
# Chunk 7
data(RandomVA1)
schema <- unlist(lapply(names(RandomVA1), tolower))
# custom negation for `in`
`%nin%` = Negate(`%in%`)
# Check which column names are not likely to be in the schema
sum((schema %nin% names(rawdata)))
# Chunk 8
vars_notin_autopsy <- setdiff(schema, names(rawdata))
print(vars_notin_autopsy)
# Chunk 9
# Create missing columns
for (i in vars_notin_autopsy){
stagedata[i] <- ""
}
# Chunk 10
stagedata <- stagedata %>%
filter(stagedata$id >= 100000 & !is.na(stagedata$id))
sum(is.na(stagedata$id))
# Chunk 11
# death age
stagedata$death_age <- as.numeric(as.character(stagedata$death_age))
stagedata %>% count(death_age > 110)
```

```
# 2 obs implausible
# dod
stagedata$dod <- as.Date(as.numeric(stagedata$dod),  origin = "1900-
01-01")
stagedata$dyear <- format(stagedata$dod,"%Y")
stagedata %>% count(dyear < 1992 | dyear > 2016)
# 1 obs outlier.
# Chunk 12
stagedata <- stagedata %>%
# Remove those whose age was not captured
filter(death_age <= 110 & !is.na(death_age) ) %>%
# exclude implausible civil reg years
filter(dyear >= 1992 & dyear <= 2016)
# rename death_age to avoid confusion later
stagedata <- rename(stagedata, age = death_age)
# convert diagnoses to upper casse
stagedata$icdmain_consensus <- toupper(stagedata$icdmain_consensus)
# Chunk 13
# create a broad category for the diagnoses
stagedata$diagnose_cat <- ""
source("./VAMachine_InputA.r")
unique(stagedata$diagnose_cat)
nrow(stagedata)
# Chunk 14
stagedata <- stagedata %>%
filter(stagedata$diagnose_cat != "Unknown")
nrow(stagedata)
# Chunk 15
clinician <- stagedata %>%
select("id", "diagnose_cat", "sex", "age") %>%
filter(stagedata$diagnose_cat %in% unique(stagedata$diagnose_cat))
nrow(clinician)  # 16063
# view(as_data_frame(clinician[, c("diagnose_cat")]))
# Chunk 16
byYear <- aggregate(stagedata$dyear,
by=list(stagedata$dyear, stagedata$sex),
FUN=length
)
# proper names for the summarized data by year
names(byYear) <- c("year", "sex", "counts")
# Graph of general mortality trends over time.
# ggplot(data=byYear[byYear$year<=2016, ],   aes(x=year, y=counts,
group=sex)) +
#     theme_bw() +
#     theme(
#         plot.title=element_text(hjust=0.5),
#         axis.text.x = element_text(angle = 90),
#         legend.position = "bottom"
#           ) +
#     labs(
#         title="Mortality trends by gender (1992-2016)",
#         x="civil registration years",
#         y = "All cause mortality"
#         ) +
#     geom_line(aes(linetype=sex, color=sex)) +
#     geom_point(aes(shape=sex, color=sex)) +
```

```
#     scale_color_brewer(palette = "Dark2")
#
#
# # save graph copies both in PDf and PNG
# ggsave("./figure/mortality_trends_gender.png")
# ggsave("./figure/mortality_trends_gender.pdf")
# Chunk 17
# WHO standard age categories
stagedata$agegroup <- ""
stagedata[stagedata$age>=0 & stagedata$age<=4, c('agegroup')] <- "0-
4"      # under 5
stagedata[stagedata$age>=5 & stagedata$age<=14, c('agegroup')] <-
"5-14"    # children
stagedata[stagedata$age>=15 & stagedata$age<=49, c('agegroup')] <-
"15-49"  # reproductive years
stagedata[stagedata$age>=50 & stagedata$age<=64, c('agegroup')] <-
"50-64"  # middle age
stagedata[stagedata$age>=65 & stagedata$age<=110, c('agegroup')] <-
"65+"   # elders
# by questionnaire categories
stagedata <- stagedata %>%
mutate(agegroup2 = case_when(neonate == "y" ~ "neonate",
infant == "y" ~ "infant",
under5 == "y" ~ "under5",
child == "y" ~ "child",
adult == "y" ~ "adult",
midage == "y" ~ "midage",
elder == "y" ~ "elder"
)
)
# Chunk 18
# stagedata$cohort <- cut(stagedata$dyear, breaks = 6, right = TRUE
)
# create simpler-to-inteprete death cohort using 3-year gap
intervals apart from the period 1992-1997 (earlier years)
stagedata <- stagedata %>%
mutate(cohort = case_when(dyear>=1992 & dyear<=1997 ~ "1992-1997",
dyear>=1998 & dyear<=2000 ~ "1998-2000",
dyear>=2001 & dyear<=2003 ~ "2001-2003",
dyear>=2004 & dyear<=2006 ~ "2004-2006",
dyear>=2007 & dyear<=2009 ~ "2007-2009",
dyear>=2010 & dyear<=2012 ~ "2010-2012",
dyear>=2013 & dyear<=2016 ~ "2013-2016")
)
# Chunk 19
byCohortAgegp <- aggregate(
stagedata$agegroup,
by = list(stagedata$cohort, stagedata$agegroup),
FUN=length
)
# define proper names for the new dataset
names(byCohortAgegp) <- c("cohort", "agegroup", "counts")
# Reorder the factor variable to align legend
byCohortAgegp$agegroup <- factor(
byCohortAgegp$agegroup, levels = c("0-4", "5-14", "15-49", "50-64",
"65+"))
```

```
# Chunk 20
# ggplot(byCohortAgegp,  aes(x=cohort, y=counts))+
#     geom_col(aes(fill=agegroup), width=0.7)+
#     theme_bw() +
#     labs(
#         title="Mortality trends by age group (1992-2016)",
#         x="civil registration years",
#         y = "All cause mortality"
#         ) +
#     theme(
#         plot.title = element_text(hjust=0.5),
#         axis.text.x = element_text(angle = 0),
#         legend.position = "bottom"
#         ) +
#     scale_fill_brewer(palette="Spectral")
#
#
# # save the graphs both as pdf and png
# ggsave("./figure/mortality_trends_agegp_img1.png")
# ggsave("./figure/mortality_trends_agegp_img1.pdf")
# Chunk 21
# ggplot(byCohortAgegp,  aes(x=cohort, y=counts, fill=agegroup))+
#     geom_bar(position="fill", stat="identity", width=0.7) +
#     theme_bw() + #theme_classic() +
#     labs(
#         title="Mortality trends by age group (1992-2016)",
#         x="civil registration years",
#         y = "All cause mortality (percent)"
#         ) +
#     theme(
#         plot.title = element_text(hjust=0.5),
#         axis.text.x = element_text(angle = 0),
#         legend.position = "bottom"
#         ) +
#     scale_fill_brewer(palette="Spectral") # RdYIBu
#
#
# # save the graphs both as pdf and png
# ggsave("./figure/mortality_trends_agegp_img2.png")
# ggsave("./figure/mortality_trends_agegp_img2.pdf")
# Chunk 22
extra_vars <- c('age', 'agegroup', 'sex', 'dyear', 'cohort')
schema_extended <- append(schema, extra_vars)
stagedata <- select(stagedata, all_of(schema_extended))
# Also replace all na's to BLANK for conformity
for (i in names(stagedata)){
if (i!="id" & i %nin% extra_vars){
stagedata[is.na( stagedata[i] ), c(i)] <- ""
}
}
# Chunk 23
# Number of observation considered
nrow(stagedata)
#-------------- age distribution
# hist(stagedata$age, breaks=30) # the data are right skewed
# standard deviation
```

```
median(stagedata$age)
IQR(stagedata$age)
# Mean age at death general
mean(stagedata$age)
sd(stagedata$age)
# sex distribution
table(stagedata$sex)
#  men women
# 8354  7709
# pecentage men
# 8354 / 16063
# women
# 7709 / 16063
# Chunk 24
# Forge a schema with sub-population variables
data(RandomVA2)
schema_insilico <- unlist(lapply(names(RandomVA2), tolower))
insilico_data <- stagedata %>%
select(all_of(schema_insilico))
# Standardize all entries to Capital case to match VA standard
for (i in names(insilico_data)){
if (i!="id" & i %nin% extra_vars){
insilico_data[i] <- unlist(lapply(insilico_data[i], toupper))
}
}
# fill in data not collected i.e variable introduced
for (i in vars_notin_autopsy){
insilico_data[i] <- "."
}
# convert the tibble into dataframe to enable datachecks with the
algorithm
df_insilico_data <- as.data.frame(insilico_data)
# fit model
fit_insilico <- insilico(df_insilico_data,
data.type=c('WHO2012', 'WHO2016', 'PHMRC', 'customized')[1],
isNumeric = FALSE,
directory = "./insilico_output",
Nsim = 4000,
burnin = 2000,
thin = 10 ,
seed = 1,
datacheck = TRUE,
warning.write = TRUE,
auto.length = TRUE
)
# summary(fit_insilico)
# Chunk 25
csmfInsilico <- as.data.frame(getCSMF(fit_insilico))
plot( fit_insilico,
type = c("errorbar", "bar", "compare")[3],
top = 12,
causelist = NULL,
xlab = "Causes",
ylab = "CSMF",
title = "Top-12 population COD distribution (1992-2016)",
horiz = TRUE,
```

```
angle = 60,
fill = "gray",
err_width = 0.2,
err_size = 0.4,
point_size = 2,
border = "black",
bw = FALSE
)
# save the graphs both as pdf and png
ggsave("./figure/top12_COD_insilico.png")
ggsave("./figure/top12_COD_insilico.pdf")
# Chunk 26
interVA_data <- stagedata[, 1:246]
# All true values are already in lower case.
# Just fill in data not collected i.e variable introduced
for (i in vars_notin_autopsy){
interVA_data[i] <- NA
}
# formatting
for (i in names(interVA_data)){
if (i != "id" & i %nin% extra_vars){
interVA_data[( interVA_data[i] != "y" &
!is.na(interVA_data[i]) &
interVA_data[i] != ""
), c(i)
] <- ""
}
}
# Fit model.
# Remember, from the data, it is known that HIV prevalence is high;
# and malaria prevalence is low
fit_interva <- InterVA(interVA_data,
HIV = "h",
Malaria = "l",
directory = "./inverVA_output",
filename = "interVA_results",
output = "extended",
append = FALSE,
groupcode = FALSE,
replicate = FALSE,
replicate.bug1 = FALSE,
replicate.bug2 = FALSE,
write = TRUE
)
# Chunk 27
InterVAdf <- read.csv("./inverVA_output/interVA_results.csv") %>%
select(ID, CAUSE1)
# give meaningful names
names(InterVAdf) <- c("id", "predicted")
InterVAdf[InterVAdf$predicted==" ", "predicted"] <- 'UNK'
#names(interVA_results)
unique(InterVAdf$predicted)
# initialize the broad category column
InterVAdf["predicted_cat"] <- NA
source("./VAMachine_InputB.r")
unique(InterVAdf$predicted_cat)
```

```
# Chunk 28
csmfInterva <- as.data.frame(getCSMF(fit_interva))
# Chunk 29
byCoD <- aggregate(InterVAdf$predicted,
by=list(InterVAdf$predicted),
FUN=length
)
byCoD$CSMF <- round(byCoD$x / sum(byCoD$x), 4)
byCoD <- byCoD[order(byCoD$x, decreasing = TRUE), ]
# Chunk 30
byCoD_12 <-  byCoD %>%
filter(Group.1 != "UNK") %>%
top_n(12) %>%
mutate(Group.1 = fct_reorder(Group.1, CSMF))
# save a copy for reporting
write.csv(byCoD_12, "./inverVA_output/byCoD_12.csv")
# Chunk 31
ggplot(byCoD_12, aes(x=Group.1, y=CSMF)) +
geom_bar(stat="identity", fill="#f68060", alpha=.8, width=.6) +
coord_flip() +
theme_bw() +
# geom_text(aes(label=x)) +
labs(
title="Top 12 population cause of death distribution",
y="Cause Specific Mortality Fraction",
x = "Cause of death"
) +
theme(
plot.title = element_text(hjust=0.5, vjust=3.5),
axis.text.x = element_text(angle = 0),
legend.position = "none"
)
# save the graphs both as pdf and png
ggsave("./figure/top12_COD_interva.png")
ggsave("./figure/top12_COD_interva.pdf")
# Chunk 32
# introduce the sex variable from the clinician dataset
InterVAdfMerged <- merge(InterVAdf, clinician,
by.x="id", by.y="id",
all.x = TRUE)
# Get the top-12 causes from the general list of causes-per-id that
are in the top-12
top12 <- InterVAdfMerged[InterVAdfMerged$predicted %in%
byCoD_12$Group.1, ]
# compute stats
top12_age_gender <- top12 %>%
group_by(predicted, sex) %>%
summarise(
n=n(),
mean=mean(age),
sd=sd(age)
) %>%
mutate( se=sd/sqrt(n))  %>%
mutate( ic=se * qt((1-0.05)/2 + .5, n-1)) # %>% mutate(predicted =
fct_reorder(predicted, mean))
# save a copy for reporting
```

```
write.csv(byCoD_12, "./inverVA_output/top12_age_gender.csv")
# using standard deviation
ggplot(top12_age_gender, aes(x=predicted, y=mean, fill=sex)) +
theme_bw() +
geom_bar(stat="identity",
color="black",
position=position_dodge(),
alpha=0.5,
width=0.9) +
geom_errorbar(aes(ymin=mean-sd, ymax=mean+sd),
width=.1,
position=position_dodge(.9),
size=0.2) +
theme( plot.title = element_text(hjust=0.5),
axis.text.x = element_text(angle = 270),
legend.position = "top" ) +
labs( title="Cause specific average age at death by gender",
y="age at death",
x = "Cause of death" ) +
theme(axis.text.x = element_text(vjust = 0, hjust=0))
# save the graphs both as pdf and png
ggsave("./figure/mortality_by_gender_interva.png")
ggsave("./figure/mortality_by_gender_interva.pdf")
# Chunk 33
interva_clinician <- merge(x=InterVAdf, y=clinician,
by.x="id",
by.y = "id",
all.x = TRUE)
nrow(interva_clinician)
# Chunk 34
InterVAdf_ <- interva_clinician[, c("id", "predicted_cat")]
clinician_ <- interva_clinician[, c("id", "diagnose_cat")]
IntervaCCC <- getCCC(InterVAdf_, clinician_, C=NULL)
print(IntervaCCC)
# Chunk 35
# true CSMF
trueCSMF <- aggregate(clinician_$diagnose_cat,
by=list(clinician_$diagnose_cat),  FUN=length )
trueCSMF$csmfTrue <- round(trueCSMF$x / sum(trueCSMF$x), 4)
trueCSMF[order(trueCSMF$x, decreasing = TRUE), ]
# Chunk 36
# predicted CSMF by
predCMSF <- aggregate(InterVAdf_$predicted_cat,
by=list(InterVAdf_$predicted_cat), FUN=length)
predCMSF$csmfPred <- round(predCMSF$x / sum(predCMSF$x), 4)
predCMSF[order(predCMSF$x, decreasing = TRUE), ]
# Chunk 37
CSMF <- merge(predCMSF, trueCSMF, by="Group.1")
names(CSMF) <- c("COD", "Npredicted", "csmfPred", "NTrue",
"csmfTrue")
getCSMF_accuracy(CSMF$csmfPred, CSMF$csmfTrue)
```

Clustering Script

```
# -*- coding: utf-8 -*-
"""
Created on Tue Feb 14 04:03:06 2023

@author: a0056407
"""
import pandas as pd, time, os, numpy as np, warnings,
matplotlib.pyplot as plt, seaborn as sns
from sklearn import preprocessing
from sklearn.cluster import KMeans
from sklearn.metrics import pairwise_distances_argmin
from sklearn.mixture import GaussianMixture
from sklearn.cluster import SpectralClustering
from sklearn.cluster import AgglomerativeClustering
from scipy.cluster.hierarchy import dendrogram, linkage
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score, confusion_matrix,
accuracy_score
from yellowbrick.cluster import SilhouetteVisualizer

start = time.time()
startptd = time.strftime('%X %x %Z')
print('The program start time and Date','\n',startptd)

#seting up a few parameters
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
warnings.filterwarnings('ignore')

#Setting the working directory
os.chdir(r'C:\Users\a0056407\Desktop\ClusteringDataset\feb')

dfA =
pd.read_excel(r'C:\Users\a0056407\Desktop\ClusteringDataset\feb\Mich
aelVADataAugust2022.xlsx')
#Converting the columns headers to lower case
dfA.columns = dfA.columns.str.lower()
dfA['gender'] = 9
dfA['gender'][dfA['male'] == 'y'] = 1
dfA['gender'][dfA['female'] == 'y'] = 0
print(dfA['gender'].value_counts())
print(dfA['male'].value_counts())
print(dfA['female'].value_counts())
dfA["dod"] = pd.to_datetime(dfA["dod"])
dfA["yod"] = dfA['dod'].map(lambda x: x.year)
print(dfA.head())

dfB = pd.concat([dfA['household'], dfA['death_age'], dfA['gender'],
dfA['yod']], axis = 1)
print(dfB.describe())
print(dfB.info())
print(len(dfB['household'].unique()))
print(len(dfB['death_age'].unique()))
print(len(dfB['gender'].unique()))
print(len(dfB['yod'].unique()))
```

```
#age groups
bins= [0,4,14,49,64,110]
labels = ['0-4','5-14','15-49','50-64','65+']
dfB['age_group'] = pd.cut(dfB['death_age'], bins=bins,
labels=labels, right=False)
print(dfB.head())

print(dfB.groupby(['gender'])['death_age'].mean().round().to_frame()
)

sns.pairplot(dfB)
plt.savefig("pair_plot.png", dpi=1200)
plt.savefig("pair_plot.pdf", dpi=1200)
plt.show()

fig = plt.figure(figsize=(15,6))
plt.ylabel('Year of death', fontsize = 18)
plt.xlabel('Age at death', fontsize = 18)
# Basic 2D density plot
sns.kdeplot(x=dfB.death_age, y=dfB.yod)
fig.savefig("age_yod_kdeplot.png", dpi=1200)
fig.savefig("age_yod_kdeplot.pdf", dpi=1200)
plt.show()

fig = plt.figure(figsize=(15,6))
plt.ylabel('Year of death', fontsize = 18)
plt.xlabel('Age at death', fontsize = 18)
# Custom the color, add shade and bandwidth
sns.kdeplot(x=dfB.death_age, y=dfB.yod, cmap="Reds", shade=True,
bw_adjust=.5)
fig.savefig("age_yod_kdeplot_II.png", dpi=1200)
fig.savefig("age_yod_kdeplot_II.pdf", dpi=1200)
plt.show()

fig = plt.figure(figsize=(15,6))
plt.ylabel('Year of death', fontsize = 18)
plt.xlabel('Age at death', fontsize = 18)
sns.kdeplot(x=dfB.death_age, y=dfB.yod, cmap="Blues", shade=True,
thresh=0)
fig.savefig("age_yod_kdeplot_III.png", dpi=1200)
fig.savefig("age_yod_kdeplot_III.pdf", dpi=1200)
plt.show()

fig = plt.figure(figsize=(15,6))
# Use the 'hue' argument to provide a factor variable
sns.lmplot( x="death_age", y="yod", data=dfB, fit_reg=False,
hue='gender', legend=False)
plt.legend(loc='lower right')
plt.ylabel('Year of death', fontsize = 18)
plt.xlabel('Age at death', fontsize = 18)
fig.savefig("age_yod_lmplot.png", dpi=1200)
fig.savefig("age_yod_lmplot.pdf", dpi=1200)
plt.show()

fig = plt.figure(figsize=(15,6))
plt.ylabel('Year of death', fontsize = 18)
```

```
plt.xlabel('Age at death', fontsize = 18)
plt.plot('death_age', 'yod', data=dfB, linestyle='-', marker='o')
fig.savefig("age_yod_connected_scatter.png", dpi=1200)
fig.savefig("age_yod_connected_scatter.pdf", dpi=1200)
plt.show()


fig = plt.figure(figsize=(15,6))
dfB["death_age"].hist()
plt.title('Age at the time of death', fontsize = 18)
plt.ylabel('Counts', fontsize = 18)
plt.xlabel('Age at death', fontsize = 18)
plt.tight_layout()
fig.savefig("age_at_death_histogram.png", dpi=1200)
fig.savefig("age_at_death_histogram.pdf", dpi=1200)
plt.show()

# KDE plots for each species
sns.kdeplot(data=dfB['death_age'], label="age distribution",
shade=True)
# Add title
#plt.title("Distribution of age at the time of cancer diagnosis")
plt.xlabel('Age at the time of deaths')
plt.ylabel('Densities weights')
plt.savefig('age_at_death_density.png', dpi = 1200)
plt.savefig('age_at_death_density.pdf', dpi = 1200)
plt.show()

fig = plt.figure(figsize=(15,6))
sns.countplot(x="age_group", hue="gender", data=dfB)
plt.title('Age groups counts per gender', fontsize = 18)
plt.ylabel('Counts', fontsize = 18)
plt.xlabel('Age groups', fontsize = 18)
plt.tight_layout()
fig.savefig("age_group_gender_countplot.png", dpi=1200)
fig.savefig("age_group_gender_countplot.pdf", dpi=1200)
plt.show()

fig = plt.figure(figsize=(15,6))
sns.countplot(x="yod", hue="age_group", data=dfB)
plt.title('Death counts per year per age group', fontsize = 18)
plt.ylabel('Counts', fontsize = 18)
plt.xlabel('Year of death', fontsize = 18)
plt.tight_layout()
fig.savefig("age_group_yod_countplot.png", dpi=1200)
fig.savefig("age_group_yod_countplot.pdf", dpi=1200)
plt.show()

fig = plt.figure(figsize=(15,6))
sns.catplot(x="yod", hue="age_group", col="gender",data=dfB,
kind="count")
plt.title('Death counts per year per age group per gender', fontsize
= 18)
plt.ylabel('Counts', fontsize = 18)
plt.xlabel('Year of death', fontsize = 18)
plt.tight_layout()
```

```
fig.savefig("age_group_yod_gender_catplot.png", dpi=1200)
fig.savefig("age_group_yod_gender_catplot.pdf", dpi=1200)
plt.show()

fig = plt.figure(figsize=(15,6))
sns.distplot(dfB['yod'])
plt.title('Density plot for deaths over the years', fontsize = 18)
plt.ylabel('Density counts', fontsize = 18)
plt.xlabel('Year of death', fontsize = 18)
plt.tight_layout()
fig.savefig("yod_distplot.png", dpi=1200)
fig.savefig("yod_distplot.pdf", dpi=1200)
plt.show()


fig = plt.figure(figsize=(15,6))
sns.countplot(x='gender',data=dfB)
plt.title('Gender counts', fontsize = 18)
plt.ylabel('Counts', fontsize = 18)
plt.xlabel('Gender', fontsize = 18)
plt.tight_layout()
fig.savefig("gender_countplot.png", dpi=1200)
fig.savefig("gender_countplot.pdf", dpi=1200)
plt.show()

fig = plt.figure(figsize=(15,6))
sns.countplot(x="yod", hue="gender", data=dfB)
plt.title('Distribution of gender for each year of death', fontsize
= 18)
plt.ylabel('Counts', fontsize = 18)
plt.xlabel('Year of death', fontsize = 18)
plt.tight_layout()
fig.savefig("yod_gender_countplot.png", dpi=1200)
fig.savefig("yod_gender_countplot.pdf", dpi=1200)
plt.show()

fig = plt.figure(figsize=(15,6))
sns.boxplot(x="gender", y="death_age", data=dfB, showmeans=True)
plt.title('Box plot for age at death per gender', fontsize = 18)
plt.ylabel('Age at death', fontsize = 18)
plt.xlabel('Gender', fontsize = 18)
plt.tight_layout()
fig.savefig("age_gender_boxplot.png", dpi=1200)
fig.savefig("age_gender_boxplot.pdf", dpi=1200)
plt.show()


fig = plt.figure(figsize=(15,6))
sns.boxplot(x="yod", y="death_age", data=dfB, showmeans=True)
plt.title('Box plot for age at death per year', fontsize = 18)
plt.ylabel('Age at death', fontsize = 18)
plt.xlabel('Year of death', fontsize = 18)
plt.tight_layout()
fig.savefig("age_yod_boxplot.png", dpi=1200)
fig.savefig("age_yod_boxplot.pdf", dpi=1200)
plt.show()
```

```
fig = plt.figure(figsize=(15,6))
sns.lineplot(x='yod', y='death_age', data=dfB)
plt.title('Distribution of age at death', fontsize = 18)
plt.ylabel('Age at death', fontsize = 18)
plt.xlabel('Year of death', fontsize = 18)
plt.tight_layout()
fig.savefig("age_yod_lineplot.png", dpi=1200)
fig.savefig("age_yod_lineplot.pdf", dpi=1200)
plt.show()

fig = plt.figure(figsize=(15,6))
sns.lineplot(x = 'yod', y = 'death_age', data = dfB, hue = 'gender',
style = 'gender')
plt.title('Distribution of age at death per gender', fontsize = 18)
plt.ylabel('Age at death', fontsize = 18)
plt.xlabel('Year of death', fontsize = 18)
plt.tight_layout()
fig.savefig("age_yod_gender_lineplot.png", dpi=1200)
fig.savefig("age_yod_gender_lineplot.pdf", dpi=1200)
plt.show()

dfC = pd.concat([dfA['household'], dfA['death_age'], dfA['gender'],
dfA['yod']], axis = 1)
#calling the dataset
scaler = preprocessing.MaxAbsScaler().fit(dfC)
#scaler = preprocessing.PowerTransformer(method='box-cox',
standardize=False)
print(scaler)
#print(scaler.mean_)
print(scaler.scale_)
#getting the scalled data
X_scaled = scaler.transform(dfC)
print(X_scaled)
#mean and std
print(X_scaled.mean(axis=0))
print(X_scaled.std(axis=0))


#####################################################################
##############
#kMeans clusttering
#####################################################################
##############
inertias = []

for i in range(1,11):
    kmeans = KMeans(n_clusters=i)
    kmeans.fit(X_scaled)
    inertias.append(kmeans.inertia_)
    #print(kmeans.inertia_)
    #print(kmeans.cluster_centers_)
    #print(kmeans.n_iter_)
fig = plt.figure(figsize=(15,6))
plt.plot(range(1,11), inertias, marker='o')
```

```
plt.title('Elbow Joint')
plt.xlabel('Number of clusters')
plt.ylabel('Inertia')
plt.tight_layout()
fig.savefig("elbow_joint.png", dpi=1200)
fig.savefig("elbow_joint.pdf", dpi=1200)
plt.show()


kmeans = KMeans(n_clusters=5)
kmeans.fit(X_scaled)
y_kmeans = kmeans.predict(X_scaled)
fig = plt.figure(figsize=(15,6))
plt.scatter(X_scaled[:, 0], X_scaled[:, 1], c=y_kmeans, s=50,
cmap='viridis')
centers = kmeans.cluster_centers_
plt.scatter(centers[:, 0], centers[:, 1], c='black', s=200,
alpha=0.5)
plt.tight_layout()
fig.savefig("kmeans_clustters.png", dpi=1200)
fig.savefig("kmeans_clustters.pdf", dpi=1200)
plt.show()

#getting the scores
ac = accuracy_score(y_kmeans, kmeans.labels_)
print('Accuracy Score: %.3f' % ac)
score = silhouette_score(X_scaled, kmeans.labels_,
metric='euclidean')
print('KMeans Clusttering Silhouetter Score: %.3f' % score)

#silhouette visualizer for kmeans
fig, ax = plt.subplots(2, 2, figsize=(15,8))
for i in [2, 3, 4, 5]:
    #Create KMeans instance for different number of clusters
    km = KMeans(n_clusters=i, init='k-means++', n_init=10,
max_iter=100, random_state=42)
    q, mod = divmod(i, 2)
    #Create SilhouetteVisualizer instance with KMeans instance Fit
the visualizer
    visualizer = SilhouetteVisualizer(km, colors='yellowbrick',
ax=ax[q-1][mod])
    visualizer.fit(X_scaled)
#plt.xlabel('Predicted label', fontsize=18)
#plt.ylabel('True label', fontsize=18)
#plt.title('Silhouette plots for 2, 3, 4, and 5 clusters',
fontsize=18)
plt.tight_layout()
fig.savefig("kmeans_silhouette_plots.png", dpi=1200)
fig.savefig("kmeans_silhouette_plots.pdf", dpi=1200)
plt.show()


#heatmap
conf_matrix=confusion_matrix(y_kmeans, kmeans.labels_)
fig, ax = plt.subplots(figsize=(15, 15))
ax.matshow(conf_matrix, cmap=plt.cm.Blues, alpha=0.3)
```

```
for i in range(conf_matrix.shape[0]):
    for j in range(conf_matrix.shape[1]):
        ax.text(x=j, y=i,s=conf_matrix[i, j], va='center',
ha='center', size='xx-large')
plt.xlabel('Predicted label', fontsize=18)
plt.ylabel('True label', fontsize=18)
plt.title('Confusion Matrix', fontsize=18)
plt.tight_layout()
fig.savefig("kmeans_heatmap.png", dpi=1200)
fig.savefig("kmeans_heatmap.pdf", dpi=1200)
plt.show()



def find_clusters(X, n_clusters, rseed=2):
    # 1. Randomly choose clusters
    rng = np.random.RandomState(rseed)
    i = rng.permutation(X.shape[0])[:n_clusters]
    centers = X[i]
    while True:
        # 2a. Assign labels based on closest center
        labels = pairwise_distances_argmin(X, centers)
        # 2b. Find new centers from means of points
        new_centers = np.array([X[labels == i].mean(0) for i in
range(n_clusters)])
        # 2c. Check for convergence
        if np.all(centers == new_centers):
            break
        centers = new_centers
    return centers, labels

centers, labels = find_clusters(X_scaled, 4)
fig = plt.figure(figsize=(15,6))
plt.scatter(X_scaled[:, 0], X_scaled[:, 1], c=labels, s=50,
cmap='viridis')
fig.savefig("kmeans_clustters_pairwise_distances.png", dpi=1200)
fig.savefig("kmeans_clustters_pairwise_distances.pdf", dpi=1200)
plt.show()


#####################################################################
##############
#Spectral clusttering
#####################################################################
##############
model = SpectralClustering(n_clusters=5,
affinity='nearest_neighbors', assign_labels='kmeans')
labels = model.fit_predict(X_scaled)
fig = plt.figure(figsize=(15,6))
plt.scatter(X_scaled[:, 0], X_scaled[:, 1], c=labels, s=50,
cmap='viridis')
fig.savefig("spectral_clustering.png", dpi=1200)
fig.savefig("spectral_clustering.pdf", dpi=1200)
plt.show()

#getting the scores
```

```
score = silhouette_score(X_scaled, labels, metric='euclidean')
print('Spectral Clustering Silhouetter Score: %.3f' % score)

########################################################################
##############
#GMM clusttering
########################################################################
##############
gmm = GaussianMixture(n_components=5, covariance_type='full',
random_state=42).fit(X_scaled)
labels = gmm.predict(X_scaled)
fig = plt.figure(figsize=(15,6))
plt.scatter(X_scaled[:, 0], X_scaled[:, 1], c=labels, s=40,
cmap='viridis')
fig.savefig("gmm_clustering.png", dpi=1200)
fig.savefig("gmm_clustering.pdf", dpi=1200)
plt.show()

#getting the scores
score = silhouette_score(X_scaled, labels, metric='euclidean')
print('Gaussian Mixture Clustering Silhouetter Score: %.3f' % score)

########################################################################
##############
#Agglomerative clusttering
########################################################################
##############
agglo = AgglomerativeClustering(n_clusters=5, affinity='euclidean',
linkage='ward')
labels = agglo.fit_predict(X_scaled)
fig = plt.figure(figsize=(15,6))
plt.scatter(X_scaled[:, 0], X_scaled[:, 1], c=labels)
fig.savefig("agglomerative_clustering.png", dpi=1200)
fig.savefig("agglomerative_clustering.pdf", dpi=1200)
plt.show()

#getting the scores
score = silhouette_score(X_scaled, labels, metric='euclidean')
print('Agglomerative Clustering Silhouetter Score: %.3f' % score)

########################################################################
##############
#Hierarchical clusttering
########################################################################
##############
fig = plt.figure(figsize=(15,6))
linkage_data = linkage(X_scaled, method='ward', metric='euclidean')
dendrogram(linkage_data)
fig.savefig("dendrogram_clustering.png", dpi=1200)
fig.savefig("dendrogram_clustering.pdf", dpi=1200)
plt.show()

print('halting...')
stoptd = time.strftime('%X %x %Z')
print('\n','The program stop time and Date','\n',stoptd)
```

```
print('It took', (time.time()-start)/60, 'minutes to run the
script.')
```

Other scritps are available on the following repositories:

https://drive.google.com/drive/folders/1nQhZE2t2s1AI2QQthZCCK5e5Tg-gjp-W?usp=drive_link

michaelmaps/PhD_ScriptsMM: This my repository for all my PhD work (github.com)

References

[1] Serena Jeblee et al. "Automatically determining cause of death from verbal autopsy narratives". In: *BMC medical informatics and decision making* 19.1 (2019), p. 127.

[2] Erin K Nichols et al. "The WHO 2016 verbal autopsy instrument: an international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0". In: *PLoS medicine* 15.1 (2018), e1002486.

[3] Lisa-Marie Thomas, Lucia D'Ambruoso, and Dina Balabanova. "Verbal autopsy in health policy and systems: a literature review". In: *BMJ global health* 3.2 (2018), e000639.

[4] Nadia Soleman, Daniel Chandramohan, and Kenji Shibuya. "Verbal autopsy: current practices and challenges". In: *Bulletin of the World Health Organization* 84 (2006), pp. 239–245.

[5] Chabila Mapoma et al. "Ascertaining Cause of non-Facility Deaths and Associated Demographic Characteristics An Application of Verbal Autopsy in Routine Civil Registration in Lusaka District of Zambia". In: (2021).

[6] Elizabeth D Paratz et al. "A systematic review of global autopsy rates in all-cause mortality and young sudden death". In: *Heart Rhythm* (2023).

[7] Rafael Lozano et al. "Performance of physician-certified verbal autopsies: multisite validation study using clinical diagnostic gold standards". In: *Population Health Metrics* 9.1 (2011), pp. 1–13.

[8] Ghulam Mujtaba et al. "Clinical text classification research trends: systematic literature review and open issues". In: *Expert systems with applications* 116 (2019), pp. 494–520.

[9] Riccardo Miotto et al. "Deep learning for healthcare: review, opportunities and challenges". In: *Briefings in bioinformatics* 19.6 (2017), pp. 1236–1246.

[10] Barnaby C Reeves and Maria Quigley. "A review of data-derived methods for assigning causes of death from verbal autopsy data." In: *International journal of epidemiology* 26.5 (1997), pp. 1080–1089.

[11] Ghulam Mujtaba et al. "Automatic text classification of ICD-10 related CoD from complex and free text forensic autopsy reports". In: *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*. IEEE. 2016, pp. 1055–1058.

[12] Samuel Danso, Eric Atwell, and Owen Johnson. "A comparative study of machine learning methods for verbal autopsy text classification". In: *arXiv preprint arXiv:1402.4380* (2014).

[13] Suresh Kumar and Shivani Goel. "Enhancing Text Classification by Stochastic Optimization method and Support Vector Machine". In: *International Journal of Computer Science and Information Technologies* 6.4 (2015).

[14] Vandana Korde and C Namrata Mahender. "Text classification and classifiers: A survey". In: *International Journal of Artificial Intelligence & Applications* 3.2 (2012), p. 85.

[15] Ghulam Mujtaba et al. "Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study". In: *Journal of forensic and legal medicine* 57 (2018), pp. 41–50.

[16] Abraham D Flaxman et al. "Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards". In: *Population health metrics* 9.1 (2011), p. 29.

[17] Pierre Miasnikof et al. "Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths". In: *BMC medicine* 13.1 (2015), p. 286.

[18] Ghulam Mujtaba et al. “Classification of forensic autopsy reports through conceptual graph-based document representation model”. In: *Journal of biomedical informatics* 82 (2018), pp. 88–105.

[19] Andrew Boulle, Daniel Chandramohan, and Peter Weller. “A case study of using artificial neural networks for classifying cause of death from verbal autopsy”. In: *International journal of epidemiology* 30.3 (2001), pp. 515–520.

[20] John Pestian et al. “Suicide note classification using natural language processing: A content analysis”. In: *Biomedical informatics insights* 3 (2010), BII–S4706.

[21] Serena Jeblee, Mireille Gomes, and Graeme Hirst. “Multi-task learning for interpretable cause of death classification using key phrase prediction”. In: *Proceedings of the BioNLP 2018 workshop*. 2018, pp. 12–17.

[22] Tom Young et al. “Recent trends in deep learning based natural language processing”. In: *arXiv preprint arXiv:1708.02709* (2017).

[23] Omran Abdel et al. “The epidemiologic transition: a theory of the epidemiology of population change”. In: *Milbank Memorial Fund Quarterly* 49.4 (1971), pp. 509–538.

[24] Abdel R Omran. “The epidemiologic transition: a theory of the epidemiology of population change”. In: *The Milbank Quarterly* 83.4 (2005), pp. 731–757.

[25] Stuart Jay Olshansky and A Brian Ault. “The fourth stage of the epidemiologic transition: the age of delayed degenerative diseases”. In: (1986).

[26] Chodziwadziwa W Kabudula et al. “Progression of the epidemiological transition in a rural South African setting: findings from population surveillance in Agincourt, 1993–2013”. In: *BMC public health* 17.1 (2017), p. 424.

[27] Julio Frenk et al. “Health transition in middle-income countries: new challenges for health care”. In: *Health policy and planning* 4.1 (1989), pp. 29–39.

[28] Jordana Leitao et al. “Comparison of physician-certified verbal autopsy with computer-coded verbal autopsy for cause of death assignment in hospitalized patients in low-and middle-income countries: systematic review”. In: *BMC medicine* 12.1 (2014), p. 22.

[29] Sean T Green and Abraham D Flaxman. “Machine Learning Methods for Verbal Autopsy in Developing Countries.” In: *AAAI Spring Symposium: Artificial Intelligence for Development*. 2010.

[30] Tyler H McCormick et al. “Probabilistic cause-of-death assignment using verbal autopsies”. In: *Journal of the American Statistical Association* 111.515 (2016), pp. 1036–1049.

[31] Edward Fottrell and Peter Byass. “Verbal autopsy: methods in transition”. In: *Epidemiologic reviews* 32.1 (2010), pp. 38–55.

[32] Min Chen et al. “Disease prediction by machine learning over big data from healthcare communities”. In: *IEEE Access* 5 (2017), pp. 8869–8879.

[33] Sharathkumar Anbu and Bhaskarjit Sarmah. “Machine learning approach for predicting womens health risk”. In: *Advanced Computing and Communication Systems (ICACCS), 2017 4th International Conference on*. IEEE. 2017, pp. 1–4.

[34] Victor Mudenda et al. “Tuberculosis and Tuberculosis/HIV/AIDS–Associated Mortality in Africa: The Urgent Need to Expand and Invest in Routine and Research Autopsies”. In: *Journal of Infectious Diseases* 205.suppl_2 (2012), S340–S346.

[35] B Nithya and V Ilango. “Predictive analytics in health care using machine learning tools and techniques”. In: *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE. 2017, pp. 492–499.

[36] Zahid Iqbal et al. “A comparative study of machine learning techniques used in non-clinical systems for continuous healthcare of independent livings”. In: *2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*. IEEE. 2018.

[37] Anton Kocheturov, Panos M Pardalos, and Athanasia Karakitsiou. "Massive datasets and machine learning for computational biomedicine: trends and challenges". In: *Annals of Operations Research* (2018), pp. 1–30.

[38] Lei Liu. "Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning". In: *2018 International Conference on Robots & Intelligent System (ICRIS)*. IEEE. 2018.

[39] Ivo D Dinov et al. "Predictive big data analytics: a study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations". In: *PloS one* 11.8 (2016), e0157077.

[40] Abraham D Flaxman and Theo Vos. "Machine learning in population health: Opportunities and threats". In: *PLoS medicine* 15.11 (2018).

[41] Zhaodong Yan, Serena Jeblee, and Graeme Hirst. "Can Character Embeddings Improve Cause-of-Death Classification for Verbal Autopsy Narratives?" In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. 2019, pp. 234–239.

[42] Vianney Jouhet et al. "Automated classification of free-text pathology reports for registration of incident cases of cancer". In: *Methods of information in medicine* 51.3 (2012), p. 242.

[43] Wei Liang Yeow, Rohana Mahmud, and Ram Gopal Raj. "An application of case-based reasoning with machine learning for forensic autopsy". In: *Expert Systems with Applications* 41.7 (2014), pp. 3497–3505.

[44] Ghulam Mujtaba et al. "Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection". In: *PloS one* 12.2 (2017), e0170242.

[45] Christopher JL Murray et al. "Using verbal autopsy to measure causes of death: the comparative performance of existing methods". In: *BMC medicine* 12.1 (2014), pp. 1–19.

[46] Maria A Quigley, Daniel Chandramohan, and Laura C Rodrigues. "Diagnostic accuracy of physician review, expert algorithms and data-derived algorithms in adult verbal autopsies." In: *International Journal of Epidemiology* 28.6 (1999), pp. 1081–1087.

[47] Mathew A Mwanyangala et al. "Verbal autopsy completion rate and factors associated with undetermined cause of death in a rural resource-poor setting of Tanzania". In: *Population health metrics* 9.1 (2011), p. 41.

[48] Noriah Maraba et al. "Verbal autopsy-assigned causes of death among adults being investigated for TB in South Africa". In: *Transactions of the Royal Society of Tropical Medicine and Hygiene* 110.9 (2016), pp. 510–516.

[49] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[50] Juan D Rodriguez, Aritz Perez, and Jose A Lozano. "Sensitivity analysis of k-fold cross validation in prediction error estimation". In: *IEEE transactions on pattern analysis and machine intelligence* 32.3 (2010), pp. 569–575.

[51] Khaled Fawagreh, Mohamed Medhat Gaber, and Eyad Elyan. "Random forests: from early developments to recent advancements". In: *Systems Science & Control Engineering: An Open Access Journal* 2.1 (2014), pp. 602–609.

[52] Nuntaporn Klinjun, Apiradee Lim, and Kanitta Bundhamcharoen. "A logistic regression model for estimating transport accident deaths using verbal autopsy data". In: *Asia Pacific Journal of Public Health* 27.3 (2015), pp. 286–292.

[53] Kunal Rajput, Girija Chetty, and Rachel Davey. "PHIs (Protected Health Information) identification from free text clinical records based on machine learning". In: *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2017, pp. 1–9.

[54] Syed Shariyar Murtaza et al. "Automated verbal autopsy classification: using one-against-all ensemble method and Naïve Bayes classifier". In: *Gates open research* 2 (2018).

[55] William H Crown. "Potential application of machine learning in health outcomes research and some statistical cautions". In: *Value in health* 18.2 (2015), pp. 137–140.

[56] Ekkarat Boonchieng and Khanita Duangchaemkarn. "Digital disease detection: Application of machine learning in community health informatics". In: *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE. 2016, pp. 1–5.

[57] Bevan Koopman et al. "Extracting cancer mortality statistics from death certificates: A hybrid machine learning and rule-based approach for common and rare cancers". In: *Artificial intelligence in medicine* (2018).

[58] Indrajani Sutedja, Hendro Nindito, Azani Cempaka Sari, et al. "Building a mobile health application and heart diagnose to assist patients: Analysis and design". In: *Information Management and Technology (ICIMTech), 2017 International Conference on*. IEEE. 2017, pp. 19–24.

[59] Khaled Alsabti, Sanjay Ranka, and Vineet Singh. "An efficient k-means clustering algorithm". In: (1997).

[60] Frederick Livingston. "Implementation of Breiman's random forest machine learning algorithm". In: *ECE591Q Machine Learning Journal Paper* (2005).

[61] Jiafan Peng et al. "Application of machine learning method in bridge health monitoring". In: *Reliability Systems Engineering (ICRSE), 2017 Second International Conference on*. IEEE. 2017, pp. 1–7.

[62] Matthew D Byrne. "Machine Learning in Health Care". In: *Journal of PeriAnesthesia Nursing* 32.5 (2017), pp. 494–496.

[63] Bevan Koopman et al. "Automatic classification of diseases from free-text death certificates for real-time surveillance". In: *BMC medical informatics and decision making* 15.1 (2015), p. 53.

[64] Robert E Schapire. "Using output codes to boost multiclass learning problems". In: *ICML*. Vol. 97. Citeseer. 1997, pp. 313–321.

[65] Samuel J Clark. "A Guide to Comparing the Performance of VA Algorithms". In: *arXiv preprint arXiv:1802.078* (2018).

[66] Samuel J Clark, Zehang Li, and Tyler H McCormick. "Quantifying the Contributions of Training Data and Algorithm Logic to the Performance of Automated Cause-assignment Algorithms for Verbal Autopsy". In: *arXiv preprint arXiv:1803.07141* (2018).

[67] Spencer L James, Abraham D Flaxman, and Christopher JL Murray. "Performance of the Tariff Method: validation of a simple additive algorithm for analysis of verbal autopsies". In: *Population Health Metrics* 9.1 (2011), p. 31.

[68] Zehang Richard Li, Tyler H McCormick, and Samuel J Clark. "Non-confirming replication of "Performance of InSilicoVA for assigning causes of death to verbal autopsies: multisite validation study using clinical diagnostic gold standards," by Flaxman et al." In: *BMC medicine* 18.1 (2020), pp. 1–5.

[69] Nikita Desai et al. "Performance of four computer-coded verbal autopsy methods for cause of death assignment compared with physician coding on 24,000 deaths in low-and middle-income countries". In: *BMC medicine* 12.1 (2014), p. 20.

[70] Henry D Kalter, Jamie Perin, and Robert E Black. "Validating hierarchical verbal autopsy expert algorithms in a large data set with known causes of death". In: *Journal of global health* 6.1 (2016).

[71] Daniel Chandramohan, Philip Setel, and Maria Quigley. "Effect of misclassification of causes of death in verbal autopsy: can it be adjusted?" In: *International Journal of Epidemiology* 30.3 (2001), pp. 509–514.

[72] Samuel Danso, Eric Atwell, and Owen Johnson. "Linguistic and statistically derived features for cause of death prediction from verbal autopsy text". In: *Language processing and knowledge in the web*. Springer, 2013, pp. 47–60.

[73] SA Danso et al. "A semantically annotated Verbal Autopsy corpus for automatic analysis of cause of death". In: *ICAME Journal* 37 (2013).

[74] Samuel Odei Danso. "Text Analytics to Predict Time and Cause of Death from Verbal Autopsies". PhD thesis. University of Leeds, 2015.

[75] Ronan Collobert and Jason Weston. "A unified architecture for natural language processing: Deep neural networks with multitask learning". In: *Proceedings of the 25th international conference on Machine learning*. ACM. 2008, pp. 160–167.

[76] Carol Friedman et al. "Automated encoding of clinical documents based on natural language processing". In: *Journal of the American Medical Informatics Association* 11.5 (2004), pp. 392–402.

[77] Vendhan Gajalakshmi and Richard Peto. "Suicide rates in rural Tamil Nadu, South India: verbal autopsy of 39 000 deaths in 1997–98". In: *International journal of epidemiology* 36.1 (2007), pp. 203–207.

[78] Cannannore Nidhi Kamath, Syed Saqib Bukhari, and Andreas Dengel. "Comparative Study between Traditional Machine Learning and Deep Learning Approaches for Text Classification". In: *Proceedings of the ACM Symposium on Document Engineering 2018*. 2018, pp. 1–11.

[79] Wenfeng Tian, Jun Li, and Hongguang Li. "A Method of Feature Selection Based on Word2Vec in Text Categorization". In: *2018 37th Chinese Control Conference (CCC)*. IEEE. 2018, pp. 9452–9455.

[80] Long Ma and Yanqing Zhang. "Using Word2Vec to process big text data". In: *2015 IEEE International Conference on Big Data (Big Data)*. IEEE. 2015, pp. 2895–2897.

[81] Marwa Naili, Anja Habacha Chaibi, and Henda Hajjami Ben Ghezala. "Comparative study of word embedding methods in topic segmentation". In: *Procedia Computer Science* 112 (2017), pp. 340–349.

[82] Eissa M Alshari et al. "Improvement of Sentiment Analysis Based on Clustering of Word2Vec Features". In: *2017 28th International Workshop on Database and Expert Systems Applications (DEXA)*. IEEE. 2017, pp. 123–126.

[83] Nabil Alami, Mohammed Meknassi, and Noureddine En-nahnahi. "Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning". In: *Expert Systems with Applications* (2019).

[84] Alexander Kirillov et al. "Efficient likelihood learning of a generic CNN-CRF model for semantic segmentation". In: *arXiv preprint arXiv:1511.05067* (2015).

[85] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. "Ask your neurons: A neural-based approach to answering questions about images". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1–9.

[86] Yoon Kim. "Convolutional neural networks for sentence classification". In: *arXiv preprint arXiv:1408.5882* (2014).

[87] Christopher JL Murray et al. "Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies". In: *Population health metrics* 9.1 (2011), pp. 1–11.

[88] Gaoxia Jiang and Wenjian Wang. "Error estimation based on variance analysis of k-fold cross-validation". In: *Pattern Recognition* 69 (2017), pp. 94–106.

[89] Tzu-Tsung Wong. "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation". In: *Pattern Recognition* 48.9 (2015), pp. 2839–2846.

[90] Samuel Danso et al. "A verbal autopsy corpus for machine learning of cause of death". In: *Proceedings of the Corpus Linguistics Conference*. 2011.

[91] Richard Socher et al. "Recursive deep models for semantic compositionality over a sentiment treebank". In: *Proceedings of the 2013 conference on empirical methods in natural language processing*. 2013, pp. 1631–1642.

[92] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[93] Samantha Herrera et al. "A systematic review and synthesis of the strengths and limitations of measuring malaria mortality through verbal autopsy". In: *Malaria journal* 16.1 (2017), p. 421.

[94] Peter Byass et al. “Comparing verbal autopsy cause of death findings as determined by physician coding and probabilistic modelling: a public health analysis of 54 000 deaths in Africa and Asia”. In: *Journal of global health* 5.1 (2015).

[95] Maria A Quigley et al. “Validity of data-derived algorithms for ascertaining causes of adult death in two African sites using verbal autopsy”. In: *Tropical Medicine & International Health* 5.1 (2000), pp. 33–39.

[96] Kelly R Moran et al. “Bayesian hierarchical factor regression models to infer cause of death from verbal autopsy data”. In: *arXiv preprint arXiv:1908.07632* (2019).

[97] Susan Idicula-Thomas, Ulka Gawde, and Prabhat Jha. “Comparison of machine learning algorithms applied to symptoms to determine infectious causes of death in children: national survey of 18,000 verbal autopsies in the Million Death Study in India”. In: *BMC public health* 21.1 (2021), pp. 1–11.

[98] Chodziwadziwa W Kabudula et al. “Two decades of mortality change in rural northeast South Africa”. In: *Global health action* 7.1 (2014), p. 25596.

[99] Gary King, Ying Lu, et al. “Verbal autopsy methods with multiple causes of death”. In: *Statistical Science* 23.1 (2008), pp. 78–91.

[100] Mohammed J Zaki and Wagner Meira Jr. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge University Press, 2019.

[101] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive data sets*. Cambridge university press, 2020.

[102] Matej Pičulin et al. “Disease Progression of Hypertrophic Cardiomyopathy: Modeling Using Machine Learning”. In: *JMIR Medical Informatics* 10.2 (2022), e30483.

[103] Yujie Yang et al. “Accurate Prediction of Stroke for Hypertensive Patients Based on Medical Big Data and Machine Learning Algorithms: Retrospective Study”. In: *JMIR Medical Informatics* 9.11 (2021), e30277.

[104] David L Poole and Alan K Mackworth. *Artificial Intelligence: foundations of computational agents*. Cambridge University Press, 2010.

[105] Zehang Richard Li, Tyler H McCormick, and Samuel J Clark. “InterVA4: An R package to analyze verbal autopsy data”. In: *Center for Statistics and the Social Sciences, University of Washington: Vienna, Austria: R Foundation for Statistical Computing.[1044]* (2014).

[106] Christopher JL Murray et al. “Population Health Metrics Research Consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets”. In: *Population health metrics* 9.1 (2011), p. 27.

[107] Thomas G Dietterich. “Approximate statistical tests for comparing supervised classification learning algorithms”. In: *Neural computation* 10.7 (1998), pp. 1895–1923.

[108] Abdhul Ahadh, Govind Vallabhasseri Binish, and Rajagopalan Srinivasan. “Text mining of accident reports using semi-supervised keyword extraction and topic modeling”. In: *Process Safety and Environmental Protection* 155 (2021), pp. 455–465.

[109] Hsin-Min Lu, Chih-Ping Wei, and Fei-Yuan Hsiao. “Modeling healthcare data using multiple-channel latent Dirichlet allocation”. In: *Journal of biomedical informatics* 60 (2016), pp. 210–223.

[110] Spyridon Kavvadias, George Drosatos, and Eleni Kaldoudi. “Supporting topic modeling and trends analysis in biomedical literature”. In: *Journal of Biomedical Informatics* 110 (2020), p. 103574.

[111] Nidal Nasser et al. “n-Gram based language processing using Twitter dataset to identify COVID-19 patients”. In: *Sustainable Cities and Society* 72 (2021), p. 103048.

[112] Naseem Cassim et al. “Using text mining techniques to extract prostate cancer predictive information (Gleason score) from semi-structured narrative laboratory reports in the Gauteng province, South Africa”. In: *BMC Medical Informatics and Decision Making* 21.1 (2021), pp. 1–11.

[113] Mónica Marrero et al. "Named entity recognition: fallacies, challenges and opportunities". In: *Computer Standards & Interfaces* 35.5 (2013), pp. 482–489.

[114] Filipe R Lucini et al. "Text mining approach to predict hospital admissions using early medical records from the emergency department". In: *International journal of medical informatics* 100 (2017), pp. 1–8.

[115] Uzma Raja et al. "Text mining in healthcare. Applications and opportunities". In: *J Healthc Inf Manag* 22.3 (2008), pp. 52–6.

[116] Joo-Chang Kim and Kyungyong Chung. "Associative feature information extraction using text mining from health big data". In: *Wireless Personal Communications* 105.2 (2019), pp. 691–707.

[117] Michael T Mapundu et al. "Performance evaluation of machine learning and Computer Coded Verbal Autopsy (CCVA) algorithms for cause of death determination: A comparative analysis of data from rural South Africa". In: *Frontiers in Public Health* (2022).

[118] Fei Zhu et al. "Biomedical text mining and its applications in cancer research". In: *Journal of biomedical informatics* 46.2 (2013), pp. 200–211.

[119] Joeky T Senders et al. "Automating clinical chart review: an open-source natural language processing pipeline developed on free-text radiology reports from patients with glioblastoma". In: *JCO Clinical Cancer Informatics* 4 (2020), pp. 25–34.

[120] Yoonseo Park, Sewon Park, and Munjea Lee. "Analyzing Community Care Research Trends Using Text Mining". In: *Journal of Multidisciplinary Healthcare* (2022), pp. 1493–1510.

[121] Yuan Luo, Gregory Riedlinger, and Peter Szolovits. "Text mining in cancer gene and pathway prioritization". In: *Cancer informatics* 13 (2014), CIN–S13874.

[122] Renu Vyas et al. "Building and analysis of protein-protein interactions related to diabetes mellitus using support vector machine, biomedical text mining and network analysis". In: *Computational biology and chemistry* 65 (2016), pp. 37–44.

[123] Anna Korhonen et al. "Text mining for literature review and knowledge discovery in cancer risk assessment and research". In: *PloS one* 7.4 (2012), e33427.

[124] Sofia IR Conceição and Francisco M Couto. "Text Mining for Building Biomedical Networks Using Cancer as a Case Study". In: *Biomolecules* 11.10 (2021), p. 1430.

[125] Joeky T Senders et al. "Natural language processing for automated quantification of brain metastases reported in free-text radiology reports". In: *JCO Clinical Cancer Informatics* 3 (2019), pp. 1–9.

[126] Farhi Marir, Huwida Said, and Feras Al-Obeidat. "Mining the web and literature to discover new knowledge about diabetes". In: *Procedia Computer Science* 83 (2016), pp. 1256–1261.

[127] Coen Hacking et al. "Text mining in long-term care: Exploring the usefulness of artificial intelligence in a nursing home setting". In: *Plos one* 17.8 (2022), e0268281.

[128] Mansour DIOUF, Mouhamadou THIAM, and Mathieu ROCHE. "New approach to discover meaningful terms to specify cause of death from narratives verbal autopsy using TF-IDF and the LDA topic model". In: *IEEE EUROCON 2023-20th International Conference on Smart Technologies*. IEEE. 2023, pp. 502–507.

[129] Ching-Hsue Cheng and Wei-Lun Hung. "Tea in benefits of health: A literature analysis using text mining and latent Dirichlet allocation". In: *Proceedings of the 2nd International Conference on Medical and Health Informatics*. 2018, pp. 148–155.

[130] Xian Cheng, Qiang Cao, and Stephen Shaoyi Liao. "An overview of literature on COVID-19, MERS and SARS: Using text mining and latent Dirichlet allocation". In: *Journal of Information Science* 48.3 (2022), pp. 304–320.

[131] Bach Xuan Tran et al. "Modeling research topics for artificial intelligence applications in medicine: latent Dirichlet allocation application study". In: *Journal of medical Internet research* 21.11 (2019), e15511.

[132] Taylor A Breuninger et al. "Associations between habitual diet, metabolic disease, and the gut microbiota using latent Dirichlet allocation". In: *Microbiome* 9.1 (2021), pp. 1–18.

[133] Eunji Jeon, Naeun Yoon, and So Young Sohn. "Exploring new digital therapeutics technologies for psychiatric disorders using BERTopic and PatentSBERTa". In: *Technological Forecasting and Social Change* 186 (2023), p. 122130.

[134] Ileana Scarpino et al. "Investigating topic modeling techniques to extract meaningful insights in Italian long COVID narration". In: *BioTech* 11.3 (2022), p. 41.

[135] Rildo Pinto da Silva, Juliana Tarossi Pollettini, and Antonio Pazin Filho. "Unsupervised natural language processing in the identification of patients with suspected COVID-19 infection". In: *Cadernos de Saúde Pública* 39 (2023), e00243722.

[136] Christopher YK Williams et al. "Exploring patient experiences and concerns in the online Cochlear implant community: A cross-sectional study and validation of automated topic modelling". In: *Clinical Otolaryngology* 48.3 (2023), pp. 442–450.

[137] Inigo Jauregi Unanue, Ehsan Zare Borzeshi, and Massimo Piccardi. "Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition". In: *Journal of biomedical informatics* 76 (2017), pp. 102–109.

[138] Mary Adewunmi et al. "Cancer Health Disparities drivers with BERTopic modelling and PyCaret Evaluation". In: *Cancer Health Disparities* 6 (2022).

[139] Muhammad Inaam ul haq and Qianmu Li. "Revealing the trends in the academic landscape of the health care system using contextual topic modeling". In: *Data Intelligence* 5.4 (2023), pp. 923–946.

[140] Mert Karabacak and Konstantinos Margetis. "Natural Language Processing Reveals Research Trends and Topics in The Spine Journal Over Two Decades: A Topic Modeling Study". In: *The Spine Journal* (2023).

[141] Aaron Baird, Yusen Xia, and Yichen Cheng. "Consumer perceptions of telehealth for mental health or substance abuse: a Twitter-based topic modeling analysis". In: *JAMIA open* 5.2 (2022), ooac028.

[142] Fu-Chen Yang, Anthony JT Lee, and Sz-Chen Kuo. "Mining health social media with sentiment analysis". In: *Journal of medical systems* 40.11 (2016), pp. 1–8.

[143] Jeremy Watt, Reza Borhani, and Aggelos Katsaggelos. *Machine learning refined: foundations, algorithms, and applications*. Cambridge University Press, 2020.

[144] Diego Buenano-Fernandez et al. "Text mining of open-ended questions in self-assessment of university teachers: An LDA topic modeling approach". In: *IEEE Access* 8 (2020), pp. 35318–35330.

[145] Carson Sievert and Kenneth Shirley. "LDAvis: A method for visualizing and interpreting topics". In: *Proceedings of the workshop on interactive language learning, visualization, and interfaces*. 2014, pp. 63–70.

[146] Maarten Grootendorst. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794* (2022).

[147] Allison E Grubbs et al. "Use of topic modeling to assess research trends in the journal Gynecologic Oncology". In: *Gynecologic Oncology* 172 (2023), pp. 41–46.

[148] S Mifrah and EH Benlahmar. "Topic modeling coherence: A comparative study between LDA and NMF models using COVID'19 corpus". In: *International Journal of Advanced Trends in Computer Science and Engineering* (2020), pp. 5756–5761.

[149] Michael T Mapundu et al. "Explainable Stacked Ensemble Deep Learning (SEDL) Framework to Determine Cause of Death from Verbal Autopsies". In: *Machine Learning and Knowledge Extraction* 5.4 (2023), pp. 1570–1588.

[150] Chodziwadziwa W Kabudula et al. "Assessing changes in household socioeconomic status in rural South Africa, 2001–2013: a distributional analysis using household asset indicators". In: *Social indicators research* 133 (2017), pp. 1047–1073.

[151] Stephen M Tollman et al. "Implications of mortality transition for primary health care in rural South Africa: a population-based surveillance study". In: *The Lancet* 372.9642 (2008), pp. 893–901.

[152] Paolo Bailo et al. "Verbal autopsy as a tool for defining causes of death in specific healthcare contexts: study of applicability through a traditional literature review". In: *International journal of environmental research and public health* 19.18 (2022), p. 11749.

[153] J Arun Prakash et al. "Stacked ensemble learning based on deep convolutional neural networks for pediatric pneumonia diagnosis using chest X-ray images". In: *Neural Computing and Applications* (2022), pp. 1–21.

[154] Daniele Ravì et al. "Deep learning for health informatics". In: *IEEE journal of biomedical and health informatics* 21.1 (2016), pp. 4–21.

[155] Gloria Hyun-Jung Kwak and Pan Hui. "Deephealth: Deep learning for health informatics". In: *ACM Transactions on Computing for Healthcare* (2019).

[156] Siddharth Srivastava et al. "Deep learning for health informatics: Recent trends and future directions". In: *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE. 2017, pp. 1665–1670.

[157] Rui Zhao et al. "Deep learning and its applications to machine health monitoring". In: *Mechanical Systems and Signal Processing* 115 (2019), pp. 213–237.

[158] Tong Zhang and Frank J Oles. "Text categorization based on regularized linear classification methods". In: *Information retrieval* 4.1 (2001), pp. 5–31.

[159] Xiang Zhang, Junbo Zhao, and Yann LeCun. "Character-level convolutional networks for text classification". In: *arXiv preprint arXiv:1509.01626* (2015).

[160] Kamran Kowsari et al. "Hdltex: Hierarchical deep learning for text classification". In: *2017 16th IEEE international conference on machine learning and applications (ICMLA)*. IEEE. 2017, pp. 364–371.

[161] Rie Johnson and Tong Zhang. "Effective use of word order for text categorization with convolutional neural networks". In: *arXiv preprint arXiv:1412.1058* (2014).

[162] Ozan Irsoy and Claire Cardie. "Opinion mining with deep recurrent neural networks". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 720–728.

[163] Duyu Tang, Bing Qin, and Ting Liu. "Document modeling with gated recurrent neural network for sentiment classification". In: *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015, pp. 1422–1432.

[164] Fagui Liu et al. "Combining attention-based bidirectional gated recurrent neural network and two-dimensional convolutional neural network for document-level sentiment classification". In: *Neurocomputing* 371 (2020), pp. 39–50.

[165] Monalisa Ghosh and Goutam Sanyal. "Document modeling with hierarchical deep learning approach for sentiment classification". In: *Proceedings of the 2nd International Conference on Digital Signal Processing*. 2018, pp. 181–185.

[166] Jiacheng Xu et al. "Cached long short-term memory neural networks for document-level sentiment classification". In: *arXiv preprint arXiv:1610.04989* (2016).

[167] Hamed Jelodar et al. "Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach". In: *IEEE Journal of Biomedical and Health Informatics* 24.10 (2020), pp. 2733–2742.

[168] Zichao Yang et al. “Hierarchical attention networks for document classification”. In: *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016, pp. 1480–1489.

[169] Alberto Blanco et al. “Extracting Cause of Death from Verbal Autopsy with Deep Learning interpretable methods”. In: *IEEE Journal of Biomedical and Health Informatics* (2020).

[170] Mehmet Kivrak, Emek Guldogan, and Cemil Colak. “Prediction of death status on the course of treatment in SARS-COV-2 patients with deep learning and machine learning methods”. In: *Computer Methods and Programs in Biomedicine* 201 (2021), p. 105951.

[171] Mojtaba Heidarysafa et al. “Analysis of Railway Accidents’ Narratives Using Deep Learning”. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2018, pp. 1446–1453.

[172] Khalid El Asnaoui. “Design ensemble deep learning model for pneumonia disease classification”. In: *International Journal of Multimedia Information Retrieval* 10.1 (2021), pp. 55–68.

[173] Shanjiang Tang et al. “EDL-COVID: ensemble deep learning for COVID-19 case detection from chest x-ray images”. In: *IEEE Transactions on Industrial Informatics* 17.9 (2021), pp. 6539–6549.

[174] Mudasir A Ganaie et al. “Ensemble deep learning: A review”. In: *Engineering Applications of Artificial Intelligence* 115 (2022), p. 105151.

[175] Asma Baccouche et al. “Ensemble deep learning models for heart disease classification: A case study from Mexico”. In: *Information* 11.4 (2020), p. 207.

[176] Hui Wen Loh et al. “Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)”. In: *Computer Methods and Programs in Biomedicine* (2022), p. 107161.

[177] Yiming Zhang, Ying Weng, and Jonathan Lund. “Applications of explainable artificial intelligence in diagnosis and surgery”. In: *Diagnostics* 12.2 (2022), p. 237.

[178] Abdul Rehman Javed et al. “Toward explainable AI-empowered cognitive health assessment”. In: *Frontiers in Public Health* 11 (2023).

[179] Samanta Knapič et al. “Explainable artificial intelligence for human decision support system in the medical domain”. In: *Machine Learning and Knowledge Extraction* 3.3 (2021), pp. 740–770.

[180] Chirag Shah. *A hands-on introduction to data science*. Cambridge University Press, 2020.

[181] David H Wolpert. “Stacked generalization”. In: *Neural networks* 5.2 (1992), pp. 241–259.

[182] Michel Jose Anzanello and Flavio Sanson Fogliatto. “Learning curve models and applications: Literature review and research directions”. In: *International Journal of Industrial Ergonomics* 41.5 (2011), pp. 573–583.

[183] Derek Hoiem et al. “Learning Curves for Analysis of Deep Networks”. In: *arXiv preprint arXiv:2010.11029* (2020).

[184] Yoshua Bengio et al. “A neural probabilistic language model”. In: *Journal of machine learning research* 3.Feb (2003), pp. 1137–1155.

[185] Marie Uncovska et al. “Rating analysis and BERTopic modeling of consumer versus regulated mHealth app reviews in Germany”. In: *NPJ Digital Medicine* 6.1 (2023), p. 115.

[186] G Catanuto et al. “Text mining and word embedding for classification of decision making variables in breast cancer surgery”. In: *European Journal of Surgical Oncology* (2022).

[187] Junchi Zhang, Mengchi Liu, and Yue Zhang. “Topic-informed neural approach for biomedical event extraction”. In: *Artificial intelligence in medicine* 103 (2020), p. 101783.

[188] Juan Antonio Lossio-Ventura et al. “Evaluation of clustering and topic modeling methods over health-related tweets and emails”. In: *Artificial Intelligence in Medicine* 117 (2021), p. 102096.

[189] Dominic D Martinelli. “Evolution of Alzheimer’s disease research from a health-tech perspective: Insights from text mining”. In: *International Journal of Information Management Data Insights* 2.2 (2022), p. 100089.

[190] Bin He, Yi Guan, and Rui Dai. "Classifying medical relations in clinical text via convolutional neural networks". In: *Artificial intelligence in medicine* 93 (2019), pp. 43–49.

[191] Yanshan Wang et al. "A comparison of word embeddings for the biomedical natural language processing". In: *Journal of biomedical informatics* 87 (2018), pp. 12–20.

[192] Maté Ongenaert et al. "PubMeth: a cancer methylation database combining text-mining and expert annotation". In: *Nucleic acids research* 36.suppl_1 (2007), pp. D842–D846.

[193] Gabriela Jurca et al. "Integrating text mining, data mining, and network analysis for identifying genetic breast cancer trends". In: *BMC research notes* 9.1 (2016), pp. 1–35.

[194] Yi-Wen Hsiao and Tzu-Pin Lu. "Text-mining in cancer research may help identify effective treatments". In: *Translational Lung Cancer Research* 8.Suppl 4 (2019), S460.

[195] Nigel Collier. "Uncovering text mining: A survey of current work on web-based epidemic intelligence". In: *Global public health* 7.7 (2012), pp. 731–749.

[196] Lukasz Aleksandrowicz et al. "Performance criteria for verbal autopsy-based systems to estimate national causes of death: development and application to the Indian Million Death Study". In: *BMC medicine* 12.1 (2014), pp. 1–14.

[197] Muhamad Rifki Taufik et al. "Predicting TB Death Using Logistic Regression and Decision Tree on VA Data". In: ().

[198] Bruno Samways dos Santos et al. "Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018". In: *Computers & Industrial Engineering* 138 (2019), p. 106120.

[199] Ken Obenson and George Enow Orock. "An overview of the challenges facing death investigation systems in certain resource limited countries". In: *Journal of forensic and legal medicine* 50 (2017), pp. 58–62.

[200] Ian Riley. "Computer-based analysis of verbal autopsies: revolution or evolution?" In: *Population health metrics* 9.1 (2011), p. 26.

[201] Yura Lee et al. "Validation for Accuracy of Cancer Diagnosis in Electronic Medical Records Using a Text Mining Method." In: *Studies in health technology and informatics* 216 (2015), pp. 882–882.

[202] Xolani Dastile, Turgay Celik, and Moshe Potsane. "Statistical and machine learning models in credit scoring: A systematic literature survey". In: *Applied Soft Computing* (2020), p. 106263.

[203] Òscar Romero Llombart. "Using machine learning techniques for sentiment analysis". In: ().

[204] Jeffrey Pennington, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[205] Ramzan Talib et al. "Text mining: techniques, applications and issues". In: *International Journal of Advanced Computer Science and Applications* 7.11 (2016), pp. 414–418.

[206] Sonali Vijay Gaikwad, Archana Chaugule, and Pramod Patil. "Text mining methods and techniques". In: *International Journal of Computer Applications* 85.17 (2014).

[207] Yang Li et al. "Learning word representations for sentiment analysis". In: *Cognitive Computation* 9.6 (2017), pp. 843–851.

[208] Aparup Khatua, Apalak Khatua, and Erik Cambria. "A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks". In: *Information Processing & Management* 56.1 (2019), pp. 247–257.

[209] Nachiket Tapas et al. "Prediction of cardiac arrest recurrence using ensemble classifiers". In: *Sādhanā* 42.7 (2017), pp. 1135–1141.

[210] Kuteesa R Bisaso et al. "A survey of machine learning applications in HIV clinical research and care". In: *Computers in biology and medicine* 91 (2017), pp. 366–371.

[211] Xianyong Yi et al. “Design and implementation of Word2Vec parallel algorithm based on HPC”. In: *2017 Chinese Automation Congress (CAC)*. IEEE. 2017, pp. 585–590.

[212] Chodziwadziwa W Kabudula et al. “The promise of record linkage for assessing the uptake of health services in resource constrained settings: a pilot study from South Africa”. In: *BMC medical research methodology* 14.1 (2014), p. 71.

[213] Jane Joubert et al. “Record-linkage comparison of verbal autopsy and routine civil registration death certification in rural north-east South Africa: 2006–09”. In: *International journal of epidemiology* 43.6 (2014), pp. 1945–1958.

[214] Chodziwadziwa W Kabudula et al. “Evaluation of record linkage of mortality data between a health and demographic surveillance system and national civil registration system in South Africa”. In: *Population Health Metrics* 12.1 (2014), p. 23.

[215] Martha Anker. “The effect of misclassification error on reported cause-specific mortality fractions from verbal autopsy.” In: *International journal of epidemiology* 26.5 (1997), pp. 1090–1096.

[216] Zhiheng Huang, Wei Xu, and Kai Yu. “Bidirectional LSTM-CRF models for sequence tagging”. In: *arXiv preprint arXiv:1508.01991* (2015).

[217] Guillaume Lample et al. “Neural architectures for named entity recognition”. In: *arXiv preprint arXiv:1603.013* (2016).

[218] Guibin Chen et al. “Ensemble application of convolutional and recurrent neural networks for multi-label text categorization”. In: *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE. 2017, pp. 2377–2383.

[219] Zahra Shams Khoozani and Ram Gopal Raj. “TF-IDF-Based Automated Application for classification Forensic Autopsy Reports to Identification of Cause of Death (CoD)”. In: ().

[220] Samath D Dharmaratne et al. “Opportunities and challenges for verbal autopsy in the national Death Registration System in Sri Lanka: past and future”. In: *Population health metrics* 9.1 (2011), p. 21.

[221] E Moses, P Melwin, and S Vigneshwari. “Instant answering for health care system by machine learning approach”. In: *Circuit, Power and Computing Technologies (ICCPCT), 2016 International Conference on*. IEEE. 2016, pp. 1–7.

[222] Ghassan Azar et al. “Intelligent data mining and machine learning for mental health diagnosis using genetic algorithm”. In: *Electro/Information Technology (EIT), 2015 IEEE International Conference on*. IEEE. 2015, pp. 201–206.

[223] Lei Ma et al. “Deep learning in remote sensing applications: A meta-analysis and review”. In: *ISPRS journal of photogrammetry and remote sensing* 152 (2019), pp. 166–177.

[224] Laith Hussain-Alkhateeb et al. “Enhancing the value of mortality data for health systems: adding Circumstances Of Mortality CATegories (COMCATs) to deaths investigated by verbal autopsy”. In: *Global health action* 12.1 (2019), p. 1680068.

[225] Marina Skurichina and Robert PW Duin. “Bagging, boosting and the random subspace method for linear classifiers”. In: *Pattern Analysis & Applications* 5.2 (2002), pp. 121–135.

[226] Johan AK Suykens and Joos Vandewalle. “Least squares support vector machine classifiers”. In: *Neural processing letters* 9.3 (1999), pp. 293–300.

[227] Esteban Alfaro, Matias Gamez, Noelia Garcia, et al. “Adabag: An R package for classification with boosting and bagging”. In: *Journal of Statistical Software* 54.2 (2013), pp. 1–35.

[228] Ioannis E Livieris et al. “Ensemble deep learning models for forecasting cryptocurrency time-series”. In: *Algorithms* 13.5 (2020), p. 121.

[229] Dilan Lasantha, Sugandima Vidanagamachchi, and Sam Nallaperuma. “Deep learning and ensemble deep learning for circRNA-RBP interaction prediction in the last decade: A review”. In: *Engineering Applications of Artificial Intelligence* 123 (2023), p. 106352.

[230] Ammar Mohammed and Rania Kora. “A comprehensive review on ensemble deep learning: Opportunities and challenges”. In: *Journal of King Saud University-Computer and Information Sciences* (2023).

[231] Jeremy Watt, Reza Borhani, and Aggelos K Katsaggelos. “Machine learning refined”. In: (2019).

[232] Alexander Partin et al. “Learning curves for drug response prediction in cancer cell lines”. In: *BMC bioinformatics* 22.1 (2021), pp. 1–18.

[233] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.

[234] Michael Tonderai Mapundu et al. “Overview of Statistical and Machine Learning Techniques for Determining Causes of Death from Verbal Autopsies: A Systematic Literature Review”. In: (2020).

[235] C Chabila Mapoma et al. “Application of verbal autopsy in routine civil registration in Lusaka District of Zambia”. In: *BMC Health Services Research* 21.1 (2021), pp. 1–11.

[236] Kelly R Moran et al. “Bayesian hierarchical factor regression models to infer cause of death from verbal autopsy data”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* (2021).