

https://dspace.org/

Predicting in-hospital mortality in heart failure patients using machine learning

Mpanya, Dineo

2023-05
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/44901
Mpanya, Dineo. (2023). Predicting in-hospital mortality in heart failure patients using machine learning. [PhD thesis, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/44901

Downloaded from a DSpace repository. https://dspace.org/

PREDICTING IN-HOSPITAL MORTALITY IN HEART FAILURE PATIENTS USING MACHINE LEARNING



DINEO MPANYA

Student number: 878824

Original published work submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy.

May, 2023

DECLARATION

DECLARATION BY STUDENT

I, Dineo Mpanya, student number 878824 declare that this thesis is my own work and that I contributed to research findings published in the articles stated below which are included in my thesis.

My role entailed data cleaning, data analysis, data visualisation, writing and editing of all drafts of the manuscripts.

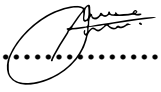
Signature of Student ……..……..

Date 30 May 2023**……**

DECLARATION BY SUPERVISORS

I, Hopewell Ntsinjana, the primary supervisor declare that the student, Dineo Mpanya reviewed the literature, analysed the data, and worked independently on the Python script for building predictive models using machine learning. She also wrote the first draft for all manuscripts appearing in this thesis.

Name of Primary Supervisor ……Prof Hopewell Ntsinjana……

Signature of Primary Supervisor …………...

Date ……30 MAY 2023……..…….

I, Turgay Celik, the co-supervisor declare that the student, Dineo Mpanya reviewed the literature, analysed the data, and worked independently on the Python script for building predictive models using machine learning. She also wrote the first draft for all manuscripts appearing in this thesis.

Name of Co-Supervisor ……..……

Signature of Co-Supervisor ……..……..

Date ……..…….

CO-AUTHORS' AGREEMENT FOR WRITE-UP WITH PUBLISHED ARTICLES

By signing this declaration, the co-authors listed below agree to the use of the article by the student as part of her thesis. The co-authors confirm that their contribution entailed supervision (Ntsinjana, H and Celik, T), addressing technical queries (Celik, T) and reviewing manuscripts (Ntsinjana, H, Celik, T, and Klug, E) drafted by Dineo Mpanya.

1. Mpanya, D., Celik, T., Klug, E., Ntsinjana, H. Machine learning and statistical methods for predicting mortality in heart failure. *Heart Fail Rev* **26**, 545–552 (2021). https://doi.org/10.1007/s10741-020-10052-y

2. Mpanya, D., Celik, T., Klug, E., Ntsinjana, H. Predicting mortality and hospitalization in heart failure using machine learning: A systematic literature review, IJC Heart & Vasculature, Volume 34, 2021, 100773. https://doi.org/10.1016/j.ijcha.2021.100773.

3. Mpanya, D.; Celik, T.; Klug, E.; Ntsinjana, H. Clustering of Heart Failure Phenotypes in Johannesburg Using Unsupervised Machine Learning. *Appl. Sci.* 2023, *13*, 1509.

4. Mpanya, D.; Celik, T.; Klug, E.; Ntsinjana, H. Predicting in-hospital all-cause mortality in heart failure using machine learning. Frontiers in Cardiovascular Medicine, Volume 9, 2023. https://doi.org/10.3389/fcvm.2022.1032524

Authors	Name	Signature	Date
1st author	Dineo Mpanya		31 January 2023
2nd author	Turgay Celik		1 Feb. 2023
3rd author	Eric Klug		31 January 2023
4th author	Hopewell Ntsinjana		31st JANUARY 2023

Copyright permission waived since all manuscripts were published as open access articles.

DEDICATION

I dedicate this work to my parents, Bafanyana and Duduzile Mpanya. Thank you for your love and support.

I am grateful for the guidance and support provided by my husband.

Thank you to my siblings Nandi, Palesa, Naledi, and Neo for the encouragement.

Thank you to my supervisors Professor Hopewell Ntsinjana and Professor Turgay Celik. I would not have been able to complete this project without your support.

Thank you Professor Eric Klug for the valuable comments and guidance.

I acknowledge the role played by Professor Mike Sathekge in supporting me throughout my journey as a PhD student.

PRESENTATIONS ARISING FROM THIS PROJECT

1. Predicting in-hospital all-cause mortality in heart failure using machine learning. University of the Witwatersrand Faculty of Health Sciences Biennial Research Day. 15 September 2022. **Winner: Best student oral presentation**

2. Predicting in-hospital all-cause mortality in heart failure using machine learning. South African Heart Congress. Sandton. October 2022

3. Predicting in-hospital all-cause mortality in heart failure using machine learning. Association of Medical Council of Africa. Sun City

PUBLICATIONS ARISING FROM THIS PROJECT

1. Mpanya, D., Celik, T., Klug, E., Ntsinjana, H. Machine learning and statistical methods for predicting mortality in heart failure. *Heart Fail Rev* **26**, 545–552 (2021). https://doi.org/10.1007/s10741-020-10052-y

2. Mpanya, D., Celik, T., Klug, E., Ntsinjana, H. Predicting mortality and hospitalization in heart failure using machine learning: A systematic literature review, IJC Heart & Vasculature, Volume 34, 2021, 100773. https://doi.org/10.1016/j.ijcha.2021.100773.

3. Mpanya, D.; Celik, T.; Klug, E.; Ntsinjana, H. Clustering of Heart Failure Phenotypes in Johannesburg Using Unsupervised Machine Learning. *Appl. Sci.* **2023**, *13*, 1509. https://doi.org/10.3390/app1303150

4. Mpanya, D.; Celik, T.; Klug, E.; Ntsinjana, H. Predicting in-hospital all-cause mortality in heart failure using machine learning. Frontiers in Cardiovascular Medicine, Volume 9, 2023. https://doi.org/10.3389/fcvm.2022.1032524

ABSTRACT

The age of onset and causes of heart failure differ between high-income and low-and-middle-income countries (LMIC). Heart failure patients in LMIC also experience a higher mortality rate. Innovative ways that can risk stratify heart failure patients in this region are needed. The aim of this study was to demonstrate the utility of machine learning in predicting all-cause mortality in heart failure patients hospitalised in a tertiary academic centre. Six supervised machine learning algorithms were trained to predict in-hospital all-cause mortality using data from 500 consecutive heart failure patients with a left ventricular ejection fraction (LVEF) less than 50%. The mean age was 55.2 ± 16.8 years. There were 271 (54.2%) males, and the mean LVEF was 29 ± 9.2%. The median duration of hospitalisation was 7 days (interquartile range: 4–11), and it did not differ between patients discharged alive and those who died. After a prediction window of 4 years (interquartile range: 2–6), 84 (16.8%) patients died before discharge from the hospital. The area under the receiver operating characteristic curve was 0.82, 0.78, 0.77, 0.76, 0.75, and 0.62 for random forest, logistic regression, support vector machines (SVM), extreme gradient boosting, multilayer perceptron (MLP), and decision trees, and the accuracy during the test phase was 88, 87, 86, 82, 78, and 76% for random forest, MLP, SVM, extreme gradient boosting, decision trees, and logistic regression. The support vector machines were the best performing algorithm, and furosemide, beta-blockers, spironolactone, early diastolic murmur, and a parasternal heave had a positive coefficient with the target feature, whereas coronary artery disease, potassium, oedema grade, ischaemic cardiomyopathy, and right bundle branch block on electrocardiogram had negative coefficients. Despite a small sample size, supervised machine learning algorithms successfully predicted all-cause mortality with modest accuracy. The SVM model will be externally validated using data from multiple cardiology centres in South Africa before developing a uniquely African risk prediction tool that can potentially transform heart failure management through precision medicine.

ACKNOWLEDGEMENTS

This project would not have been seen to completion without the financial support from The Bongani Mayosi Netcare Clinical Scholarship, the Discovery Academic Fellowship *[Grant No. 03902]*, the Carnegie Corporation of New York [*Grant No. b8749*], the South African Heart Association, and the University of the Witwatersrand Chancellor's Female Academic Leaders Fellowship (NPC) Grant.

I would like to thank the Division of Cardiology at Charlotte Maxeke Johannesburg Academic Hospital for allowing me to use their database of patients admitted in their wards. Thank you to Geethen Singh, Marheb Badianyama, Michaela Gough, Yemurai Muchenge, Jenna Damelin, Thakgalo Thibela, and Mpumelelo Ngadla for assisting with data capturing and verification. Thank you Ntuthuko Mngomezulu for guiding me while learning to use Python.

TABLE OF CONTENTS

Page

Declaration by student …..……. …… ii

Declaration by supervisors …….…… iii

Co-authors' agreement for write-up with published articles…… iv

Dedication…… v

Presentations arising from this project……. vi

Publications arising from this project……... vi

Abstract……..… vii

Acknowledgements…….. viii

Table of Contents……. ix

List of Figures ……..... xii

List of abbreviations…… xii

Overview of the thesis…….. xii

CHAPTER 1 - INTRODUCTION 1

1.1 Study Background 1

1.2 Study Rationale 3

1.3 References 4

CHAPTER 2 - LITERATURE SURVEY: MACHINE LEARNING AND STATISTICAL METHODS FOR PREDICTING MORTALITY IN HEART FAILURE 6

2.1 Introduction 6

2.2 Current Knowledge 6

CHAPTER 3 – MANUSCRIPT ONE: SYSTEMATIC REVIEW OF MACHINE LEARNING MODELS PREDICTING MORTALITY AND REHOSPITALISATION IN HEART FAILURE..... 15

3.1 Introduction 15

3.2 Current Knowledge 15
3.3 Aims and Objectives 15
CHAPTER 4 – MANUSCRIPT TWO: CLUSTERING OF HEART FAILURE PHENOTYPES IN JOHANNESBURG USING UNSUPERVISED MACHINE LEARNING 25
4.1 Introduction 25
4.2 Current Knowledge 25
4.3 Aims and Objectives 25
CHAPTER 5 – MANUSCRIPT THREE: PREDICTING ALL-CAUSE MORTALITY IN HEART FAILURE USING MACHINE LEARNING 41
5.1 Introduction 41
5.2 Current Knowledge 41
5.3 Aims and Objectives 41
CHAPTER 6 – DISCUSSION AND CONCLUSIONS 54
6.1 Discussion 54
6.2 Limitations 57
6.3 Recommendations 58
6.4 Future Direction 61
6.5 Conclusions 62
6.6 References 62
All References 66
Appendices 77
A Excluded Studies: Systematic Review of Machine Learning Models 77
B Clinical Parameters Extracted From the Electronic Health Record System 78
C Predicting 30-Day All-Cause Rehospitalisation 79
D Study Protocol 81
E Ethics Certificate 105
F Turn-It-In Report For Plagiarism 106

List of Figures

Figure 6.1 Schematic diagram showing the proposed data laboratory for an efficient machine learning predictive model, starting and ending with the data collection and capturing phase

LIST OF ABBREVIATIONS

ANN	artificial neural networks
ANOVA	analysis of variance
ARB	angiotensin receptor blocker
AUC	area under the curve
AUROC	area under the receiver operating characteristic curve
BIRCH	Balanced Iterative Reducing and Clustering hierarchies
BP	blood pressure
CAD	coronary artery disease
CART	classification and regression tree
CI	confidence interval
CMJAH	Charlotte Maxeke Johannesburg Academic Hospital
CNN	convolutional neural networks
CPXR	contrast pattern aided logistic regression
CVD	cardiovascular
DBSCAN	density-based spatial clustering of applications with noise
Duration	duration of hospitalisation
ECG	electrocardiogram
ED	end diastolic
eGFR	estimated glomerular filtration rate
EHR	electronic health records
GBM	gradient-boosted model

HDL	high-density lipoprotein
HF	heart failure
HFmEF	heart failure with mid-range ejection fraction
HFpEF	heart failure with preserved ejection fraction
HFrEF	heart failure with reduced ejection fraction
HIC	high-income country
HR	hazard ratio
ICD-10	tenth revision of the international classification of diseases
IG	information gain
IHD	ischaemic heart disease
INTER-CHF	International Congestive Heart Failure
IQR	interquartile ranges
JVP	jugular venous pressure
K^+	potassium
KNN	K-Nearest Neighbours
LASSO	least absolute shrinkage and selection operator
LBBB	left bundle branch block
LDL	low-density lipoprotein
LMIC	low-and-middle-income country
LR	linear regression
LVEF	left ventricular ejection fraction
LVIDd	left ventricular internal diameter in diastole
LVIDs	left ventricular internal diameter in systole
MAGGIC	Meta-Analysis Global Group in Chronic Heart Failure
MDRD	modification of diet in renal disease
ML	machine learning

MLP	multi-layer perceptron
MRI	magnetic resonance imaging
NB	Naïve Bayes
NHLS	National Health Laboratory Service
NYHA	New York Heart Association
OPTICS	Ordering Points To Identify the Clustering Structure
PMR	Pravin Manga Registry
PRAISE	Prospective Randomized Amlodipine Survival Evaluation
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RBBB	right bundle branch block
ReLU	rectified linear unit
ROC	receiver operating characteristic
S2	second heart sound
S3	third heart sound
S4	fourth heart sound
SD	standard deviation
SGD	stochastic gradient descent
SHAP	Shapley additive explanations
SHF	Seattle Heart Failure
SSA	sub-Saharan Africa
SVM	support vector machines
TAN	tree augmented Bayesian network
TB	tuberculosis
USA	United States of America
XGB	extreme gradient boosting
XGBoost	extreme gradient boosting

OVERVIEW OF THE THESIS

This thesis comprises of the following main chapters:

a) CHAPTER 1: INTRODUCTION
b) CHAPTER 2: LITERATURE REVIEW
c) CHAPTER 3: MANUSCRIPT 1
d) CHAPTER 4 : MANUSCRIPT 2
e) CHAPTER 5 : MANUSCRIPT 3
f) CHAPTER 6: DISCUSSION AND CONCLUSIONS

CHAPTER 1 introduces the thesis by briefly outlining the study background, focusing on heart failure and predictive modelling using machine learning.
CHAPTER 2 is a narrative review that describes statistical and machine learning techniques for building risk prediction models and the clinical utility of risk prediction models in heart failure, with a focus on machine learning techniques.
CHAPTER 3 is an original research article, a systematic review of studies using machine learning algorithms to predict all-cause mortality and rehospitalisation in heart failure.
CHAPTER 4 describes the baseline clinical characteristics of heart failure patients, followed by the clustering of patients hospitalised with heart failure using unsupervised machine learning algorithms.
CHAPTER 5 describes the application of six supervised machine learning algorithms in creating a model predicting in-hospital all-cause mortality.
CHAPTER 6 contains the discussion, conclusions, and recommendations for future studies.

CHAPTER 1 - INTRODUCTION

1.1 Study Background

Heart failure is a rapidly growing public health problem associated with the increased utilisation of the health care budget, with the lifetime cost of treating heart failure patients estimated to be United States (US) dollars $126 819 per patient (Lesyuk et al., 2018). The mean cost of each heart failure-related hospitalisation is estimated to be US$23 000 (Wang et al., 2010). In Nigeria, the total cost of inpatient and outpatient heart failure-related care is approximately US$50 859. There are no population studies conducted in South Africa reporting the cost associated with treating heart failure patients. Despite the widespread introduction of novel pharmacological agents, advancement in open heart surgery techniques and development of life-prolonging intra-cardiac devices, heart failure patients, in particular, those with a reduced left ventricular ejection fraction still experience persistent clinical symptoms that lead to an overall decreased quality of life (Dokainish et al., 2016, Dokainish et al., 2017a).

Heart failure is a term used to define the final common pathway in almost all cardiovascular diseases, where the heart fails to pump sufficient blood to major organs in the body (Hilfiker-Kleiner et al., 2006). Essentially, any condition affecting the pericardial sac, coronary circuit, heart muscle, heart valves or the heart's conduction system can lead to heart failure. There are three different classes of heart failure, distinctly categorised based on the left ventricular ejection fraction (LVEF). Once the LVEF is less than or equal to 40% and clinical signs and symptoms suggestive of venous congestion and/or fluid overload are apparent, the patient is classified as having heart failure with reduced ejection fraction (HFrEF). In comparison, patients with a LVEF equal to or greater than 50% are classified as having heart failure with preserved ejection fraction (HFpEF) and those with a LVEF between 40 – 50% are classified as having heart failure with a mid-range ejection fraction (HFmEF) (Bozkurt, 2021).

In South Africa, the common causes of heart failure are hypertension, dilated cardiomyopathy, rheumatic heart disease, human immunodeficiency virus (HIV)-associated heart disease, coronary

artery disease and tuberculosis pericarditis (Ntusi and Mayosi, 2009). In a study involving 109 patients with acute heart failure treated in a tertiary academic centre in Cape Town, the in-hospital mortality rate was 8%, and after a follow-up period of six months, 26.1% of patients died (Szymanski, 2018). Dokainish and colleagues studied heart failure patients residing in Africa and found a higher all-cause mortality rate of 34% (95% confidence interval: 30.2 – 37.4) within one year of follow-up (Dokainish et al., 2017b). The high mortality rate calls for innovative strategies that aim to risk stratify patients with heart failure, ensuring that high-risk individuals are identified timeously and subjected to interventions that delay disease progression. Risk predicting models are valuable adjuncts when managing patients with cardiovascular disease. Both statistical and machine learning techniques are used to create risk predicting models (Mpanya et al., 2020). While statistics focus on making inferences based on the sample studied, machine learning focuses on learning patterns in datasets and subsequently making predictions based on the learnt observations. Machine learning algorithms have the advantage of improving the performance metrics of the model over time as more data is added.

The Seattle Heart Failure (SHF) model is one of the best performing models created using Cox proportional hazards regression. The aim was to predict survival in heart failure. The data used to build the SHF model was obtained from 1125 outpatients recruited in the Prospective Randomised Trial of Amlodipine versus Placebo (PRAISE 1). The model was prospectively validated on a new cohort comprising 9 942 participants. Although some of the data used to validate the model originated from heart failure registries in Italy and the USA, a more substantial fraction of the data was obtained from outpatient clinical trial participants. However, this model cannot be generalised to hospitalised patients or those with life-threatening conditions such as renal failure since the data used to create the model was obtained from carefully screened trial participants without significant comorbidities (Levy et al., 2006). Another limiting factor is that most participants (64%) had heart failure secondary to ischaemic heart disease. In contrast, the cause of heart failure in sub-Saharan Africa is predominantly non-ischaemic (up to 75%), although risk factors for ischaemic heart disease are becoming more prevalent in sub-Saharan Africa (Gallagher et al., 2018).

Other risk predicting models created using data from patients with chronic heart failure include the Candesartan in Heart Failure: Assessment of Reduction in Mortality and Morbidity (CHARM)

Risk Score, the Controlled Rosuvastatin Multinational Trial in Heart Failure (CORONA) Risk Score, the Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico-Heart Failure (GISSI-HF) model, the Meta-analysis Global Group in Chronic Heart Failure (MAGGIC) Risk Score, and a risk score based on data collected from the Prospective comparison of Angiotensin Receptor-neprilysin inhibitor (ARNI) with Angiotensin-converting enzyme inhibitor (ACEI) to Determine Impact on Global Mortality and Morbidity in Heart Failure (PARADIGM) trial (Wehbe et al., 2020). These risk scores were all created using multivariable logistic regression analysis on data of patients residing in high-income countries, where heart failure is predominantly caused by ischaemic heart disease. Furthermore, data used in some of these models was obtained from carefully selected trial patients and, therefore, may not accurately represent the disease status of a heart failure patient in the general population.

Machine learning algorithms analyse and learn patterns in datasets. Data is transformed and placed in a two-dimensional or three-dimensional space, such that each data point can be perceived as a vector, comprising size or weight and direction or coordinates. In addition, the performance of algorithms tends to improve over time as the algorithm is exposed to a larger dataset. In our setting, most heart failure patients still have limited access to heart transplant services and intracardiac devices such as cardiac resynchronisation defibrillator therapy. The first human heart transplant in South Africa was performed on 3 December 1967 (Brink, 2017). Despite this, heart transplant services are only available in the private sector. In addition, novel pharmacological therapies proven to lower the risk of mortality such as sodium glucose co-transporters are not available in the public sector. As such, innovative ways of identifying patients at high risk for mortality are still required.

1.2 Study Rationale

In South Africa, there are currently no risk calculators that estimate the risk of mortality in patients with heart failure. Risk tools are population specific and are influenced by the underlying cause of heart failure, the availability of guideline-recommended therapy and referral pathways. Therefore, the primary aim of the study is to predict all-cause in-hospital mortality in patients with HFrEF using supervised machine learning algorithms. The objectives of the study are:

a) To describe the baseline demographic and clinical characteristics of heart failure patients.

b) To compare supervised machine learning algorithms to traditional statistical methods for predicting all-cause mortality in heart failure patients.

c) To create a risk model predicting hospitalisation due to any cause after initial hospitalisation in patients with heart failure.

d) To externally validate the model with the best performing metrics for predicting in-hospital mortality and the risk of hospitalisation in heart failure patients. The model will be validated by comparing the actual and predicted risk of in-hospital mortality and hospitalisation after one year of follow-up.

In conclusion, this research will address the lack of risk calculators derived from uniquely African data. Risk calculators will be used to estimate the risk of all-cause mortality in heart failure patients. This will be achieved by training six machine learning algorithms to predict mortality using a dataset of patients hospitalised with heart failure at the Charlotte Maxeke Johannesburg Academic Hospital. Currently, there are no population-based studies comparing outcomes in heart failure patients risk stratified based on risk scores with those managed based on standard care. The next section of the thesis is a narrative review describing how risk models are created using machine learning and statistical methods.

1.3 References

1. Dokainish, H., Teo, K., Zhu, J., Roy, A., Alhabib, K. F., Elsayed, A., et al. 2017a. Global mortality variations in patients with heart failure: results from the International Congestive Heart Failure (INTER-CHF) prospective cohort study. Lancet Glob Health, 5**,** e665-e672.
2. Dokainish, H., Teo, K., Zhu, J., Roy, A., Alhabib, K. F., Elsayed, A., et al. 2016. Heart Failure in Africa, Asia, the Middle East and South America: The INTER-CHF study. Int J Cardiol, 204**,** 133-41.
3. Gallagher, J., Mcdonald, K., Ledwidge, M. & Watson, C. J. 2018. Heart Failure in Sub-Saharan Africa. Card Fail Rev, 4**,** 21-24.

4. Hilfiker-Kleiner, D., Landmesser, U. & Drexler, H. 2006. Molecular Mechanisms in Heart Failure: Focus on Cardiac Hypertrophy, Inflammation, Angiogenesis, and Apoptosis. JACC, 48**,** A56-A66.
5. Bozkurt B, Coats AJ, Tsutsui H, et al. Universal Definition and Classification of Heart Failure: A Report of the Heart Failure Society of America, Heart Failure Association of the European Society of Cardiology, Japanese Heart Failure Society and Writing Committee of the Universal Definition of Heart Failure. J Card Fail. 2021;S1071-9164(21)00050-6. doi:10.1016/j.cardfail.2021.01.022
6. Lesyuk, W., Kriza, C. & Kolominsky-Rabas, P. 2018. Cost-of-illness studies in heart failure: a systematic review 2004-2016. BMC Cardiovasc Disord, 18**,** 74.
7. Levy, W. C., Mozaffarian, D., Linker, D. T., Sutradhar, S. C., Anker, S. D., Cropp, A. B., et al. 2006. The Seattle Heart Failure Model: prediction of survival in heart failure. Circulation, 113**,** 1424-33.
8. Mpanya, D., Celik, T., Klug, E. & Ntsinjana, H. 2020. Machine learning and statistical methods for predicting mortality in heart failure. Heart Fail Rev.
9. Ntusi, N. B. & Mayosi, B. M. 2009. Epidemiology of heart failure in sub-Saharan Africa. Expert Rev Cardiovasc Ther, 7**,** 169-80.
10. Ogah, O. S., Stewart, S., Onwujekwe, O. E., Falase, A. O., Adebayo, S. O., Olunuga, T., et al. 2014. Economic burden of heart failure: investigating outpatient and inpatient costs in Abeokuta, Southwest Nigeria. PLoS One, 9**,** e113032.
11. Szymanski, P. B., M; Mayosi, B 2018. Clinical characteristics and causes of heart failure, adherence to treatment guidelines, and mortality of patients with acute heart failure: Experience at Groote Schuur Hospital, Cape Town, South Africa. S Afr Med J, 108**,** 94-98.
12. Wang, G., Zhang, Z., Ayala, C., Wall, H. K. & Fang, J. 2010. Costs of heart failure-related hospitalisations in patients aged 18 to 64 years. Am J Manag Care, 16**,** 769-76.
13. Wehbe, R. M., Khan, S. S., Shah, S. J. & Ahmad, F. S. 2020. Predicting High-Risk Patients and High-Risk Outcomes in Heart Failure. Heart Fail Clin, 16**,** 387-407.
14. Brink, J, Pennel, T, Seele, K, & Zilla, P. 2017. The world's first human-to-human heart transplant at Groote Schuur Hospital: 50 years later. S Afr Med J, 107(12), 1035-1036.

CHAPTER 2 - LITERATURE SURVEY: MACHINE LEARNING AND STATISTICAL METHODS FOR PREDICTING MORTALITY IN HEART FAILURE

2.1 Introduction

This chapter contains a narrative review of techniques used to create machine learning and statistical models predicting mortality in heart failure. The first section introduces the concept of risk prediction, followed by a comprehensive description of statistical methods such as linear and logistic regression and Cox proportional hazards regression. The section on machine learning techniques discusses classification algorithms such as decision trees, random forests, support vector machines, K-nearest neighbours, naïve Bayes, and artificial networks, followed by techniques for evaluating model performance, limitations, future recommendations, and conclusions.

2.2 Current Knowledge

Both machine learning and statistical methods are used to build predictive models. Machine learning models learn patterns in the dataset and use the knowledge gained to predict outcomes when new data is entered into the model.

This is a published manuscript in the journal titled Heart Failure Reviews. The citation details are Mpanya, D., Celik, T., Klug, E., Ntsinjana, H. Machine learning and statistical methods for predicting mortality in heart failure. Heart Fail Rev 26, 545–552 (2021). https://doi.org/10.1007/s10741-020-10052-y. The next chapter contains the results of a systematic literature search of machine learning models predicting mortality and rehospitalisation in heart failure patients.

Heart Failure Reviews
https://doi.org/10.1007/s10741-020-10052-y

Machine learning and statistical methods for predicting mortality in heart failure

Dineo Mpanya[1,5] · Turgay Celik[2,5] · Eric Klug[3] · Hopewell Ntsinjana[4]

Accepted: 27 October 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Heart failure is a debilitating clinical syndrome associated with increased morbidity, mortality, and frequent hospitalization, leading to increased healthcare budget utilization. Despite the exponential growth in the introduction of pharmacological agents and medical devices that improve survival, many heart failure patients, particularly those with a left ventricular ejection fraction less than 40%, still experience persistent clinical symptoms that lead to an overall decreased quality of life. Clinical risk prediction is one of the strategies that has been implemented for the selection of high-risk patients and for guiding therapy. However, most risk predictive models have not been well-integrated into the clinical setting. This is partly due to inherent limitations, such as creating risk predicting models using static clinical data that does not consider the dynamic nature of heart failure. Another limiting factor preventing clinicians from utilizing risk prediction models is the lack of insight into how predictive models are built. This review article focuses on describing how predictive models for risk-stratification of patients with heart failure are built.

Keywords Heart failure · Predict · Models · Machine learning · Deep learning

Abbreviations

ANN	Artificial neural networks
AUC	Area under the curve
CI	Confidence interval
INTER-CHF	International Congestive Heart Failure
KNN	K-nearest neighbors
LR	Linear regression
LVEF	Left ventricular ejection fraction
NB	Naïve Bayes
NYHA	New York Heart Association
PRAISE	Prospective Randomized Amlodipine Survival Evaluation
SHF	Seattle Heart Failure
SVM	Support vector machine

✉ Dineo Mpanya
Dineo.Mpanya@wits.ac.za

[1] Department of Internal Medicine, Division of Cardiology, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand and the Charlotte Maxeke Johannesburg Academic Hospital, 17 Jubilee Road, Parktown, Johannesburg, Gauteng 2193, South Africa

[2] Faculty of Engineering and Built Environment, School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa

[3] Netcare Sunninghill, Sunward Park Hospitals and Division of Cardiology, Department of Internal Medicine, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand and the Charlotte Maxeke Johannesburg Academic Hospital, Johannesburg, South Africa

[4] Department of Paediatrics and Child Health, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[5] Institute of Data Science , University of the Witwatersrand , Johannesburg, South Africa

Introduction

Heart failure is estimated to affect at least 64.3 million individuals globally [1]. Despite the availability of therapeutic agents and medical devices that aim to prolong survival, many heart failure patients still experience progressive disease and eventually death, particularly in developing countries. The International Congestive Heart Failure (INTER-CHF) study recruited 5823 patients with heart failure and studied them over 1-year [2]. Forty-two percent of these patients were in New York Heart Association (NYHA) functional class III or IV. In the entire study cohort, the 1-year all-cause mortality was 16.5% (95% CI 15.4–17.6). However, patients residing in

Published online: 09 November 2020

Africa had a higher mortality of 34% (95% CI 30.2–37.4) within 1 year of follow-up [2]. The high mortality rate emphasizes the need for new, innovative strategies such as predictive models that aim to risk-stratify patients with heart failure, ensuring that high-risk individuals are identified early and subjected to interventions that delay disease progression.

Both statistical and recently, machine learning techniques are used to create risk predicting models. There is a significant overlap between traditional statistical methods and machine learning algorithms, and one method cannot be deemed superior to the other. While the focus with statistics is to make inferences based on the sample studied, machine learning focuses on learning patterns in datasets and subsequently making predictions based on the learned observations. This review aims to describe what predictive models are, how they are created, and their strengths and weaknesses.

Risk prediction models

Before describing what predictive models are, we will define the verb "predict" and the noun "model." The Oxford English dictionary defines "predict" as "stating that a specified event will happen in the future." A model is defined as a "simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions" [3].

Predictive models are built using both statistical methods and machine learning techniques. However, a paradigm shift has occurred, where machine learning algorithms are becoming popular because of their ability to consider complex associations between clinical parameters. The sophisticated mathematical formulas used in machine learning for automated medical image interpretation have also led to increased confidence in using machine learning algorithms in the health sector. Furthermore, a wider variety of machine learning algorithms is available for classification and regression tasks, allowing the user to handpick the best performing model.

Heart failure predictive models created with statistical methods

Rahimi et al. identified 64 multivariate statistical models used in heart failure patients published between January 1995 and March 2013 [4]. Another systematic review by Di Tanna et al. found 58 multivariate statistical models predicting outcomes in patients with heart failure, published between March 2013 and May 2018 [5]. Seventeen of the models identified by Di Tanna et al. predicted all-cause mortality, nine predicted cardiovascular death, and the rest predicted heart failure hospitalization and composite endpoints [5]. The next section will describe the techniques used to create multivariate linear and logistic regression models.

Linear regression

Linear regression (LR) is used to predict a numeral or continuous outcome. In simple linear regression, the dependent variable (outcome) is featured in the y-axis and the independent variable in the x-axis (Fig. 1). The dependent variable only changes when there is an adjustment of the values in the x-axis. A scatterplot is used to establish a relationship between the dependent and independent variables. A correlation coefficient is then calculated if a linear relationship exists between the variables in the x and y coordinates. The regression line or straight line, is then plotted through the points [6]. The straight line is given by the equation: $y = a + \beta \times x$, where y is the dependent variable; a is the y-intersect of the line; β is the slope of the line, and x is the explanatory or independent variable [6]. The predicted value (y), is found by sketching a line that originates in the x-axis, connects to the regression line, and ultimately a point in the y-axis (Fig. 1). For example, the weight, plotted in the x-axis, will predict a numerical outcome such as height (y). Multiple linear regression analysis involves adding two or more predictor variables concurrently and adjusting their respective regression coefficients [6]. This statistical method is used to determine the relative contribution of each of the predictors. The formula for multiple linear regression is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots . \beta_k x_k + \varepsilon$, wher y is the dependent variable, x is the independent variable, β_0 is the y-intercept, ε the residual, and β_k is the slope coefficient for each predictor variables [7].

Logistic regression

Logistic regression predicts a binary outcome by estimating the probability of an output belonging to a particular class. When building a predictive model using logistic regression, the first step is to utilize the univariable analysis to discover the unadjusted association between each variable and the predicted class [8]. Categorical and continuous variables, usually with a p value less than 0.25, are selected and included in the multivariable analysis [9].

The second step entails placing all the selected variables with a p value less than 0.25 in the multivariable logistic regression model. In the final multivariable regression model, variables with a p value more than 0.05 are subsequently eliminated. In the third step, smoothed scatter plots are created to assess the relationship between continuous variables and the logit scale outcome. Step four entails evaluating for potential interactions between the selected predictor variables. When building a model predicting mortality, the predicted variable will be plotted in the y-axis and classified into two categories: mortality ($y = 1$) and no mortality ($y = 0$). Predictor variables such as age, left ventricular ejection fraction (LVEF), NYHA, and serum sodium levels will be placed in the x-axis. The log odds or logit of mortality will be calculated

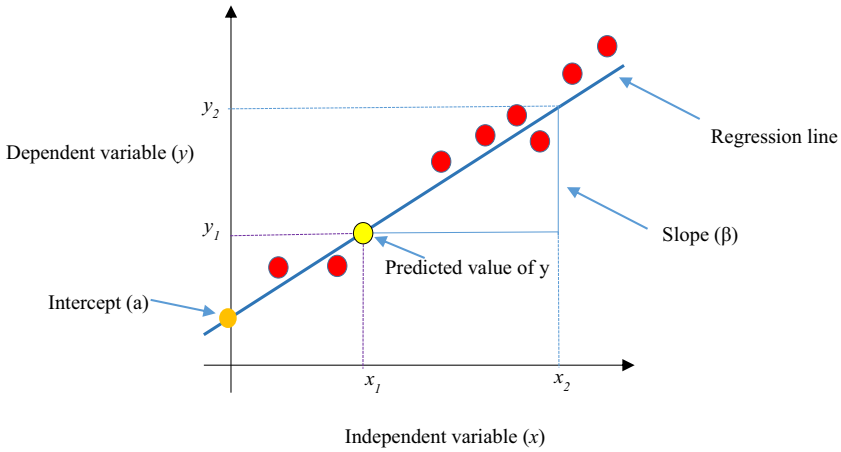


Fig. 1 Linear regression model

using the equation: $\log(p/1\text{-}p) = a + b \times x$, where p is the probability, a is the intercept, b is the regression coefficient of x, and x is the predictor variable [10].

Cox regression

The Seattle Heart Failure (SHF) model is one of the best performing models built using data from 1125 heart failure patients. A Cox regression model was used to select predictors. The predicted versus actual 1-year survival rates in the derivation cohort were 73.4% versus 74.3%. In the five validation cohorts, the predicted versus actual survival rates were 90.5% versus 88.5%, 86.5% versus 86.5%, 83.8% versus 83.3%, 90.9% versus 91.0%, and 89.6% versus 86.7% [11]. The data used to develop the model was obtained from a registry of carefully screened patients from the Prospective Randomized Amlodipine Survival Evaluation (PRAISE) trial, possibly limiting generalizability in the wider population. The online SHF risk calculator may be cumbersome for use in a busy practice as it requires an imputation of 24 clinical and laboratory parameters when estimating the risk of one and 5-year mortality.

Cox proportional hazard is a regression model used to measure an association between survival time and predictor variables. Patients are tracked from the time they enter the study until the occurrence of an event. An event is defined as the occurrence of a disease, relapse from therapy, discharge from hospital, or death. Multiple predictors or covariates of survival are added to the model to estimate their influence on survival, analogous to a multivariate regression model. The Cox proportional hazard model highlights how survival differs between the groups of patients studied by estimating the hazard ratio, unlike the Kaplan-Meier survival curve, which only assesses whether survival between groups differs (log-rank test with a p value < 0.05).

The Cox proportional hazard model, as the name implies, assumes that the covariates remain constant over time. This assumption has been challenged by Giolo et al., who suggested that heart failure prognostication models should be built using covariates with time varying-effects [12]. In their study, 500 patients with heart failure were followed between 2002 and 2006. Survival was analyzed using Cox proportional hazard, variations of the Cox's model and the Aalen's additive model. They used the patients' age, serum sodium levels, haemoglobin, serum creatinine, and the LVEF as covariates. A high haemoglobin and LVEF were associated with a lower risk of mortality, while the impact of age and sodium remained constant over time [12].

Heart failure predictive models created with machine learning techniques

The fundamental principle of machine learning is learning from data and predicting an output, where $y = f(x)$. The input (x) is any group of features, commonly known as variables used to predict an output (y). A machine is given a learning task under any of the following conditions: supervised, unsupervised, semi-supervised, and through reinforcement learning [13]. A computer can learn from a dataset with a labeled output or outcome in supervised learning. For example, in a dataset comprising of patients with heart failure, we may include a column header, mortality as the output. Within the column "mortality," we then capture whether the patient demised or not by adding yes or no, respectively.

Unsupervised learning takes place when a computer learns from unlabelled datasets. The data is subsequently clustered into similar categories. Ahmad et al. used unsupervised learning in a large clinical data set of more than 40000 Swedish heart failure patients to separate patients into four groups [14]. After 1 year of follow-up, there was a marked difference in outcomes per cluster, with 1-year survival rates of 69%, 77%, 92%, and 93% [14].

In contrast, when the same heart failure patients were categorized based on the LVEF, only slight differences in 1-year

survival (80%, 81%, 83%, and 84%) were noted [14]. Semi-supervised learning occurs when learning takes place using a combination of labeled and unlabelled datasets [13]. In reinforcement learning, a computer learns by trial and error by performing tasks and receiving penalty and reward scores for inaccurate and accurate responses, respectively [15].

How are predictive models built with machine learning techniques?

Prior to building a predictive model using machine learning, the data must be cleaned. When cleaning continuous variables, one can create box plots and verify the values of all the outliers. The next step is to train the model. Training involves splitting the dataset, where the majority of the dataset (e.g., 70%) is used to train the model. During the training process, the selected algorithm learns patterns in the dataset. The remaining dataset is used for testing and validating the model. During the test phase, the labeled output variable, for example, "mortality" is hidden, and the machine learning algorithm attempts predicting whether the patient demised or not.

Numerous supervised learning algorithms are available and are selected based on the task at hand, be it for classification of a binary outcome or regression analysis of continuous features. Some of the commonly used machine learning algorithms for classification or regression analysis include decision trees, random forests, support vector machines (SVM), K-nearest neighbors (KNN), and Naïve Bayes (NB).

Decision trees

A decision tree is a simple tree-shaped algorithm used to predict a categorical (binary) or numerical outcome. To illustrate how a decision tree classifies data with a categorical outcome, we will use data from patients with heart failure to build a model predicting mortality, as shown in Fig. 2. Each decision tree comprises of branches, leaf, and root nodes [16]. The root node is the most superiorly located node, and the leaf or terminal node is the final node that carries the decision. Each branch of the tree represents a decision, occurrence, or reaction. Data is then partitioned into subsets containing similar values by using mathematical equations such as the Gini index, Chi-square, information gain, and a reduction in variance [17].

Based on the known predictors of mortality in heart failure patients, the decision tree algorithm can categorize a patient with heart failure into class A (high risk) and class B (low risk) (Fig. 2). Decision trees require minimal data preparation before building a predictive model, and their performance is not affected by non-linear data distribution. The main disadvantage is overfitting, which occurs when a model demonstrates high-accuracy levels during training, makes inaccurate predictions during testing, or has significantly lower test accuracy than the training accuracy [18].

Random forest

A random forest algorithm is an ensemble decision tree that operates by constructing multiple decision trees, creating a forest. When building a classification model using a random forest algorithm, the first step is to select random samples from a given dataset. A decision tree is then constructed for each sample. Predicted results are obtained from each decision tree. Lastly, the most voted class from the predictions made by each decision tree is selected as the final predicted class [19]. In a regression model, where the output class is numerical, the mean or average value of the predicted output is used. The random forest algorithm is ideal for handling datasets with missing values. It also performs well on large datasets and can rank features in the order of importance. The random forest algorithm's main disadvantage is that it is computationally expensive, requiring more training time than most algorithms [18].

Support vector machines

Support vector machines (SVM) perform classification and regression analysis for both linear and non-linear data. The main advantage of SVM lies in their ability to classify more than two classes within the same dataset simultaneously. Support vector machines categorize data by focusing on the observations on the edge of each class (Fig. 3). A line or hyperplane is then placed between the classes. The maximum distance between the classes (support vectors) and the hyperplane is then chosen [20]. When the data cannot be linearly separable, the data is transformed using a kernel function from low dimension to high dimensional structure, rendering the data separable.

K-nearest neighbors

The K-nearest neighbors (KNN) algorithm does not require training. It is used for predicting a binary or continuous output. Data is separated into clusters, and the number of nearest neighbors is specified by stating the value of "K", a constant [13]. If $K = 3$, three nearest neighbors are selected based on the distance measured between the new data point and nearest neighbors. If $K = 6$, the distance between the new data and the six nearest neighbors is calculated. The new data point is then placed in the class with the majority of votes. Figure 4 depicts how a KNN algorithm classifies a new data point. The biggest challenge with using K-nearest neighbors lies in deciding on the optimal number of neighbors to consider when classifying a new data point. K-nearest neighbors also do not perform well in an imbalanced dataset, since the model will

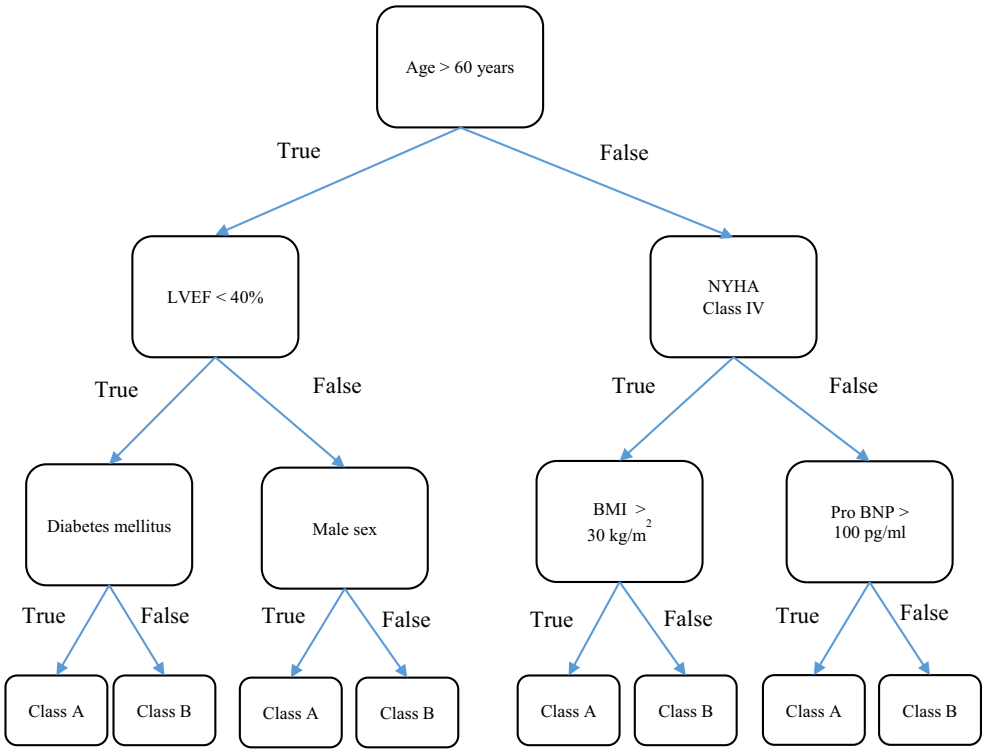


Fig. 2 A decision tree algorithm based on predictors of mortality in heart failure patients. BMI, body mass index; BNP, beta natriuretic peptide; LVEF, left ventricular ejection fraction; NYHA, New York Heart Association

give preference to the class with the higher number of observations [18].

Naïve Bayes

Naïve Bayes is a probabilistic method borrowed from statistics, based on the Bayes' theorem, which describes the probability of an event occurring based on previous knowledge of the conditions associated with that particular event [21]. For classification tasks, a model learns the probabilities of an object for belonging to a specific class. The name "naïve" is derived from the fact that a Naïve Bayes model assumes that the occurrence of a particular feature is not dependent on the occurrence of other features [22]. Pakhomov et al. used a Naïve Bayes classifier and another classification algorithm, perceptron to train models capable of identifying patients'

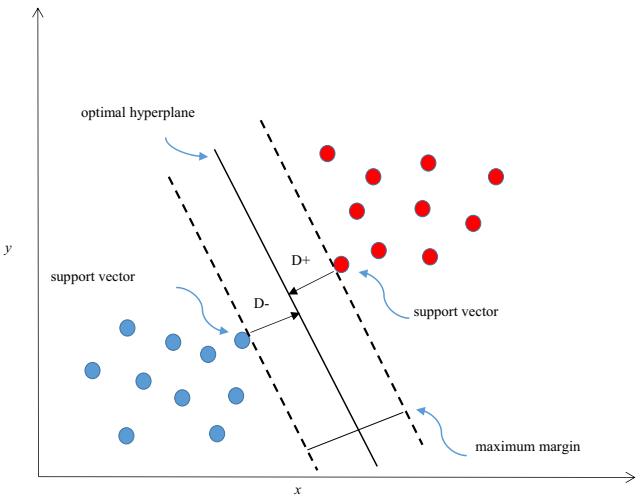


Fig. 3 Support vector machine algorithm depicting classification of linearly separable data using a decision boundary, the optimal hyperplane. D− and D+ represent the shortest distance to the closest negative and positive classes, respectively and the optimal hyperplane

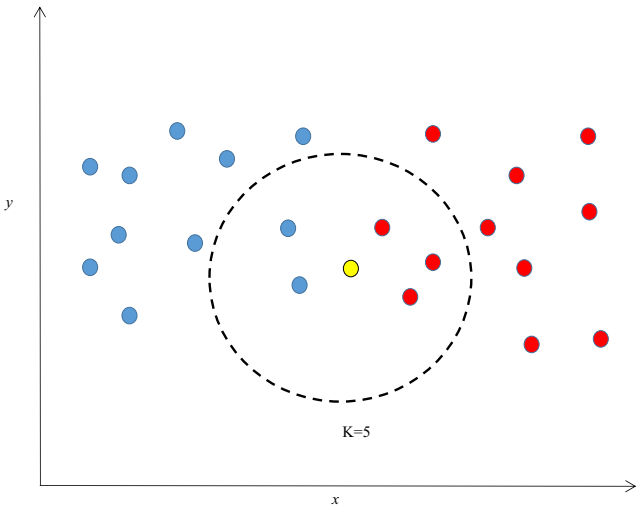


Fig. 4 K-nearest neighbor algorithm. The circle shows five data points closest to the new data point (yellow). Since the majority of the neighbors (3 out of 5) are in the red class, the new data point will be classified as belonging to the red class

clinical notes with a diagnosis of congestive cardiac failure [23]. The Naïve Bayes classifier yielded better recall on positive samples (95% vs 86%) but had low accuracy (57% vs 65%) [23]. Although the Naïve Bayes classifier is more comfortable implementing and has a lesser training period, its main limitation is the assumption that all features are mutually independent. For example, in the study by Pakhomov et al., the terms "heart" and "failure" were considered to occur independently of each other, rendering the classifier unattractive for text categorization [23].

Artificial neural networks

Artificial neural networks (ANNs) resemble neurons, the functional unit of the central nervous system. Each ANN is characterized by an input layer, a hidden layer, and an output layer [24]. The input layer represents features extracted by the model. A feature can be any continuous or categorical variable, such as the left ventricular ejection fraction, weight, or gender. The features in the input layer are used to predict an output, represented by one or more categories. The nodes displayed in the input layer communicate with each node in the hidden layer (Fig. 5). Initially, an arbitrary value between zero and one is used to grade the connection between the input and hidden layers. The weighted sum of the input layers' signal is then passed through the activation function. The sigmoid function transforms each inputs' weighted sum with a negative value to a value close to zero. Positive values are transformed into a value close to one. Most neural networks are now designed with more than one hidden layer, increasing their capacity to classify non-linear data.

Kwon et al. studied 2165 patients with acute heart failure to predict in-hospital mortality, 1-year and 3-year risk of mortality using deep neural networks, random forest, logistic regression, support vector machines, and a Bayesian network. Deep neural networks had the largest area under the curve (AUC) of 0.880, 0.782, and 0.813 for predicting in-hospital, 1-year and 3-year risk of mortality, respectively [25].

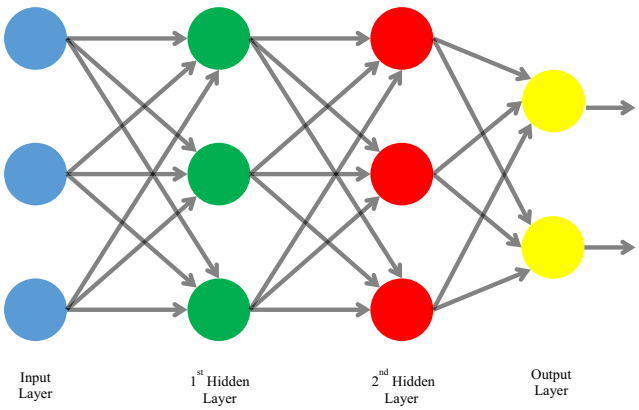


Fig. 5 An artificial neural network showing nodes in the input layer communicating with each node in the first hidden layer

Model performance

Models predicting a binary outcome

The AUC is the most common parameter used to evaluate the performance of classification models [26]. Models with good discriminatory abilities have larger AUC, generally greater than 0.70, whereas those with AUC less than 0.50 lack discriminatory abilities [26]. Model performance is also evaluated with a confusion matrix, allowing one to calculate accuracy, precision, sensitivity, and specificity. Other performance metrics include the F measure or F score, geometric mean, and logarithmic loss [27].

Models predicting a continuous outcome

The overall fit of a linear regression model is evaluated with the r-squared (r^2). Simply put, the r^2 is a measure of how much the prediction error is reduced, relative to how much potential error there is [28]. The error is estimated with indices such as the mean absolute error, mean squared error, and the root mean squared error [29].

Model flexibility

Model overfitting and underfitting are the most common problems encountered when evaluating performance. Overfitting occurs when a model shows high accuracy scores during training and low accuracy scores during validation. Model overfitting is minimized by adding more data to the training set and reducing the number of layers in the neural network. Underfitting occurs when the model fails to classify data or make predictions during the training phase. A low accuracy score and a high loss identify a model that is underfitting [30].

Limitations of statistical and machine learning predictive models

Despite numerous models predicting outcomes in patients with heart failure, only a few are easily accessible online as risk score calculators [11, 31, 32]. Although the lack of external validation of the predictive models is the driving factor for the limited availability, several inherent limitations exist for the relatively unhurried integration of predictive models into clinical practice.

Some clinicians find risk calculation cumbersome [33]. Also, most risk prediction models were designed before new heart failure drugs that further reduce mortality and the rate of heart failure hospitalizations were discovered [34, 35].

Class imbalance refers to the disproportionality between the classes of data used to train the predictive

model [36], a common problem that is not unique to medical data. When the training data with the negative outcome (e.g., dead) has significantly fewer observations than the class (e.g., alive) with the majority of observations, the classification algorithm is inclined to favor the majority class. This poses problems as the minority class, which bears the outcome of interest, will have a low accuracy score. Fortunately, it is possible to address class imbalance problems by manipulating data, algorithms, or both [36].

Conclusion

Both machine learning techniques and statistical methods should be employed when creating predictive models. Prior to clinical application, the best performing model should be externally validated in a cohort of patients not used for model derivation.

Future directions and recommendations

Unsupervised machine learning algorithms should be considered for the detection of data patterns not recognized by clinicians. Incorporating genomic data, biomarkers, imaging data, and features representing the patients' socioeconomic and psychological standing can further improve the accuracy levels of predictive models. Risk predictive models should be created with dynamic data and ideally be embedded in automated data capturing tools. The increase in training data volume will ultimately lead to a robust predictive model with higher accuracy levels.

Authors' contributions HN, TC, and DM contributed to the study conception and design. DM conducted the literature search and wrote the first draft of the manuscript. All authors (HN, TC, EK and DM) commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Dr. Dineo Mpanya is a full-time PhD Clinical Research fellow in the Division of Cardiology at the University of the Witwatersrand. Her PhD is funded by the Professor Bongani Mayosi Netcare Clinical Scholarship, the Discovery Academic Fellowship [Grant No. 03902], the Carnegie Corporation of New York [Grant No. b8749], and the South African Heart Association.

Compliance with ethical standards

Conflict of interest EK has received consulting fees from Novartis Pharmaceuticals, Pfizer, Servier, Takeda, and AstraZeneca. All other authors declare that they have no conflict of interest.

Ethical approval Not requested as the manuscript is a narrative literature review.

References

1. Lippi G, Sanchis-Gomar F (2020) Global epidemiology and future trends of heart failure. AME Med J 2020(5):15
2. Dokainish H, Teo K, Zhu J, Roy A, AlHabib KF, ElSayed A et al (2017) Global mortality variations in patients with heart failure: results from the International Congestive Heart Failure (INTER-CHF) prospective cohort study. Lancet Glob Health 5(7):e665–e672
3. Concise Oxford English Dictionary, 11th ed. 2006. Oxford University Press, New York
4. Rahimi K, Bennett D, Conrad N, Williams TM, Basu J, Dwight J, Woodward M, Patel A, McMurray J, MacMahon S (2014) Risk prediction in patients with heart failure: a systematic review and analysis. JACC Heart Fail 2(5):440–446
5. Di Tanna GL, Wirtz H, Burrows KL, Globe G (2020) Evaluating risk prediction models for adults with heart failure: A systematic literature review. PLoS ONE 15(1): e0224135. https://doi.org/10.1371/journal.pone.0224135
6. Schneider A, Hommel G, Blettner M (2010) Linear regression analysis: part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int 107(44):776–782
7. Alexopoulos EC (2010) Introduction to multivariate regression analysis. Hippokratia 14(Suppl 1):23–28
8. Zhang Z (2016) Model building strategy for logistic regression: purposeful selection. Ann Transl Med 4(6):111–111
9. Mickey RM, Greenland S (1989) The impact of confounder selection criteria on effect estimation. Am J Epidemiol 129(1):125–137
10. Pandis N (2017) Logistic regression: part 1. Am J Orthod Dentofac Orthop 151(4):824–825
11. Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, Anand I, Maggioni A, Burton P, Sullivan MD, Pitt B, Poole-Wilson PA, Mann DL, Packer M (2006) The Seattle Heart Failure Model: prediction of survival in heart failure. Circulation 113(11):1424–1433
12. Giolo SR, Krieger JE, Mansur AJ, Pereira AC (2012) Survival analysis of patients with heart failure: implications of time-varying regression effects in modeling mortality. PLoS One 7(6): e37392–e37392
13. Awad M, Khanna R (2015) Machine learning. In: Efficient learning machines: theories, concepts, and applications for engineers and system designers. Apress, Berkeley, pp 1–18
14. Ahmad T, Lund LH, Rao P, Ghosh R, Warier P, Vaccaro B et al (2018) Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. J Am Heart Assoc 7(8):e008081
15. Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS et al (2018) Current applications and future impact of machine learning in radiology. Radiology 288(2):318–328
16. Song Y-Y, Lu Y (2015) Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry 27(2): 130–135
17. Kingsford C, Salzberg SL (2008) What are decision trees? Nat Biotechnol 26(9):1011–1013
18. Uddin S, Khan A, Hossain ME, Moni MA (2019) Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19(1):281
19. Fawagreh K, Gaber MM, Elyan E (2014) Random forests: from early developments to recent advancements. Syst Sci Control Eng 2(1):602-609. https://doi.org/10.1080/21642583.2014.956265
20. Shmilovici A (2005) Support vector machines. In: Maimon O, Rokach L (eds) Data mining and knowledge discovery handbook. Springer US, Boston, pp 257–276

21. Hackenberger BK (2019) Bayes or not Bayes, is this the question? Croat Med J 60(1):50–52
22. Ali L, Khan SU, Golilarz NA, Yakubu I, Qasim I, Noor A et al (2019) A feature-driven decision support system for heart failure prediction based on statistical model and Gaussian Naive Bayes. Comput Math Methods Med 2019:6314328
23. Pakhomov SV, Buntrock J, Chute CG (2005) Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier. J Biomed Inform 38(2):145–153
24. Sengupta S, Basak S, Saikia P, Paul S, Tsalavoutis V, Atiah F, Ravi V, Peters A (2020) A review of deep learning with special emphasis on architectures, applications and recent trends. Knowl-Based Syst 194:105596
25. Kwon JM, Kim KH, Jeon KH, Lee SE, Lee HY, Cho HJ, Choi JO, Jeon ES, Kim MS, Kim JJ, Hwang KK, Chae SC, Baek SH, Kang SM, Choi DJ, Yoo BS, Kim KH, Park HY, Cho MC, Oh BH (2019) Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. PLoS One 14(7):e0219302
26. Zou KH, O'Malley AJ, Mauri L (2007) Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation 115(5):654–657
27. Powers D, Ailab (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. J Mach Learn Technol 2:2229–3981
28. Hagquist C, Stenbeck M (1998) Goodness of fit in regression analysis – R2 and G2 reconsidered. Qual Quant 32(3):229–245
29. Emmert-Streib F, Dehmer M (2019) Evaluation of regression models: model assessment, model selection and generalization error. Mach Learn Knowl Extr 1:521–551
30. Nichols JA, Herbert Chan HW, Baker MAB (2019) Machine learning: applications of artificial intelligence to imaging and diagnosis. Biophys Rev 11(1):111–118
31. Pocock SJ, Ariti CA, McMurray JJV, Maggioni A, Køber L, Squire IB et al (2012) Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. Eur Heart J 34(19):1404–1413
32. Peterson PN, Rumsfeld JS, Liang L, Albert NM, Hernandez AF, Peterson ED, Fonarow GC, Masoudi FA, American Heart Association Get With the Guidelines-Heart Failure Program (2010) A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program. Circulation 3(1):25–32
33. Eichler K, Zoller M, Tschudi P, Steurer J (2007) Barriers to apply cardiovascular prediction rules in primary care: a postal survey. BMC Fam Pract 8(1):1
34. McMurray JJV, Packer M, Desai AS, Gong J, Lefkowitz MP, Rizkala AR et al (2014) Angiotensin–neprilysin inhibition versus enalapril in heart failure. N Engl J Med 371(11):993–1004
35. McMurray JJV, Solomon SD, Inzucchi SE, Køber L, Kosiborod MN, Martinez FA, Ponikowski P, Sabatine MS, Anand IS, Bělohlávek J, Böhm M, Chiang CE, Chopra VK, de Boer RA, Desai AS, Diez M, Drozdz J, Dukát A, Ge J, Howlett JG, Katova T, Kitakaze M, Ljungman CEA, Merkely B, Nicolau JC, O'Meara E, Petrie MC, Vinh PN, Schou M, Tereshchenko S, Verma S, Held C, DeMets D, Docherty KF, Jhund PS, Bengtsson O, Sjöstrand M, Langkilde AM, DAPA-HF Trial Committees and Investigators (2019) Dapagliflozin in patients with heart failure and reduced ejection fraction. N Engl J Med 381(21):1995–2008
36. Rekha G, Tyagi AK & Reddy VK (2019) A Wide Scale Classification of Class Imbalance Problem and its Solutions: A Systematic Literature Review. J Comput Sci 15(7):886-929. https://doi.org/10.3844/jcssp.2019.886.929

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

CHAPTER 3 – MANUSCRIPT ONE: SYSTEMATIC REVIEW OF MACHINE LEARNING MODELS PREDICTING MORTALITY AND REHOSPITALISATION IN HEART FAILURE

3.1 Introduction

This section of the thesis contains an original research article in the form of a systematic literature review of machine learning models predicting mortality and rehospitalisation in heart failure. It comprises data extracted and synthesised from 30 individual studies using machine learning to predict mortality and rehospitalisation in patients with acute and chronic heart failure.

3.2 Current Knowledge

Machine learning algorithms are valuable tools for identifying high-risk patients. However, incorporating machine learning algorithms-derived risk calculators into clinical practice has been rather slow, in contrast to the relatively high and rapid usage in the banking and insurance sectors.

3.3 Aims and Objectives

This systematic review aims to identify factors restricting the integration of machine learning derived risk scores into clinical practice when treating adults with acute and chronic heart failure. The results in this chapter are accompanied by Appendix A, where I state reasons for excluding 71 studies after reviewing full-text articles. The citation for the manuscript is Mpanya, D., Celik, T., Klug, E., Ntsinjana, H. Predicting mortality and hospitalisation in heart failure using machine learning: A systematic literature review, IJC Heart & Vasculature, Volume 34, 2021, 100773. https://doi.org/10.1016/j.ijcha.2021.100773. The next chapter describes the baseline demographic and clinical profile of the heart failure cohort used in this study.

IJC Heart & Vasculature 34 (2021) 100773

Contents lists available at ScienceDirect

IJC Heart & Vasculature

journal homepage: www.journals.elsevier.com/ijc-heart-and-vasculature

Predicting mortality and hospitalization in heart failure using machine learning: A systematic literature review

Dineo Mpanya [a,e,*], Turgay Celik [b,e], Eric Klug [c], Hopewell Ntsinjana [d]

[a] Division of Cardiology, Department of Internal Medicine, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[b] School of Electrical and Information Engineering, Faculty of Engineering and Built Environment, University of the Witwatersrand, Johannesburg, South Africa
[c] Netcare Sunninghill, Sunward Park Hospitals and Division of Cardiology, Department of Internal Medicine, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand and the Charlotte Maxeke Johannesburg Academic Hospital, Johannesburg, South Africa
[d] Department of Paediatrics and Child Health, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[e] Wits Institute of Data Science, University of the Witwatersrand, Johannesburg, South Africa

ARTICLE INFO

Article history:
Received 10 January 2021
Received in revised form 11 March 2021
Accepted 23 March 2021

Keywords:
Heart failure
Risk score
Predictive modelling
Machine learning
Sub-Saharan Africa
Mortality
Hospitalization

ABSTRACT

Objective: The partnership between humans and machines can enhance clinical decisions accuracy, leading to improved patient outcomes. Despite this, the application of machine learning techniques in the healthcare sector, particularly in guiding heart failure patient management, remains unpopular. This systematic review aims to identify factors restricting the integration of machine learning derived risk scores into clinical practice when treating adults with acute and chronic heart failure.
Methods: Four academic research databases and Google Scholar were searched to identify original research studies where heart failure patient data was used to build models predicting all-cause mortality, cardiac death, all-cause and heart failure-related hospitalization.
Results: Thirty studies met the inclusion criteria. The selected studies' sample size ranged between 71 and 716 790 patients, and the median age was 72.1 (interquartile range: 61.1–76.8) years. The minimum and maximum area under the receiver operating characteristic curve (AUC) for models predicting mortality were 0.48 and 0.92, respectively. Models predicting hospitalization had an AUC of 0.47 to 0.84. Nineteen studies (63%) used logistic regression, 53% random forests, and 37% of studies used decision trees to build predictive models. None of the models were built or externally validated using data originating from Africa or the Middle-East.
Conclusions: The variation in the aetiologies of heart failure, limited access to structured health data, distrust in machine learning techniques among clinicians and the modest accuracy of existing predictive models are some of the factors precluding the widespread use of machine learning derived risk calculators.
© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Predictive analytics is applied across many industries, typically for insurance underwriting, credit risk scoring and fraud detection [1-3]. Both statistical methods and machine learning algorithms are used to create predictive models [4]. In heart failure, machine learning algorithms create risk scores estimating the likelihood of a heart failure diagnosis and the probability of outcomes such as all-cause mortality, cardiac death and hospitalization [5-13].

Clinicians treating heart failure patients may underestimate or overestimate the risk of complications and may battle with dose titration, failing to reach target dosages when prescribing oral medication such as beta-blockers [14,15]. Despite these challenges, risk calculators are still not widely used to guide the management of heart failure patients. Most clinicians find risk calculation time consuming and are not convinced of the value of the information derived from predictive models [15,16]. Moreover, the lack of integration of risk scores predicting heart failure outcomes into management guidelines may diminish clinicians' confidence when using risk calculators. Also, clinicians may question the integrity of unsupervised machine learning and deep learning methods since algorithms single-handedly select features (*predictors*) without human input.

Machine learning and its subtype, deep learning, have shown an impressive performance in medical image analysis and interpretation [17]. Convolutional neural networks (CNN) were trained to classify chest radiographs as pulmonary tuberculosis (TB) or nor-

* Corresponding author.
E-mail address: Dineo.Mpanya@wits.ac.za (D. Mpanya).

https://doi.org/10.1016/j.ijcha.2021.100773
2352-9067/© 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

mal using chest radiographs from 685 patients. The ensemble of CNN's performed well with an area under the receiver operating characteristic curve (AUC) of 0.99 [17]. These impressive results have resulted in the commercialization of chest x-ray interpretation software [18]. The availability of such software can play a critical role in remote areas with limited or no access to radiologists, as CNN can potentially identify subtle manifestations of TB on chest radiographs, leading to prompt initiation therapy, curbing further transmission of TB. Amid these capabilities, the uptake of machine learning techniques in the healthcare sector remains limited. This systematic review aims to identify models predicting mortality and hospitalization in heart failure patients and discuss factors that restrict the widespread clinical use of risk scores created with machine learning algorithms.

2. Methods

2.1. Search strategy for identification of relevant studies

A systematic literature search was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Literature searches were conducted in MEDLINE, Google Scholar, Springer Link, Scopus, and Web of Science. The search string contained the following termi-

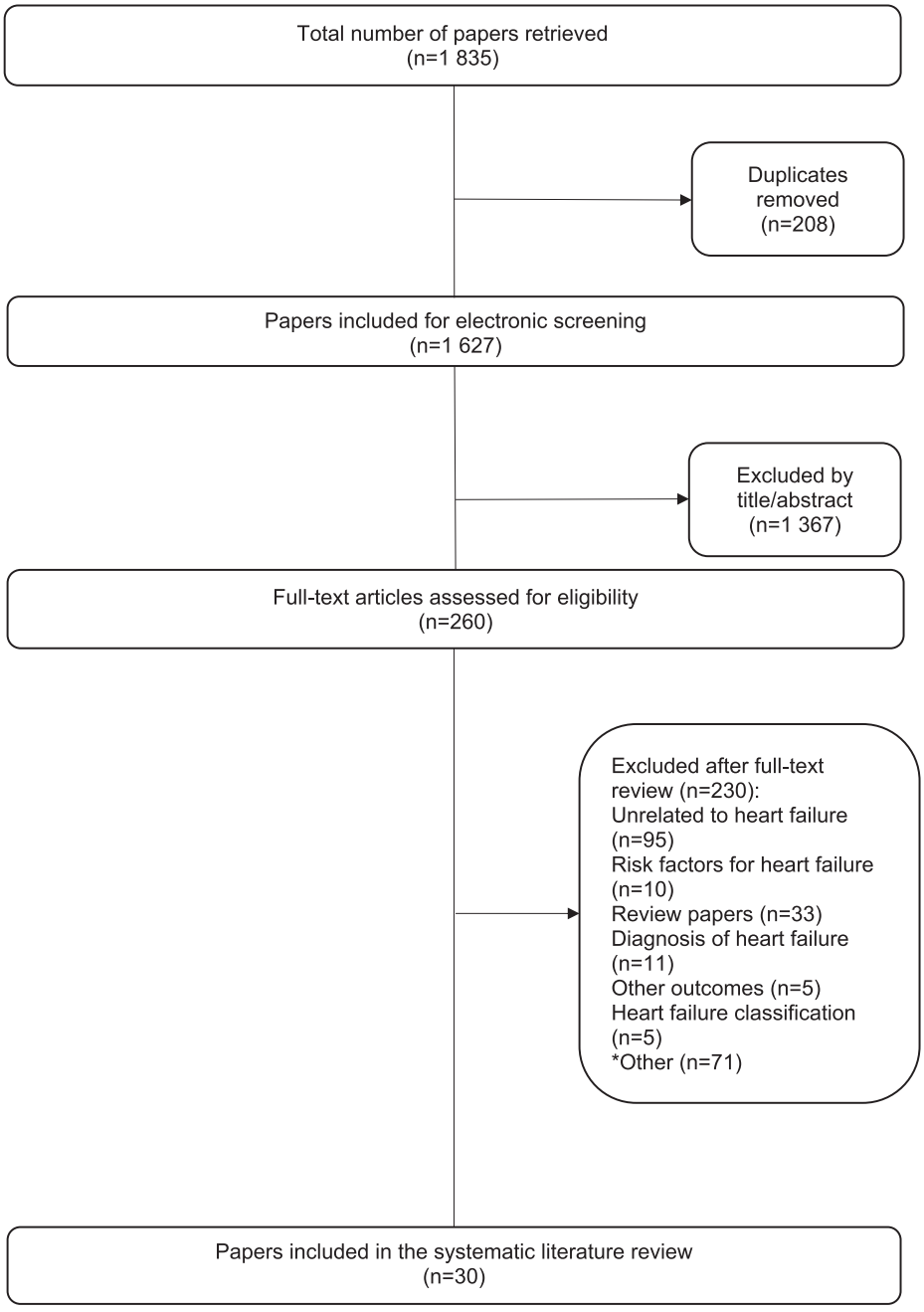


Fig. 1. Flow chart of the systematic literature search.

nology: (Mortality OR Death OR Readmission OR Hospitalization) AND (Machine Learning OR Deep Learning) AND (Heart Failure OR Heart Failure, Diastolic OR Heart Failure, Systolic).

2.2. Review methods and selection criteria

Studies reported in languages other than English were not included. A single reviewer screened titles, abstracts and full-text articles and made decisions regarding potential eligibility. Studies were eligible if they reported models predicting all-cause or cardiac mortality or all-cause or heart failure-related hospitalization in heart failure patients. Models included in the study were created using machine learning algorithms and/or deep learning. We did not include studies using solely logistic regression for a classification task. Logistic regression analysis is a machine learning algorithm borrowed from traditional statistics. When logistic regression is used as a machine learning algorithm, the algorithm is initially trained to identify clinical data patterns using a dataset with labelled classes, a process known as supervised learning. After that, the logistic regression algorithm attempts to classify new data into two or more categories based on "posteriori knowledge."

2.3. Data extraction

The following items were extracted: study region, data collection period, sample size, age, gender, cause of heart failure (ischaemic vs non-ischaemic), predictor variables, handling of missing data, internal and external validation, all-cause mortality and cardiovascular death rate, all-cause hospitalization rate and performance metrics (sensitivity, accuracy, AUC or c-statistics and F-score). Summary statistics were generated with STATA MP version 13.0 (StataCorp, Texas).

3. Results

3.1. The review process

The initial search yielded 1 835 research papers. After screening titles and abstracts, 1 367 did not meet the inclusion criteria. Excluded papers were predominantly theoretical reviews and conference papers in the field of computer science. Two hundred and sixty full-text articles were assessed for eligibility. A further 230 studies were excluded, leaving thirty papers legible for analysis (Fig. 1). Reasons for excluding 230 studies are provided as **supplementary data**.

3.2. Characteristics of the included studies

The source of data in the majority of the studies were electronic health records (EHR) (n = 16), followed by claims data (n = 5), trial data (n = 3), registry (n = 3) and data obtained from research cohorts (n = 3). Data was collected from hospitalized patients in twelve studies. The sample size in the predictive models ranged between 71 and 716 790, with the smallest sample size used to predict survival in patients with advanced heart failure managed with second-generation ventricular assist devices [19]. Within the 30 studies, twelve studies created models predicting mortality. Another 13 studies predicted hospitalization, and five studies predicted both mortality and hospitalization. The data used to create predictive models was collected between 1993 and 2017 (Table 1**).** Of the 30 included studies, 22 included data originating from North America, seven from Asia and six from Europe. There were no studies conducted in Africa or Middle-East (Fig. 2).

3.3. Clinical characteristics of patients with heart failure

The majority of studies reported the patients' age (93%) and gender (87%). The median age was 72.1 (61.1–76.85) years. Between 14.0 and 83.9% of the extracted studies' participants had ischaemic heart disease (Table 2). In total, 30% of studies mentioned Black patients. Between 0.95% and 100% of the individuals were Black, with one study enrolling only African American males with heart failure [20].

3.4. Machine learning algorithms

Only eight (27%) studies used a single algorithm to build a predictive model. Nineteen studies (63%) used logistic regression, 53% random forests, and 36% of studies used decision trees to create predictive models. The rest of the algorithms are depicted in Fig. 3.

3.5. Predictors

Twelve (36.4%) studies did not report on the number of predictors or features used. The number of predictors in the identified studies were between 8 and 4 205. Some authors only mentioned the number of predictors and did not list them. Age, gender, diastolic blood pressure, left ventricular ejection fraction (LVEF), estimated glomerular filtration rate, haemoglobin, serum sodium, and blood urea nitrogen were some of the predictors of mortality identified in the extracted studies [10,11,13]. Predictors of hospitalization included ischaemic cardiomyopathy, age, LVEF, hypotension, haemoglobin, creatinine, and potassium serum levels [7].

3.6. Model development, internal and external validation

When creating a predictive model using machine learning, data is generally partitioned into three or four datasets. In the studies extracted, between 60 and 80% of the data was used for training models, while the rest was used for testing and/or internally validating the models. Although the data on model validation was scanty, external validation was explicitly mentioned in two studies. None of the models were externally validated using data originating from Africa or the Middle-East.

3.7. Model performance and evaluation metrics

Parameters used to evaluate model performance were the confusion matrix, reporting sensitivity, specificity, positive and negative predictive value, accuracy, and precision. Most studies also reported the f-score, AUC, concordance statistic (C-statistic), and recall. The minimum and maximum AUC for models predicting mortality were 0.477 and 0.917, and models predicting hospitalization had an AUC between 0.469 and 0.836 (Table 3).

4. Discussion

This systematic review highlights several factors that restrict the use of risk scores created with machine learning algorithms in the clinical setting. The existence of clinical information with prognostic significance such as the New York Heart Association functional class in the free-text format in EHR systems may result in models with low predictive abilities if such critical data is omitted when building predictive models. Fortunately, newer emerging techniques such as bidirectional long short-term memory with a conditional random fields layer have been introduced to remedy the problem of free-text in EHR [21,22].

Risk scores derived from heart failure patients residing in North America or Europe may not be suitable for application in low and

Table 1
Characteristics of the included studies.

Study ID	Data collection period	No. of patients	Setting	Data source	No. of features	Primary outcome assessed
Adler, E.D (2019) [10]	2006–2017	5 822	Inpatient and outpatient	EHR and Trial	8	All-cause mortality
Ahmad, T (2018) [30]	2000–2012	44 886	Inpatient and outpatient	Registry	8	1-year all-cause mortality
Allam, A (2019) [31]	2013	272 778	Inpatient	Claims dataset	50	30-day all-cause readmission
Angraal, S (2020) [13]	2006–2013	1 767	Inpatient	Trial	26	All-cause mortality and HF hospitalization
Ashfaq, A (2019) [32]	2012–2016	7 655	Inpatient and outpatient	EHR		30-day all-cause readmission
Awan, SE (2019) [33]	2003–2008	10 757	Inpatient and outpatient	EHR	47	30-day HF-related readmission and mortality
Chen, R (2019) [34]	2014–2017	98	Inpatient	Prospective Clinical and MRI	32	Cardiac death, heart transplantation and HF-related hospitalization
Chicco, D (2020) [11]	2015	299	Inpatient	Medical records	13	One year survival
Chirinos, J (2020) [35]	2006–2012	379	Inpatient	Trial	48	Risk of all-cause death or heart failure-related hospital admission
Desai, R.J (2020) [6]	2007–2014	9 502	Inpatient and outpatient	Claims data and EHR	62	All-cause mortality and HF hospitalization, total costs for hospitalization, outpatient visits, and medication
Frizzell, J.D (2017) [36]	2005–2011	56 477	Inpatient	Registry and claims data		All-cause readmission 30-days after discharge
Gleeson, S (2017) [37]	2010–2015	295	Inpatient	Echo database & EHR	291	All-cause mortality and heart failure admissions
Golas, S.B (2018) [12]	2011–2015	11 510	Inpatient and outpatient	EHR	3 512	All-cause 30-day readmission, healthcare utilization cost
Hearn, J (2018) [38]	2001–2017	1 156		EHR and Cardiopulmonary stress test data		All-cause mortality
Hsich, E (2011) [9]	1997–2007	2 231		Cardiopulmonary stress test data	39	All-cause mortality
Jiang, W (2019) [39]	2013–2015	534	Inpatient	EHR	57	30-day readmission
Kourou, K (2016) [19]		71		Pre and post-operative data	48	1-year all-cause mortality
Krumholz, H (2019) [40]	2013–2015	716 790	Inpatient	Claims dataset		All-cause death within 30-days of admission
Kwon, J (2019) [5]	2016–2017	2 165 (training dataset)	Inpatient	Registry		12 and 36-month in-hospital mortality
Liu, W (2020) [41]		303 233 (heart failure)	Inpatient	Readmission database		Admission 3H myocardial infarction, congestive heart failure and pneumonia 30-day readmission
Lorenzoni, G (2019) [7]	2011–2015	380	Inpatient	Research data		Hospitalization among patients with heart failure
Maharaj, S.M (2018) [42]	2015	1 778	Inpatient	EHR	56	30-day readmission
McKinley, D (2019) [20]	2012–2015	132	Inpatient	EHR	29	All-cause readmission within 30-days
Miao, F (2017) [43]	2001–2007	8 059		Public database	32	1-year in-hospital mortality
Nakajima, K (2020) [24]	2005–2016	526		Multicentre database	13	2-year life-threatening arrhythmic events and heart failure death
Shameer, K (2016) [44]		1 068	Inpatient	EHR	4 205	30-day readmission
Shams, I (2015) [45]	2011–2012	1 674	Inpatient	EHR		30-day readmission
Stampehl, M (2020) [46]	2010–2014	206 644	Inpatient	EHR		30-day and one-year post-discharge all-cause mortality
Taslimitehrani, V (2016) [47]	1993–2013	5 044	Inpatient	EHR	43	1,2 and 5-year survival after HF diagnosis
Turgeman, L (2016) [27]	2006–2014	4 840	Inpatient	EHR		Readmission

CVD = cardiovascular disease; EHR = electronic health record; HF = heart failure; MRI = magnetic resonance imaging.

middle-income countries (LMIC). In high income countries (HIC), the predominant cause of heart failure is ischaemic heart disease (IHD), whereas, in sub-Saharan Africa, hypertension is still the leading cause of heart failure [23]. Also, healthcare services' availability and efficiency differ significantly between countries, suggesting that algorithms trained using data from HIC should be retrained using local data before adopting risk calculators.

Despite the endemicity of heart failure in LMIC, risk scores derived from patients residing in LMIC are scanty or non-existent. The lack of EHR systems, registries, and pooled data from multicentre studies is responsible for the absence of risk scores derived from patients in LMIC. If digital structured health data were available in LMIC, models predicting outcomes could be created instead of extrapolating from studies conducted in HIC. The absence of structured health data in LMIC resulted in the underrepresentation of this population in the training and test datasets included in this systematic review.

The AUC was one of the most commonly reported performance metric in the extracted studies. The highest AUC for models predicting mortality was 0.92, achieved by the random forest algorithm in a study by Nakajima et al., where both clinical and physiological imaging data were used to train algorithms [24]. A model with an AUC equal to or below 0.50 is unable to discriminate between classes. One might as well toss a coin when making pre-

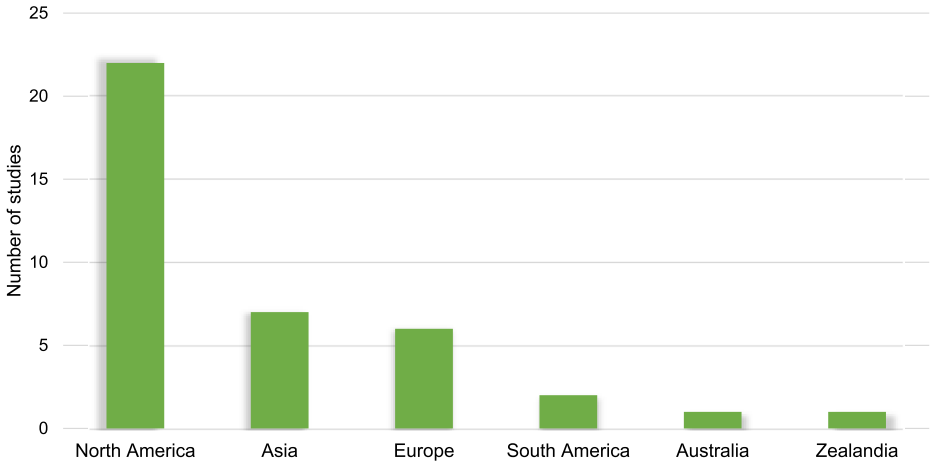


Fig. 2. Study population region.

Table 2
Characteristics of heart failure patients included in the 30 models predicting mortality and hospitalization.

First Author (year)	Study Region	No. of patients	% Black	Age	% male	% Hypertension	% IHD
Adler, E.D (2019) [10]	USA and Europe	5 822		60.3			
Ahmad, T (2018) [30]	Europe	44 886		73.2	63		
Allam, A (2019)[31]	USA and Europe	272 778		73 ± 14	51		
Angraal, S (2020)[13]	USA, Canada, Brazil, Argentina, Russia, Georgia	1 767		72 (64–79)	50		
Ashfaq, A (2019) [32]	Europe	7 655		78.8	57		
Awan, SE (2019) [33]	Australia	10 757		82 ± 7.6	49	67	55
Chen, R (2019) [34]	China	98		47 ± 14	79	23	
Chicco, D (2020) [34]	Pakistan	299		40–95*	65		
Chirinos, J (2020) [35]	USA, Canada, Russia	379	7.4	70 (62–77)	53.5	94.5	30.6
Desai, R.J (2020) [6]	USA	9 502	5.1	78 ± 8	45	87.1	22
Frizzell, J.D (2017) [36]	USA	56 477	10	80 (74–86)	45.5	75.7	58
Gleeson, S (2017) [37]	New Zealand	295		62	74	43	
Golas, S.B (2018) [12]	USA	11 510	7.9	75.7 (64–85)	52.8		
Hearn, J (2018) [38]	Canada	1 156		54	74.6		
Hsich, E (2011) [9]	USA	2 231		54 ± 11	73		41
Jiang, W (2019) [39]	USA	534	28	74.8	46		
Kourou, K (2016) [19]	Belgium	71		48.07 ± 14.82	80.3		
Krumholz, H (2019) [40]	USA	716 790	11.3	81.1 ± 8.4	45.6		
Kwon, J (2019) [5]	Asia	2 165		69.8	59.7		
Liu, W (2019) [41]	USA	303 233		72.5	50.9		
Lorenzoni, G (2019) [7]	Italy	380		78 (72–83)	42.9		18.9
Maharaj, S.M (2018) [42]	USA	1 778	0.95	72.3 ± 12.1	97.6		14
McKinley, D (2019) [20]	USA	132	100	59.25	100	91	
Miao, F (2017) [43]	USA	8 059		73.7	54	25	23.2
Nakajima, K (2020) [24]	Japan	526		66 ± 14	72	53	37
Shameer, K (2016) [44]	USA	1 068					
Shams, I (2015) [45]	USA	1 674	70.4	69.9	96		
Stampehl, M (2020) [46]	USA	206 644	12.6	80.5 ± 11.2	38.3	96.5	0.4
Taslimitehrani, V (2016) [47]	USA	5 044		78 ± 10	52	81	70.2
Turgeman, L (2016) [27]	USA	4 840		69.3 ± 11.02		96.5	84.9

Age showed as mean ± standard deviation, median (25th-75th percentile interquartile range) or minimum and maximum value.* IHD: ischaemic heart disease; USA: United States of America.

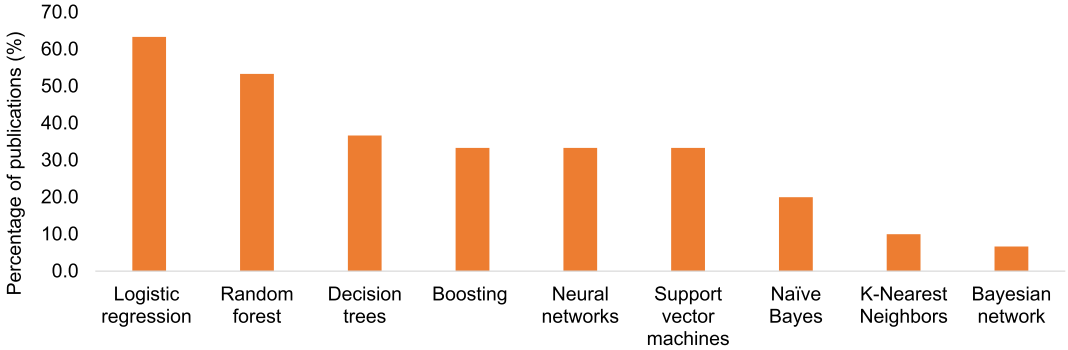


Fig. 3. Number of studies using machine learning algorithms.

Table 3
Performance metrics of algorithms predicting mortality and hospitalization in heart failure.

Author	Algorithms	Sensitivity	Accuracy	AUC (mortality)	AUC (Hospitalization)	F-score
Adler, E.D (2019) [10]	Boosted decision trees			0.88 (0.85–0.90)		
Ahmad, T (2018) [30]	Random forest			0.83		
Allam, A (2019) [31]	Recurrent neural network				0.64 (0.640–0.645)	
	Logistic regression l_2-norm regularization (LASSO)				0.643 (0.640–0.646)	
Angraal, S (2020) [13]	Logistic regression			0.66 (0.62–0.69)	0.73 (0.66–0.80)	
	Logistic regression with LASSO regularization			0.65 (0.61–0.70)	0.73 (0.67–0.79)	
	Gradient descent boosting			0.68 (0.66–0.71)	0.73 (0.69–0.77)	
	Support vector machines (linear kernel)			0.66 (0.60–0.72)	0.72 (0.63–0.81)	
	Random forest			0.72 (0.69–0.75)	0.76 (0.71–0.81)	
Ashfaq, A (2019) [32]	Long Short-Term Memory (LSTM) neural network				0.77	0.51
Awan, SE (2019) [33]	Multi-layer perceptron (MLP)	48.4		0.62		
Chen, R (2019) [34]	Naïve Bayes		0.827	0.855		
	Naïve Bayes + IG		0.857	0.887		
	Random forest		0.817	0.890		
	Random forest + IG		0.827	0.877		
	Decision trees (bagged)		0.827	0.852		
	Decision trees (bagged) + IG		0.816	0.847		
	Decision trees (boosted)		0.735	0.705		
	Decision trees (boosted) + IG		0.806	0.797		
Chicco, D (2020) [11]	Random forest		0.740	0.800		0.547
	Decision tree		0.737	0.681		0.554
	Gradient boosting		0.738	0.754		0.527
	Linear regression		0.730	0.643		0.475
	One rule		0.729	0.637		0.465
	Artificial neural network		0.680	0.559		0.483
	Naïve Bayes		0.696	0.589		0.364
	SVM (radial)		0.690	0.749		0.182
	SVM (linear)		0.684	0.754		0.115
	K-nearest neighbors		0.624	0.493		0.148
Chirinos, J (2020) [35]	Tree-based pipeline optimizer				0.717 (0.643–0.791)	
Desai, R.J (2020) [6]	Logistic regression (traditional)			0.749 (0.729–0.768)	0.738 (0.711–0.766)	
	LASSO			0.750 (0.731–0.769)	0.764 (0.738–0.789)	
	CART			0.700 (0.680–0.721)	0.738 (0.710–0.765)	
	Random forest			0.757 (0.739–0.776)	0.764 (0.738–0.790)	
	GBM			0.767 (0.749–0.786)	0.778 (0.753–0.802)	
Frizzell, J.D (2017) [36]	Random forest				0.607	
	GBM				0.614	
	TAN				0.618	
	LASSO				0.618	
	Logistic regression				0.624	
Gleeson, S (2017) [37]	Decision trees			0.7505		
Golas, S.B (2018) [12]	Logistic regression		0.626		0.664	0.435
	Gradient boosting		0.612		0.650	0.425
	Maxout networks		0.645		0.695	0.454
	Deep unified networks		0.646		0.705	0.464
Hearn, J (2018) [38]	Staged LASSO			0.827 (0.785–0.867)		
	Staged neural network			0.835 (0.795–0.880)		
	LASSO (breath-by-breath)			0.816 (0.767–0.866)		
	Neural network (breath-by-breath)			0.842 (0.794–0.882)		
Hsich, E (2011) [9]	Random survival forest			0.705		
	Cox proportional hazard			0.698		
Jiang, W (2019) [39]	Logistic and beta regression (ML)				0.73	
Kourou, K (2016) [19]	Naïve Bayes	85		0.86		
	Bayesian network	85.9		0.596		
	Adaptive boosting	78		0.74		
	Support vector machines	90		0.74		
	Neural networks	87		0.845		
	Random forest	75		0.65		
Krumholz, H (2019) [40]	Logistic regression (ML)			0.776		
Kwon, J (2019) [5]	Deep learning			0.813 (0.810–0.816)		
	Random forest			0.696 (0.692–0.700)		
	Logistic regression			0.699 (0.695–0.702)		
	Support vector machine			0.636 (0.632–0.640)		
	Bayesian network			0.725 (0.721–0.728)		
Liu, W (2019) [41]	Logistic regression				0.580 (0.578–0.583)	
	Gradient boosting				0.602 (0.599–0.605)	
	Artificial neural networks				0.604 (0.602–0.606)	
Lorenzoni, G (2019) [7]	GLMN	77.8	0.812		0.86	
	Logistic regression	54.7	0.589		0.646	
	CART	44.3	0.635		0.586	
	Random forest	54.9	0.726		0.691	
	Adaptive Boosting	57.3	0.671		0.644	
	Logitboost	66.7	0.625		0.654	

Table 3 (*continued*)

Author	Algorithms	Sensitivity	Accuracy	AUC (mortality)	AUC (Hospitalization)	F-score
	Support vector machines	57.3	0.699		0.695	
	Artificial neural networks	61.6	0.682		0.677	
Maharaj, S.M (2018) [42]	Boosted tree				0.719	
	Spike and slab regression				0.621	
McKinley, D (2019) [20]	K-nearest neighbor		0.773		0.768	
	K-nearest neighbor (randomized)		0.477		0.469	
	Support vector machines		0.545		0.496	
	Random forest		0.682		0.616	
	Gradient boosting machine		0.614		0.589	
	LASSO		0.614		0.576	
Miao, F (2017) [43]	Random survival forest			0.804		
	Random survival forest (improved)			0.821		
Nakajima, K (2020) [24]	Logistic regression			0.898		
	Random forest			0.917		
	GBT			0.907		
	Support vector machine			0.910		
	Naïve Bayes			0.875		
	k-nearest neighbors			0.854		
Shameer, K (2016) [44]	Naïve Bayes		0.832		0.78	
Shams, I (2015) [45]	Phase type Random forest	91.95			0.836	0.892
	Random forest	88.43			0.802	0.865
	Support vector machine	86.16			0.775	0.857
	Logistic regression	83.40			0.721	0.833
	Artificial neural network	82.39			0.704	0.823
Stampehl, M (2020) [46]	CART					
	Logistic regression					
	Logistic regression (stepwise)			0.74		
Taslimitehrani, V (2016) [47]	CPXR(Log)		0.914			
	Support vector machine		0.75			
	Logistic regression		0.89			
Turgeman, L (2016) [27]	Naïve Bayes	48.9			0.676	
	Logistic regression	28.1			0.699	
	Neural network	8.9			0.639	
	Support vector machine	23.0			0.643	
	C5 (ensemble model)	43.5			0.693	
	CART (boosted)	22.6			0.556	
	CART (bagged)	9.0			0.579	
	CHAID Decision trees (boosted)	30.3			0.691	
	CHAID Decision trees (bagged)	10.5			0.707	
	Quest decision tree (boosted)	20.3			0.487	
	Quest decision tree (bagged)	7.2			0.579	
	Naïve network + Logistic regression	38.2			0.653	
	Naïve network + Neural network	26.3			0.635	
	Naïve network + SVM	35.8			0.649	
	Logistic regression + Neural network	16.8			0.59	
	Logistic regression + SVM	26.2			0.607	
	Neural network + SVM	16.5			0.577	

AUC: area under the receiver operating characteristic curve; CART: classification and regression tree; CPXR: contrast pattern aided logistic regression; GBM: gradient-boosted model; HR: hazard ratio; IG: information gain; LASSO: least absolute shrinkage and selection operator; ML: machine learning; SVM: support vector machine; TAN: tree augmented Bayesian network. The AUC is displayed under both the mortality and hospitalization column if the authors did not specify the outcome predicted.

dictions. Some of the reasons for the modest performance metrics demonstrated by machine learning algorithms include a training dataset with excessive missing data or few predictors, absence of ongoing partnership between clinicians and data scientists and class imbalance. In most instances, when handling healthcare data, the negative class tends to outnumber positive classes. The learning environment is rendered unfavourable since there are fewer positive observations or patterns for an algorithm to learn from. For example, when predicting mortality, the class with patients that demised is frequently smaller than the class with alive patients.

Models with perfect precision and recall have an F-measure, also known as the F-Score or F1 Score, equal to one [25]. Sensitivity, also known as recall, measures a proportion of positive classes accurately classified as positive [26]. Machine learning algorithms in the extracted studies had a sensitivity rate between 7.2 and 91.9%. The low sensitivity, reported by Turgeman and May, improved to 43.5% when they used an ensemble method to combine multiple predictive models to produce a single model [27].

Although the random forest algorithm appeared to have the highest predictive abilities in most studies, one cannot conclude that it should be the algorithm of choice whenever one attempts to create a predictive model. The random forest algorithm's main advantage is that it is an ensemble-based classifier that takes random samples of data, exposing them to multiple decision tree algorithms. Decision trees are intuitive and interpretable and can immediately suggest why a patient is stratified into a high-risk category, hence guiding subsequent risk reduction interventions. The interpretability of decision trees is a significant advantage in contrast to deep learning methodologies such as artificial neural networks with a "black box" nature. Once random samples of data have been exposed to multiple decision tree algorithms, the decision trees' ensemble identifies the class with the highest number of votes when making predictions. Random forests also perform well in large datasets with missing data, a common finding when handling healthcare data, and can rank features (*predictors*) in the order of importance, based on predictive powers [28].

Predictors of mortality identified by machine learning algorithms in the extracted studies were explainable and included features such as the LVEF, hypotension, age and blood urea nitrogen levels. Whether these predictors should be considered significant risk factors for all heart failure, irrespective of genetic makeup, is debatable. The youngest patient in the studies reviewed was 40 years old, but most of the patients included in the predictive models were significantly older, with a median age of 72 years. Risk scores derived from older patients may reduce the applicability of the existing risk calculators in the sub-Saharan African (SSA) context, considering that patients with heart failure in SSA are generally a decade younger [29].

Geographically unique heart failure aetiologies and diverse clinical presentations call for predictive models that incorporate genomic, clinical and imaging data. We recommend that clinicians treating heart failure patients focus on establishing structured EHR systems and comparing outcomes such as mortality and hospitalization in patients managed with and without risk scores. Clinicians without access to EHR systems should carefully study the cohort used to create risk scores before implementing risk scores in their clinical practice.

5. Limitations

This systematic literature review has several limitations. The systematic literature search was conducted by a single reviewer, predisposing the review to selection bias. We only included original research studies published after 2009. The rationale for including studies published in the past 11 years was to avoid including studies where rule-based expert systems were used instead of newer machine learning techniques. Although the data used to create predictive models was grossly heterogeneous, a meta-analytic component as part of the review would have provided a broader perspective on machine learning algorithms' performance metrics when predicting heart failure patient outcomes.

6. Conclusion

The variation in the aetiologies of heart failure, limited access to structured health data, distrust in machine learning techniques among clinicians and the modest accuracy of predictive models are some of the factors precluding the widespread use of machine learning derived risk calculators.

7. Grant support

The study did not receive financial support. The primary author Dr Dineo Mpanya is a full-time PhD Clinical Research fellow in the Division of Cardiology, Department of Internal Medicine at the University of the Witwatersrand. Her PhD is funded by the Professor Bongani Mayosi Netcare Clinical Scholarship, the Discovery Academic Fellowship *(Grant No. 039023)*, the Carnegie Corporation of New York (*Grant No. b8749*) and the South African Heart Association.

Declaration of Competing Interest

All authors take responsibility for all aspects of the reliability and freedom from bias of the data presented and their discussed interpretation.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijcha.2021.100773.

References

[1] N. Boodhun, M. Jayabalan, Risk prediction in life insurance industry using supervised learning algorithms, Compl. Intell. Syst. 4 (2018) 145–154.
[2] F. Carcillo, Y.-A. Le Borgne, O. Caelen, G. Bontempi, Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization, Int. J. Data Sci. Analyt. 5 (2018) 285–300.
[3] S. Moradi, Rafiei F. Mokhatab, A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks, Financ. Innovat. 5 (2019) 15.
[4] D. Mpanya, T. Celik, E. Klug, H. Ntsinjana, Machine learning and statistical methods for predicting mortality in heart failure, Heart Fail Rev. (2020).
[5] J.M. Kwon, K.H. Kim, K.H. Jeon, S.E. Lee, H.Y. Lee, H.J. Cho, et al., Artificial intelligence algorithm for predicting mortality of patients with acute heart failure, PLoS One. 14 (2019) e0219302.
[6] R.J. Desai, S.V. Wang, M. Vaduganathan, T. Evers, S. Schneeweiss, Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes, JAMA Network Open. 3 (2020) e1918962.
[7] G. Lorenzoni, S. Santo Sabato, C. Lanera, D. Bottigliengo, C. Minto, H. Ocagli, et al., Comparison of machine learning techniques for prediction of hospitalization in heart failure patients, J. Clin. Med. 2019;8.
[8] S. Blecker, D. Sontag, L.I. Horwitz, G. Kuperman, H. Park, A. Reyentovich, et al., Early identification of patients with acute decompensated heart failure, J. Card Fail. 24 (2018) 357–362.
[9] E. Hsich, E.Z. Gorodeski, E.H. Blackstone, H. Ishwaran, M.S. Lauer, Identifying important risk factors for survival in patient with systolic heart failure using random survival forests, Circulat.: Cardiovas. Qual. Outcomes 4 (2011) 39–45.
[10] E.D. Adler, A.A. Voors, L. Klein, F. Macheret, O.O. Braun, M.A. Urey, et al., Improving risk prediction in heart failure using machine learning, Eur. J. Heart Fail. 22 (2020) 139–147.
[11] D. Chicco, G. Jurman, Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone, BMC Med. Inf. Decis. Making 20 (2020).
[12] S.B. Golas, T. Shibahara, S. Agboola, H. Otaki, J. Sato, T. Nakae, et al., A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data, BMC Med. Inform. Decis. Mak. 18 (2018) 44.
[13] S. Angraal, B.J. Mortazavi, A. Gupta, R. Khera, T. Ahmad, N.R. Desai, et al., Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction, JACC Heart Fail. 8 (2020) 12–21.
[14] M. Gheorghiade, N.M. Albert, A.B. Curtis, J. Thomas Heywood, M.L. McBride, P.J. Inge, et al., Medication dosing in outpatients with heart failure after implementation of a practice-based performance improvement intervention: findings from IMPROVE HF, Congest Heart Fail. 18 (2012) 9–17.
[15] B. Hanratty, D. Hibbert, F. Mair, C. May, C. Ward, S. Capewell, et al., Doctors' perceptions of palliative care for heart failure: focus group study, BMJ 325 (2002) 581–585.
[16] K. Eichler, M. Zoller, P. Tschudi, J. Steurer, Barriers to apply cardiovascular prediction rules in primary care: a postal survey, BMC Family Pract. 8 (2007) 1.
[17] P. Lakhani, B. Sundaram, Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks, Radiology 284 (2017) 574–582.
[18] F.A. Khan, A. Majidulla, G. Tavaziva, A. Nazish, S.K. Abidi, A. Benedetti, et al., Chest x-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: a prospective study of diagnostic accuracy for culture-confirmed disease, Lancet Digit Health. 2 (2020) e573–e581.
[19] K. Kourou, G. Rigas, K.P. Exarchos, Y. Goletsis, T.P. Exarchos, S. Jacobs, et al., Prediction of time dependent survival in HF patients after VAD implantation using pre- and post-operative data, Comput. Biol. Med. 70 (2016) 99–105.
[20] D. McKinley, P. Moye-Dickerson, S. Davis, A. Akil, Impact of a pharmacist-led intervention on 30-Day readmission and assessment of factors predictive of readmission in African American men with heart failure, Am. J. Men's Health. 13 (2019).
[21] A.N. Jagannatha, H. Yu, Bidirectional RNN for medical event detection in electronic health records, Proc. Conf. 2016 (2016) 473–482.
[22] A.N. Jagannatha, H. Yu, Structured prediction models for RNN based sequence labeling in clinical text, Proc. Conf. Empir. Methods Nat. Lang. Process. 2016 (2016) 856–865.
[23] V.N. Agbor, M. Essouma, N.A.B. Ntusi, U.F. Nyaga, J.J. Bigna, J.J. Noubiap, Heart failure in sub-Saharan Africa: a contemporaneous systematic review and meta-analysis, Int. J. Cardiol. 257 (2018) 207–215.
[24] K. Nakajima, T. Nakata, T. Doi, H. Tada, K. Maruyama, Machine learning-based risk model using 123I-metaiodobenzylguanidine to differentially predict modes of cardiac death in heart failure, J. Nucl. Cardiol.: Off. Publ. Am. Soc. Nucl. Cardiol. May (2020).
[25] G. Hripcsak, A.S. Rothschild, Agreement, the f-measure, and reliability in information retrieval, J. Am. Med. Informat. Assoc.: JAMIA 12 (2005) 296–298.
[26] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PLoS One. 10 (2015) e0118432.
[27] L. Turgeman, J.H. May, A mixed-ensemble model for hospital readmission, Artif. Intell. Med. 72 (2016) 72–82.
[28] S. Uddin, A. Khan, M.E. Hossain, M.A. Moni, Comparing different supervised machine learning algorithms for disease prediction, BMC Med. Inf. Decis. Making 19 (2019).

[29] G.S. Bloomfield, F.A. Barasa, J.A. Doll, E.J. Velazquez, Heart failure in sub-Saharan Africa, Curr. Cardiol. Rev. 9 (2013) 157–173.
[30] T. Ahmad, L.H. Lund, P. Rao, R. Ghosh, P. Warier, B. Vaccaro, et al., Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients, J. Am. Heart Assoc. 7 (2018).
[31] A. Allam, M. Nagy, G. Thoma, M. Krauthammer, Neural networks versus Logistic regression for 30 days all-cause readmission prediction, Sci. Rep. 9 (2019).
[32] A. Ashfaq, A. Sant'Anna, M. Lingman, S. Nowaczyk, Readmission prediction using deep learning on electronic health records, J. Biomed. Inform. 97 (2019).
[33] S.E. Awan, M. Bennamoun, F. Sohel, F.M. Sanfilippo, B.J. Chow, G. Dwivedi, Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death, PLoS One. 14 (2019) e0218760.
[34] R. Chen, A. Lu, J. Wang, X. Ma, L. Zhao, W. Wu, et al., Using machine learning to predict one-year cardiovascular events in patients with severe dilated cardiomyopathy, Eur. J. Radiol. 117 (2019) 178–183.
[35] J.A. Chirinos, A. Orlenko, L. Zhao, M.D. Basso, M.E. Cvijic, Z. Li, et al., Multiple plasma biomarkers for risk stratification in patients with heart failure and preserved ejection fraction, J. Am. Coll. Cardiol. 75 (2020) 1281–1295.
[36] J.D. Frizzell, L. Liang, P.J. Schulte, C.W. Yancy, P.A. Heidenreich, A.F. Hernandez, et al., Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches, JAMACardiol. 2 (2017) 204–209.
[37] S. Gleeson, Y.W. Liao, C. Dugo, A. Cave, L. Zhou, Z. Ayar, et al., ECG-derived spatial QRS-T angle is associated with ICD implantation, mortality and heart failure admissions in patients with LV systolic dysfunction, PLoS One. 12 (2017) e0171069.
[38] J. Hearn, H.J. Ross, B. Mueller, C.P. Fan, E. Crowdy, J. Duhamel, et al., Neural networks for prognostication of patients with heart failure, Circulation Heart Failure 11 (2018) e005193.
[39] W. Jiang, S. Siddiqui, S. Barnes, L.A. Barouch, F. Korley, D.A. Martinez, et al., Readmission risk trajectories for patients with heart failure using a dynamic prediction approach: Retrospective study, J. Med. Int. Res. 21 (2019).
[40] H.M. Krumholz, A.C. Coppi, F. Warner, E.W. Triche, S.X. Li, S. Mahajan, et al., Comparative effectiveness of new approaches to improve mortality risk models from medicare claims data, JAMA Network Open. 2 (2019).
[41] W. Liu, C. Stansbury, K. Singh, A.M. Ryan, D. Sukul, E. Mahmoudi, et al., Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding, PLoS ONE. 15 (2020).
[42] S.M. Mahajan, A.S. Mahajan, R. King, S. Negahban, Predicting risk of 30-day readmissions using two emerging machine learning methods, Stud. Health Technol. Inform. 250 (2018) 250–255.
[43] F. Miao, Y.P. Cai, Y.X. Zhang, X.M. Fan, Y. Li, Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest, IEEE Access 6 (2018) 7244–7253.
[44] K. Shameer, K.W. Johnson, A. Yahi, R. Miotto, L.I. Li, D. Ricks, et al., Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using mount sinai heart failure cohort, Pac. Symp. Biocomput. 22 (2017) 276–287.
[45] I. Shams, S. Ajorlou, K. Yang, A predictive analytics approach to reducing 30-day avoidable readmissions among patients with heart failure, acute myocardial infarction, pneumonia, or COPD, Health Care Manage. Sci. 18 (2015) 19–34.
[46] M. Stampehl, H.S. Friedman, P. Navaratnam, P. Russo, S. Park, E.N. Obi, Risk assessment of post-discharge mortality among recently hospitalized Medicare heart failure patients with reduced or preserved ejection fraction, Curr. Med. Res. Opin. 36 (2020) 179–188.
[47] V. Taslimitehrani, G. Dong, N.L. Pereira, M. Panahiazar, J. Pathak, Developing EHR-driven heart failure risk prediction models using CPXR(Log) with the probabilistic loss function, J. Biomed. Inform. 60 (2016) 260–269.

CHAPTER 4 – MANUSCRIPT TWO: CLUSTERING OF HEART FAILURE PHENOTYPES IN JOHANNESBURG USING UNSUPERVISED MACHINE LEARNING

4.1 Introduction

This chapter introduces the cohort of heart failure patients hospitalised between 2009 and 2018 and describes the baseline demographics. Also, unsupervised machine learning algorithms are used to cluster heart failure patients. The rationale for clustering heart failure patients was to establish heart failure phenotypes and identify a cluster of patients with high-risk clinical features. Seven unsupervised machine learning algorithms autonomously clustered heart failure patients into various groups.

4.2 Current Knowledge

Despite cardiovascular diseases being common in sub-Saharan Africa (SSA), only a few studies have been published describing the spectrum of heart disease in this region. The onset of heart failure in SSA occurs at a younger age.

4.3 Aims and Objectives

The study aimed to describe the baseline demographic and clinical characteristics of heart failure patients hospitalised in a tertiary academic centre in Johannesburg. The citation details are Mpanya, D.; Celik, T.; Klug, E.; Ntsinjana, H. Clustering of Heart Failure Phenotypes in Johannesburg Using Unsupervised Machine Learning. Appl. Sci. 2023, 13, 1509. The next chapter contains results obtained after building a model predicting mortality in heart failure using machine learning algorithms and a traditional statistical method, logistic regression.

applied sciences

MDPI

Article

Clustering of Heart Failure Phenotypes in Johannesburg Using Unsupervised Machine Learning

Dineo Mpanya [1,2,*], Turgay Celik [2,3], Eric Klug [1,4] and Hopewell Ntsinjana [5]

[1] Division of Cardiology, Department of Internal Medicine, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg 2193, South Africa
[2] Wits Institute of Data Science, University of the Witwatersrand, Johannesburg 2000, South Africa
[3] School of Electrical and Information Engineering, Faculty of Engineering and Built Environment, University of the Witwatersrand, Johannesburg 2000, South Africa
[4] Netcare Sunninghill, Sunward Park Hospitals and the Charlotte Maxeke Johannesburg Academic Hospital, Johannesburg 2196, South Africa
[5] Department of Paediatrics and Child Health, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg 2193, South Africa
* Correspondence: dineo.mpanya@wits.ac.za; Tel.: +27-11-488-3611; Fax: +27-11-642-9041

Abstract: Background: The diagnosis and therapy of heart failure are guided mainly by a single imaging parameter, the left ventricular ejection fraction (LVEF). Recent studies have reported on the value of machine learning in characterising the various phenotypes of heart failure patients. Therefore, this study aims to use unsupervised machine learning algorithms to phenotype heart failure patients into different clusters using multiple clinical parameters. Methods: Seven unsupervised machine learning clustering algorithms were used to cluster heart failure patients hospitalised with acute and chronic heart failure. Results: The agglomerative clustering algorithm identified three clusters with a silhouette score of 0.72. Cluster 1 (uraemic cluster) comprised 229 (36.0%) patients with a mean age of 56.2 ± 17.2 years and a serum urea of 14.5 ± 31.3 mmol/L. Cluster 2 (hypotensive cluster) comprised 117 (18.4%) patients with a minimum systolic and diastolic blood pressure of 91 and 60 mmHg, respectively. In cluster 3 (congestive cluster), patients predominantly had symptoms of fluid overload, and 93 (64.6%) patients had ascites. Among the 636 heart failure patients studied, the median LVEF was 32% (interquartile range: 25–45), and the rate of in-hospital all-cause mortality was 14.5%. Systolic and diastolic blood pressure, age, and the LVEF had the most substantial impact on discriminating between the three clusters. Conclusions: Clinicians without access to echocardiography could potentially rely on blood pressure measurements and age to risk stratify heart failure patients. However, larger prospective studies are mandatory for the validation of these clinical parameters.

Keywords: unsupervised machine learning; heart failure; mortality; cluster; Africa

Citation: Mpanya, D.; Celik, T.; Klug, E.; Ntsinjana, H. Clustering of Heart Failure Phenotypes in Johannesburg Using Unsupervised Machine Learning. *Appl. Sci.* **2023**, *13*, 1509. https://doi.org/10.3390/app13031509

Academic Editors: Josué Álvarez Borrego and Juan A. Gómez-Pulido

Received: 22 November 2022
Revised: 29 December 2022
Accepted: 18 January 2023
Published: 23 January 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

Unsupervised machine learning algorithms autonomously learn patterns in a dataset without human guidance or input. For example, clustering is an unsupervised machine learning technique that groups data points with similar attributes. A wide spectrum of algorithms are used to cluster data using dimension reduction, partitioning, hierarchical, density-based, and model-based methods [1]. Unsupervised machine algorithms may reveal risk factors in patients with complex clinical syndromes such as heart failure, a clinical syndrome characterised by a constellation of symptoms and signs of fluid overload, mainly in patients with underlying cardiovascular disease. The global prevalence and incidence of heart failure continue to rise to the extent that heart failure has been declared a global pandemic [2]. Therefore, innovative strategies that may assist in improving the management of heart failure patients should be sought.

The current approach to the diagnosis, risk stratification, management, and prognosis of heart failure is fundamentally determined by a single imaging parameter, the left ventricular ejection fraction (LVEF) [3]. The LVEF is a measurement that estimates the fraction of blood volume ejected by the heart's left ventricle with each contraction [4]. Among a wide range of non-invasive cardiac imaging modalities, the echocardiogram is one of the widely applied techniques for measuring the LVEF. However, the results obtained are generally influenced by the experience of the echocardiogram operator and the patient's body habitus. In addition, the echocardiography machine is not readily accessible in most hospitals in low-and-middle-income countries (LMICs). Therefore, its use is mainly restricted to the private sector and state-owned hospitals affiliated with academic institutions.

A recent systematic review identified 81 articles reporting the utility of both supervised and unsupervised machine learning algorithms in heart failure. Of the articles reviewed, 11 (13%) studies used classification or clustering algorithms, and 25 (31%) used machine learning for the detection of the onset of heart failure, 18 (21%) for predicting rehospitalisation, 20 (26%) for predicting mortality, and the rest (6%) for predicting outcomes and response to implantable cardiac devices [5]. Another systematic review identified 27 studies using unsupervised machine learning for heart failure subtype definition [6]. However, there are limited or no papers published from LMICs describing heart failure phenotypes using machine learning algorithms. This study aims to describe baseline demographic and clinical characteristics of heart failure patients hospitalised in a tertiary-level hospital in Johannesburg, South Africa. Furthermore, this study describes the utility of unsupervised machine learning in clustering heart failure patients using a constellation of clinical parameters, instead of a single imaging parameter, the LVEF.

2. Methods

2.1. Study Design and Participants

Data were exported from the PMRCardio database, a Microsoft Structured Query Language Server Management Studio (version 15.0.18330.0) that stores clinical data of patients hospitalised in cardiology wards at the Charlotte Maxeke Johannesburg Academic Hospital (CMJAH, Johannesburg, South Africa). The CMJAH is a quaternary-level state-owned institution equipped with a cardiac intensive care unit, a cardiac catheterisation and electrophysiology laboratory, and outpatient services. All acutely and chronically ill patients with cardiovascular diseases are hospitalised in cardiac wards. Their admission data are stored digitally in the PMRCardio database by registrars or residents rotating in the cardiology wards. Clinical parameters available in the dataset (features) include demographic data, clinical history and examination findings, laboratory indices, electrocardiogram (ECG), echocardiogram, angiography, and in-hospital all-cause mortality status.

The tenth revision of the international classification of diseases (ICD-10) code was used to select patients diagnosed with heart failure. The following ICD-10 diagnoses were used to select patients included in the analysis: "heart failure unspecified" or "congestive heart failure or "left ventricular failure", or "dilated cardiomyopathy", or "ischaemic cardiomyopathy." Heart failure patients who are 18 years of age and older with a documented left ventricular ejection fraction (LVEF) hospitalised between October 2009 and December 2018 were included.

Data exported from the database were merged into a single Microsoft Excel Sheet. All retrospective patient data used in the final analysis were further verified by comparing hard copies of medical records with electronic data. Patients were subsequently stratified into three groups based on the baseline LVEF: heart failure with reduced ejection fraction—HFrEF (LVEF < 40%), heart failure with midrange ejection fraction—HFmEF (LVEF = 40–49%), and heart failure with preserved ejection fraction—HFpEF (LVEF ≥ 50%) [3].

Admission biochemical results were obtained from the National Health Laboratory Service (NHLS, Johannesburg, South Africa). Approval to conduct the study was received

from the University of the Witwatersrand Human Research Ethics Committee (clearance certificate number: M190515). Permission was also obtained from NHLS senior authorities.

2.2. Descriptive Statistics

Descriptive statistics were used to compare clinical parameters in the LVEF categories and clusters. Categorical features are expressed as numbers and percentages and were compared using the Chi-square test, and continuous features with a normal distribution were expressed as mean and standard deviation. The median and interquartile ranges (IQR) were used for continuous features with a non-normal distribution. We compared continuous features in each LVEF category and cluster using the one-way analysis of variance (ANOVA). Confidence intervals were calculated at 95% interval levels, and differences were considered statistically significant at a p-value <0.05. The Tukey test (or Tukey procedure), also called Tukey's Honest Significant Difference post hoc analysis, was performed if ANOVA showed statistically significant differences when comparing the means in the LVEF categories and clusters. Python version 3.10.0 and Jupyter notebook (version 6.0.3) were used for descriptive data analysis and for the clustering of heart failure patients using various unsupervised machine learning algorithms. Kaplan–Meier survival curves were plotted to evaluate survival in each cluster.

2.3. Data Pre-Processing for Clustering of Heart Failure Patients

Packages and libraries required for data analysis and visualisation were imported into a Jupyter notebook (version 6.0.3). These included sklearn (version 1.0.2), pandas (version 1.3.5), numpy (version 1.21.6), matplotlib (version 3.1.2), seaborn (version 0.11.2), statsmodels (version 0.13.2), lifelines (version 0.27.0), and researchpy (version 0.3.2). Features with more than 20% missing values, patient identifiers, and column data without variation were removed from the dataset. The remaining columns with missing data were imputed using the mean or median. Categorical features with missing values were ascribed with zero. The data were subsequently scaled, standardised, and normalised prior to the clustering of patients. The Jupyter notebook script is attached as a Supplementary file.

2.4. Cluster Analysis

Clustering is a form of unsupervised machine learning where data are categorised into groups based on similarities within data points in each group. Before clustering the heart failure data, the feature "mortality" that characterises the vitality status of each patient upon discharge (dead or alive) from the dataset was removed. The rationale for removing the outcome "mortality" lies behind the premise that unsupervised machine learning clusters unlabelled data without a clearly defined outcome. As a result, mortality did not influence the creation of clusters. Clustering was conducted using a dataset with and without the LVEF. This was carried out to evaluate the true importance of the LVEF in categorising heart failure patients.

Seven clustering algorithms were used to cluster heart failure patients: a partitioning method (K-prototype and K-means), a hierarchical method (agglomerative and Balanced Iterative Reducing and Clustering Hierarchies—BIRCH), density-based (Ordering Points To Identify the Clustering Structure—OPTICS and density-based spatial clustering of applications with noise—DBSCAN), and a model-based clustering method, a Gaussian mixture model.

Distance-based algorithms use K-means, K-modes, or both to estimate the distance between data points. Agglomerative clustering, also referred to a "bottom-up" approach, initially treats each observation (patient) as a single cluster. The algorithm then successively agglomerates pairs of clusters until a single cluster is formed.

The DBSCAN algorithm groups data points close to each other based on the density of data points in a particular region, marking outliers as points in a low-density region. The Gaussian mixture model uses multiple Gaussian distributions to identify various clusters, with the assumption that each modal distribution represents a particular structure.

Using the dataset that includes the LVEF, agglomerative clustering was chosen as the best algorithm for clustering heart failure patients since it achieved the highest silhouette score. Using the ward linkage method, the agglomerative clustering algorithm identified three clusters with a silhouette score of 0.72 (Table 1). A silhouette score measures the goodness of fit of clusters by measuring the distance between points within a cluster and spacing between each cluster. This score ranges between −1 and +1, where the best score, +1, suggests that clusters are far from each other and −1 infers that clusters are poorly describing patterns in the dataset. A dendrogram was also plotted to represent the clusters visually. Shapley additive explanations (SHAP) was used to identify and rank features that played a prominent role in the segmentation of the heart failure cohort into three clusters using the agglomerative clustering algorithm. Both the DBSCAN and OPTICS algorithms were unsuccessful in clustering the data.

Table 1. Number of clusters vs. silhouette score for the algorithms used in the study.

Algorithm	2 Clusters	3 Clusters	4 Clusters	5 Clusters	6 Clusters
K-prototype	0.580	0.100	0.120	0.120	0.120
K-means	0.036	0.042	0.048	0.048	0.048
Agglomerative	0.640	0.720	0.680	0.540	0.580
BIRCH	0.030	0.020	0.022	0.023	0.026
Gaussian mixture model	0.580	0.580	0.570	0.550	0.460

BIRCH = Balanced Iterative Reducing and Clustering Hierarchies. Ordering Points To Identify the Clustering Structure (OPTICS) and density-based spatial clustering of applications with noise (DBSCAN) algorithms were not included since they were unsuccessful in clustering the data. DBSCAN and OPTICS algorithms generated an optimum number of clusters as one and zero, respectively.

After removing the LVEF from the dataset, the Gaussian mixture model had the highest silhouette score of 0.30. This silhouette score corresponded to three clusters. These clusters were subsequently compared with the LVEF categories (HFrEF, HFmEF, and HFpEF). The SHAP values for the Gaussian mixture model were also generated.

3. Results

3.1. Baseline Demographics and Clinical Characteristics

Between October 2009 and December 2018, there were 10737 admissions, of which 4730 were heart failure-related (Figure 1). After excluding patients with incomplete data, particularly the LVEF, the final heart failure cohort comprised 636 patients, of which 352 (55.3%) were males. The mean age in the entire cohort was 55.6 ± 16.9 years, and patients with HFmEF were older, with a mean age of 58.0 ± 17.1 years. The baseline median LVEF in the entire study cohort was 32% (25–45), 409 (64.3%) had a LVEF less than 40% (HFrEF), 91 (14.3%) had a LVEF between 40 and 49% (HFmEF), and 136 (21.4%) had a LVEF ≥ 50% (HFpEF). Heart failure patients had a mean systolic and diastolic pressure of 128 ± 30.1 mmHg and 82 ± 19.7 mmHg, respectively. The rest of the baseline clinical characteristics in the study cohort are depicted in Table 2.

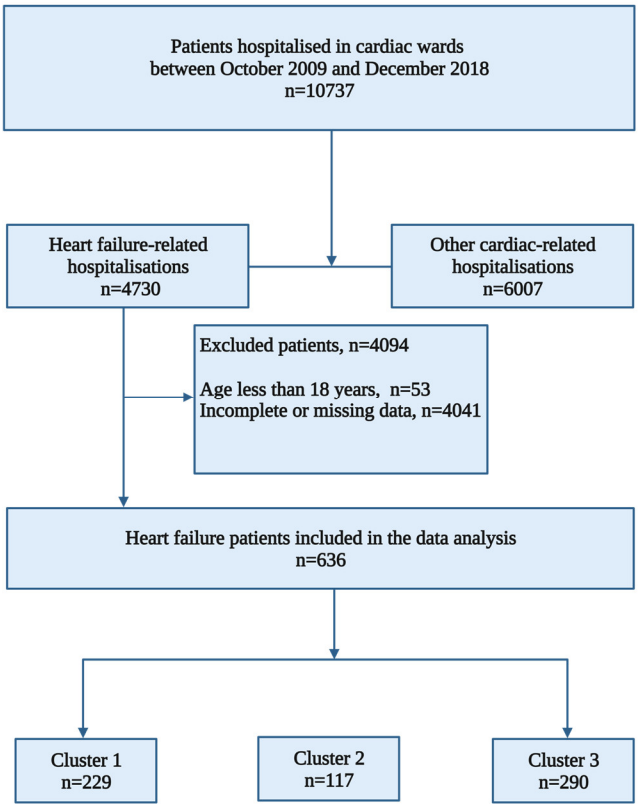


Figure 1. Flow chart showing heart failure patient selection for clustering.

Table 2. Baseline demographic and clinical characteristics of the study patients stratified according to the left ventricular ejection fraction.

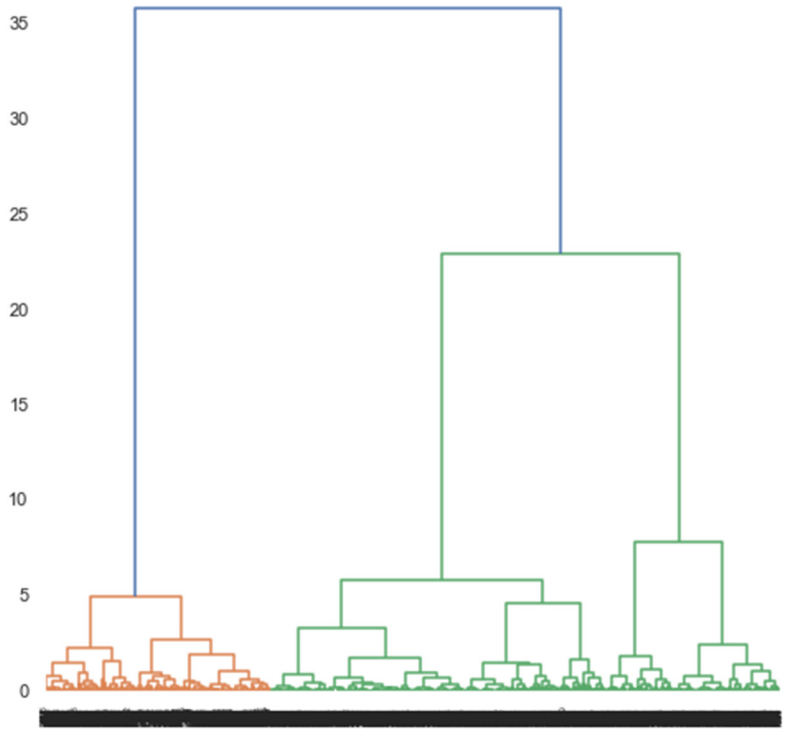
Clinical Parameters	All Patients (*n* = 636)	HFrEF (*n* = 409)	HFmEF (*n* = 91)	HFpEF (*n* = 136)	*p*-Value
Age (years), mean	55.6 ± 16.9	54.6 ± 16.9	58.0 ± 17.1	56.9 ± 16.4	0.1183
Male sex, n (%)	352 (55.3)	218 (61.9)	53 (15.1)	81 (23.0)	0.3720
Ethnicity, n (%)					
African	380 (59.7)	251 (66.1)	49 (12.9)	80(21.1)	0.4038
Caucasian	154 (24.2)	93 (60.4)	31 (20.1)	30 (19.5)	0.0595
Indian	72 (11.3)	46 (63.9)	8 (11.1)	18 (25.0)	0.5831
Mixed ancestry	29 (4.6)	19 (65.5)	3 (10.3)	7 (24.1)	0.7992
Clinical examination					
Lung crepitations	392 (61.6)	251 (64.0)	60 (15.3)	81 (20.7)	0.6153
Ascites	144 (22.6)	90 (62.5)	20 (13.9)	34 (23.6)	0.7597
Cyanosed	51 (8.0)	32 (62.5)	2 (13.9)	17 (23.6)	0.7597
Hepatomegaly	249 (39.2)	173 (69.5)	36 (14.5)	40 (16.1)	0.0284
Peripheral oedema	436 (68.6)	282 (64.7)	58 (13.3)	96 (22.0)	0.5298
S3	324 (50.9)	227 (70.1)	32 (9.9)	65 (20.1)	0.0015
JVP elevated	493(77.5)	326(66.1)	69(14.0)	98(19.9)	0.1653
Pulse	100 ± 24.3	102 ± 25.1	101 ± 16.8	94 ± 25.1	0.1324
Systolic BP, mmHg	128 ± 30.1	125 ± 28.8	136 ± 35.7	135 ± 27.5	0.0002
Diastolic BP, mmHg	82 ± 19.7	83 ± 20.4	85 ± 20.5	78 ± 16.1	0.0318
Laboratory tests					
Haemoglobin, g/dL	12.9 ± 2.5	13.3 ± 2.4	12.4 ± 2.9	12.5 ± 2.8	0.0016
Sodium, mmol/L	139 ± 5.4	139 ± 5.5	138 ± 5.1	140 ± 5.2	0.1415

Potassium, mmol/L	4.4 ± 0.7	4.4 ± 0.7	4.3 ± 0.9	4.4 ± 0.8	0.5181
Urea, mmol/L, IQR	8.5 (6.0–12.5)	12.1 ± 16.9	16.5 ± 43.3	13.2 ± 20.8	0.2692
eGFR, mL/min, IQR	66.1 (41.2–85.0)	66.7 ± 27.8	73.5 ± 35.5	63.7 ± 37.3	0.4604
Echocardiography					
LVIDd, cm	5.6 ± 1.1	6.0 ± 0.9	5.3 ± 1.1	4.5 ± 1.1	<0.001
LVIDs, cm	4.6 ± 1.3	5.2 ± 1.0	4.3 ± 0.8	3.1 ± 1.3	<0.001

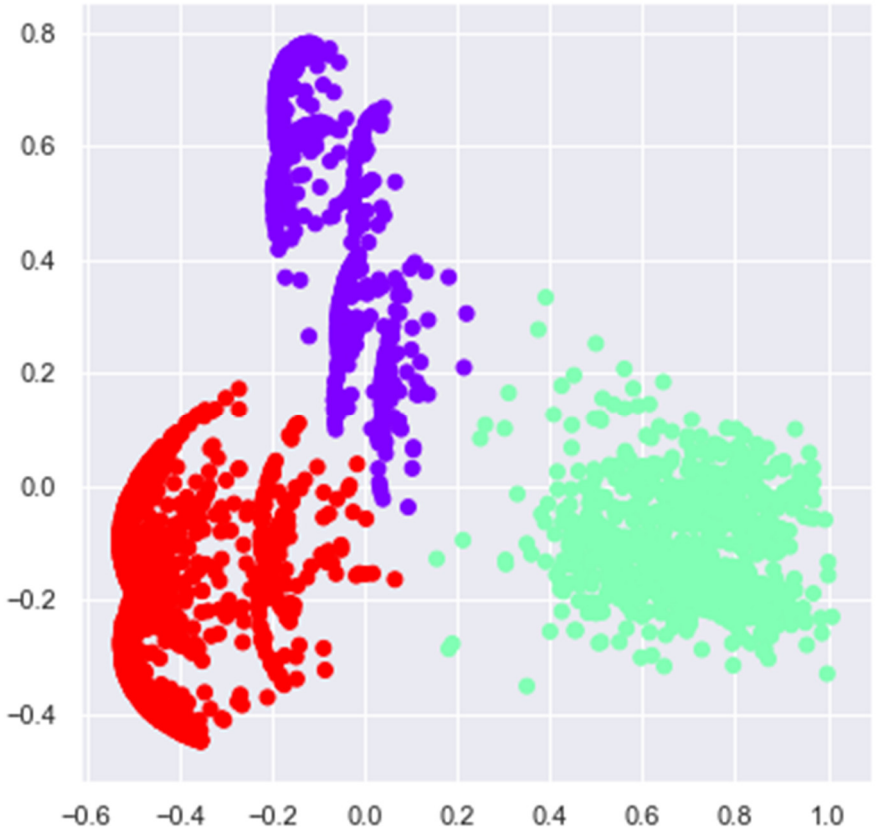
Data are presented as mean ± standard deviation for continuous features with a normal distribution, and median and interquartile ranges (IQR) (p25–p75) when the distribution is skewed. Categorical data are represented as absolute numbers and percentages. BP = blood pressure; eGFR = estimated glomerular filtration rate using the modification of diet in renal disease (MDRD) formula; HFrEF = heart failure with reduced ejection fraction (LVEF < 40%); HFmEF = heart failure with midrange ejection fraction (LVEF = 40–49%); HFpEF = heart failure with preserved ejection fraction (LVEF ≥ 50%); JVP = jugular venous pressure; LVEF = left ventricular ejection fraction; LVIDd = left ventricular internal diameter in end-diastole; LVIDs = left ventricular internal diameter in end-systole; S3 = third heart sound.

3.2. Cluster Analysis

The agglomerative clustering hierarchical algorithm identified three clusters of heart failure patients (Figure 2). Cluster 1 (*uraemic cluster*) comprised 229 (36.0%) patients with a mean age of 56.2 ± 17.2 years (Table 3). Patients in Cluster 1 had a relatively higher mean LVEF of 45 ± 17.6 ($p < 0.001$). Forty-six (50.5%) patients with HFmEF were in cluster 1 and 83 (61.0%) patients with HFpEF were also in cluster 1. The patients in the *uraemic cluster* also had a smaller mean left ventricular internal diameter in end-systole (LVIDs) of 4.1 ± 1.3 cm ($p < 0.001$) and mean left ventricular internal diameter end-diastole (LVIDd) of 5.2 ± 1.1 cm.



(**A**)



(**B**)

Figure 2. Clusters of heart failure patients identified by agglomerative clustering. (**A**) Dendrogram showing three clusters. Cluster 1 (orange), cluster 2 and 3 (green). Three clusters were selected based on the silhouette score of 0.7, which is a measure of goodness of fit of patients into each cluster. The x-axis shows various clusters, and the y-axis shows the average distance between clusters. (**B**) A two-dimensional scatterplot of the three clusters of heart failure patients identified by the agglomerative clustering algorithm. The x- and y-axis represent principal component 1 and 2, respectively. Each component represents a constellation of clinical features that were used to cluster data points.

Table 3. Baseline demographic and clinical characteristics of the study patients stratified according to clusters.

Clinical Parameters	All Patients (*n* = 636)	Cluster 1 Uraemic (*n* = 229)	Cluster 2 Hypotensive (*n* = 117)	Cluster 3 Congestive (*n* = 290)	*p*-Value
Age (years), mean	55.6 ± 16.9	56.2 ± 17.2	54.3 ± 15.9	55.6 ± 16.9	0.7642
Male sex, n (%)	352 (55.3)	128 (36.4)	66 (18.8)	158 (44.9)	0.9189
Ethnicity, n (%)					
African	380 (59.7)	132 (34.7)	68 (17.9)	180 (47.4)	0.5485
Caucasian	154 (24.2)	57 (37.0)	25 (16.2)	72 (46.8)	0.7286
Indian	72 (11.3)	28 (38.9)	17 (23.6)	27 (37.5)	0.2787
Mixed Ancestry	29 (4.6)	12 (41.4)	7 (24.1)	10 (34.5)	0.4468
Clinical examination					
Lung crepitations	392 (61.6)	122 (31.1)	67 (17.1)	203 (51.8)	0.0002
Ascites	144 (22.6)	23 (16.0)	28 (19.4)	93 (64.6)	<0.0001
Cyanosed	51 (8.0)	9 (17.6)	10 (19.6)	32 (62.7)	0.0122
Hepatomegaly	249 (39.2)	65 (26.1)	44 (17.7)	140 (56.2)	<0.0001
Peripheral oedema	436 (68.6)	115 (26.4)	82 (18.8)	239 (54.8)	<0.0001
S3	324 (50.9)	102 (31.5)	66 (20.4)	156 (48.1)	0.0474

JVP elevated	493 (77.5)	148 (30.0)	94 (19.1)	251 (50.9)	<0.001
Systolic BP, mmHg	128 ± 30.1	135 ± 28.6	117 ± 26.5	128 ± 26.2	0.0135
Diastolic BP, mmHg	82 ± 19.7	83 ± 17.3	76 ± 16.8	84 ± 18.8	0.7012
Laboratory tests					
Haemoglobin, g/dL	12.9 ± 2.5	12.7 ± 2.4	13.2 ± 2.3	13.1 ± 2.4	0.0986
Sodium, mmol/L	139 ± 5.4	139 ± 4.7	138 ± 4.9	140 ± 5.8	0.4697
Potassium, mmol/L	4.4 ± 0.7	4.3 ± 0.6	4.5 ± 0.8	4.4 ± 0.8	0.1063
Urea, mmol/L, IQR	8.5 (6.0–12.5)	14.5 ± 31.3	14.3 ± 16.0	11.2 ± 16.4	0.0917
Echocardiography					
LVEF, %	32 (25–45)	45 ± 17.6	31 ±13.8	33 ± 14.7	<0.001
LVIDd, cm	5.6 ± 1.1	5.2 ± 1.1	5.8 ± 1.1	5.9 ± 1.0	<0.001
LVIDs, cm	4.6 ± 1.3	4.1 ± 1.3	5.0 ± 1.0	4.9 ± 1.1	<0.001

Data are presented as mean ± standard deviation for continuous features with a normal distribution, and median and interquartile ranges (IQR) (p25–p75) when the distribution is skewed. Categorical data are represented as absolute numbers and percentages. BP = blood pressure; eGFR = estimated glomerular filtration rate using the modification of diet in renal disease (MDRD) formula; JVP = jugular venous pressure; LVEF = left ventricular ejection fraction; LVIDd = left ventricular internal diameter in end-diastole; LVIDs = left ventricular internal diameter in end-systole; S3 = third heart sound. Pulse, pro brain natriuretic peptide, and estimated glomerular filtration rate not included in the table since the number of missing values >20%.

Cluster 2 (hypotensive) comprised 117 (18.4%) patients with a mean age of 54.3 ± 15.9 years. Patients in cluster 2 had a lower mean systolic and diastolic blood pressure of 117 ± 26.5 (p = 0.0135) and 76 ± 16.8 mmHg, respectively. The minimum systolic and diastolic blood pressure were 91 and 60 mmHg, respectively. In cluster 3 (congestive cluster), there was a predominance of patients with symptoms of fluid overload, and 93 (64.6%) patients had ascites. A total of 216 (52.8%) of patients with HFrEF were in cluster 3. The post hoc analyses for all statistically significant values are depicted in Tables 4 and 5.

Table 4. Post hoc analysis (showing *p*-values) of clinical parameters in the study patients stratified according to the left ventricular ejection fraction.

Clinical Parameters	All Patients (*n* = 636)	HFrEF vs. HFmEF	HFrEF vs. HFpEF	HFmEF vs. HFpEF
Systolic BP, mmHg	128 ± 30.1	0.0049	0.0022	0.9000
Diastolic BP, mmHg	82 ± 19.7	0.6592	0.0626	0.0468
Haemoglobin, g/dL	12.9 ± 2.5	0.0142	0.0156	0.9000
LVIDd, cm	5.6 ± 1.1	0.0010	0.0010	0.0010
LVIDs, cm	4.6 ± 1.3	0.0010	0.0010	0.0010

BP = blood pressure; LVIDd = left ventricular internal diameter end-diastole; LVIDs = left ventricular internal diameter end-systole.

Table 5. Post hoc analysis (showing *p*-values) of clinical parameters in the study patients stratified according to clusters.

Clinical Parameters	All Patients (*n* = 636)	Cluster 1 vs. Cluster 2	Cluster 1 vs. Cluster 3	Cluster 2 vs. Cluster 3
Systolic BP, mmHg	128 ± 30.1	0.001	0.0152	0.001
LVEF, %	32 (25–45)	0.001	0.001	0.5583
LVIDd, cm	5.6 ± 1.1	0.001	0.001	0.7457
LVIDs, cm	4.6 ± 1.3	0.001	0.001	0.9000

BP = blood pressure; LVIDd = left ventricular internal diameter end-diastole; LVIDs = left ventricular internal diameter end-systole.

The SHAP values were used to estimate the contribution of each feature in the clustering of heart failure patients. Both low and high values of systolic blood pressure, age, the LVEF, and diastolic blood pressure were the top four features that played a role in discriminating between the three clusters (Figure 3). The mean SHAP value of the systolic blood pressure (SHAP value = 0.36) was approximately 2.6 times higher than that of the LVEF (SHAP value = 0.14).

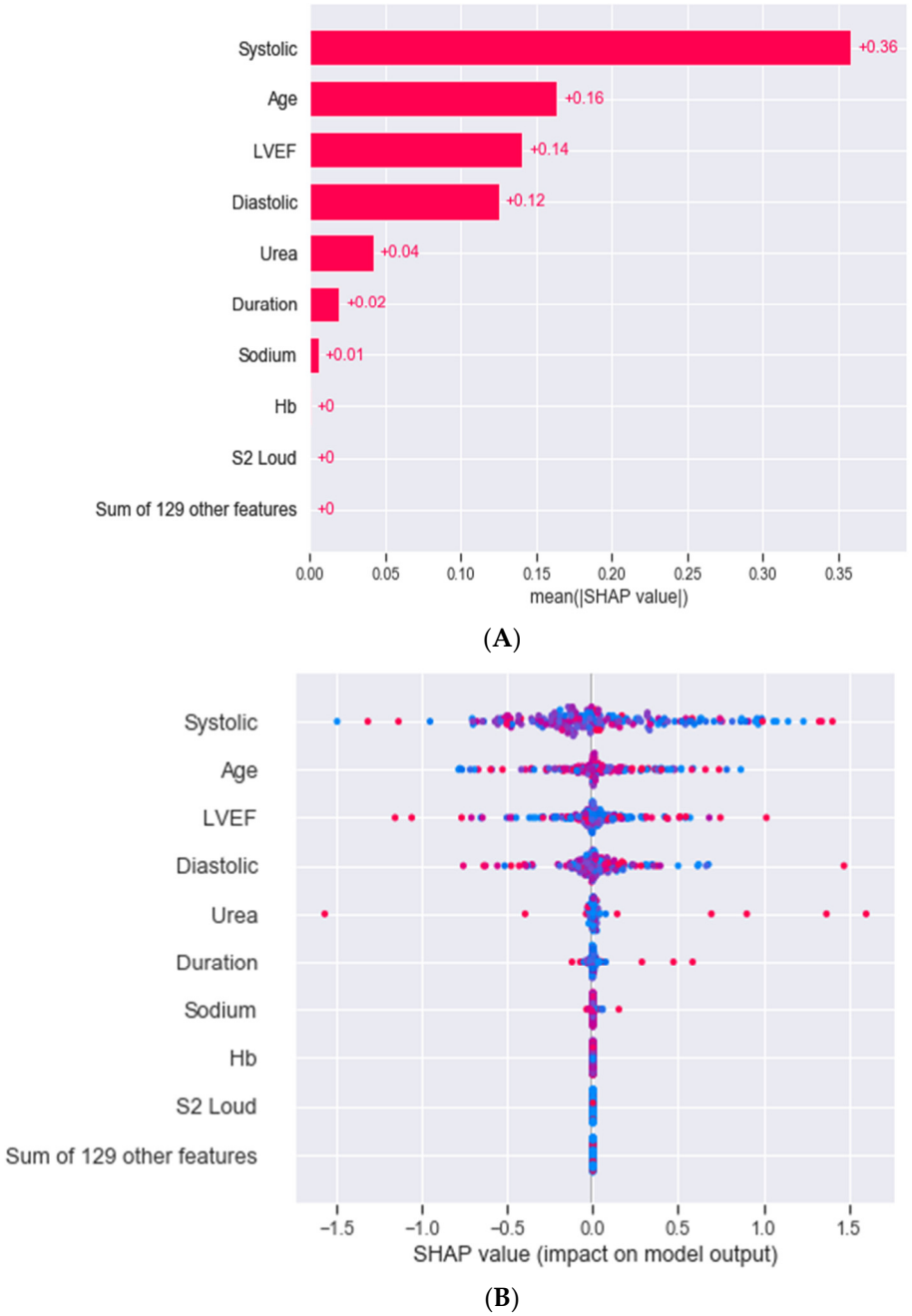


Figure 3. Features identified with Shapley additive explanations (SHAP) that played a role in the segmentation of the heart failure cohort into three clusters using the agglomerative clustering algorithm. (**A**) Bar plot depicting the importance of each feature. The higher the mean weight, the higher the contribution of the feature. (**B**) Bee swarm plot showing the contribution of the top eight features. Each dot represents a single patient. Red and blue refer to high and low feature values, respectively. Both high and low values of the systolic and diastolic blood pressure, age, and LVEF contributed towards clustering patients. Hb = haemoglobin; LVEF = left ventricular ejection fraction; S2 = second heart sound.

3.3. Outcomes

Among the 636 patients hospitalised with heart failure, 92 patients died. The rate of in-hospital all-cause mortality was 14.5%. The rate of death was the highest in patients with HFrEF (63%), followed by patients with HFpEF (19.6%) and HFmEF (17.4%). The median length of stay in the hospital in the entire heart failure cohort was 6 (IQR: 4–11) days, and it did not differ based on the LVEF category (p = 0.4441).

Cluster analysis revealed a higher mortality rate (39.1%) in patients in cluster 3 (congested cluster), followed by cluster 2, the hypotensive cluster (38%), and cluster 1 (22.8%), the uraemic cluster. Patients in cluster 3 were hospitalised for 9.0 ± 7.8 days, whereas patients in cluster 1 and 2 were both hospitalised for a mean duration of 9.6 ± 8.9 and 8.6 ± 7.7 days, respectively. Survival was shorter in the congestive cluster, and none of the patients in this cluster survived beyond 42 days (log rank test: *p*-value 0.39) (Figure 4).

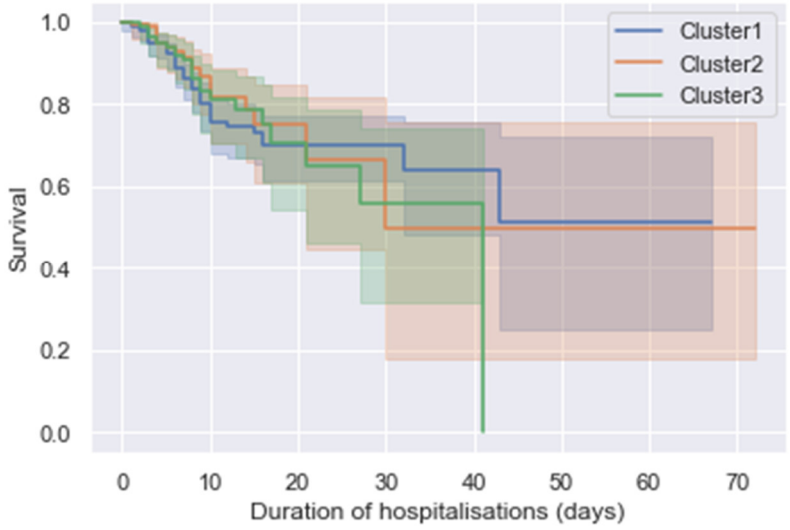


Figure 4. Kaplan–Meier survival curves for the three clusters of heart failure patients. The *x*-axis shows the duration of hospitalisation in days, and the survival rate is on the *y*-axis. Patients in cluster 3 (congested cluster) did not survive beyond 42 days, while the survival rate in cluster 1 and 2 is comparable.

3.4. Cluster Analysis on a Dataset without the Left Ventricular Ejection Fraction

Using a dataset without the LVEF, the algorithm that achieved the highest silhouette score of 0.37 was the Gaussian mixture model with seven clusters (Table 6). The most distinct clinical findings were venous hypertension in cluster 1, with 100% of patients in the cluster hospitalised with raised jugular venous pressure.

Table 6. Baseline demographic and clinical characteristics of the study patients stratified according to clusters identified by the Gaussian mixture model using the dataset without the left ventricular ejection parameter.

Clinical Parameters	All Patients (*n* = 636)	Cluster 1 (*n* = 38)	Cluster 2 (*n* = 56)	Cluster 3 (*n* = 118)	Cluster 4 (*n* = 134)	Cluster 5 (*n* = 49)	Cluster 6 (*n* = 123)	Cluster 7 (*n* = 118)	*p*-Value
Age (years), mean	55.6 ± 16.9	53.3 ± 19.4	50.9 ± 16.3	53.7 ± 17.6	57.3 ± 16.9	58.0 ± 13.9	56.0 ± 15.7	56.9 ± 17.4	0.0283
Male sex, n (%)	352 (55.3)	15 (39.5)	30 (53.6)	69 (58.5)	74 (55.2)	32 (65.3)	70 (56.9)	62 (52.5)	0.3323
Ethnicity, n (%)									
African	380 (59.7)	27 (71.1)	33 (58.9)	69 (58.5)	82 (61.2)	30 (61.2)	75 (61.0)	64 (54.2)	0.6979
Caucasian	154 (24.2)	5 (13.2)	12 (21.4)	32 (27.1)	32 (23.9)	14 (28.6)	26 (21.1)	33 (28.0)	0.4979
Indian	72 (11.3)	5 (13.2)	9 (16.1)	6 (5.1)	16 (11.9)	4 (8.2)	17 (13.8)	15 (12.7)	0.2779

Mixed Ancestry	29 (4.6)	1 (2.6)	2 (3.6)	11 (9.3)	3 (2.2)	1 (2.0)	5 (4.1)	6 (5.1)	0.1671
Clinical examination									
Lung crepitations	392 (61.6)	32 (84.2)	31 (55.4)	92 (78.0)	85 (63.4)	21 (42.9)	77 (62.6)	54 (45.8)	<0.0001
Ascites	144 (22.6)	12 (31.6)	11 (19.6)	37 (31.4)	40 (29.9)	2 (4.1)	34 (27.6)	8 (6.8)	<0.0001
Cyanosed	51 (8.0)	2 (5.3)	0 (0.0)	9 (7.6)	16 (11.9)	2 (4.1)	17 (13.8)	5 (4.2)	0.0092
Hepatomegaly	249 (39.2)	17 (44.7)	9 (16.1)	61 (51.7)	65 (48.5)	10 (20.4)	58 (47.2)	29 (24.6)	<0.0001
Peripheral oedema	436 (68.6)	34 (89.5)	31 (55.4)	95 (80.5)	104 (77.6)	18 (36.7)	105 (85.4)	49 (41.5)	<0.0001
S3	324 (50.9)	20 (52.6)	37 (66.1)	77 (65.3)	70 (52.2)	23 (46.9)	54 (43.9)	43 (36.4)	0.0001
JVP elevated	493 (77.5)	38 (100.0)	40 (71.4)	102 (86.4)	108 (80.6)	33 (67.3)	108 (87.8)	64 (54.2)	<0.001
Systolic BP, mmHg	128 ± 30.1	131 ± 23.8	120 ± 27.3	136 ± 28.7	129 ± 25.9	136 ± 29.8	119 ± 26.4	132 ± 27.5	0.4155
Diastolic BP, mmHg	82 ± 19.7	86 ± 17.0	76 ± 16.1	86 ± 18.8	85 ± 19.3	86 ± 15.5	76 ± 17.5	82 ± 17.0	0.0759
Laboratory tests									
Haemoglobin, g/dL	12.9 ± 2.5	13.4 ± 2.2	13.1 ± 2.2	12.3 ± 2.6	13.1 ± 2.5	12.8 ± 2.4	13.4 ± 2.4	13.0 ± 2.3	0.2322
Sodium, mmol/L	139 ± 5.4	140 ± 4.1	138 ± 4.4	140 ± 5.3	138 ± 5.8	140 ± 4.2	138 ± 5.8	139 ± 5.0	0.0147
Potassium, mmol/L	4.4 ± 0.7	4.3 ± 0.6	4.3 ± 0.6	4.3 ± 0.6	4.4 ± 0.7	4.3 ± 0.7	4.6 ± 0.9	4.3 ± 0.6	0.1096
Urea, mmol/L	8.5 (6.0–12.5)	8.5 ± 4.2	13.9 ± 17.4	11.3 ± 16.5	10.2 ± 5.4	12.0 ± 14.7	15.2 ± 25.5	16.8 ± 39.8	0.0194
Echocardiography									
LVIDd, cm	5.6 ± 1.1	6.1 ± 0.9	5.6 ± 1.2	5.4 ± 0.9	5.8 ± 1.1	5.2 ± 1.1	5.9 ± 1.0	5.3 ± 1.2	0.1748
LVIDs, cm	4.6 ± 1.3	5.2 ± 0.9	4.8 ± 1.1	4.4 ± 1.1	4.8 ± 1.2	3.9 ± 1.4	5.0 ± 1.0	4.3 ± 1.5	0.0536

Data are presented as mean ± standard deviation for continuous features with a normal distribution, and median and interquartile ranges (IQR) (p25–p75) when the distribution is skewed. Categorical data are represented as absolute numbers and percentages. BP= blood pressure; eGFR = estimated glomerular filtration rate using the modification of diet in renal disease (MDRD) formula; JVP = jugular venous pressure; LVIDd = left ventricular internal diameter in end-diastole; LVIDs = left ventricular internal diameter in end-systole; S3 = third heart sound. Pulse, pro brain natriuretic peptide, and estimated glomerular filtration rate not included in the table since the number of missing values >20%.

Cluster 6 (hyponatraemic cluster) had the lowest mean serum sodium of 138 ± 5.8 (confidence interval: 137–138) mmol/L, while cluster 7 (hyperuraemic cluster) had 118 patients and the highest mean urea of 16.8 ± 39.8 mmol/L. Clinical parameters that had a significant impact on model output included systolic blood pressure (mean SHAP value = 0.27), diastolic blood pressure (mean SHAP value = 0.16), and age (mean SHAP value = 0.14). Serum sodium levels were the only clinical parameter that showed statistical significance on the post hoc analysis, confirming low sodium levels in cluster 6.

4. Discussion

In this study, we used retrospective data of consecutive patients hospitalised with new-onset heart failure symptoms or decompensated chronic heart failure to describe demographic and clinical characteristics and categorise heart failure phenotypes using unsupervised machine algorithms. The mean age in our cohort was 55.6 ± 16.9 years, and 14.5% of all patients hospitalised with the diagnosis of heart failure died.

Heart failure is commonly associated with poorer outcomes in LMICs, and the onset of heart failure occurs at a younger age of 53 years, affecting economic productivity in this

region [7,8]. The most plausible explanations for poorer outcomes include, but are not limited to, the unavailability of life-prolonging therapy such as cardiac resynchronisation defibrillators and heart transplant services. Furthermore, our hospital is a quartenary-level institution, implying that there might be a delay in the referral of patients, resulting our patients being hospitalised with advanced stages of heart failure. The younger age of onset may be partly related to the dominance of non-ischaemic cardiomyopathy, possibly with a genetic aetiology, considering that in sub-Saharan Africa (SSA), the most common cardiovascular diseases are still hypertensive heart disease, cardiomyopathies, rheumatic heart disease, and congenital heart disease [7,8], unlike in high-income countries, where ischaemic heart disease is the primary cause of heart failure.

After subjecting our heart failure dataset to unsupervised machine learning algorithms, the agglomerative clustering algorithm identified three clusters: the uraemic cluster, with relatively high urea levels; the hypotensive cluster, with lower systolic and diastolic blood pressure; and the congested cluster, with a high burden of heart failure symptoms and signs of fluid overload. The uraemic cluster had slightly older patients, potentially with underlying acute kidney injury or chronic renal dysfunction.

The heart is responsible for maintaining adequate perfusion to all organs. Low systolic and diastolic pressure were a dominant finding in cluster 2, the hypotensive cluster. The reduction in myocardial contractility in this cluster likely caused a decrease in cardiac output, systemic vasoconstriction, and hypotension in the hypotensive cluster [9]. This finding resonates with the importance of preventing hypotension in heart failure patients, considering that hypoperfusion may lead to cardiogenic shock and subsequent mortality. The in-hospital all-cause mortality rate in the hypotensive cluster was 38%, similar to the congestive cluster, with the highest mortality rate of 39.1%. It is well established that patients who present with severe symptoms of acute decompensated heart failure experience poor outcomes, suggesting the need to expedite their management, irrespective of the LVEF.

Most studies have confirmed the utility of supervised machine learning algorithms for classification tasks [10–18]. However, few studies have since been published on phenotyping heart failure patients using unsupervised machine learning algorithms [19–25]. In a paper summarising 40 studies on heart failure subtype definition using unsupervised machine learning, hierarchical clustering was used in 25 studies. Most clusters were derived based on heart failure symptoms only (n = 20) [6]. Harada et al. used K-Means clustering to stratify heart failure patients with an LVEF above 40% and found four distinct groups. They concluded that older patients might be at a higher risk of right ventricular afterload mismatch and renal dysfunction [26]. In addition, in Russia, 25952 heart failure patients were clustered using K-Means, and 15 distinct heart failure phenotypes were identified. The clusters identified in the study from Russia were characterised and labelled based on comorbidities and causes of heart failure only.

Clustering algorithms have the potential to unravel distinct phenotypes of heart failure that may assist clinicians in triaging or risk profiling patients, therefore positively influencing the management of heart failure patients. Our study is unique since it offers insights into the role of machine learning in better understanding heart failure phenotypes in our region. The systolic blood pressure had a 2.6 times higher impact on differentiating between clusters, followed by the age at the time of admission with a mean SHAP value of 0.16, and the LVEF had a lower SHAP value of 0.14. These findings suggest that blood pressure and age may play a role in the diagnosis, risk stratification, and management of heart failure patients in centres without access to echocardiography. Despite removing the LVEF from the dataset, both systolic and diastolic blood pressure still had a significant impact on model output.

Each institution should attempt to cluster its heart failure patients and discover new clinical patterns. For example, our study's findings could motivate for the design and implementation of protocols that guide the management of heart failure patients, specifically

with hyperuricaemia, hypotension, and severe symptoms of fluid overload in our institution. However, we are mindful of the fact that clusters formed will differ across regions, partly due to the variation in the causes of heart failure and the availability of clinical parameters used to create clusters.

Newer unsupervised machine learning techniques should take data analysis to the next level by creating a code that allows algorithms to interrogate the data and predict the risk of complications. Future studies should use larger, prospective cohorts of patients and subsequently train and test the performance of clustering algorithms using new data that were not used to create original clusters of patients. Most importantly, the prognostic value of the unsupervised machine learning algorithms should be verified by measuring clinical outcomes and comparing them with outcomes obtained from patients classified solely with the LVEF.

A larger dataset would have provided more perspective into the characteristics of patients in each cluster. The study is also limited by its retrospective nature and was a single-centre study. As such, the findings from this study may not be generalisable to all heart failures treated in LMICs. In addition, pertinent features such as the cause of mortality were not available in the dataset. Most clinical history findings were documented in free text format and were not analysed. Furthermore, a significant number of patients diagnosed with heart failure with missing echocardiogram findings were not included in the analysis. Moreover, discharge oral medications were not captured for all patients, and were not included in the final analysis. Overall, the lack of electronic health records and dedicated data capturers are major hindrances in LMICs, such that patients may only partially benefit from the capabilities of artificial intelligence.

5. Conclusions

Agglomerative clustering is an unsupervised machine learning algorithm that identified three clusters of heart failure phenotypes in Johannesburg. The most important features that had an impact on the creation of clusters were blood pressure, age, and the LVEF. Clinicians without access to echocardiography could rely on blood pressure measurements and age to risk stratify and identify heart failure patients that require referral to centres equipped with cardiology services. However, larger prospective studies are mandatory for validating the role of these clinical parameters.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/app13031509/s1, The Jupyter notebook script.

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

Funding: D.M. was a full-time PhD student funded by the Bongani Mayosi Netcare Clinical Scholarship, the Discovery Academic Fellowship (Grant No. 03902), the Carnegie Corporation of New York (Grant No. b8749), the South African Heart Association, and the University of the Witwatersrand Chancellor's Female Academic Leaders Fellowship (NPC) Grant.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the University of the Witwatersrand Human Research Ethics Committee (protocol code M190515, 18/07/2019).

Informed Consent Statement: Patient consent was waived since the study entailed a retrospective chart review.

Data Availability Statement: Data is unavailable due to ethical restrictions.

Acknowledgments: The authors would like to thank Pravin Manga for implementing the PMRCardio database; the staff members in the Division of Cardiology at the CMJAH; Marheb Badianyama,

Jenna Damelin, Ashlee Firer, Thakgalo Thibeng, Yemurai Muchenge, Michaela Gough, and Mpumelelo Ngadla, who assisted with verifying patient data before the analysis; and Karl Basson, Geethan Singh, and Ntuthuko Mngomezulu for assisting with the technical data queries.

Conflicts of Interest: EK has received speaker honorarium, travel assistance from Astra Zeneca, Servier, Novartis, and has been the head of ethics and guideline committee of the South African Heart Association. The rest of the authors declare no other competing interests.

References

1. Jovel, J.; Greiner, R. An Introduction to Machine Learning Approaches for Biomedical Research. *Front. Med.* **2021**, *8*, 771607. https://doi.org/10.3389/fmed.2021.771607.
2. Ponikowski, P.; Anker, S.D.; AlHabib, K.F.; Cowie, M.R.; Force, T.L.; Hu, S.; Jaarsma, T.; Krum, H.; Rastogi, V.; Rohde, L.E.; et al. Heart failure: Preventing disease and death worldwide. *ESC Heart Fail.* **2014**, *1*, 4–25. https://doi.org/10.1002/ehf2.12005.
3. Ponikowski, P.; Voors, A.A.; Anker, S.D.; Bueno, H.; Cleland, J.G.; Coats, A.J.; Falk, V.; González-Juanatey, J.R.; Harjola, V.-P.; Jankowska, E.A.; et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur. J. Heart Fail.* **2016**, *18*, 891–975. https://doi.org/10.1002/ejhf.592.
4. Kosaraju, A.; Goyal, A.; Grigorova, Y.; Makaryus, A.N. Left Ventricular Ejection Fraction. In *StatPearls. Treasure Island (FL)*; StatPearls Publishing; 2022 January. Available online: https://www.ncbi.nlm.nih.gov/books/NBK459131/ (accessed on July 2022)
5. Jasinska-Piadlo, A.; Bond, R.; Biglarbeigi, P.; Brisk, R.; Campbell, P.; McEneaney, D. What can machines learn about heart failure? A systematic literature review. *Int. J. Data Sci. Anal.* **2022**, *13*, 163–183.
6. Banerjee, A.; Chen, S.; Fatemifar, G.; Zeina, M.; Lumbers, R.T.; Mielke, J.; Gill, S.; Kotecha, D.; Freitag, D.F.; Denaxas, S.; et al. Machine learning for subtype definition and risk prediction in heart failure, acute coronary syndromes and atrial fibrillation: Systematic review of validity and clinical utility. *BMC Med.* **2021**, *19*, 85. https://doi.org/10.1186/s12916-021-01940-7.
7. Dokainish, H.; Teo, K.; Zhu, J.; Roy, A.; AlHabib, K.F.; ElSayed, A.; Palileo-Villaneuva, L.; Lopez-Jaramillo, P.; Karaye, K.; Yusoff, K.; et al. Heart Failure in Africa, Asia, the Middle East and South America: The INTER-CHF study. *Int. J. Cardiol.* **2016**, *204*, 133–141. https://doi.org/10.1016/j.ijcard.2015.11.183.
8. Ajayi, A.A.; Sofowora, G.G.; Ladipo, G.O. Explaining Heart Failure Hyper-mortality in Sub Saharan Africa: Global Genomic and Environmental Contribution Review. *J. Natl. Med. Assoc.* **2020**, *112*, 141–157.
9. van Diepen, S.; Katz, J.N.; Albert, N.M.; Henry, T.D.; Jacobs, A.K.; Kapur, N.K.; Kilic, A.; Menon, V.; Ohman, E.M.; Sweitzer, N.K.; et al. Contemporary Management of Cardiogenic Shock: A Scientific Statement from the American Heart Association. *Circulation* **2017**, *136*, e232–e268. https://doi.org/10.1161/CIR.0000000000000525.
10. Kwon, J.M.; Kim, K.H.; Jeon, K.H.; Park, J. Deep learning for predicting in-hospital mortality among heart disease patients based on echocardiography. *Echocardiography* **2019**, *36*, 213–218. https://doi.org/10.1111/echo.14220.
11. Segar, M.W.; Jaeger, B.C.; Patel, K.V.; Nambi, V.; Ndumele, C.E.; Correa, A.; Butler, J.; Chandra, A.; Ayers, C.; Rao, S.; et al. Development and Validation of Machine Learning-Based Race-Specific Models to Predict 10-Year Risk of Heart Failure: A Multi-Cohort Analysis. *Circulation* **2021**, *143*, 2370–2383. https://doi.org/10.1161/CIRCULATIONAHA.120.053134.
12. Su, Z.; Brecht, T.; O'Donovan, F.; Boussios, C.; Menon, V.; Gliklich, R.; Fonarow, G. Machine Learning Enhanced Predictions of Hospital Readmission or Death in Heart Failure. *Circulation* **2017**, *136*, A16941.
13. Ahmad, T.; Lund, L.H.; Rao, P.; Ghosh, R.; Warier, P.; Vaccaro, B.; Dahlström, U.; O'Connor, C.M.; Felker, G.M.; Desai, N.R. Machine Learning Methods Improve Prognostication, Identify Clinically Distinct Phenotypes, and Detect Heterogeneity in Response to Therapy in a Large Cohort of Heart Failure Patients. *J. Am. Heart Assoc.* **2018**, *7*, e008081. https://doi.org/10.1161/JAHA.117.008081.
14. Shameer, K.; Johnson, K.W.; Yahi, A.; Miotto, R.; Li, L.I.; Ricks, D.; Jebakaran, J.; Kovatch, P.; Sengupta, P.P.; Gelijns, S.; et al. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: A case-study using mount sinai heart failure cohort. *Pac. Symp. Biocomput.* **2017**, *22*, 276–287. https://doi.org/10.1142/9789813207813_0027.
15. Golas, S.B.; Shibahara, T.; Agboola, S.; Otaki, H.; Sato, J.; Nakae, T.; Hisamitsu, T.; Kojima, G.; Felsted, J.; Kakarmath, S.; et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: A retrospective analysis of electronic medical records data. *BMC Med. Inform. Decis. Mak.* **2018**, *18*, 44. https://doi.org/10.1186/s12911-018-0620-z.
16. Miao, F.; Cai, Y.P.; Zhang, Y.X.; Fan, X.M.; Li, Y. Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest. *IEEE Access* **2018**, *6*, 7244–7253. https://doi.org/10.1109/ACCESS.2018.2789898.
17. Meng, F.; Zhang, Z.; Hou, X.; Qian, Z.; Wang, Y.; Chen, Y.; Wang, Y.; Zhou, Y.; Chen, Z.; Zhang, X.; et al. Machine learning for prediction of sudden cardiac death in heart failure patients with low left ventricular ejection fraction: Study protocol for a retroprospective multicentre registry in China. *BMJ Open* **2019**, *9*, e023724. https://doi.org/10.1136/bmjopen-2018-023724.
18. Liu, W.; Stansbury, C.; Singh, K.; Ryan, A.M.; Sukul, D.; Mahmoudi, E.; Waljee, A.; Zhu, J.; Nallamothu, B.K. Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding. *PLoS ONE* **2020**, *15*, e0221606. https://doi.org/10.1371/journal.pone.0221606.

19. Ahmad, T.; Pencina, M.J.; Schulte, P.J.; O'Brien, E.; Whellan, D.J.; Pina, I.L.; Kitzman, D.W.; Lee, K.L.; O'Connor, C.M.; Felker, G.M. Clinical implications of chronic heart failure phenotypes defined by cluster analysis. *J. Am. Coll. Cardiol.* **2014**, *64*, 1765–1774. https://doi.org/10.1016/j.jacc.2014.07.979.
20. Ieva, F.; Paganoni, A.M.; Pietrabissa, T. Dynamic clustering of hazard functions: An application to disease progression in chronic heart failure. *Health Care Manag. Sci.* **2017**, *20*, 353–364. https://doi.org/10.1007/s10729-016-9357-3.
21. Moser, D.K.; Lee, K.S.; Wu, J.-R.; Mudd-Martin, G.; Jaarsma, T.; Huang, T.-Y.; Fan, X.; Strömberg, A.; Lennie, T.A.; Riegel, B. Identification of symptom clusters among patients with heart failure: An international observational study. *Int. J. Nurs. Stud.* **2014**, *51*, 1366–1372.
22. Bose, E.; Radhakrishnan, K. Using Unsupervised Machine Learning to Identify Subgroups among Home Health Patients with Heart Failure Using Telehealth. *Comput. Inform. Nurs.* **2018**, *36*, 242–248. https://doi.org/10.1097/CIN.0000000000000423.
23. Segar, M.W.; Patel, K.V.; Ayers, C.; Basit, M.; Tang, W.H.W.; Willett, D.; Berry, J.; Grodin, J.L.; Pandey, A. Phenomapping of patients with heart failure with preserved ejection fraction using machine learning-based unsupervised cluster analysis. *Eur. J. Heart Fail.* **2020**, *22*, 148–158. https://doi.org/10.1002/ejhf.1621.
24. Nouraei, H.; Nouraei, H.; Rabkin, S.W. Comparison of Unsupervised Machine Learning Approaches for Cluster Analysis to Define Subgroups of Heart Failure with Preserved Ejection Fraction with Different Outcomes. *Bioengineering* **2022**, *9*, 175. https://doi.org/10.3390/bioengineering9040175.
25. Perry, A.; Loh, F.; Adamo, L.; Zhang, K.W.; Deych, E.; Foraker, R.; Mann, D.L. Unsupervised cluster analysis of patients with recovered left ventricular ejection fraction identifies unique clinical phenotypes. *PLoS ONE* **2021**, *16*, e0248317. https://doi.org/10.1371/journal.pone.0248317.
26. Harada, D.; Asanoi, H.; Noto, T.; Takagawa, J. Different Pathophysiology and Outcomes of Heart Failure with Preserved Ejection Fraction Stratified by K-Means Clustering. *Front. Cardiovasc. Med.* **2020**, *7*, 607760. https://doi.org/10.3389/fcvm.2020.607760.

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.

CHAPTER 5 – MANUSCRIPT THREE: PREDICTING ALL-CAUSE MORTALITY IN HEART FAILURE USING MACHINE LEARNING

5.1 Introduction

In this chapter, 500 heart failure patients with a left ventricular ejection fraction (LVEF) less than 50% were used to train six machine learning algorithms to predict mortality. Features used to train algorithms are in Appendix B.

5.2 Current Knowledge

Machine learning algorithms learn complex patterns in the dataset before making predictions, and can risk stratify patients, hence identifying high-risk patients who are candidates for preventative therapy.

5.3 Aims and Objectives

The aim of the study was to predict in-hospital all-cause mortality in patients with a LVEF < 50%. In addition, the study objectives were to compare supervised machine learning algorithms to traditional statistical methods for predicting all-cause mortality in heart failure patients and to predict 30-day all-cause readmission using machine learning algorithms.

Due to the small number of patients readmitted within 30 days, the algorithms failed to learn from the data and were overfitting. Results obtained after training algorithms to predict rehospitalisation have been attached as Appendix C.

Citation details: Mpanya, D.; Celik, T.; Klug, E.; Ntsinjana, H. Predicting in-hospital all-cause mortality in heart failure using machine learning. Frontiers in Cardiovascular Medicine, Volume 9, 2023. https://doi.org/10.3389/fcvm.2022.1032524. The final chapter contains the discussion, recommendations and conclusions.

frontiers | Frontiers in Cardiovascular Medicine

TYPE Original Research
PUBLISHED 11 January 2023
DOI 10.3389/fcvm.2022.1032524

Check for updates

OPEN ACCESS

EDITED BY
Zhenjie Yao,
Institute of Microelectronics (CAS), China

REVIEWED BY
Binhua Wang,
Disaster Medicine Research Center of Hospital, China
Xinxin Zhang,
Dalian Medical University, China

*CORRESPONDENCE
Dineo Mpanya
✉ dineo.mpanya@wits.ac.za

SPECIALTY SECTION
This article was submitted to General Cardiovascular Medicine, a section of the journal Frontiers in Cardiovascular Medicine

RECEIVED 30 August 2022
ACCEPTED 23 December 2022
PUBLISHED 11 January 2023

CITATION
Mpanya D, Celik T, Klug E and Ntsinjana H (2023) Predicting in-hospital all-cause mortality in heart failure using machine learning. *Front. Cardiovasc. Med.* 9:1032524. doi: 10.3389/fcvm.2022.1032524

COPYRIGHT
© 2023 Mpanya, Celik, Klug and Ntsinjana. 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.

Predicting in-hospital all-cause mortality in heart failure using machine learning

Dineo Mpanya[1,2]*, Turgay Celik[2,3], Eric Klug[4] and Hopewell Ntsinjana[5]

[1]Division of Cardiology, Department of Internal Medicine, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa, [2]Wits Institute of Data Science, University of the Witwatersrand, Johannesburg, South Africa, [3]School of Electrical and Information Engineering, Faculty of Engineering and Built Environment, University of the Witwatersrand, Johannesburg, South Africa, [4]Netcare Sunninghill, Sunward Park Hospitals and Division of Cardiology, Department of Internal Medicine, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa, [5]Department of Paediatrics and Child Health, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

Background: The age of onset and causes of heart failure differ between high-income and low-and-middle-income countries (LMIC). Heart failure patients in LMIC also experience a higher mortality rate. Innovative ways that can risk stratify heart failure patients in this region are needed. The aim of this study was to demonstrate the utility of machine learning in predicting all-cause mortality in heart failure patients hospitalised in a tertiary academic centre.

Methods: Six supervised machine learning algorithms were trained to predict in-hospital all-cause mortality using data from 500 consecutive heart failure patients with a left ventricular ejection fraction (LVEF) less than 50%.

Results: The mean age was 55.2 ± 16.8 years. There were 271 (54.2%) males, and the mean LVEF was 29 ± 9.2%. The median duration of hospitalisation was 7 days (interquartile range: 4–11), and it did not differ between patients discharged alive and those who died. After a prediction window of 4 years (interquartile range: 2–6), 84 (16.8%) patients died before discharge from the hospital. The area under the receiver operating characteristic curve was 0.82, 0.78, 0.77, 0.76, 0.75, and 0.62 for random forest, logistic regression, support vector machines (SVM), extreme gradient boosting, multilayer perceptron (MLP), and decision trees, and the accuracy during the test phase was 88, 87, 86, 82, 78, and 76% for random forest, MLP, SVM, extreme gradient boosting, decision trees, and logistic regression. The support vector machines were the best performing algorithm, and furosemide, beta-blockers, spironolactone, early diastolic murmur, and a parasternal heave had a positive coefficient with the target feature, whereas coronary artery disease, potassium, oedema grade, ischaemic cardiomyopathy, and right bundle branch block on electrocardiogram had negative coefficients.

Conclusion: Despite a small sample size, supervised machine learning algorithms successfully predicted all-cause mortality with modest accuracy. The SVM model will be externally validated using data from multiple cardiology centres in South Africa before developing a uniquely African risk prediction tool that can potentially transform heart failure management through precision medicine.

KEYWORDS

machine learning, heart failure, mortality, predict, Africa

Highlights

- Models predicting outcomes in heart failure are built mainly from data originating in high-income countries, where ischaemic heart disease predominates.
- In this study, we used data of in-hospital heart failure patients to train machine learning algorithms to predict all-cause mortality.
- This is the first sub-Saharan African model demonstrating machine learning algorithms' utility in making predictions.
- In addition, the study shows that the performance of machine learning algorithms is sensitive to the sample size and the credibility of the data used to train algorithms.

1. Introduction

Patients diagnosed with heart failure experience a high mortality rate, with 30–50% of patients demising within 5 years from diagnosis (1). This is despite the introduction of novel pharmacological and intracardiac device therapy. Furthermore, due to the high cost associated with the implantation of device therapy and transplant services, only a select few patients in sub-Saharan Africa (SSA) can access these life-saving interventions. As such, there is a need to risk stratify patients with precision, ensuring that the right patient receives the right therapy at the right time.

Risk calculators are derived from predictive modeling, and they are used to estimate the risk of outcomes such as mortality and rehospitalisation in heart failure. Most readily available risk calculators were created using data predominantly from high-income countries (HIC). For example, the Meta-Analysis Global Group in Chronic Heart Failure (MAGGIC) risk calculator, created from a statistical Poisson model, estimates the risk of one and 3-year mortality in heart failure (2). Another widely available risk calculator predicting in-hospital all-cause mortality was derived from the Get with the Guidelines Heart Failure data of multiple sites in HIC, using a multivariable logistic regression model (3). To the best of our knowledge, risk calculators derived from SSA do not exist, and existing heart failure predictive models have not been validated using data originating from SSA.

Predictive models are population-specific and are prone to inherent biases such as the availability of therapy, referral pathways, genotype, and varying population data used to create these models. Furthermore, most of the available risk calculators were derived using data from heart failure patients in whom the primary cause of heart failure is ischaemic heart disease, unlike in SSA, where the predominant cause of heart failure is non-ischaemic heart disease. Therefore, there is a need to create uniquely African predictive models, mainly using machine learning techniques, since machine learning algorithms are capable of learning from a labelled dataset prior to making predictions. Also, the accuracy of predictions may improve over time as the algorithms are exposed to a larger dataset. This study aims to use machine learning algorithms to predict all-cause mortality in heart failure patients hospitalised in a tertiary academic centre in Johannesburg, South Africa.

2. Materials and methods

2.1. Study design and participants

Patient data was exported from the PMRCardio database, a Microsoft Structured Query Language Server Management Studio version 15.0.18330.0, that stores clinical data of patients hospitalised in cardiology wards. The Charlotte Maxeke Johannesburg Academic Hospital (CMJAH) is a tertiary-level state-owned institution equipped with general cardiac wards, a cardiac intensive unit, a catheterisation and electrophysiological laboratory, and outpatient clinics.

All acutely and chronically ill patients with cardiovascular diseases are hospitalised in dedicated cardiac wards. Their admission data is stored digitally in the PMRCardio database by trainee physicians rotating in the cardiology wards. Clinical parameters available in the dataset (features) include demographic data, clinical history and examination

findings, laboratory indices, electrocardiogram (ECG) data, echocardiography data, angiography data, the status of the patient upon discharge (alive or dead), the date of subsequent admission, and discharge oral medication (**Supplementary material**). In our institution, heart failure patients are routinely prescribed guideline-directed medical therapy.

After exporting data from the PMRCardio database, data were merged into a single Microsoft Excel Sheet and imported into Python software, version 3.10.0. The international classification of diseases, tenth revision (ICD-10) code, was used to select patients diagnosed with heart failure due to any aetiology. The following ICD-10 diagnoses were used to select patients included in the analysis: "heart failure unspecified," "congestive heart failure," "left ventricular failure," "dilated cardiomyopathy," or "ischaemic cardiomyopathy." Only heart failure patients 18 years of age and older with a left ventricular ejection fraction (LVEF) < 50% hospitalised between 2009 and 2018 were included in the dataset used to build predictive models. The rationale for only including patients with a LVEF < 50% is that in our institution, patients with a LVEF between 40 and 49% (*heart failure with a mid-range ejection fraction)* are managed similarly to those with LVEF < 40%.

All retrospective patient data used in the final analysis was further prospectively verified by comparing hard copies of medical records with electronic data. Patients without a documented LVEF were excluded from the analysis. Biochemical data captured at the time of admission was obtained from the National Health Laboratory Service (NHLS). Approval to conduct the study was received from the University of the Witwatersrand Human Research Ethics Committee (Clearance certificate number: M190515). Permission to conduct the study was also obtained from NHLS senior authorities. Informed consent was not obtained from the patients as this was a retrospective chart review. The study protocol conformed to the ethical guidelines of the 1975 Declaration of Helsinki as reflected in *a priori* approval by the institution's human research committee.

2.2. Data cleaning and pre-processing

Packages and libraries required for data analysis were imported into Python software, version 3.10.0. The following software, packages and libraries were installed: Jupyter notebook version 6.0.3, scikit-learn version 1.02, statsmodels version 0.13.2, seaborn version 0.11.2, numpy version 1.21.6, scipystats version 1.5.4, matplotlib version 3.1.3, and pandas version 1.3.5. Features with more than 20% missing values, patient identifiers, and column data without variation were removed from the dataset. The remaining continuous features with missing data were imputed using the mean or median value. Categorical features were coded as 0 for no and 1 for yes, while blank or missing values were ascribed with zero. Data used to train machine learning algorithms was normalised, transformed and scaled between 0 and 1. The rationale for scaling the data is to standardise the weight of each feature. For example, sodium levels equal to 120 will have more weight than a potassium level of 5.0.

2.3. Exploratory data analysis and hyperparameter optimisation

Exploratory data analyses were carried out visually and quantitatively. Data were partitioned into training and test datasets, with 70% of the data used for training. Grid searching

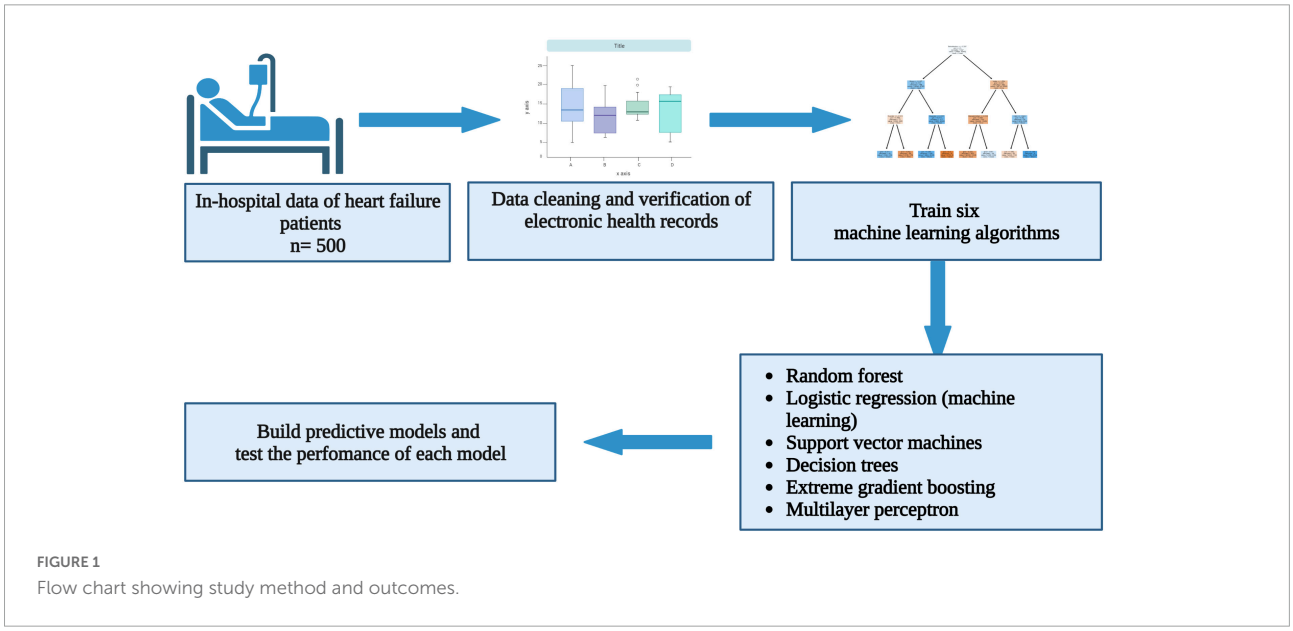


FIGURE 1
Flow chart showing study method and outcomes.

was implemented for all machine learning models, where parameters associated with the best performance metrics are discovered and selected during model hyper-tuning. Since the data is unbalanced with more patients in the survivors' class, cost-sensitive learning was applied at a ratio of 0.14: 0.86, where a higher weight was allocated to the minority class (*dead = 1*).

All demographic and clinical parameters captured during arrival at the hospital were used as features. A filter method was then used to identify features that are positively or negatively correlated with all-cause mortality. Filtering was done using correlation coefficients. To ascertain feature-to-target correlation, a correlation coefficient was estimated for each feature. The threshold was set at 0.5. The lowest and highest values were furosemide ($r = -0.277395$, $p < 0.001$) and potassium ($r = 0.180126$, $p < 0.001$). Since all features were weakly correlated with the target value, none of the features were removed. To build predictive models, we only used features captured during the first hospitalisation. As such, the model with the best performance metrics will be used to predict the risk of in-hospital mortality using patient data obtained at the time of arrival at the cardiology wards at the CMJAH.

In this study, the prediction window is 4 years (interquartile range: 2–6 years). The prediction window starts from the time of hospitalisation (index date), where the risk of death is assessed, and ends at the time when a patient dies while hospitalised. The EHR system only stores a single entry of data. As such, there is no time window. For example, although clinical parameters are measured several times while the patient is hospitalised, only the first measurement (*baseline*) was captured in the EHR system. The parameters presented in the paper were collected from the baseline index admission for each of the 500 patients. The following six algorithms were used to build models predicting all-cause in-hospital mortality.

2.3.1. Decision trees

Decision trees implement a sequential decision process based on whether the conditions set are true or false.

TABLE 1 Baseline demographic and clinical characteristics of all heart failure patients.

Clinical parameters	All patients (*n* = 500)	Died (*n* = 84)	Alive (*n* = 416)	*P*-value
Age (years)	55.2 ± 16.8	57.3 ± 18.6	54.8 ± 16.6	0.2162
Male	271 (54.2)	47 (55.9)	224 (53.8)	0.1249
Vital signs				
Pulse, bpm	102 ± 23.8	95 ± 25.5	103 ± 23.5	0.1963
Systolic BP, mmHg	127 ± 30.4	121 ± 31.7	128 ± 30.1	0.0588
Diastolic BP, mmHg	83 ± 20.4	83 ± 21.2	84 ± 20.3	0.6691
Laboratory tests				
Haemoglobin, g/dL	13.1 ± 2.5	13.3 ± 2.2	13.1 ± 2.5	0.4467
Sodium, mmol/l	139 ± 5.5	139 ± 5.5	139 ± 5.5	0.5220
Potassium, mmol/l	4.3 (3.9–4.7)	4.4 (3.9–4.7)	4.3 (3.9–4.7)	0.6610
Urea	8.6 (6.2–12.5)	8.05 (6–11.9)	8.7 (6.2–12.7)	0.2874
Creatinine	104 (80–137)	98 (81–124)	105 (80–142)	0.3774
eGFR, ml/min	65.7 (44.5–85.0)	68.0 (37.0–84.8)	65.4 (46.0– 85.0)	0.6116
Pro BNP, ng/l	8,541 (4,813–14,780)	7,592 (4,402–12,734)	8,709 (4,813–14,998)	0.3267
Troponin	23.5 (0.15–66.0)	4.06 (0.054–77.0)	24.5 (0.17–65.0)	0.3427
Total cholesterol	3.5 (2.8–4.1)	3.3 (2.8–4.1)	3.5 (2.8–4.2)	0.9705
LDL	2.1 (1.5–2.7)	2.1 (1.7–2.5)	2.1 (1.5–2.7)	0.9618
HDL	0.85 (0.6–1.1)	0.9 (0.6–1.2)	0.8 (0.6–1.1)	0.6493
Echocardiogram				
LVEF, %	29 ± 9.2	30 ± 9.2	29 ± 9.2	0.7356
LVIDd, cm	5.9 ± 1.0	5.9 ± 0.1	5.8 ± 0.9	0.8655
LVIDs, cm	5.1 (4.5–5.6)	5.0 (4.3–5.7)	5.1 (4.5–5.6)	0.5306
Length of stay (days)	7 (4–11)	7 (4–10)	7 (4–11)	0.5796

Data are represented as mean and standard deviation (SD) for continuous features with a normal distribution and median and interquartile ranges (p25–p75) when the distribution is skewed. Categorical data are represented as absolute numbers and percentages. BP, blood pressure; eGFR, estimated glomerular filtration rate; HDL, high-density lipoprotein; LDL, low-density lipoprotein; LVEF, left ventricular ejection fraction; LVIDd, left ventricular internal diameter at diastole; LVIDs, left ventricular internal diameter at systole.

The decision tree algorithm was trained, and the tree was plotted based on grid search results by placing the criterion as Gini, maximum depth at five, and the minimum sample split at eight.

2.3.2. Random forest

A random forest classifier uses multiple decision trees, and each of the decision trees outputs a prediction (dead vs. alive). To estimate the final output, the random forest algorithm then counts the number of votes for each class predicted by each decision tree. Grid search identified the following parameters to be responsible for best model performance: the number of estimators = 1 757, minimum samples split = 5, minimum samples leaf = 2, maximum depth = 150, and criterion = entropy.

2.3.3. Support vector machines

Support vector machines classify data by creating a line or hyperplane between the classes. The maximum distance between the classes (support vectors) and the hyperplane is then chosen. After applying grid search, the best parameters for creating a model were a linear kernel, gamma = 0.001, and $C = 10$. The "C" is a parameter that controls the effect of support vectors.

2.3.4. Logistic regression

For the machine learning logistic regression model, the coefficient of each feature was obtained after training and testing models' performance. A conventional statistical univariate and multivariate logistic regression model was also built to compare predictors of mortality identified by machine learning algorithms with those extracted from the statistical model. Clinical parameters with a p-value less than 0.05 were selected and included in the final multivariate logistic regression model to identify independent predictors of all-cause mortality. Confidence interval percentiles were set between 0.025 and 0.975.

2.3.5. Extreme gradient boosting

Extreme gradient boosting is an ensemble of weak algorithms; usually, decision trees added to the model sequentially. Feature importance is calculated based on the number of times the feature appears in isolated trees.

2.3.6. Multilayer perceptron

The MLP network consists of one input layer, hidden layers, and an output layer. Grid search was used to find the best model parameters. The search for optimal parameters was performed with the following limits: 0–0.4 dropout rate, batch size between 10 and 60, and epochs

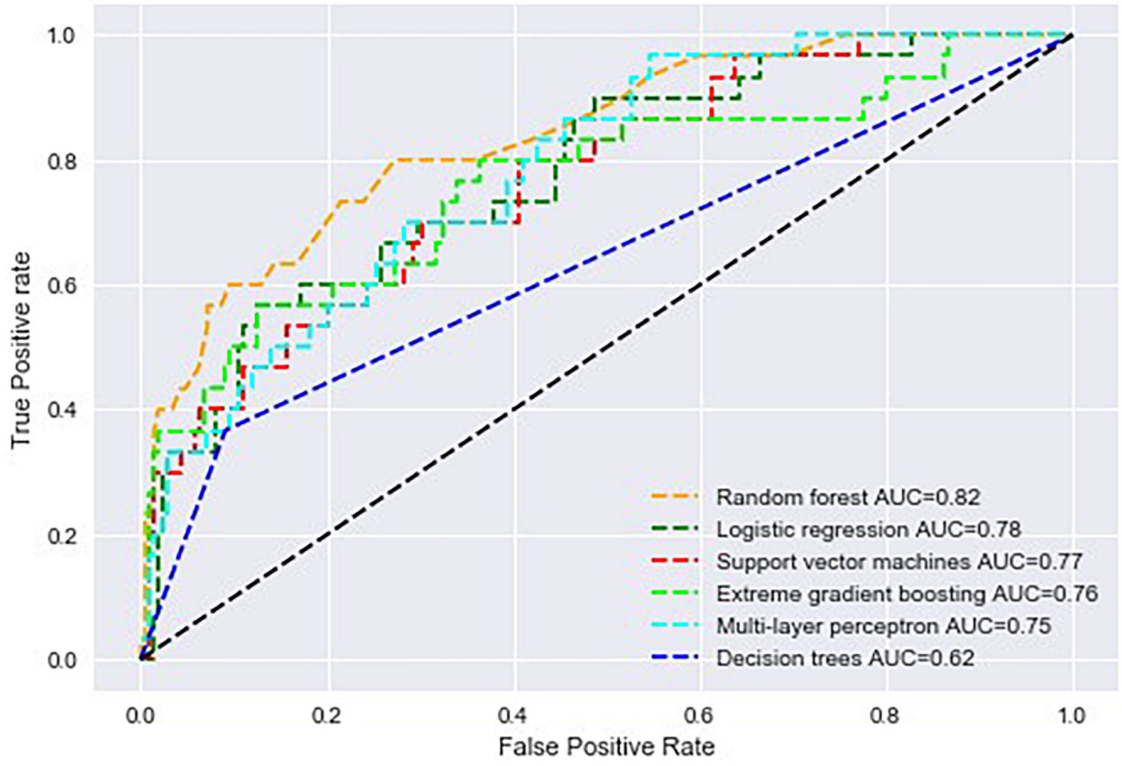


FIGURE 2
Area under the receiver operating characteristic curve of all machine learning algorithms used to build predictive models in the study.

between 1 and 150, and the activation function was chosen as the rectified linear unit (ReLU). The learning rate was set at 0.01. The hyperparameters resulting in the best performing networks were as follows: a multi-layer network with 124 nodes using the ReLU activation function, batch size of 60 and 20 epochs. Shapley additive explanations (SHAP) values were used to identify features influencing the construction of the model.

Six supervised machine learning algorithms were trained to predict all-cause in-hospital mortality in heart failure patients with a LVEF less than 50% using 123 features. The performance metrics of each algorithm were summarised using the confusion

TABLE 2 Performance of supervised machine learning algorithms in all-cause mortality prediction on the PMRCardio dataset.

Algorithm	Accuracy	Precision	Recall	f1-score
Random forest				
Training: Survived	1.00	1.00	1.00	1.00
Mortality		1.00	1.00	1.00
Test: Survived	0.88	0.89	1.00	0.94
Mortality		0.75	0.10	0.18
Weighted training score		1.00	1.00	1.00
Weighted test score		0.87	0.88	0.84
Logistic regression (ML)				
Training: Survived	0.78	0.96	0.78	0.86
Mortality		0.39	0.83	0.53
Test: Survived	0.76	0.94	0.78	0.85
Mortality		0.29	0.63	0.40
Weighted training score		0.88	0.78	0.81
Weighted test score		0.86	0.76	0.79
Support vector machines				
Training: Survived	0.91	0.91	0.99	0.95
Mortality		0.92	0.44	0.60
Test: Survived	0.86	0.92	0.92	0.92
Mortality		0.45	0.43	0.44
Weighted training score		0.91	0.91	0.90
Weighted test score		0.86	0.86	0.86
Extreme gradient boosting				
Training: Survived	0.85	0.94	0.89	0.91
Mortality		0.50	0.67	0.57
Test: Survived	0.82	0.92	0.87	0.90
Mortality		0.36	0.50	0.42
Weighted training score		0.88	0.85	0.86
Weighted test score		0.85	0.82	0.84
Decision trees				
Training: Survived	0.86	0.99	0.84	0.91
Mortality		0.50	0.94	0.66
Test: Survived	0.78	0.95	0.79	0.86
Mortality		0.32	0.70	0.44
Weighted training score		0.92	0.86	0.87
Weighted test score		0.87	0.78	0.81

ML, machine learning. Performance metrics for the multilayer perceptron (MLP) model not included as the metrics reported in this table are not routinely generated with MLP model.

matrix to calculate the accuracy, recall, precision, and F1 score, except for the MLP, where only the accuracy was reported since the rest of the performance metrics are not routinely available after training MLP.

Features identified by the algorithms were ranked, weighed and compared with predictors identified by logistic regression using the conventional statistical method. An area under the receiver operating characteristic (ROC) curve was used to evaluate the ability of each model to discriminate between the negative (*alive = 0*) and positive (*dead = 1*) classes.

2.4. Biostatistics

Categorical data are reported as absolute numbers and percentages, while the mean with standard deviation (±) and median with 25–75th percentile interquartile ranges were used to summarise numerical data with a normal and non-normal distribution, respectively. Normality was assessed visually with histograms and quantitatively with the Shapiro–Wilk and skewness tests. The Pearson chi-square and unpaired *t*-tests were used to compare categorical and normally distributed numerical data, respectively. In contrast, the Wilcoxon signed-rank test was used to compare numerical data with a non-normal distribution between patients that survived and died. A *p*-value less than 0.05 represented a statistically significant difference in the distribution of the features between survivors and non-survivors.

3. Results

3.1. Demographic and clinical findings

Between January 2009 and December 2018, there were 4,730 acute and chronic heart failure-related hospitalisations. After excluding patient records with missing LVEF and data that could not be verified using original hand-written patient records and biochemical data captured in the NHLS website, a total of 500 heart failure patients with a LVEF < 50% were included in the descriptive cohort analysis (**Figure 1**). The mean age was 55.2 ± 16.8 years, and 271 (54.2%) of these patients were males. The mean LVEF was 29 ± 9.2%. In this cohort, there were 300 (60.0%), 124 (24.8%), 54 (10.8%), and 22 (4.4%) Black, White, Indian, and Mixed ancestry patients, respectively. The rest of the baseline clinical characteristics of the study cohort are depicted in **Table 1**.

There were 159 (31.8%) patients with hypertension and 268 (53.6%) patients with dilated cardiomyopathy. A total of 209 (41.8%) patients had hepatomegaly, 110 (22.0%) had ascites, 350 (55.0%) had mitral regurgitation, 340 (68.0%) had bipedal oedema, and 395 (79.0%) had an elevated jugular venous pressure on clinical examination. The mean resting pulse rate

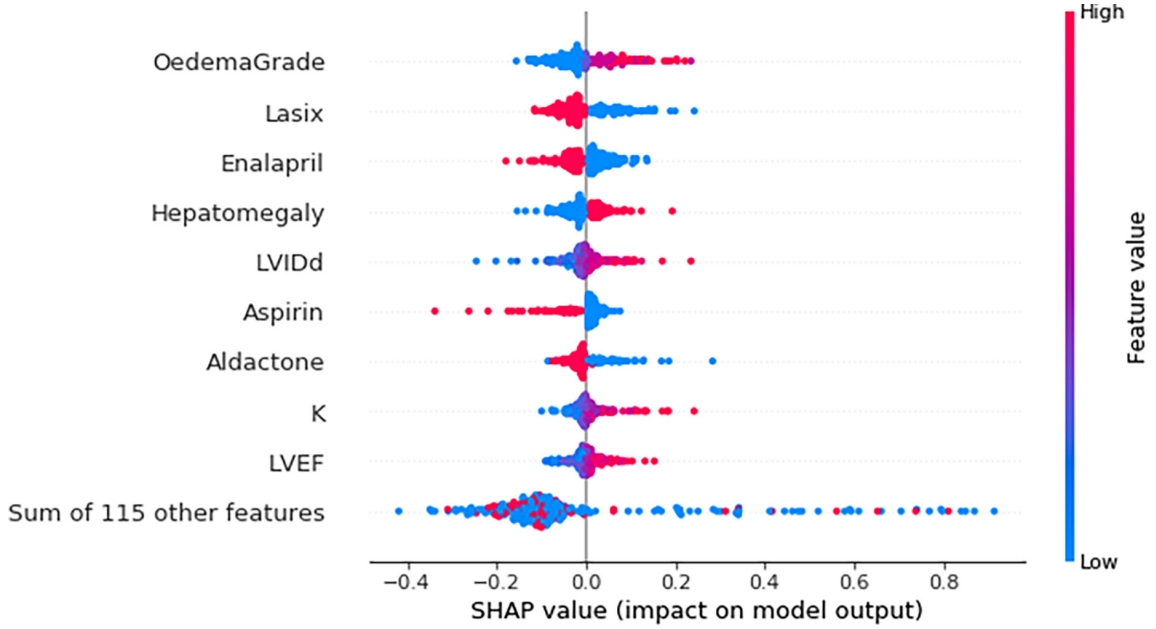


FIGURE 3
Beeswarm plot showing the impact of top nine feature on the multilayer perceptron (MLP) model output based on Shapley additive explanations (SHAP) values. K, potassium; LVEF, left ventricular ejection fraction; LVIDd, left ventricular internal diameter in diastole. Each dot represents a single patient. Red and blue refer to a high and low feature value, respectively. For example, both high and low potassium levels were associated with the risk of death. However, higher potassium levels (red) were strongly associated with the risk of mortality.

was 102 ± 23 beats per minute. Upon discharge from the hospital, there were 299 (59.8%) patients on oral beta-blocker therapy, and angiotensin-converting enzyme (ACE) inhibitors, furosemide and spironolactone were taken by 201 (40.2%), 300 (60.0%), 362 (72.4%) patients, respectively.

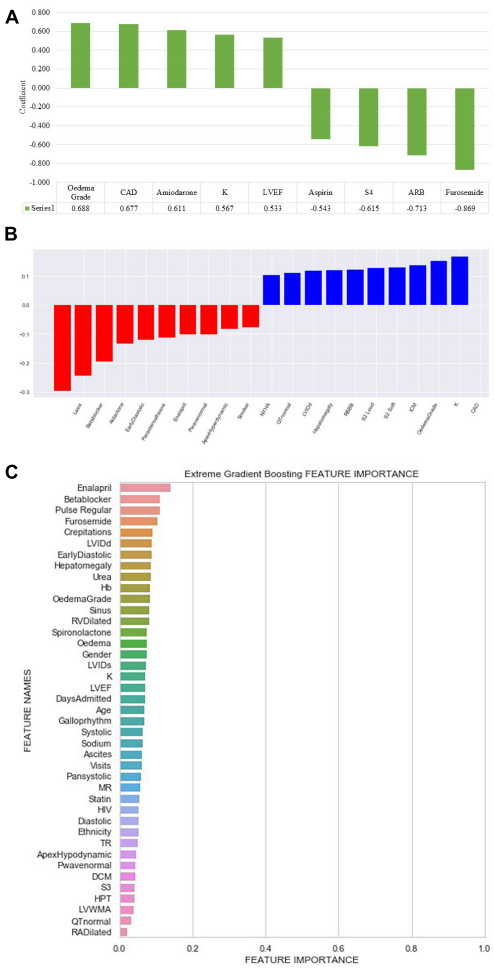


FIGURE 4
Feature importance and weight identified by machine learning algorithms. **(A)** Logistic regression (ML): ARB, angiotensin receptor blocker; CAD, coronary artery disease; K, potassium; LVEF, left ventricular ejection fraction; S4, fourth heart sound. **(B)** Support vector machine: CAD, coronary artery disease; ICM, ischaemic cardiomyopathy; K, potassium; LVIDd, left ventricular internal diameter in diastole; NYHA, New York Heart Association; RBBB, right bundle branch block; S2, second heart sound; Early diastolic, early diastolic murmur. **(C)** Extreme gradient boosted trees (XGB) classifier ranked features based on how many times a feature is used to split the data across all trees. Crepitations, lung crepitations on chest auscultation; Days admitted, duration of hospitalisation; DCM, dilated cardiomyopathy; Diastolic, diastolic blood pressure; Early diastolic, early diastolic murmur; Hb, haemoglobin; HIV, human immunodeficiency virus; HPT, hypertension; LVEF, left ventricular ejection fraction; LVIDd, left ventricular internal diameter in diastole; LVIDs, left ventricular internal diameter in systole; LVWMA, left ventricular wall motion abnormality; MR, mitral regurgitation; Pan systolic, pan systolic murmur; RA, right atrium; RV, right ventricle; S3, third heart sound; Sinus, sinus rhythm; Systolic, systolic blood pressure; TR, tricuspid regurgitation; visits, number of hospitalisations.

3.2. Outcomes

In the selected 500 patients hospitalised with heart failure, 84 (16.8%) patients died before discharge from the hospital. The median duration of hospitalisation was 7 days (interquartile range: 4–11), and it did not differ between patients discharged alive and those who died. Among the 416 patients discharged alive, 29 (7.0%) patients were readmitted within 30 days, 42 (10.1%) beyond 30 days, and 340 were not readmitted.

3.3. Performance metrics

Multilayer perceptron had an accuracy of 87% during training and testing. The AUROC curve was 0.82, 0.78, 0.77, 0.76, 0.75, and 0.62 for random forest, logistic regression (ML), SVM, XGBoost, MLP, and decision trees, respectively (Figure 2). Support vector machines achieved an accuracy of 86% during the test phase, followed by XGBoost, decision trees and logistic regression at 82, 78, and 76%, respectively (Table 2). The recall rate was above 50% in all algorithms, except for SVM, with a recall rate of 44 and 43%, respectively, during training and testing using data of patients that died while hospitalised. Logistic regression had the lowest precision of 39 and 29% during training and testing.

3.4. Feature ranking

3.4.1. Multilayer perceptron

The top nine features identified by the MLP model were peripheral oedema grade, furosemide, enalapril, hepatomegaly, LVIDd, aspirin, spironolactone, serum potassium, and the LVEF (Figure 3).

3.4.2. Logistic regression (ML)

Peripheral oedema grade had a coefficient of 0.688. Furosemide, angiotensin receptor blockers, a fourth heart sound on auscultation and aspirin, negatively correlated with all-cause mortality (Figure 4).

3.4.3. Support vector machines

The following top five features had a positive coefficient with the target feature: furosemide, beta-blockers, spironolactone, early diastolic murmur, and a parasternal heave, whereas coronary artery disease, potassium, oedema grade, ischaemic cardiomyopathy, and right bundle branch block on electrocardiogram had negative coefficients. The rest of the features are depicted in Figure 4.

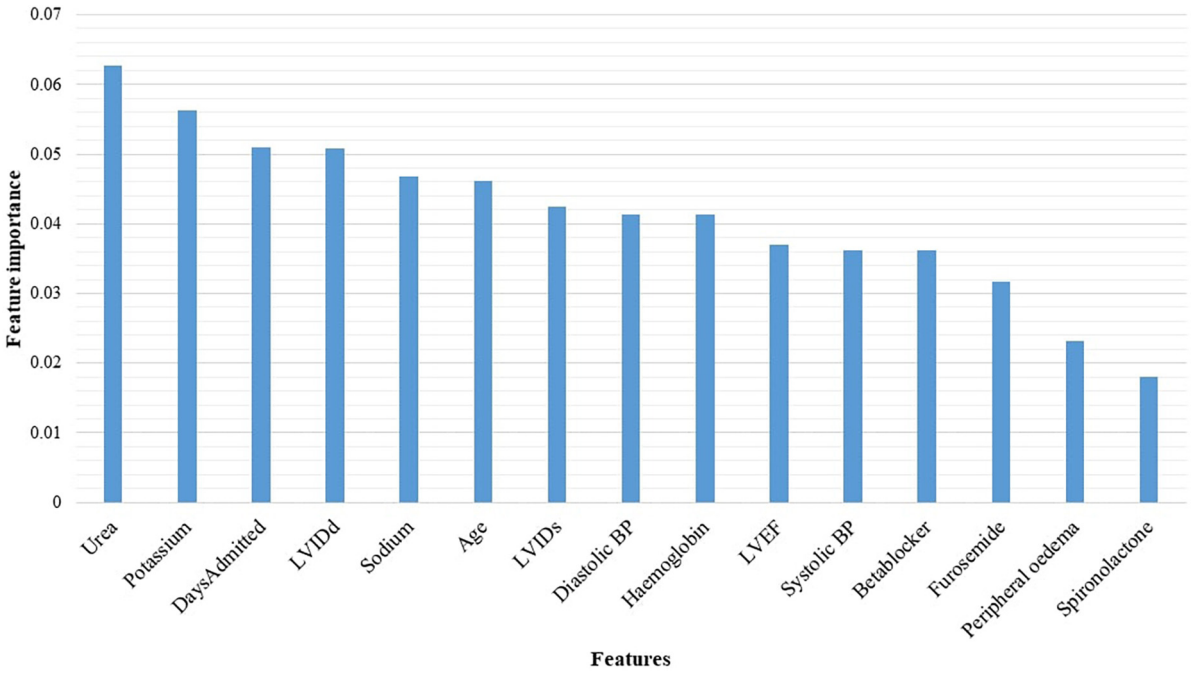


FIGURE 5
Features identified by the random forest algorithm. BP, blood pressure; LVEF, left ventricular ejection fraction; LVIDd, left ventricular internal diameter in diastole; LVIDs, left ventricular internal diameter in systole.

3.4.4. Extreme gradient boosting

Enalapril was the most common feature used to split decision trees and was used 20 times, followed by oral beta-blocker therapy, elevated jugular venous pressure, furosemide, and CAD (**Figure 4**).

3.4.5. Decision trees

Decision trees identified beta blockers as the root node (Gini = 0.5). In addition, the following features were placed as leaf nodes: systolic and diastolic blood pressure, urea, duration, the number of hospitalisations, and CAD.

3.4.6. Random forest

Urea, potassium, duration of admission, LVIDd, serum sodium levels, age, LVIDs, diastolic blood pressure, haemoglobin, and the LVEF, were identified as the top 10 features discriminating the two classes (Alive vs. Dead) (**Figure 5**).

3.4.7. The conventional statistical logistic regression model

In the multivariate logistic regression model, the following variables were independent predictors of all-cause mortality: diastolic blood pressure, coronary artery disease, enalapril, spironolactone, furosemide, sodium, and potassium (**Table 3**).

4. Discussion

This study used data from 500 patients with a LVEF less than 50% on the index or baseline hospitalisation to train six supervised machine learning algorithms to predict all-cause mortality. Although the random forest algorithm achieved a higher AUROC curve, the striking difference in training and testing performance metrics suggests that the model is overfitting data. Support vector machines had an AUC of 0.77, and an accuracy of 91 and 86% during training and testing, making it an ideal model for validation. The most plausible explanation for the SVM showing a desirable performance lies in the ability of the algorithm to process and model complex and non-linear data by creating a decision boundary that separates classes.

The all-cause mortality and 30-day rehospitalisation rate in 500 patients admitted with heart failure were 16.8 and 7.0%, respectively, while the median duration of hospitalisation was 7 (interquartile range: 4–11) days. In another tertiary-level hospital in South Africa, the in-hospital mortality rate among patients with acute heart failure was 8.4% and the mean duration of hospitalisation was 9 ± 12 days (1). Similarly, in a multicentre study conducted across Africa involving 169 heart failure patients recruited in South Africa, the 1-year all-cause mortality was 11.8% (2). In a meta-analysis involving 67,255 heart failure patients hospitalised in Australia, the pooled 30-day all-cause mortality was 8% (3). Similarly, in

TABLE 3 Conventional statistical multivariate logistic regression model predicting all-cause mortality.

	Univariate regression			Multivariate regression		
	Coefficient	*P*-value	CI	Coefficient	*P*-value	CI
Female	0.4330	0.043	0.013–0.853			
Black race	0.5169	0.028	0.055–0.978			
Gallop rhythm	0.5399	0.045	0.012–1.068			
ED murmur	−0.5496	0.012	−0.977 to 0.122			
Diastolic BP	−0.0356	0.000	−0.049 to −0.022	−0.0239	0.002	−0.039 to −0.009
CAD	0.7432	0.022	0.105–1.381	0.9490	0.033	0.078–1.820
Spironolactone	−1.0043	0.000	−1.412 to 0.596	−0.6674	0.012	−1.190 to −0.145
Enalapril	−1.3472	0.000	−1.845 to 0.849	−0.6440	0.044	−1.270 to −0.018
Aspirin	−0.6101	0.037	−1.183 to −0.037			
Furosemide	−1.5959	0.000	−2.022 to −1.170	−1.5685	0.000	−2.142 to −0.995
Statins	−0.6847	0.012	−1.216 to −0.153			
Oedema grade	0.2503	0.002	0.089 to 0.411			
Sodium	−0.0782	0.000	−0.115 to −0.041	−0.0288	0.001	−0.045 to −0.012
Potassium	0.6609	0.000	0.398–0.924	0.5366	0.001	0.230–0.843
Urea	0.0095	0.032	0.001–0.018			
RBBB	0.9321	0.007	0.259–1.606			
LVIDd	0.3835	0.000	0.175–0.592			
LVIDs	0.3337	0.002	0.122–0.545			
Length of stay (days)	0.0257	0.012	0.006–0.046			

CAD, coronary artery disease; BP, blood pressure; CI, confidence interval (0.025–0.975); ED, end diastolic; RBBB, right bundle branch block; LVIDd, left ventricular internal diameter in diastole; LVIDs, left ventricular internal diameter in systole.

another systematic review of research studies conducted in low-and-middle-income countries (LMIC), the in-hospital all-cause mortality was 8%, with a confidence interval of 6–10% (4). The most plausible explanation for our cohort's higher in-hospital all-cause mortality is the delay in presentation to the hospital. As a tertiary academic centre, most patients referred to our centre are initially managed in primary and secondary level institutions and eventually referred to our centre in advanced stages of heart failure.

Despite the variation in the aetiology of heart failure between patients residing in LMIC and HIC, and limited access to life-prolonging intracardiac device therapy, particularly in patients attending state-owned treatment facilities, we found similar predictors of all-cause mortality using machine learning algorithms and a conventional statistical logistic regression model. After training the SVM, furosemide, beta-blockers, spironolactone, early diastolic murmur and a parasternal heave, yielded positive coefficients, whereas coronary artery disease, potassium, oedema grade, ischaemic cardiomyopathy, and right bundle branch block on electrocardiogram had negative coefficients. The support vector machines and the machine learning logistic regression algorithm identified similar predictors as the MLP model. Some of the predictors of all-cause mortality in heart failure reported in the literature include age, sex, diastolic blood pressure, LVEF, serum sodium and creatinine, estimated glomerular filtration rate, haemoglobin, haematocrit, and blood urea nitrogen (5–7). Predictors of mortality specifically in patients with a LVEF above 50% include blood urea nitrogen and the body mass index (6).

Healthcare databases are notoriously famous for housing imbalanced datasets. Class imbalance refers to the disproportionality between the data classes used to train the predictive model (8), a common problem that is not unique to medical data. For example, when the training data with the negative outcome (*dead*) has significantly fewer observations than the majority class, the classification algorithm favours the majority class, by focusing the learning process on the majority class. Our dataset had fewer entries attributed to the negative outcome and was imbalanced. This accounts for the suboptimal performance shown by the algorithms when learning and making predictions specifically for the minority class. However, instead of under-sampling the majority class or generating new synthetic values to increase the number of samples in the minority class, we used class weights to excessively penalise

false negative and false positive results from the minority class (*dead*).

Six supervised machine learning algorithms were successfully trained to predict mortality despite an imbalanced and relatively small dataset. The success implies that data quality is also essential to consider before building predictive models. The project's next phase is to increase the sample size by collecting data from multiple cardiac centres around sub-Saharan Africa. Adding new data that was not previously used to train or test the algorithms tends to improve the models' performance metrics. Once the performance of the classifier has been validated, a risk calculator will be developed and implemented for use in Africa.

The performance of machine learning algorithms is commonly attributed to the size of the dataset. Indeed, the higher the number of observations the better the performance metrics. Some authors have attempted to derive an equation for calculating a suitable sample size for machine learning algorithms by calculating the error rate of prediction (9, 10). In our opinion, the performance of machine learning algorithms is also influenced by the number of features (clinical parameters) used, the ratio of observations within the classes, and the credibility of the data. In our study, 500 patients with 123 features were used to train algorithms. Consider a model created with over a million patients, but with only five features. Its performance might be suboptimal. An ideal sample size calculation should incorporate the number of observations and features and a measure of data credibility. In general, the error rate should decrease as the size of the training dataset increases. The AUROC curve for the machine learning models can be improved by increasing the size of the training dataset. Other techniques for improving the AUC such as feature normalization and scaling, setting class weights, and grid search were applied in this study.

Machine learning algorithms cannot process data with empty fields. As such innovative ways of handling missing data have been introduced. These include deleting incomplete records, filling incomplete fields with the mean, median or mode, and predicting missing values using machine learning algorithms and applying the last observation carried forward method (11). In this study, we removed clinical parameters with more than 20% missing values and imputed missing values using the mean or median value of each feature. The rationale for using the mean is that overall the mean value (Gaussian distribution) of each feature does not change. However, this method does not factor for the covariance between features.

This was a single-centre study conducted in a tertiary centre since the CMJAH division of cardiology is the only department equipped with an electronic health record system. As such, the findings from this study may not be generalisable to all heart failure patients treated in SSA. In addition, most pertinent clinical parameters such as comorbidities, causes of heart failure, angiographic, and ECG findings were available in the database but underreported, as some of the data were captured in the free text format and could not be extracted for analysis and incorporation into the predictive models. Also, we excluded a significant amount of patient data due to the unavailability of handwritten patient records or objective clinical evidence supporting the heart failure diagnosis. This was a major limitation, since handwritten patient records are routinely removed from the CMJAH after 5 years from the time of initial admission, due to the unavailability of storage facilities. Furthermore, pertinent clinical data, such as the presence of cardiogenic shock, and the administration of ionotropes and vasopressors, were documented in the EHR system in free-text format and could not be extracted. Nevertheless, the study provides proof of concept for developing a heart failure model derived from LMIC predicting all-cause mortality.

5. Conclusion

Despite a small sample size, machine learning algorithms successfully predicted all-cause mortality with modest accuracy. The SVM model will be externally validated using data from multiple cardiology centres in South Africa before developing a uniquely African risk prediction tool.

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 Human Research Ethics Committee (Clearance certificate number: M190515). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

DM: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualization, roles/writing—original draft, and writing—review and editing. TC: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, roles/writing—original draft, and writing—review and editing. EK: methodology, validation, roles/writing—original draft,

and writing—review and editing. HN: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, visualization, roles/writing—original draft, and writing—review and editing. All authors contributed to the article and approved the submitted version.

Funding

DM was a full-time Ph.D. student funded by the Professor Bongani Mayosi Netcare Clinical Scholarship, the Discovery Academic Fellowship [Grant No. 03902], the Carnegie Corporation of New York [Grant No. b8749], the South African Heart Association, and the University of the Witwatersrand Chancellor's Female Academic Leaders Fellowship (NPC) Grant.

Acknowledgments

The authors would like to thank Professor Pravin Manga for implementing the PMRCardio database, the staff members in the Division of Cardiology at the CMJAH and Drs. Marheb Badianyama, Jenna Damelin, Ashlee Firer, Thakgalo Thibeng, Yemurai Muchenge, Michaela Gough, and Mpumelelo Ngadla who assisted with verifying patient data before the analysis, and Mr. Karl Basson, Geethan Singh, and Ntuthuko Mngomezulu for assisting with the technical data queries.

Conflict of interest

EK has received speaker honorarium, travel assistance from Astra Zeneca, Servier, Novartis, and has been the head of Ethics and Guideline Committee of the South African Heart Association.

The remaining 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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2022.1032524/full#supplementary-material

References

1. Szymanski PZ, Badri M, Mayosi BM. Clinical characteristics and causes of heart failure, adherence to treatment guidelines, and mortality of patients with acute heart failure: experience at Groote Schuur hospital, Cape Town, South Africa. *S Afr Med J.* (2018) 108:94–8. doi: 10.7196/SAMJ.2017.v108i2.12519

2. Karaye KM, Dokainish H, ElSayed A, Mondo C, Damasceno A, Sliwa K, et al. Clinical profiles and outcomes of heart failure in five African countries: results from INTER-CHF study. *Glob Heart.* (2021) 16:50. doi: 10.5334/gh.940

3. Al-Omary MS, Davies AJ, Evans TJ, Bastian B, Fletcher PJ, Attia J, et al. Mortality and readmission following hospitalisation for heart failure in Australia: a systematic review and meta-analysis. *Heart Lung Circ.* (2018) 27:917–27. doi: 10.1016/j.hlc.2018.01.009

4. Callender T, Woodward M, Roth G, Farzadfar F, Lemarie JC, Gicquel S, et al. failure care in low- and middle-income countries: a systematic review and meta-analysis. *PLoS Med.* (2014) 11:e1001699. doi: 10.1371/journal.pmed.1001699

5. Adler ED, Voors AA, Klein L, Macheret F, Braun OO, Urey MA, et al. Improving risk prediction in heart failure using machine learning. *Eur J Heart Fail.* (2020) 22:139–47.

6. Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, et al. Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. *JACC Heart Fail.* (2020) 8:12–21.

7. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inf Decis Mak.* (2020) 20:16. doi: 10.1186/s12911-020-1023-5

8. Gillala R, Amit Kumar T, Reddy VK, Wide Scale A. Classification of class imbalance problem and its solutions: a systematic literature review. *J Comput Sci.* (2019) 15:886–929.

9. Balki I, Amirabadi A, Levman J, Martel AL, Emersic Z, Meden B, et al. Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. *Can Assoc Radiol J.* (2019) 70:344–53. doi: 10.1016/j.carj.2019.06.002

10. Dobbin KK, Zhao Y, Simon RM. How large a training set is needed to develop a classifier for microarray data? *Clin Cancer Res.* (2008) 14:108–14. doi: 10.1158/1078-0432.CCR-07-0443

11. Kumar, S. *7 Ways to Handle Missing Values in Machine Learning.* (2020). Available online at: https://www.towardsdatascience.com (accessed July 1, 2021).

CHAPTER 6 – DISCUSSION AND CONCLUSIONS

This chapter summarises the main study findings, limitations, future directions, and recommendations. Finally, the last section narrates the conclusions drawn from this research.

6.1 Discussion

Heart failure has been declared a global pandemic due to the high mortality rate, particularly in low-and-middle-income countries (LMICs). Globally, heart failure has a five-year mortality between 25-50% (Dokainish et al., 2017a). The high mortality rate is observed despite the continuous introduction of novel pharmacological and intracardiac device therapies. Unfortunately, life-prolonging heart transplant services are still not widely available in South Africa. As we strive to introduce treatment that reduces morbidity and mortality associated with heart failure, innovative techniques such as risk prediction may assist in managing heart failure, facilitating better allocation of disease-modifying therapy to high-risk individuals.

As part of this thesis, a systematic literature review was conducted, and 30 original research studies predicting mortality and/or rehospitalisation using machine learning were identified (Mpanya et al., 2021). The sample size ranged between 71 and 716 790. Twelve models predicted mortality, 13 rehospitalisations, and five predicted both. However, none of the models were created or validated using data from sub-Saharan Africa or regions where heart failure is predominantly caused by non-ischaemic heart diseases.

The baseline demographic and clinical characteristics of 636 consecutive patients hospitalised with new-onset heart failure symptoms or decompensated chronic heart failure were described, and heart failure phenotypes were identified using the left ventricular ejection fraction (LVEF) and unsupervised machine algorithms. The mean age in our cohort was 55.6 ± 16.9 years, and 16.8% of all patients hospitalised with the diagnosis of heart failure died over the study period. Heart failure is generally associated with poorer outcomes in LMICs, and the onset of heart failure occurs at a younger age of 53 years, affecting economic productivity in this region (Dokainish et al., 2016, Ajayi et al., 2020). The most plausible explanations for poorer

outcomes include, but are not limited to the unavailability of life-prolonging therapy such as cardiac resynchronisation defibrillators and heart transplant services. Furthermore, there might be a delay in the referral of patients from primary and secondary level hospitals to our centre, causing our patients to be at advanced stages of the disease by the time they are eventually treated at our centre.

The younger age of onset of heart failure in sub-Saharan Africa (SSA) may be partly related to the dominance of non-ischaemic cardiomyopathy, possibly with a genetic aetiology, considering that the most common cardiovascular diseases in SSA are still hypertensive heart disease, cardiomyopathies, rheumatic heart disease and congenital heart disease (Ajayi et al., 2020, Dokainish et al., 2016), unlike in high-income countries, where ischaemic heart disease is the primary cause of heart failure. Also, perhaps the limited access to life-prolonging pharmacological and non-pharmacological may speed up the progression to heart failure, where patients eventually present with a LVEF < 50% earlier than expected.

Clustering algorithms have the potential to unravel distinct phenotypes of heart failure that may assist clinicians in triaging or risk profiling patients, hence influencing the management of heart failure patients. Our study is unique since it offers insights into the role of machine learning in better understanding heart failure phenotypes in our region. After subjecting the heart failure cohort to unsupervised machine learning algorithms, the agglomerative clustering algorithm identified three clusters, the *uraemic cluster*, with relatively higher urea levels, the *hypotensive cluster,* with lower systolic and diastolic blood pressure and the *congested cluster,* with a high burden of heart failure symptoms and signs of fluid overload.

The systolic blood pressure had a 2.6 times higher impact when differentiating between clusters, followed by the age at the time of admission with a mean SHAP value of 0.16, and the LVEF had a lower SHAP value of 0.14. These findings suggest that blood pressure and age may play a role in the diagnosis, risk stratification, and management of heart failure patients in centres without access to echocardiography. Most studies have confirmed the utility of supervised machine learning algorithms for classification tasks (Kwon et al., 2019, Segar et al., 2021, Su et al., 2017, Ahmad et al., 2018, Shameer et al., 2017, Golas et al., 2018, Miao et al., 2018, Meng et al., 2019, Liu et al., 2020). However, only a few studies have since been published on phenotyping heart failure patients using unsupervised machine learning

algorithms (Ahmad et al., 2014, Ieva et al., 2017, Moser et al., 2014, Bose and Radhakrishnan, 2018, Segar et al., 2020, Nouraei et al., 2022, Perry et al., 2021).

The primary aim of the study was to use machine learning algorithms to predict in-hospital all-cause mortality in heart failure patients. Six supervised machine learning algorithms were trained to predict all-cause mortality and identify predictors. The best-performing algorithm was the random forest algorithm, based on its ability to discriminate between patients that lived and died, with an area under the receiver operating characteristic curve (AUC) of 0.82. This is unsurprising because random forests perform well when exposed to a robust dataset, unlike individual decision trees, which showed the lowest AUC of 0.62, and an accuracy of 86% and 78% during training and testing, respectively, suggesting that they were overfitting the data. Decision trees performed poorly, likely because of the dataset's complexity with many features. Support vector machines (SVM) had the highest accuracy of 91% and 86% during training and testing, respectively. Perhaps the main reason for the impressive performance lies in the ability of the SVM to handle non-linearly separable data using a kernel.

Features or predictors identified by the six algorithms included coronary artery disease, ischaemic cardiomyopathy, peripheral oedema grade, elevated jugular venous pressure, hepatomegaly, early diastolic murmur, parasternal heave, a fourth heart sound on auscultation, beta-blockers, furosemide, enalapril, spironolactone, angiotensin receptor blockers, aspirin, serum potassium, urea, right bundle branch block, systolic and diastolic blood pressure, left ventricular ejection fraction, left ventricular internal diameter in diastole, duration and number of hospitalisations. There was much overlap between predictors identified by algorithms. Also, the predictors identified by algorithms were similar to those identified using a traditional statistical logistic regression model.

The advantage of using machine learning over statistical methods is related to the ability of machine learning algorithms to increase the accuracy of predictions and discriminatory ability between classes when exposed to new data. Also, machine learning employs many different algorithms for classification and discrimination between classes, unlike statistical methods where the choice may be between logistic regression and Cox proportional hazards regression to identify outcome predictors. Furthermore, machine learning algorithms can manipulate the dimensionality of the data into various forms, ensuring that similar data points are grouped in

the same class. Due to the small number of patients rehospitalised within 30 days, it was impossible to predict all-cause rehospitalisation. During training, algorithms had 30 samples to learn from, hence the poor performance. During training, the algorithms only had 21 observations to learn from before making predictions, and only 13 were available for testing algorithms. As such, the models created by the algorithms were overfitting and biased toward learning from the cohort of patients that were admitted beyond 30 days and those that were not admitted. As such, predictors of 30-day all-cause rehospitalisation were not extracted. Another objective of this research was to validate the performance of the selected machine learning model using new data, ideally from a different region and subsequent follow-up of these patients for one year. This will be the project's next phase, requiring collaboration with multiple hospitals equipped with basic cardiology services such as echocardiography.

6.2 Limitations

This was a single-centre study conducted in an academic centre, the Division of Cardiology at the Charlotte Maxeke Johannesburg Academic Hospital, one of the few centres equipped with an electronic health record system. As such, the findings from this study may not be generalisable to all heart failure patients treated in sub-Saharan Africa. In addition, most pertinent clinical parameters such as comorbidities, causes of heart failure, angiographic data, and electrocardiogram findings were available in the database but underreported. Some data was captured in the free-text format and could not be extracted for analysis and incorporation into the predictive models. Also, we excluded a significant amount of patient data due to the unavailability of patient records or objective clinical evidence supporting the heart failure diagnosis. Furthermore, details on managing heart failure patients, such as the presence of cardiogenic shock and administration of inotropes and vasopressors, were unavailable. Nevertheless, the study provides proof of concept for developing a uniquely African heart failure model predicting all-cause mortality.

6.3 Recommendations

For efficient implementation of machine learning tasks, clinicians should collaborate with data scientists and establish data archiving systems that follow the outline below in Figure 6.1.

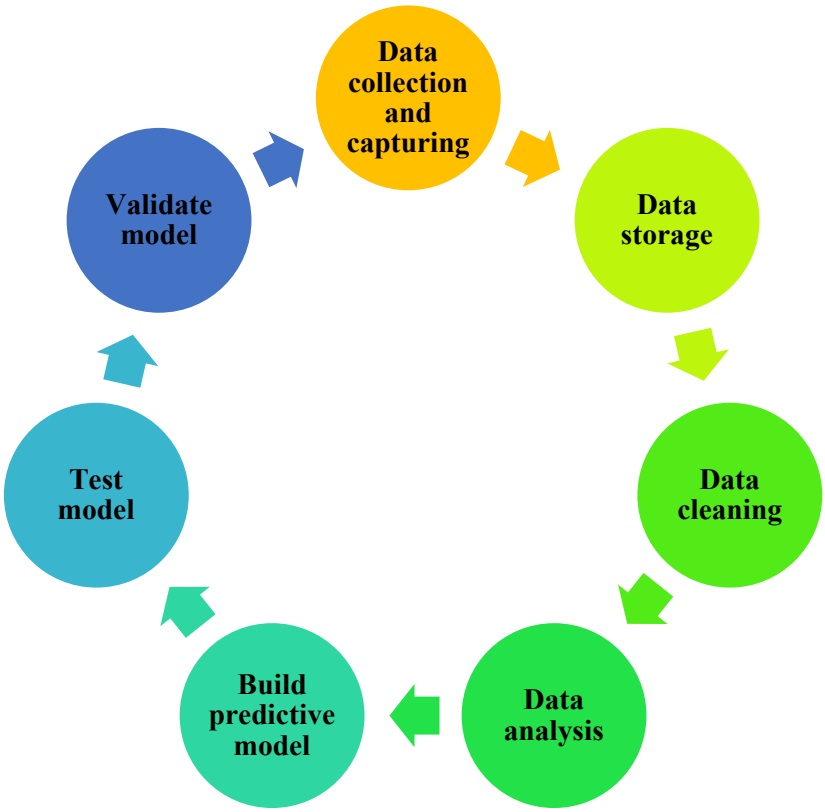


Figure 6.1 Schematic diagram showing the proposed data laboratory for an efficient machine learning predictive model, starting and ending with the data collection and capturing phase.

Data should be collected from multiple sites and deposited in a central location whenever possible. Each of the sites or hubs should establish databases and build predictive machine learning models using the following recommendations:

6.3.1 Data collection and capturing

a) Data should be collected by data capturers with a medical background. These could be clinical associates, medical officers, etc.
b) Data collection systems should be user-friendly and efficient. The proposed design or outlook should be similar to a patient's clinical file.
c) A quality control officer should evaluate the quality of the data captured and place a stamp of approval for each data entry.
d) There should be no empty fields on the data set.
e) The data should be accessible for both clinical and research purposes.
f) One should be able to query the database using simple code such as "filter by heart failure AND age<40 AND betablocker=yes."

6.3.2 Data storage

a) The data storage facility should be cloud-based
b) It should function in the absence of electricity (off the grid)
c) It should continuously backup data in a different platform
d) It should store binary data and raw imaging data in pixel form

6.3.3 Data cleaning

a) Create a box plot and detect outliers when dealing with continuous features
b) When cleaning continuous features, verify all outliers
c) Ideally, data cleansing should occur daily or at least weekly
d) Delimiters should be used for all continuous features. This will assist in querying all data entries perceived as outliers.

6.3.4 Data analysis

a) Use both machine learning and statistical methods to analyse data
b) Use multiple algorithms, including deep learning, for data analysis
c) Data visualisation is critical and quantifiable parameters such as the skewness test, and

Shapiro-Wilk test should be used to evaluate normality in the data distribution.

d) Histograms of each continuous feature must be plotted to assess the Gaussian distribution.

6.3.5 Building a predictive model

a) Models should have the capacity to learn from normal reference values.
b) Continuously update the model after adding new data, at least bi-annually.
c) Train algorithms using at least 60 - 70% of the dataset.

6.3.6 Testing the model

a) Model performance should be tested by calculating the true positive and true negative rates.
b) An area under the receiver operating characteristic curve should be plotted for all models.
c) Accuracy should also be checked, along with the F-score.
d) A dataset that was not used to train the model should be used as test data.

6.3.7 Validating the model

a) The model should be validated using a dataset not used to train or test the model.
b) Outcomes should be measured and compared in patients managed using risk calculators and standard care.
c) Data collection should be part of the clinical workflow rather than just a tool for driving the research agenda.

6.4 Future Direction

The field of data science is soaring, with the introduction of newer techniques such as quantum machine learning, federated learning, and natural language processing (Esteva et al., 2019, Xu et al., 2021, Zeguendry et al., 2023). With the advancements in technology and the increasing amount of data available, machine learning algorithms will continue to improve and become more sophisticated. This will lead to more accurate predictions and better decision-making. Quantum machine learning is an emerging field that has the potential to revolutionise the way we process and analyse data. By leveraging the principles of quantum mechanics, quantum machine learning algorithms can solve complex problems that are beyond the capabilities of classical computers. This technology has already shown promising results in areas such as drug discovery, image recognition, and financial modeling (Zeguendry et al., 2023). Federated learning, which allows multiple parties to collaborate on machine learning without sharing their data, is one potential solution to the privacy concern (Xu et al., 2021).

Natural language processing has the ability to analyse and understand human text and speech (Esteva et al., 2019). As machine learning continues to evolve, its impact will be felt across a wide range of industries. However, as with any new technology, it is important to consider the ethical implications of its use and ensure that it is being developed in a responsible manner.

6.5 Conclusions

I predicted all-cause mortality in hospitalised heart failure patients using six machine learning algorithms in this research. The random forest algorithm had better discriminatory abilities, shown by the area under the curve of 0.82 and a higher accuracy score of 88% during the test phase. Seven unsupervised machine learning algorithms were used to cluster heart failure patients, and three clusters were identified, suggesting that heart failure patients could be phenotyped beyond the LVEF. Overall, heart failure patients in our cohort were younger than their counterparts in high-income countries, and 64.3% had a LVEF less than 40%. The performance of machine learning algorithms was influenced by factors such as the sample size, number of features, and data credibility. In contrast to traditional statistical methods, machine learning algorithms identified more predictors of mortality. Due to a small number of patients rehospitalised within 30 days, creating a model predicting rehospitalisation was not feasible. The next phase of this project is to increase the size of the training set and validate model performance using new data collected from other sites. A risk calculator predicting the risk of all-cause mortality will then be created using an algorithm with the best performance metrics.

6.6 References

1. Ahmad, T., Lund, L. H., Rao, P., Ghosh, R., Warier, P., Vaccaro, B., et al. 2018. Machine Learning Methods Improve Prognostication, Identify Clinically Distinct Phenotypes, and Detect Heterogeneity in Response to Therapy in a Large Cohort of Heart Failure Patients. Journal of the American Heart Association, 7**,** e008081.
2. Ahmad, T., Pencina, M. J., Schulte, P. J., O'brien, E., Whellan, D. J., Pina, I. L., et al. 2014. Clinical implications of chronic heart failure phenotypes defined by cluster analysis. J Am Coll Cardiol, 64**,** 1765-74.

3. Ajayi, A. A., Sofowora, G. G. & Ladipo, G. O. 2020. Explaining Heart Failure Hyper-mortality in Sub Saharan Africa: Global Genomic and Environmental Contribution Review. Journal of the National Medical Association, 112**,** 141-157.
4. Bose, E. & Radhakrishnan, K. 2018. Using Unsupervised Machine Learning to Identify Subgroups Among Home Health Patients With Heart Failure Using Telehealth. Comput Inform Nurs, 36**,** 242-248.
5. Dokainish, H., Teo, K., Zhu, J., Roy, A., Alhabib, K. F., Elsayed, A., et al. 2017a. Global mortality variations in patients with heart failure: results from the International Congestive Heart Failure (INTER-CHF) prospective cohort study. Lancet Glob Health, 5**,** e665-e672.
6. Dokainish, H., Teo, K., Zhu, J., Roy, A., Alhabib, K. F., Elsayed, A., et al. 2017b. Global mortality variations in patients with heart failure: results from the International Congestive Heart Failure (INTER-CHF) prospective cohort study. The Lancet Global Health, 5**,** e665-e672.
7. Dokainish, H., Teo, K., Zhu, J., Roy, A., Alhabib, K. F., Elsayed, A., et al. 2016. Heart Failure in Africa, Asia, the Middle East and South America: The INTER-CHF study. Int J Cardiol, 204**,** 133-41.
8. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., Depristo, M., Chou, K., et al. 2019. A guide to deep learning in healthcare. Nat Med, 25**,** 24-29.
9. Gallagher, J., Mcdonald, K., Ledwidge, M. & Watson, C. J. 2018. Heart Failure in Sub-Saharan Africa. Card Fail Rev, 4**,** 21-24.
10. Golas, S. B., Shibahara, T., Agboola, S., Otaki, H., Sato, J., Nakae, T., et al. 2018. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC Med Inform Decis Mak, 18**,** 44.
11. Hilfiker-Kleiner, D., Landmesser, U. & Drexler, H. 2006. Molecular Mechanisms in Heart Failure: Focus on Cardiac Hypertrophy, Inflammation, Angiogenesis, and Apoptosis. JACC, 48**,** A56-A66.
12. Ieva, F., Paganoni, A. M. & Pietrabissa, T. 2017. Dynamic clustering of hazard functions: an application to disease progression in chronic heart failure. Health Care Manag Sci, 20**,** 353-364.

13. Kwon, J. M., Kim, K. H., Jeon, K. H. & Park, J. 2019. Deep learning for predicting in-hospital mortality among heart disease patients based on echocardiography. Echocardiography, 36**,** 213-218.
14. Lesyuk, W., Kriza, C. & Kolominsky-Rabas, P. 2018. Cost-of-illness studies in heart failure: a systematic review 2004-2016. BMC Cardiovasc Disord, 18**,** 74.
15. Levy, W. C., Mozaffarian, D., Linker, D. T., Sutradhar, S. C., Anker, S. D., Cropp, A. B., et al. 2006. The Seattle Heart Failure Model: prediction of survival in heart failure. Circulation, 113**,** 1424-33.
16. Liu, W., Stansbury, C., Singh, K., Ryan, A. M., Sukul, D., Mahmoudi, E., et al. 2020. Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding. PLoS ONE, 15.
17. Meng, F., Zhang, Z., Hou, X., Qian, Z., Wang, Y., Chen, Y., et al. 2019. Machine learning for prediction of sudden cardiac death in heart failure patients with low left ventricular ejection fraction: study protocol for a retroprospective multicentre registry in China. BMJ Open, 9**,** e023724.
18. Miao, F., Cai, Y. P., Zhang, Y. X., Fan, X. M. & Li, Y. 2018. Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest. IEEE Access, 6**,** 7244-7253.
19. Moser, D. K., Lee, K. S., Wu, J.-R., Mudd-Martin, G., Jaarsma, T., Huang, T.-Y., et al. 2014. Identification of symptom clusters among patients with heart failure: An international observational study. International Journal of Nursing Studies, 51**,** 1366-1372.
20. Mpanya, D., Celik, T., Klug, E. & Ntsinjana, H. 2020. Machine learning and statistical methods for predicting mortality in heart failure. Heart Fail Rev.
21. Mpanya, D., Celik, T., Klug, E. & Ntsinjana, H. 2021. Predicting mortality and hospitalisation in heart failure using machine learning: A systematic literature review. Int J Cardiol Heart Vasc, 34**,** 100773.
22. Nouraei, H., Nouraei, H. & Rabkin, S. W. 2022. Comparison of Unsupervised Machine Learning Approaches for Cluster Analysis to Define Subgroups of Heart Failure with Preserved Ejection Fraction with Different Outcomes. Bioengineering (Basel), 9.

23. Ntusi, N. B. & Mayosi, B. M. 2009. Epidemiology of heart failure in sub-Saharan Africa. Expert Rev Cardiovasc Ther, 7**,** 169-80.
24. Perry, A., Loh, F., Adamo, L., Zhang, K. W., Deych, E., Foraker, R., et al. 2021. Unsupervised cluster analysis of patients with recovered left ventricular ejection fraction identifies unique clinical phenotypes. PLoS One, 16**,** e0248317.
25. Segar, M. W., Jaeger, B. C., Patel, K. V., Nambi, V., Ndumele, C. E., Correa, A., et al. 2021. Development and Validation of Machine Learning-Based Race-Specific Models to Predict 10-Year Risk of Heart Failure: A Multi-Cohort Analysis. Circulation.
26. Segar, M. W., Patel, K. V., Ayers, C., Basit, M., Tang, W. H. W., Willett, D., et al. 2020. Phenomapping of patients with heart failure with preserved ejection fraction using machine learning-based unsupervised cluster analysis. European Journal of Heart Failure, 22**,** 148-158.
27. Shameer, K., Johnson, K. W., Yahi, A., Miotto, R., Li, L. I., Ricks, D., et al. 2017. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: A case-study using mount sinai heart failure cohort. Pac Symp Biocomput, 22**,** 276-287.
28. Su, Z., Brecht, T., O'donovan, F., Boussios, C., Menon, V., Gliklich, R., et al. 2017. Machine Learning Enhanced Predictions of Hospital Readmission or Death in Heart Failure. Circulation, 136.
29. Szymanski, P. B., M; Mayo, Bm 2018. Clinical characteristics and causes of heart failure, adherence to treatment guidelines, and mortality of patients with acute heart failure: Experience at Groote Schuur Hospital, Cape Town, South Africa. S Afr Med J, 108**,** 94-98.
30. Wang, G., Zhang, Z., Ayala, C., Wall, H. K. & Fang, J. 2010. Costs of heart failure-related hospitalisations in patients aged 18 to 64 years. Am J Manag Care, 16**,** 769-76.
31. Wehbe, R. M., Khan, S. S., Shah, S. J. & Ahmad, F. S. 2020. Predicting High-Risk Patients and High-Risk Outcomes in Heart Failure. Heart Fail Clin, 16**,** 387-407.
32. Xu, J., Glicksberg, B. S., Su, C., Walker, P., Bian, J. & Wang, F. 2021. Federated Learning for Healthcare Informatics. J Healthc Inform Res, 5**,** 1-19.

33. Zeguendry, A., Jarir, Z. & Quafafou, M. 2023. Quantum Machine Learning: A Review and Case Studies. Entropy (Basel), 25.

All References

1. Adler, E. D., Voors, A. A., Klein, L., Macheret, F., Braun, O. O., Urey, M. A., et al. 2020. Improving risk prediction in heart failure using machine learning. European Journal of Heart Failure, 22, 139-147.
2. Agbor, V. N., Essouma, M., Ntusi, N. a. B., Nyaga, U. F., Bigna, J. J. & Noubiap, J. J. 2018. Heart failure in sub-Saharan Africa: A contemporaneous systematic review and meta-analysis. Int J Cardiol, 257, 207-215.
3. Ahmad, T., Lund, L. H., Rao, P., Ghosh, R., Warier, P., Vaccaro, B., et al. 2018. Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. Journal of the American Heart Association, 7.
4. Ahmad, T., Pencina, M. J., Schulte, P. J., O'brien, E., Whellan, D. J., Pina, I. L., et al. 2014. Clinical implications of chronic heart failure phenotypes defined by cluster analysis. J Am Coll Cardiol, 64, 1765-74.
5. Ajayi, A. A., Sofowora, G. G. & Ladipo, G. O. 2020. Explaining Heart Failure Hyper-mortality in Sub Saharan Africa: Global Genomic and Environmental Contribution Review. Journal of the National Medical Association, 112, 141-157.
6. Alexopoulos, E. C. 2010. Introduction to multivariate regression analysis. Hippokratia, 14, 23-28.
7. Ali, L., Khan, S. U., Golilarz, N. A., Yakubu, I., Qasim, I., Noor, A., et al. 2019. A Feature-Driven Decision Support System for Heart Failure Prediction Based on Statistical Model and Gaussian Naive Bayes. Computational and Mathematical Methods in Medicine, 2019, 6314328.
8. Allam, A., Nagy, M., Thoma, G. & Krauthammer, M. 2019. Neural networks versus Logistic regression for 30 days all-cause readmission prediction. Scientific Reports, 9.

9. Al-Omary, M. S., Davies, A. J., Evans, T. J., Bastian, B., Fletcher, P. J., Attia, J., et al. 2018. Mortality and Readmission Following Hospitalisation for Heart Failure in Australia: A Systematic Review and Meta-Analysis. Heart Lung Circ, 27, 917-927.
10. Angraal, S., Mortazavi, B. J., Gupta, A., Khera, R., Ahmad, T., Desai, N. R., et al. 2020. Machine Learning Prediction of Mortality and Hospitalisation in Heart Failure With Preserved Ejection Fraction. JACC Heart Fail, 8, 12-21.
11. Ashfaq, A., Sant'anna, A., Lingman, M. & Nowaczyk, S. 2019. Readmission prediction using deep learning on electronic health records. Journal of Biomedical Informatics, 97.
12. Awad, M. & Khanna, R. 2015. Machine Learning. Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Berkeley, CA: Apress.
13. Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F. M., Chow, B. J. & Dwivedi, G. 2019. Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death. PLoS One, 14, e0218760.
14. Balki, I., Amirabadi, A., Levman, J., Martel, A. L., Emersic, Z., Meden, B., et al. 2019. Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review. Can Assoc Radiol J, 70, 344-353.
15. Banerjee, A., Chen, S., Fatemifar, G., Zeina, M., Lumbers, R. T., Mielke, J., et al. 2021. Machine learning for subtype definition and risk prediction in heart failure, acute coronary syndromes and atrial fibrillation: systematic review of validity and clinical utility. BMC Med, 19, 85.
16. Blecker, S., Sontag, D., Horwitz, L. I., Kuperman, G., Park, H., Reyentovich, A., et al. 2018. Early Identification of Patients With Acute Decompensated Heart Failure. J Card Fail, 24, 357-362.
17. Bloomfield, G. S., Barasa, F. A., Doll, J. A. & Velazquez, E. J. 2013. Heart failure in sub-Saharan Africa. Current cardiology reviews, 9, 157-173.
18. Boodhun, N. & Jayabalan, M. 2018. Risk prediction in life insurance industry using supervised learning algorithms. Complex & Intelligent Systems, 4, 145-154.

19. Bose, E. & Radhakrishnan, K. 2018. Using Unsupervised Machine Learning to Identify Subgroups Among Home Health Patients With Heart Failure Using Telehealth. Comput Inform Nurs, 36, 242-248.
20. Callender, T., Woodward, M., Roth, G., Farzadfar, F., Lemarie, J. C., Gicquel, S., et al. 2014. Heart failure care in low- and middle-income countries: a systematic review and meta-analysis. PLoS Med, 11, e1001699.
21. Carcillo, F., Le Borgne, Y.-A., Caelen, O. & Bontempi, G. 2018. Streaming active learning strategies for real-life credit card fraud detection: assessment and visualisation. International Journal of Data Science and Analytics, 5, 285-300.
22. Chen, R., Lu, A., Wang, J., Ma, X., Zhao, L., Wu, W., et al. 2019. Using machine learning to predict one-year cardiovascular events in patients with severe dilated cardiomyopathy. European Journal of Radiology, 117, 178-183.
23. Chicco, D. & Jurman, G. 2020. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making, 20.
24. Chirinos, J. A., Orlenko, A., Zhao, L., Basso, M. D., Cvijic, M. E., Li, Z., et al. 2020. Multiple Plasma Biomarkers for Risk Stratification in Patients With Heart Failure and Preserved Ejection Fraction. Journal of the American College of Cardiology, 75, 1281-1295.
25. Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A. E., Pianykh, O. S., et al. 2018. Current Applications and Future Impact of Machine Learning in Radiology. Radiology, 288, 318-328.
26. Concise Oxford English Dictionary. In: SOANES, C. S., A; HAWKER, S (ed.). New York: Oxford University Press. 2006
27. Desai, R. J., Wang, S. V., Vaduganathan, M., Evers, T. & Schneeweiss, S. 2020. Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes. JAMA network open, 3, e1918962.
28. Di Tanna, G. L., Wirtz, H., Burrows, K. L. & Globe, G. 2020. Evaluating risk prediction models for adults with heart failure: A systematic literature review. PLoS ONE, 15.

29. Dobbin, K. K., Zhao, Y. & Simon, R. M. 2008. How large a training set is needed to develop a classifier for microarray data? Clin Cancer Res, 14, 108-14.
30. Dokainish, H., Teo, K., Zhu, J., Roy, A., Alhabib, K. F., Elsayed, A., et al. 2016. Heart Failure in Africa, Asia, the Middle East and South America: The INTER-CHF study. Int J Cardiol, 204, 133-41.
31. Dokainish, H., Teo, K., Zhu, J., Roy, A., Alhabib, K. F., Elsayed, A., et al. 2017a. Global mortality variations in patients with heart failure: results from the International Congestive Heart Failure (INTER-CHF) prospective cohort study. Lancet Glob Health, 5, e665-e672.
32. Eichler, K., Zoller, M., Tschudi, P. & Steurer, J. 2007. Barriers to apply cardiovascular prediction rules in primary care: a postal survey. BMC Family Practice, 8, 1.
33. Emmert-Streib, F. & Dehmer, M. 2019. Evaluation of Regression Models: Model Assessment, Model Selection and Generalization Error. Mach. Learn. Knowl. Extr., 1, 521-551.
34. Fawagreh, K., Gaber, M. M. & Elyan, E. 2014. Random forests: from early developments to recent advancements. Systems Science & Control Engineering, 2, 602-609.
35. Frizzell, J. D., Liang, L., Schulte, P. J., Yancy, C. W., Heidenreich, P. A., Hernandez, A. F., et al. 2017. Prediction of 30-Day All-Cause Readmissions in Patients Hospitalised for Heart Failure: Comparison of Machine Learning and Other Statistical Approaches. JAMA Cardiol, 2, 204-209.
36. Gheorghiade, M., Albert, N. M., Curtis, A. B., Thomas Heywood, J., Mcbride, M. L., Inge, P. J., et al. 2012. Medication dosing in outpatients with heart failure after implementation of a practice-based performance improvement intervention: findings from IMPROVE HF. Congest Heart Fail, 18, 9-17.
37. Gillala, R., Amit Kumar, T. & Reddy, V. K. 2019. A Wide Scale Classification of Class Imbalance Problem and its Solutions: A Systematic Literature Review. Journal of Computer Science, 15.
38. Giolo, S. R., Krieger, J. E., Mansur, A. J. & Pereira, A. C. 2012. Survival analysis of patients with heart failure: implications of time-varying regression effects in modeling mortality. PloS one, 7, e37392-e37392.

39. Gleeson, S., Liao, Y. W., Dugo, C., Cave, A., Zhou, L., Ayar, Z., et al. 2017. ECG-derived spatial QRS-T angle is associated with ICD implantation, mortality and heart failure admissions in patients with LV systolic dysfunction. PLoS One, 12, e0171069.
40. Golas, S. B., Shibahara, T., Agboola, S., Otaki, H., Sato, J., Nakae, T., et al. 2018. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC Med Inform Decis Mak, 18, 44.
41. Hackenberger, B. K. 2019. Bayes or not Bayes, is this the question? Croatian medical journal, 60, 50-52.
42. Hagquist, C. & Stenbeck, M. 1998. Goodness of Fit in Regression Analysis – R2 and G2 Reconsidered. Quality and Quantity, 32, 229-245.
43. Hanratty, B., Hibbert, D., Mair, F., May, C., Ward, C., Capewell, S., et al. 2002a. Doctors' perceptions of palliative care for heart failure: focus group study. BMJ, 325, 581-5.
44. Harada, D., Asanoi, H., Noto, T. & Takagawa, J. 2020. Different Pathophysiology and Outcomes of Heart Failure With Preserved Ejection Fraction Stratified by K-Means Clustering. Front Cardiovasc Med, 7, 607760.
45. Hearn, J., Ross, H. J., Mueller, B., Fan, C. P., Crowdy, E., Duhamel, J., et al. 2018. Neural Networks for Prognostication of Patients With Heart Failure. Circulation. Heart failure, 11, e005193.
46. Hripcsak, G. & Rothschild, A. S. 2005. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc, 12, 296-8.
47. Hsich, E., Gorodeski, E. Z., Blackstone, E. H., Ishwaran, H. & Lauer, M. S. 2011. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. Circulation: Cardiovascular Quality and Outcomes, 4, 39-45.
48. Ieva, F., Paganoni, A. M. & Pietrabissa, T. 2017. Dynamic clustering of hazard functions: an application to disease progression in chronic heart failure. Health Care Manag Sci, 20, 353-364.
49. Jagannatha, A. N. & Yu, H. 2016a. Bidirectional RNN for Medical Event Detection in Electronic Health Records. Proc Conf, 2016, 473-482.

50. Jagannatha, A. N. & Yu, H. 2016b. Structured prediction models for RNN based sequence labeling in clinical text. Proc Conf Empir Methods Nat Lang Process, 2016, 856-865.
51. Jasinska-Piadlo, A., Bond, R., Biglarbeigi, P., Brisk, R., Campbell, P. & Mceneaneny, D. 2022. What can machines learn about heart failure? A systematic literature review. Int J Data Sci Anal, 13, 163–183.
52. Jiang, W., Siddiqui, S., Barnes, S., Barouch, L. A., Korley, F., Martinez, D. A., et al. 2019. Readmission risk trajectories for patients with heart failure using a dynamic prediction approach: Retrospective study. Journal of Medical Internet Research, 21.
53. Karaye, K. M., Dokainish, H., Elsayed, A., Mondo, C., Damasceno, A., Sliwa, K., et al. 2021. Clinical Profiles and Outcomes of Heart Failure in Five African Countries: Results from INTER-CHF Study. Glob Heart, 16, 50.
54. Khan, F. A., Majidulla, A., Tavaziva, G., Nazish, A., Abidi, S. K., Benedetti, A., et al. 2020. Chest x-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: a prospective study of diagnostic accuracy for culture-confirmed disease. Lancet Digit Health, 2, e573-e581.
55. Kingsford, C. & Salzberg, S. L. 2008. What are decision trees? Nature biotechnology, 26, 1011-1013.
56. Kourou, K., Rigas, G., Exarchos, K. P., Goletsis, Y., Exarchos, T. P., Jacobs, S., et al. 2016. Prediction of time dependent survival in HF patients after VAD implantation using pre- and post-operative data. Computers in Biology and Medicine, 70, 99-105.
57. Krumholz, H. M., Coppi, A. C., Warner, F., Triche, E. W., Li, S. X., Mahajan, S., et al. 2019. Comparative Effectiveness of New Approaches to Improve Mortality Risk Models from Medicare Claims Data. JAMA Network Open, 2.
58. Kwon, J. M., Kim, K. H., Jeon, K. H. & Park, J. 2019. Deep learning for predicting in-hospital mortality among heart disease patients based on echocardiography. Echocardiography, 36, 213-218.
59. Kwon, J. M., Kim, K. H., Jeon, K. H., Lee, S. E., Lee, H. Y., Cho, H. J., et al. 2019. Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. PLoS One, 14, e0219302.

60. Lakhani, P. & Sundaram, B. 2017. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology, 284, 574-582.
61. Levy, W. C., Mozaffarian, D., Linker, D. T., Sutradhar, S. C., Anker, S. D., Cropp, A. B., et al. 2006. The Seattle Heart Failure Model: prediction of survival in heart failure. Circulation, 113, 1424-33.
62. Lippi, G. & Sanchis-Gomar, F. 2020. Global epidemiology and future trends of heart failure. AME Medical Journal, 5.
63. Liu, W., Stansbury, C., Singh, K., Ryan, A. M., Sukul, D., Mahmoudi, E., et al. 2020. Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding. PLoS ONE, 15.
64. Lok, L. 2022. Decision Trees, Random Forests, and Gradient Boosting: What's the Difference? [Online]. Available: https://towardsdatascience.com. (Accessed on 31 May 2022) [Accessed].
65. Lorenzoni, G., Santo Sabato, S., Lanera, C., Bottigliengo, D., Minto, C., Ocagli, H., et al. 2019. Comparison of Machine Learning Techniques for Prediction of Hospitalisation in Heart Failure Patients. Journal of Clinical Medicine, 8.
66. Mahajan, S. M., Mahajan, A. S., King, R. & Negahban, S. 2018. Predicting Risk of 30-Day Readmissions Using Two Emerging Machine Learning Methods. Stud Health Technol Inform, 250, 250-255.
67. Mckinley, D., Moye-Dickerson, P., Davis, S. & Akil, A. 2019. Impact of a Pharmacist-Led Intervention on 30-Day Readmission and Assessment of Factors Predictive of Readmission in African American Men With Heart Failure. American Journal of Men's Health, 13.
68. Mcmurray, J. J. V., Packer, M., Desai, A. S., Gong, J., Lefkowitz, M. P., Rizkala, A. R., et al. 2014. Angiotensin–Neprilysin Inhibition versus Enalapril in Heart Failure. New England Journal of Medicine, 371, 993-1004.
69. Mcmurray, J. J. V., Solomon, S. D., Inzucchi, S. E., Køber, L., Kosiborod, M. N., Martinez, F. A., et al. 2019. Dapagliflozin in Patients with Heart Failure and Reduced Ejection Fraction. New England Journal of Medicine, 381, 1995-2008.

70. Meng, F., Zhang, Z., Hou, X., Qian, Z., Wang, Y., Chen, Y., et al. 2019. Machine learning for prediction of sudden cardiac death in heart failure patients with low left ventricular ejection fraction: study protocol for a retroprospective multicentre registry in China. BMJ Open, 9, e023724.
71. Miao, F., Cai, Y. P., Zhang, Y. X., Fan, X. M. & Li, Y. 2018. Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest. IEEE Access, 6, 7244-7253.
72. Mickey, R. M. & Greenland, S. 1989. The impact of confounder selection criteria on effect estimation. Am J Epidemiol, 129, 125-37.
73. Moradi, S. & Mokhatab Rafiei, F. 2019. A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks. Financial Innovation, 5, 15.
74. Moser, D. K., Lee, K. S., Wu, J.-R., Mudd-Martin, G., Jaarsma, T., Huang, T.-Y., et al. 2014. Identification of symptom clusters among patients with heart failure: An international observational study. International Journal of Nursing Studies, 51, 1366-1372.
75. Mpanya, D., Celik, T., Klug, E. & Ntsinjana, H. 2020. Machine learning and statistical methods for predicting mortality in heart failure. Heart Fail Rev.
76. Nakajima, K., Nakata, T., Doi, T., Tada, H. & Maruyama, K. 2020. Machine learning-based risk model using 123I-metaiodobenzylguanidine to differentially predict modes of cardiac death in heart failure. Journal of nuclear cardiology : official publication of the American Society of Nuclear Cardiology.
77. Nichols, J. A., Herbert Chan, H. W. & Baker, M. a. B. 2019. Machine learning: applications of artificial intelligence to imaging and diagnosis. Biophysical reviews, 11, 111-118.
78. Nouraei, H., Nouraei, H. & Rabkin, S. W. 2022. Comparison of Unsupervised Machine Learning Approaches for Cluster Analysis to Define Subgroups of Heart Failure with Preserved Ejection Fraction with Different Outcomes. Bioengineering (Basel), 9.
79. Pakhomov, S. V., Buntrock, J. & Chute, C. G. 2005. Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier. Journal of Biomedical Informatics, 38, 145-153.

80. Pandis, N. 2017. Logistic regression: Part 1. Am J Orthod Dentofacial Orthop, 151, 824-825.
81. Perry, A., Loh, F., Adamo, L., Zhang, K. W., Deych, E., Foraker, R., et al. 2021. Unsupervised cluster analysis of patients with recovered left ventricular ejection fraction identifies unique clinical phenotypes. PLoS One, 16, e0248317.
82. Peterson, P. N., Rumsfeld, J. S., Liang, L., Albert, N. M., Hernandez, A. F., Peterson, E. D., et al. 2010. A Validated Risk Score for In-Hospital Mortality in Patients With Heart Failure From the American Heart Association Get With the Guidelines Program. Circulation: Cardiovascular Quality and Outcomes, 3, 25-32.
83. Pocock, S. J., Ariti, C. A., Mcmurray, J. J. V., Maggioni, A., Køber, L., Squire, I. B., et al. 2012. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. European Heart Journal, 34, 1404-1413.
84. Ponikowski, P., Anker, S. D., Alhabib, K. F., Cowie, M. R., Force, T. L., Hu, S., et al. 2014. Heart failure: preventing disease and death worldwide. ESC Heart Fail, 1, 4-25.
85. Powers, D. & Ailab 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol, 2, 2229-3981.
86. Rahimi, K., Bennett, D., Conrad, N., Williams, T. M., Basu, J., Dwight, J., et al. 2014. Risk prediction in patients with heart failure: a systematic review and analysis. JACC Heart Fail, 2, 440-6.
87. Saito, T. & Rehmsmeier, M. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One, 10, e0118432.
88. Schneider, A., Hommel, G. & Blettner, M. 2010. Linear regression analysis: part 14 of a series on evaluation of scientific publications. Deutsches Arzteblatt international, 107, 776-782.
89. Segar, M. W., Jaeger, B. C., Patel, K. V., Nambi, V., Ndumele, C. E., Correa, A., et al. 2021. Development and Validation of Machine Learning-Based Race-Specific Models to Predict 10-Year Risk of Heart Failure: A Multi-Cohort Analysis. Circulation.
90. Segar, M. W., Patel, K. V., Ayers, C., Basit, M., Tang, W. H. W., Willett, D., et al. 2020. Phenomapping of patients with heart failure with preserved ejection fraction

using machine learning-based unsupervised cluster analysis. European Journal of Heart Failure, 22, 148-158.

91. Sengupta, S., Basak, S., Saikia, P., Paul, S., Tsalavoutis, V., Atiah, F., et al. 2020. A review of deep learning with special emphasis on architectures, applications and recent trends. Knowledge-Based Systems, 194, 105596.
92. Shameer, K., Johnson, K. W., Yahi, A., Miotto, R., Li, L. I., Ricks, D., et al. 2017. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: A case-study using mount sinai heart failure cohort. Pac Symp Biocomput, 22, 276-287.
93. Shams, I., Ajorlou, S. & Yang, K. 2015. A predictive analytics approach to reducing 30-day avoidable readmissions among patients with heart failure, acute myocardial infarction, pneumonia, or COPD. Health Care Management Science, 18, 19-34.
94. Shmilovici, A. 2005. Support Vector Machines. In: MAIMON, O. & ROKACH, L. (eds.) Data Mining and Knowledge Discovery Handbook. Boston, MA: Springer US.
95. Song, Y.-Y. & Lu, Y. 2015. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27, 130-135.
96. Stampehl, M., Friedman, H. S., Navaratnam, P., Russo, P., Park, S. & Obi, E. N. 2020. Risk assessment of post-discharge mortality among recently hospitalised Medicare heart failure patients with reduced or preserved ejection fraction. Current Medical Research and Opinion, 36, 179-188.
97. Su, Z., Brecht, T., O'donovan, F., Boussios, C., Menon, V., Gliklich, R., et al. 2017. Machine Learning Enhanced Predictions of Hospital Readmission or Death in Heart Failure. Circulation, 136.
98. Szymanski, P. Z., Badri, M. & Mayosi, B. M. 2018. Clinical characteristics and causes of heart failure, adherence to treatment guidelines, and mortality of patients with acute heart failure: Experience at Groote Schuur Hospital, Cape Town, South Africa. S Afr Med J, 108, 94-98.
99. Taslimitehrani, V., Dong, G., Pereira, N. L., Panahiazar, M. & Pathak, J. 2016. Developing EHR-driven heart failure risk prediction models using CPXR(Log) with the probabilistic loss function. Journal of Biomedical Informatics, 60, 260-269.

100. Turgeman, L. & May, J. H. 2016. A mixed-ensemble model for hospital readmission. Artificial Intelligence in Medicine, 72, 72-82.
101. Uddin, S., Khan, A., Hossain, M. E. & Moni, M. A. 2019. Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19.
102. Van Diepen, S., Katz, J. N., Albert, N. M., Henry, T. D., Jacobs, A. K., Kapur, N. K., et al. 2017. Contemporary Management of Cardiogenic Shock: A Scientific Statement From the American Heart Association. Circulation, 136, e232-e268.
103. Zhang, Z. 2016. Model building strategy for logistic regression: purposeful selection. Annals of translational medicine, 4, 111-111.
104. Zou, K. H., O'malley, A. J. & Mauri, L. 2007. Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. Circulation, 115, 654-657.

Appendices

A Excluded Studies: Systematic Review of Machine Learning Models

Original research studies unrelated to heart failure (n=95), studies designed for the identification of risk factors for heart failure (n=10), prediction of outcomes other than mortality or rehospitalisation (n=5), narrative review paper related to heart failure (n=10), review paper unrelated to heart failure (n=23), diagnosis of heart failure using an electrocardiogram (n=2), evaluation of response of cardiac resynchronisation therapy (n=1), heart failure validation study (n=1), evaluation of predictors (n=1), limited or no clinical data related to heart failure patients (n=8), abstract/poster only (n=14), study protocol (n=1), heart failure classification (n=5), cardiovascular disease classification (n=3), rehospitalisation of all conditions with no isolated data on heart failure patients (n=12), detection of heart failure (n=9), predicting heart failure (n=6), home monitoring of heart failure (n=1), distinguishing heart failure patients from normal subjects (n=3), tele monitoring in heart failure (n=2), regression analysis with traditional statistical methods only (n=12), models created with Cox proportional hazard for survival analysis in heart failure (n=4), Bowtie model for palliative care (n=1), study using mathematical formula not classified as a machine learning algorithm (n=1).

B Clinical Parameters Extracted From the Electronic Health Record System

	Categorical variables	Continuous variables
Demographic data	Gender, ethnicity, international classification of diseases, tenth revision (ICD-10) code,	Age
Comorbidities	Hypertension, diabetes mellitus, smoking, dyslipidaemia, family history of coronary artery disease, human immunodeficiency virus (HIV) and chronic kidney disease	
Clinical examination findings	Lung crepitations, ascites, elevated jugular venous pressure, pedal oedema, displaced apex beat,	Systolic and diastolic blood pressure
Electrocardiography	Rhythm, P wave, Q wave, bundle branch block, ST segment	Heart rate, PR interval, QT interval
Echocardiography		Left ventricular internal diameter at diastole (LVIDd), Left ventricular internal diameter at systole (LVIDs), Left ventricular ejection fraction (LVEF) and Left atrial size
Diseased vessel on coronary angiography	Left main artery, left anterior descending artery, right coronary artery, circumflex artery and diagonal artery	
Laboratory results		Troponin I, haemoglobin, sodium, potassium, urea, creatinine, International Normalized Ratio (INR), total cholesterol, low-density lipoprotein (LDL), high-density lipoprotein (HDL), glucose, glycated haemoglobin (HBA1c) and pro-brain natriuretic peptide (proBNP)
Oral medication	Aspirin, beta-blockers, furosemide, angiotensin converting enzyme (ACE) inhibitors, calcium antagonists and statins	
Outcome	Mortality	Length of hospital stay

C Predicting 30-Day All-Cause Rehospitalisation

Appendix C consists of results obtained after building a model predicting 30-day all-cause rehospitalisation in heart failure using supervised machine learning algorithms. The 30-day all-cause rehospitalisation rate among patients with a left ventricular ejection fraction (LVEF) < 50% was 4.3%, with 34 patients admitted within 30 days and 50 patients beyond 30 days. Patients readmitted within 30 days were younger, with a mean age of 48.9 ± 14.8 vs 54.9 ± 16.3, p=0.032. The rest of the clinical parameters did not differ significantly between patients readmitted within 30 days, those admitted beyond 30 days, and those not admitted. The table below shows the performance metrics of machine learning algorithms when predicting 30-day all-cause rehospitalisation. The MLP showed an accuracy of 0.95 and 0.94 during training and testing, respectively.

Performance metrics of supervised machine learning algorithms predicting 30-day all-cause rehospitalisation in heart failure patients with a left ventricular ejection fraction less than 50%

Algorithm	**Accuracy**	**Precision**	**Recall**	**f1-score**
Logistic regression (ML)				
Training: Not rehospitalised	0.89	1.00	0.88	0.94
Rehospitalised		0.24	0.95	0.39
Test: Not rehospitalised	0.79	0.95	0.82	0.88
Rehospitalised		0.09	0.31	0.14
Weighted training score		0.97	0.89	0.92
Weighted test score		0.91	0.79	0.84
Support vector machines				
Training: Not rehospitalised	0.98	0.98	1.00	0.99

	Rehospitalisation		1.00	0.48	0.65
Test:	Not rehospitalised	0.92	0.94	0.97	0.96
	Rehospitalisation		0.00	0.00	0.00
Weighted training score			0.98	0.98	0.98
Weighted test score			0.89	0.92	0.90
Extreme Gradient Boosting					
Training:	Not rehospitalised	0.93	0.99	0.93	0.96
	Rehospitalisation		0.32	0.81	0.46
Test:	Not rehospitalised	0.90	0.97	0.93	0.95
	Rehospitalisation		0.27	0.46	0.34
Weighted training score			0.97	0.93	0.94
Weighted test score			0.93	0.90	0.92
Decision Trees					
Training:	Not rehospitalised	0.90	1.00	0.90	0.95
	Rehospitalisation		0.28	1.00	0.44
Test:	Not rehospitalised	0.84	0.96	0.87	0.91
	Rehospitalisation		0.14	0.38	0.21
Weighted training score			0.97	0.90	0.93
Weighted test score			0.92	0.84	0.87

ML: machine learning

Predicting In-hospital Mortality in Heart Failure Patients using Machine Learning

STUDY PROTOCOL

Student name: Dr Dineo Mpanya

Student number: 878824

Degree: Doctor of Philosophy (PhD)

Institution: University of the Witwatersrand

Supervisor: Professor Hopewell Ntsinjana

MBBCh, FC Paed (SA), Cert Cardiol (Paed) SA, PhD

Honorary Associate Professor, Department of Paediatrics and Child Health, School of Clinical Medicine, University of the Witwatersrand

Head of Paediatric Cardiology, Nelson Mandela Children's Hospital

Co-supervisor: Professor Turgay Celik

PhD

Professor in Computer Science and Applied Mathematics

University of the Witwatersrand

TABLE OF CONTENTS

1. List of abbreviations 3
2. Introduction 4
3. Study Aim 11
4. Study Objectives 11
5. Methods 12
6. Data Analysis 14
7. Ethics 14
8. Problems and Limitations 15
9. Funding 16
10. Timing 17
11. References 19
12. Appendix A 24
13. Appendix B 25

LIST OF ABBREVIATIONS

ACE	Angiotensin converting enzyme
ANN	Artificial neural network
BIOSTAT-CHF	Biology study to tailored treatment in chronic heart failure cohort
CMJAH	Charlotte Maxeke Johannesburg Academic Hospital
EHR	Electronic Health Records
ESC	European Society of Cardiology
HBA_1C	Glycated haemoglobin
HFpEF	Heart failure with a preserved left ventricular ejection fraction
HFrEF	Heart failure with a reduced left ventricular ejection fraction
HIV	Human Immunodeficiency Virus
HDL	High density lipoprotein
HPCSA	Health Professions Council of South Africa
ICD	International Classification of Disease
INR	International Normalized Ratio
k-NN	k nearest neighbour
LDL	Low density lipoprotein
LVEF	Left ventricular ejection fraction
LVIDd	Left ventricular internal diameter at diastole
LVIDs	Left ventricular internal diameter at systole
LR	Linear regression
NHLS	National Health Laboratory Service
NYHA	New York Heart Association
PRAISE	Prospective Randomized trial of Amlodipine versus placebo
ProBNP	Pro brain natriuretic peptide
RF	Random forest
RNN	Recurrent neural network
SHF	Seattle Heart Failure
SVM	Support vector machines
USA	United States of America

INTRODUCTION

Heart failure is a rapidly growing public health problem that is associated with a high morbidity and mortality rate with as many as 37.7 million people globally suffering from this pandemic [1]. The European Society of Cardiology (ESC) defines heart failure as a clinical syndrome characterised by symptoms such as shortness of breath, persistent coughing, wheezing and fatigue that may be accompanied by an increased heart rate, limb oedema, pulmonary crackles and raised jugular venous pressure [2]. However, in the early stages of heart failure, symptoms may not be apparent.

Heart failure is a common pathophysiological pathway for all cardiovascular diseases initiated by an insult or injury to the heart muscles. The insult may be congenital or acquired for example, hypertension, obstructive coronary artery disease and valvular heart diseases [3]. Heart failure has traditionally been thought to result from impaired left ventricular contractile function. However, almost half of patients with heart failure have a preserved left ventricular ejection fraction (HFpEF) [4]. In a meta-analysis evaluating the prognostic significance of HFpEF in 24 501 patients over a mean follow up period of 47 months, 2 468 (32.1%) patients diagnosed with HFpEF died compared with 6 831 (40.6%) deaths among the 16 813 patients with heart failure with a reduced left ventricular ejection fraction (HFrEF) [5].

Over the past few decades, advancements in the pharmacological and non-pharmacological management of heart failure have occurred. Despite these interventions, there are still a lot of unmet needs since a significant proportion of patients with HFrEF experience progressive disease and eventually mortality. Risk prediction models have been proposed as adjuncts in

the management of heart failure [6]. The rationale behind the creation of risk models is to identify high-risk individuals who might benefit from early preventative strategies, aggressive therapy or closer follow up. Nonetheless, the incorporation of risk prediction models into most heart failure management guidelines has been rather slow. This is partly because most risk prediction models have not been externally validated and the focus has been primarily on the creation of models with high accuracy and probability scores, rather than on studying the influence of risk models on decision-making and health outcomes. Risk models are population specific and cannot be generalizable to any population, unless if validated in a cohort not used to create the model. To the best of my knowledge, there are no risk models predicting heart failure-related mortality or rehospitalisation created using data originating in Africa. Furthermore, none of the existing risk prediction models have been validated in any of the African countries.

A risk prediction model is a mathematical model used to estimate the risk of future outcomes, based on one or more underlying characteristics or predictors. In some instances, terminology such as "score", "tool" or "index" often replaces the term model. Risk models are not only used to predict future events, but they are also used to estimate the risk of having a particular disease. Both traditional statistical methods and machine learning algorithms are used to create risk prediction models. They both employ mathematical formulas to establish a relationship or a pattern between variables. However, there are slight differences in their methodological approach. The primary focus with traditional statistical methods is to conduct inference about a sample or population parameters, whereas machine learning focuses on learning from data and improving the accuracy of predictions by minimizing the calculated error [7].

Machine learning refers to the ability of a computer to analyse data and learn patterns without any form of explicit programming [8]. Machine learning techniques have not only gained interest in the commercial industry but also in the healthcare sector. A plethora of research papers have been published in the application of machine learning algorithms for the detection and classification of heart failure, severity estimation and prediction of adverse events such as mortality and rehospitalisation [9-14]. There are four main types of machine learning algorithms, supervised, unsupervised, semi-supervised and reinforcement learning [15].

In supervised learning, a computer is given a task of learning from a dataset with a labelled output or outcome. For example, an output may be represented as "heart failure" or "no heart failure." In contrast, unsupervised learning uses unlabelled datasets. A computer analyses data, learns patterns and subsequently clusters the variables into categories based on the estimated probability of each variable belonging to a particular cluster, while in reinforcement learning, a computer learns by trial and error, with penalty and reward scores awarded for inaccurate and accurate responses, respectively [16]. Semi-supervised learning occurs when learning takes place using a combination of labelled and unlabelled datasets. Some of the common supervised machine learning algorithms include decision trees, random forests, naïve Bayes classifier, support vector machines (SVM), k-nearest neighbors (k-NN) and artificial neural networks (ANN) [17].

Decision trees have been labelled as the best “off-the-shelf” algorithm for prediction as they are among the simple algorithms that are commonly used. A decision tree is constructed with a set of nodes split into three disjoint sets known as the decision, chance and terminal nodes. In a decision node, an action is selected from one of the two edges or sides of the node. In a

chance node, a reaction is selected randomly while terminal nodes represent the end of the sequence in the decision problem [18]. Decisions tree algorithms are ideal for use when a dataset contains many observations and fewer variables [19]. Random forests can be perceived as an aggregate of simple decision trees, where instead of selecting all observations in the dataset, random samples are selected and incorporated into the model.

Support vector machines classify data by placing variables in a higher-dimensional space, generally variables with a non-linear relationship to each other. After that, a "hyperplane" is created that separates the data into categories, while K- nearest neighbors use a simple integer "k" to classify data. For instance, if k=3, the distance between the new observation and the *three* nearest variables will be determined. The new observation is then added to the class with the majority of nearest neighbors [19]. In instances where classification algorithms have a high error rate (weak learner), boosting algorithms are used to minimise the error rate. The boosting algorithm is a machine learning technique that reduces the bias and variance of the predictor model by transforming a set of weak classifiers, gradually boosting them into stronger classifiers [20].

With the increased complexity of data, deep learning algorithms such as ANN have proven to outperform traditional statistical methods and most of the machine learning algorithms. Deep learning is becoming popular as a subtype of machine learning that is responsible for image and speech recognition software as well as autonomously driving cars. Deep learning is inspired by the structure and function of the neuron, the functional unit of the central nervous system. Choi et al. used recurrent neural networks (RNN), a type of ANN for the prediction of the initial diagnosis of heart failure. The RNN algorithm was superior to supervised machine learning algorithms [21].

There is no consensus on the best method for the selection of predictor variables when creating heart failure models predicting mortality or rehospitalisation. In a meta-analysis evaluating factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalisation in patients with heart failure, Ouwerkerk et al. identified 117 models created with both traditional statistics and machine learning algorithms that predict mortality and hospitalisation in patients with chronic heart failure and acute decompensated heart failure [22]. Overall, there were 249 variables used as predictors (see appendix A). The most commonly used variables with the highest predictive values were blood urea nitrogen and sodium levels [22]. Other persistently identified predictors of mortality in patients with heart failure reported in the literature are age, gender, blood pressure, renal function, left ventricular ejection fraction and New York Heart Association functional class [14]. The models predicting mortality had good discriminatory powers compared to those predicting hospitalisation. The most plausible explanation for such an observation is that hospitalisation may be associated with parameters that cannot be easily measured, such as knowledge, attitude, beliefs and health-seeking behaviour.

The mean cost of each heart failure-related hospitalisation has been estimated to be $23 000 (United States dollars) [23]. Early identification of individuals at risk of rehospitalisation will allow the implementation of strategies that prevent or reduce the frequency of rehospitalisation. In a prospective study evaluating the predictors of rehospitalisation in patients admitted with heart failure in Nigeria, multivariable regression analysis identified female gender and a body mass index less than 19 kilograms per metre squared (kg/m^2) as independent correlates of rehospitalisation during the 6-month follow-up period [24]. The mean duration of the hospital stay was ten days, and the most common reason for rehospitalisation was worsening heart failure symptoms, reported in 75% of the patients

studied. The readmission rates were lower (12%) in this study, in contrast to findings from the developed world with rehospitalisation rates as high as 23% [25]. The authors attributed lower readmission rates to a younger age in their cohort and strikingly different underlying causes of heart failure, compared to the population in the developed world [24].

The Seattle Heart failure (SHF) model is one of the best performing models created using cox proportional hazard regression. The aim was to predict survival in heart failure. The data used to build the SHF model was obtained from 1125 outpatients recruited in the Prospective Randomized Trial of Amlodipine versus Placebo (PRAISE 1). The model was prospectively validated on a new cohort, comprising of 9942 participants. Although some of the data used to validate the model originated from heart failure registries in Italy and the United States of America (USA), a more substantial fraction of the data was obtained from outpatient clinical trial participants. The correlation between predicted and actual survival was 97%. However, this model cannot be generalized to hospitalised patients or those with life-threatening conditions such as renal failure, since the data used to create the model was obtained from carefully screened trial participants without significant co-morbidities [26]. Another limiting factor is the major proportion of the participants (64%) with heart failure caused by ischaemic heart disease. In contrast, the cause of heart failure in sub-Saharan Africa is predominantly non-ischaemic (up to 75%), although risk factors for ischaemic heart disease are becoming more prevalent in sub-Saharan Africa [3].

The ability of the model to generalise accurately is reflected by its performance when externally validated using data from a cohort of patients not previously used to train, test or internally validate the model. Data from the Biology Study to Tailored Treatment in Chronic Heart Failure (BIOSTAT-CHF) cohort comprising of 2516 patients from 11 European

countries with worsening symptoms of heart failure and a left ventricular ejection fraction of 40% or less was used to train a model to predict all-cause mortality and first heart failure-related hospitalisation. The model used to predict events was externally validated on a cohort of 1 738 patients from six centres in Scotland. Model accuracy was 73%, 69% and 71% when predicting mortality, hospitalisation and both, respectively [27].

Model overfitting and underfitting are the most common problems encountered when evaluating model performance. Underfitting occurs when the model fails to accurately classify data or make predictions using data that it was trained on. A low accuracy score and a high loss identify a model that is underfitting. Overfitting occurs when a model shows high accuracy levels during training and low accuracy levels during validation. Bootstrapping and bagging are two methods that attempt to eliminate overfitting. Bootstrapping involves taking a random sample of the observations instead of selecting all observations in the dataset, while bagging entails randomly selecting a sample of variables [19].

Some authors still argue that machine learning techniques hold modest accuracy when compared to traditional statistical methods and further state that most of the published studies failed to perform a head to head comparison of machine learning algorithms to statistical methods [28]. Mortazavi et al. demonstrated the efficacy of machine learning algorithms over conventional statistical methods in predicting heart failure readmissions. In this study, predictive models using Linear Regression (LR) and Poisson Regression as well as machine learning techniques such as random forest (RF), Boosting and support vector machine algorithms (SVM) were used to predict 30-day and 180-day all-cause and heart failure related readmission. The best performing algorithm for the prediction of 30-day all-cause readmission was the random forest, providing an 18% improvement over LR [29]. This study

shows that complex learning methods may produce comparable prediction results, thus if we employ recently developed advanced machine learning methods, we may see further improvements in the accuracy scores. My research project will facilitate precision medicine by risk stratification of patients with heart failure using risk calculators, such that aggressive therapy is initiated early and optimally in high risk patients.

AIM OF THE STUDY

The primary aim of the study is to predict all-cause in-hospital mortality in patients with HFrEF using supervised machine learning algorithms.

STUDY OBJECTIVES

a) To describe the baseline demographic and clinical characteristics in patients with heart failure.
b) To compare supervised machine learning algorithms to traditional statistical methods for the prediction of all-cause mortality in heart failure patients.
c) To create a risk model predicting hospitalisation due to any cause after initial hospitalisation in patients with heart failure.
d) To externally validate the model with the best performing metrics for predicting in-hospital mortality and the risk of hospitalisation in heart failure patients. The model will be validated by comparing actual and predicted risk of in-hospital mortality and hospitalisation after one year of follow up.

METHODS

Data

Data will be mined from discharge summaries captured in the Electronic Health Record (EHR) system. The study population will comprise of patients admitted in the cardiology wards at Charlotte Maxeke Johannesburg Academic Hospital (CMJAH) between January 2009 and December 2018.

Inclusion criteria

The data from all patients who are 18 years and older will be included. Patients will be separated into two groups, those with and without the diagnosis of heart failure based on the International Classification of Disease, tenth revision; ICD-10: 2016, code I42.0 and I50. The diagnosis of heart failure will be verified by correlating the ICD-10 code with the clinical history and examination findings. Patients with the diagnosis of heart failure will be subsequently categorized into two distinct groups based on the baseline resting left ventricular ejection fraction (less than 50% and greater or equal to 50%).

Exclusion criteria

The following patients will be excluded from the analysis:

a) Patients without a diagnosis
b) Patients with a terminal illness (e.g. malignancy)
c) Patients with missing clinically significant information (e.g. LVEF)

Step 1: Data preparation

a) Retrieve the data from the Electronic Health Record system in .xls (excel) format

b) Dedicated data cleaning software, such as Tableau, will be used.

    - Duplicate entries will be removed.
    - Records that have too many blank or empty entries will be dropped or handled appropriately.

Step 2: Identification of variables available from the EHR (feature selection):

Please see appendix B

Step 3: Building and training the model

a) Data from heart failure patients with a LVEF<50% will be used to build the model.

b) Split the dataset into respective sets: training, test and validation sets:

    - training set: data from Jan 2009 to December 2016 (80%)
    - testing set: data from Jan 2009 to December 2016 (20%)
    - validation set: data from Jan 2017 to December 2018 (100%)

c) Use the training datasets to create a predictor model using supervised machine learning algorithms and traditional statistical methods:

    - Logistic regression – *use both statistical method and machine learning algorithm*
    - Artificial Neural Networks
    - Support vector machine
    - Decision trees
    - Boosting

Step 4: Evaluate machine learning model's performance:

Use the test dataset to evaluate and compare different machine learning model performances based on the following measures: precision, recall and the F1 score.

Step 5: Evaluate model created with Logistic Regression:

The best performing model will be assessed by taking into consideration parameters such as the r squared value.

Step 6: Validation of research findings

a) Validate the best performing machine learning model using data that was not used to train or test the model (data from Jan 2017 to December 2018).
b) Evaluate model's performance

DATA ANALYSIS

Machine learning and statistical models will be generated with the latest version of R statistical software (R studio). Logistic regression analysis (univariable and multivariable) will be used to identify predictors of mortality and re-hospitalisation. Supervised machine learning algorithms will be used to categorize data.

ETHICS

Ethics clearance has been requested from the University of the Witwatersrand's Human Research Ethics Committee (medical).

PROBLEMS AND LIMITATIONS

This project will suffer all problems and limitations associated with retrospective data analysis. I will attempt to retrieve missing data by reviewing in and out-patient medical records as well as other sources of patient data such as the NHLS (National Health Laboratory Service) website.

FUNDING

	Income	Expenditure
Discovery Foundation Award R350 000: Jan-Dec 2019 The outstanding amount will be received in Jan 2020.	800 000	
PhD registration (cost for 2020)		30 000
HPCSA annual fee (cost for 2020)		2 700
Tableau software		5 000
iMac Pro 27-inch 4.2GHz (For handling computer science software with large datasets)		88 000
Data cleaning and verification: Assistants (3 @ R 10 000 per month), retrieve hard copies, call patients, retrieve mortality data at the Department of Home Affairs		180 000
SA Heart Congress 2019: o Registration o Workshop for non-cardiologists		 6 200 650
ESC Congress 2020 o Registration o Flights o Accommodation		 12 000 17 000 22 400
Heart Failure Congress 2020 o Registration o Flights o Accommodation		 10 320 15 000 13 000
International Congress on Machine Learning and Application 2020 o Registration o Flights o Accommodation		 4 200 18 000 8 500
Online Data Science Courses		1 600
Stipend (Jan- Dec 2020) R40 000/month		480 000
TOTAL as of April 2019		914 570
Outstanding Funding		**R 564 570**

Additional funding will be requested from the following organization:
South African Heart Association (SA Heart)
The Professor Bongani Mayosi Netcare Clinical Scholarship
South African Medical Association (SAMA)
Medical Research Council (SA)

TIMING

	June 2018	July	Aug	Sept	Oct	Nov	Dec	Jan 2019	Feb	Mar	Apr	May	June	July	Aug
Literature review															
Protocol development															
Protocol Assessment															
Ethics application															
Systematic review															
Data cleaning and verification															
Data analysis															

	Sept	Oct	Nov	Dec	Jan 2020	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov
Manuscript 1															
Manuscript 2															
Manuscript 3															

REFERENCES

1. Ziaeian B, Fonarow GC. Epidemiology and aetiology of heart failure. Nat Rev Cardiol. 2016;13(6):368-78.

2. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JG, Coats AJ, et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur J Heart Fail. 2016;18(8):891-975.

3. Gallagher J, McDonald K, Ledwidge M, Watson CJ. Heart Failure in Sub-Saharan Africa. Card Fail Rev. 2018;4(1):21-4.

4. Hayley BD, Burwash IG. Heart failure with normal left ventricular ejection fraction: role of echocardiography. Curr Opin Cardiol. 2012;27(2):169-80.

5. Somaratne JB, Berry C, McMurray JJ, Poppe KK, Doughty RN, Whalley GA. The prognostic significance of heart failure with preserved left ventricular ejection fraction: a literature-based meta-analysis. Eur J Heart Fail. 2009;11(9):855-62.

6. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE, Jr., Drazner MH, et al. 2013 ACCF/AHA guideline for the management of heart failure: executive summary: a report of the American College of Cardiology Foundation/American Heart Association Task Force on practice guidelines. Circulation. 2013;128(16):1810-52.

7. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods. 2018;15(4):233-4.

8. N DSDAPAR. Applications of Artificial Intelligence in Machine Learning: Review and Prospect. International Journal of Computer Applications 2015;115(9).

9. Zaninovic Jurjevic T, Dvornik S, Kovacic S, Matana Kastelan Z, Brumini G, Matana A, et al. A simple prognostic model for assessing in-hospital mortality risk in patients with acutely decompensated heart failure. Acta Clin Belg. 2018:1-8.

10. Sanchez-Martinez S, Duchateau N, Erdei T, Kunszt G, Aakhus S, Degiovanni A, et al. Machine Learning Analysis of Left Ventricular Function to Characterize Heart Failure With Preserved Ejection Fraction. Circ Cardiovasc Imaging. 2018;11(4):e007138.

11. Karauzum K, Karauzum I, Ural D, Baydemir C, Aktas M, Celikyurt U, et al. A simple discharge risk model for predicting 1-year mortality in hospitalised acute decompansated heart failure patients with reduced ejection fraction. Acta Cardiol. 2018;73(2):164-70.

12. Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC Med Inform Decis Mak. 2018;18(1):44.

13. Tripoliti EE, Papadopoulos TG, Karanasiou GS, Naka KK, Fotiadis DI. Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques. Comput Struct Biotechnol J. 2017;15:26-47.

14. Sahle BW, Owen AJ, Chin KL, Reid CM. Risk Prediction Models for Incident Heart Failure: A Systematic Review of Methodology and Model Performance. J Card Fail. 2017;23(9):680-7.

15. Hadi Hormozi EH, and Hamed Rahimi Nohooji The Classification of the Applicable Machine Learning Methods in Robot Manipulators International Journal of Machine Learning and Computing. 2012;2(5):560-3.

16. MathWorks. 2018. Available from: https://www.mathworks.com/discovery/supervised-learning.html?s_tid=srchtitle. Accessed 21 January 2019.

17. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial Intelligence in Precision Cardiovascular Medicine. J Am Coll Cardiol. 2017;69(21):2657-64.

18. Kaminski B, Jakubczyk M, Szufel P. A framework for sensitivity analysis of decision trees. Cent Eur J Oper Res. 2018;26(1):135-59.

19. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial Intelligence in Cardiology. J Am Coll Cardiol. 2018;71(23):2668-79.

20. Chen P, Pan C. Diabetes classification model based on boosting algorithms. BMC Bioinformatics. 2018;19(1):109.

21. Choi E SA, Stewart WF and Sun J. Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association. 2017;24(2):361-70.

22. Ouwerkerk W, Voors AA, Zwinderman AH. Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalisation in patients with heart failure. JACC Heart Fail. 2014;2(5):429-36.

23. Wang G, Zhang Z, Ayala C, Wall HK, Fang J. Costs of heart failure-related hospitalisations in patients aged 18 to 64 years. Am J Manag Care. 2010;16(10):769-76.

24. Ogah OS, Stewart S, Falase AO, Akinyemi JO, Adegbite GD, Alabi AA, et al. Predictors of rehospitalisation in patients admitted with heart failure in Abeokuta, Nigeria: data from the Abeokuta heart failure registry. J Card Fail. 2014;20(11):833-40.

25. Ross JS, Chen J, Lin Z, Bueno H, Curtis JP, Keenan PS, et al. Recent national trends in readmission rates after heart failure hospitalisation. Circ Heart Fail. 2010;3(1):97-103.

26. Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, et al. The Seattle Heart Failure Model: prediction of survival in heart failure. Circulation. 2006;113(11):1424-33.

27. Voors AA, Anker SD, Cleland JG, Dickstein K, Filippatos G, van der Harst P, et al. A systems BIOlogy Study to TAilored Treatment in Chronic Heart Failure: rationale, design, and baseline characteristics of BIOSTAT-CHF. Eur J Heart Fail. 2016;18(6):716-26.

28. Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. PLoS One. 2018;13(3):e0194889.

29. Mortazavi BJ, Downing NS, Bucholz EM, Dharmarajan K, Manhapra A, Li SX, et al. Analysis of Machine Learning Techniques for Heart Failure Readmissions. Circ Cardiovasc Qual Outcomes. 2016;9(6):629-40.

APPENDIX A

TABLE 1 The Most Frequently Used Variables, Along With the Number of Times Used in the Different Models, and Their Predictive Scores

		Odds Ratio			Hazard Ratio		
	n	z-Score	Mean	95% CI	z-Score	Mean	95% CI
CHF and ADHF							
Age, yrs	78	17,656	1.04	1.03 to 1.04	20,728	1.06	1.06 to 1.07
Sex	53	15,542	1.12	1.11 to 1.14	10,813	1.22	1.19 to 1.26
Systolic blood pressure	51	31,799	1.30	1.29 to 1.32	19,353	1.16	1.15 to 1.18
Sodium	46	33,502	1.41	1.39 to 1.43	30,343	1.09	1.08 to 1.09
Diabetes	41	15,849	1.14	1.12 to 1.15	21,227	1.44	1.41 to 1.47
Creatinine	38	34,191	1.12	1.11 to 1.13	10,411	1.07	1.05 to 1.08
New York Heart Association functional class	35	15,109	4.14	3.96 to 4.33	17,683	1.41	1.37 to 1.45
Blood urea nitrogen	28	60,698	2.28	2.26 to 2.31	12,495	1.07	1.06 to 1.09
Ejection fraction	23				12,843	1.08	1.07 to 1.10
Hemoglobin	23	15,346	1.12	1.11 to 1.14	8,526	1.08	1.06 to 1.10
(N-terminal pro) B-type natriuretic peptide	23				15,327	1.44	1.39 to 1.49
CHF							
Age, yrs	75	17,566	1.04	1.03 to 1.04	20,650	1.06	1.06 to 1.07
Sex	53	15,542	1.12	1.11 to 1.14	10,813	1.22	1.19 to 1.26
Systolic blood pressure	46	54,716	1.16	1.15 to 1.16	18,647	1.18	1.16 to 1.19
Sodium	44	33,502	1.41	1.39 to 1.43	36,527	1.07	1.06 to 1.07
Diabetes	41	15,849	1.14	1.12 to 1.15	21,227	1.44	1.41 to 1.47
Creatinine	37	34,191	1.12	1.11 to 1.13	10,411	1.07	1.05 to 1.08
New York Heart Association functional class	34	15,109	4.14	3.96 to 4.33	17,506	1.41	1.37 to 1.44
Blood urea nitrogen	25	61,497	2.37	2.34 to 2.39	12,014	1.07	1.06 to 1.08
Ejection fraction	23				12,843	1.08	1.07 to 1.10
(N-terminal pro) B-type natriuretic peptide	23				27,196	1.11	1.1 to 1.12
Etiology	22				6,795	1.23	1.17 to 1.29
ADHF							
Systolic blood pressure	7	2,181	1.22	1.04 to 1.41	0.378	1.28	-0.01 to 2.57
Age, yrs	6	8,454	1.46	1.37 to 1.55	0.254	1.27	-0.57 to 3.11
Blood urea nitrogen	4	0.250	1.26	-0.55 to 3.07	0.339	1.33	-0.32 to 2.98
Heart failure admissions	4	12,443	1.44	1.39 to 1.5	0.737	2.00	
Sodium	4	0.530	1.53	-0.04 to 3.1	1,663	1.33	0.99 to 1.67
Dementia/Alzheimer disease or senility	3	6,306	2.42	2.15 to 2.70			
Cancer	2	0.827	1.86	0.39 to 3.33			
Cerebrovascular disease	2	0.410	1.43	-0.28 to 3.14			
Creatinine	2						
Sex	2	0.636	1.63	0.13 to 3.13	0.438	1.33	

ADHF = acute decompensated heart failure; CHF = chronic heart failure; CI = confidence interval.

Ouwerkerk W, Voors AA, Zwinderman AH. Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalisation in patients with heart failure. JACC Heart Fail. 2014;2(5):429-36.

APPENDIX B

	Categorical variables	Continuous variables
Demographic data	Gender, ethnicity	Age
Co-morbidities	Hypertension, diabetes mellitus, smoking, dyslipidaemia, family history of coronary artery disease, human immunodeficiency virus (HIV) and chronic kidney disease	
Clinical examination findings	Lung crepitations, ascites, elevated jugular venous pressure, pedal oedema, displaced apex beat, New York Heart Association (NYHA) functional class	Systolic and diastolic blood pressure
Electrocardiography	Rhythm, P wave, Q wave, bundle branch block, ST segment	Heart rate, PR interval, QT interval
Echocardiography		Left ventricular internal diameter at diastole (LVIDd), Left ventricular internal diameter at systole (LVIDs), Left ventricular ejection fraction (LVEF) and Left atrial size
Diseased vessel on coronary angiography	Left main artery, left anterior descending artery, right coronary artery, circumflex artery and diagonal artery	
Laboratory results		Troponin I, haemoglobin, sodium, potassium, urea, creatinine, International Normalized Ratio (INR), total cholesterol, low-density lipoprotein (LDL), high-density lipoprotein (HDL), glucose, glycated haemoglobin (HBA1c) and pro-brain natriuretic peptide (proBNP)
Oral medication	Aspirin, beta-blockers, furosemide, angiotensin converting enzyme (ACE) inhibitors, calcium antagonists and statins	
Outcome	Mortality Hospitalisation	Length of hospital stay

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

R14/49 Dr D Mpanya

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M190515

NAME: (Principal Investigator)	Dr D Mpanya
DEPARTMENT:	School of Clinical Medicine Department of Cardiology Charlotte Maxeke Johannesburg Academic Hospital
PROJECT TITLE:	Predicting in-hospital mortality in heart failure patients using machine learning
DATE CONSIDERED:	2019/05/31
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Professsors H Ntsinjana and T Celik
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	2019/07/18

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 3rd Floor, Phillip Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

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 submit details to the Committee. **I agree to submit a yearly progress report**. When a funder requires annual re-certification, the application date will be one year after the date when the study was initially reviewed. In this case, the study was initially reviewed in **May** and will therefore reports and re-certification will be due early in the month of **May** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

Date

PLEASE QUOTE THE CLEARANCE CERTIFICATE NUMBER IN ALL ENQUIRIES

Turnitinnarrative.docx

by Dineo Mpanya

Submission date: 01-Feb-2023 02:09PM (UTC+0200)
Submission ID: 2004011699
File name: Turnitinnarrative.docx (25.41K)
Word count: 3285
Character count: 18359

Abstract

Heart failure is a debilitating clinical syndrome associated with increased morbidity, mortality and frequent hospitalisation, leading to increased utilisation of the health care budget. Despite the exponential growth in the introduction of pharmacological agents and medical devices that improve survival, many heart failure patients, in particular, those with a left ventricular ejection fraction less than 40% still experience persistent clinical symptoms that lead to an overall decreased quality of life. Clinical risk prediction is one of the strategies that has been implemented for the selection of high-risk patients and for guiding therapy. However, most risk predictive models have not been well-integrated into the clinical setting. This is partly due to inherent limitations such as the creation of risk predicting models using static clinical data that does not consider the dynamic nature of heart failure. This review article focuses on describing how predictive models for risk-stratification of patients with heart failure are built.

Keywords

Heart failure, predict, models, machine learning, deep learning

Introduction

Heart failure is estimated to affect at least 64.3 million individuals globally.[1] Despite the availability of therapeutic agents and medical devices that aim to prolong survival, many heart failure patients still experience progressive disease and eventually death. The International Congestive Heart Failure (INTER-CHF) study recruited 5 823 patients with heart failure and studied them over 1-year.[2] Forty-two per cent of these patients were in New York Heart

Association (NYHA) functional class III or IV. In the entire study cohort, the 1-year all-cause mortality was 16.5% (95% CI 15.4-17.6). However, patients residing in Africa had a higher mortality of 34% (95% CI 30.2-37.4) within one year of follow-up.[3] The high mortality rate calls for new, innovative strategies such as predictive models, that aim to risk-stratify patients with heart failure, ensuring that high-risk individuals are identified early and subjected to interventions that delay disease progression.

Risk Prediction Models

Before we describe what predictive models are, we will start by defining the verb "predict" and the noun "model." The Oxford English dictionary defines "predict" as "stating that a specified event will happen in the future." A model is defined as a "simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions."[4]

Predictive models are built using both statistical methods and machine learning techniques. However, a paradigm shift has occurred, where machine learning algorithms are becoming popular because of their ability to take into consideration complex associations between clinical parameters. The sophisticated mathematical formulas used in machine learning for automated medical image interpretation have also led to increased confidence in using machine learning algorithms in the health sector. Further, a wider variety of machine learning algorithms are available for classification and regression tasks, allowing the user to handpick the best performing model.

Heart failure predictive models created with statistical methods

A systematic review by Rahimi and colleagues identified 64 multivariate statistical models used in heart failure patients published between January 1995 and March 2013.[5] Another systematic review by Di Tanna et al. identified 58 multivariate statistical models predicting outcomes in patients with heart failure, published between March 2013 and May 2018.[6] Seventeen of the models identified by Di Tanna et al. predicted all-cause mortality, nine predicted cardiovascular death, and the rest predicted heart failure hospitalisation and composite endpoints.[6] The next section will describe techniques used to create multivariate linear and logistic regression models.

Linear Regression

Linear regression (LR) is used to predict a numeral or continuous outcome. In simple linear regression, the dependent variable is featured in the y-axis and the independent variable in the x-axis (**Fig. 1**). The dependent variable only changes when there is an adjustment of the values in the x-axis. A scatterplot is used to establish a relationship between the dependent and independent variables. If a linear relationship exists between the variables in the x and y-coordinates, a correlation coefficient is calculated. A best-fitting straight line, the regression line is then plotted through the points.[7] The straight line is given by the equation $Y=a + \beta * X$, where Y is the dependent variable; a: the y-intersect of the line; β: the slope of the line and X the explanatory or independent variable.[7]

Multiple linear regression analysis involves adding two or more predictor variables concurrently and adjusting their respective regression coefficients.[7] This statistical method is used to determine the relative contribution of *each* of the predictors. The formula for multiple linear regression is $Y= \beta_0 + \beta_1X_1+ \beta_2X_2+ \beta_3X^3.... \beta_kX_k+\varepsilon$, where Y is the dependent variable, X is the

independent variable, β_0 is the Y-intercept, ε the residual and β_k the slope coefficient for each predictor variables.[8]

Logistic Regression

Logistic regression predicts a binary outcome by estimating the probability of an output belonging to a particular class. When building a predictive model using logistic regression, the first step is to utilise the univariable analysis to discover the unadjusted association between each variable and the predicted class.[9] Categorical and continuous variables, usually with a p-value less than 0.25, are selected and included in the multivariable analysis.[10]

The second step entails placing all the selected variables with a p-value less than 0.25 in the multivariable logistic regression model. In the final multivariable regression model, variables with a p-value more than 0.05 are subsequently eliminated. In the third step, smoothed scatter plots are created to assess the relationship between continuous variables and the outcome on the logit scale. Step four entails evaluating for potential interactions between the selected predictor variables.

Cox regression

The Seattle Heart Failure (SHF) model is one of the best performing models built using data from 1 125 heart failure patients. A Cox regression model was used to predict 1-year survival. The predicted versus actual 1-year survival rates in the derivation cohort were 73.4% versus 74.3%. In the five validation cohorts, the predicted versus actual survival rates were 90.5% versus 88.5%, 86.5% versus 86.5%, 83.8% versus 83.3%, 90.9% versus 91.0%, and 89.6%

versus 86.7%.[11] The data used to develop the model was obtained from a registry of carefully screened patients from the Prospective Randomized Amlodipine Survival Evaluation (PRAISE) trial, possibly limiting generalizability in the wider population. The online SHF risk calculator may be cumbersome for use in a busy practice as it requires imputation of 24 clinical and laboratory parameters when estimating the risk of one and five-year mortality.

Cox proportional hazard is a regression model used to measure an association between survival time and predictor variables. Patients are tracked from the time they enter the study, until the occurrence of an event. An event is defined as the occurrence of a disease, relapse from therapy, discharge from hospital or death. Multiple predictors or covariates of survival are added to the model to estimate their influence on survival, analogous to a multivariate regression model. The Cox proportional hazard model informs about ways in which survival differs between the groups of patients studied by estimating the hazard ratio, unlike the Kaplan-Meier survival curve, which only assesses whether survival between groups differs (log-rank test with a p-value < 0.05).

The Cox proportional hazard model, as the name implies, assumes that the covariates remain constant over time. This assumption has been challenged by Giolo et al., who suggested that heart failure prognostication models should be built using covariates with time varying-effects.[12] In their study, five hundred patients with heart failure were followed between 2002 and 2006. Survival was analysed by using Cox proportional hazard, variations of the Cox's model and the Aalen's additive model. As covariates, they used the patients' age, serum sodium levels, haemoglobin, serum creatinine and the left ventricular ejection fraction (LVEF). High

haemoglobin and high LVEF were both associated with a lower risk of mortality, while the impact of age and sodium remain constant over time.[12]

Heart failure predictive models created with machine learning techniques

The fundamental principle of machine learning is learning from data and predicting an output, where y=f(x). The input (x) is any group of features, commonly known as variables that are used to predict an output (y). A machine is given a task of learning under any of the following conditions: supervised, unsupervised, semi-supervised and through reinforcement learning.[13] In supervised learning, a computer has a task of learning from a dataset with a labelled output or outcome. For example, in a dataset comprising of patients with heart failure, we may include a column header, mortality as the output. Within the column "mortality", we then capture whether the patient demised or not by adding yes or no, respectively.

Unsupervised learning takes place when a computer learns from unlabelled datasets. The data is subsequently clustered into similar categories. Ahmad et al. used unsupervised learning in a large clinical data set of more than 40 000 Swedish heart failure patients to separate patients into four groups.[14] After one year of follow up, there was a marked difference in outcomes per cluster, with 1-year survival rates of 69%, 77%, 92% and 93%.[15]

In contrast, when the same heart failure patients were categorised based on the LVEF, only slight differences in 1-year survival (80%, 81%, 83% and 84%) were noted.[15] Semi-supervised learning occurs when learning takes place using a combination of labelled and unlabelled datasets.[13] In reinforcement learning, a computer learns by trial and error by performing tasks and

subsequently receiving penalty and reward scores for inaccurate and accurate responses, respectively.[16]

How are predictive models built with machine learning techniques?

Prior to building a predictive model using machine learning, the data must be cleaned. When cleaning continuous variables, one can create box plots and verify values of all the outliers. The next step is to train the model. Training involves splitting the dataset, where the majority of the dataset (e.g. 70%) is used to train the model. During the training process, the selected algorithm learns patterns in the dataset. The remaining dataset is used for testing and validating the model. During the test phase, the labelled output variable, for example, "mortality" is hidden, and the machine learning algorithm attempts predicting whether the patient demised or not.

Numerous supervised learning algorithms are available and are selected based on the task at hand, be it for classification of a binary outcome or regression analysis of continuous features. Some of the commonly used machine learning algorithms for classification or regression analysis include decision trees, random forests, support vector machines (SVM), K-Nearest Neighbours (KNN) and Naïve Bayes (NB).

Decision Trees

A decision tree is a simple tree-shaped algorithm used to predict a categorical (binary) or numerical outcome. To illustrate how a decision tree classifies data with a categorical outcome, we will use data from patients with heart failure to build a model predicting mortality, as shown in **Fig. 2.** Each decision tree comprises of branches, leaf and root nodes.[17] The root node is the

most superiorly located node, and the leaf or terminal node is the final node that carries the decision. Each branch of the tree represents a decision, occurrence or reaction. Data is then partitioned into subsets that contain instances with similar values by using mathematical equations such as the Gini index, Chi-square, information gain and a reduction in variance.[18]

Based on the known predictors of mortality in heart failure patients, the decision tree algorithm can categorise a patient with heart failure into class A (high risk) and class B (low risk) (**Fig.2**). Decision trees require minimal data preparation before building a predictive model, and their performance is not affected by non-linear data distribution. The main disadvantage is overfitting, which occurs when a model demonstrated high accuracy levels during training and makes inaccurate predictions during testing, or has significantly lower test accuracy than the training accuracy.[19]

Random Forest

A random forest algorithm is an ensemble decision trees that operates by constructing multiple decision trees, therefore creating a forest. The first step when building a classification model using a random forest algorithm is to select random samples from a given dataset. A decision tree is then constructed for each sample. Predicted results are obtained from each decision tree. Lastly, the most voted class from the predictions made by each decision tree is selected as the final predicted class.[20] In the case of a regression model, where the output class is numerical, the mean or average value of the predicted output is used. The random forest algorithm is ideal for handling datasets with missing values. It also performs well on large datasets and can rank

features in the order of importance. The main disadvantage of the random forest algorithm is that it is computationally expensive, requiring more training time than other algorithms.[19]

Support Vector Machines

Support Vector Machines (SVM) perform classification and regression analysis for both linear and non-linear data. The main advantage of SVM lies in their ability to classify more than two classes within the same dataset simultaneously. Support vector machines categorise data by focusing on the observations on the edge of each class (**Fig. 3**). A line or hyperplane is then placed between the classes. The maximum distance between the classes (support vectors) and the hyperplane is then chosen.[21] In instances where the data cannot be linearly separable, the data is transformed using a kernel function from low dimension to high dimensional structure, rendering the data separable.

K-Nearest Neighbours

The K-Nearest Neighbours (KNN) algorithm does not require training. It is used for predicting a binary or continuous output. Data is separated into clusters, and the number of nearest neighbours is specified by stating the value of "K", a constant.[13] If K=3, three nearest neighbours are selected based on the distance measured between the new data point and nearest neighbours. If K=6, the distance between the new data and the six nearest neighbours is calculated. The new data point is then placed in the class with the majority of votes. **Fig. 4** depicts how a KNN algorithm classifies a new data point. The biggest challenge with using K-Nearest Neighbours lies in deciding on the optimal number of neighbours to consider when classifying a new data point. K-Nearest Neighbours also do not perform well in an imbalanced

dataset, since the model will give preference to the class with the higher number of observations.[19]

Naïve Bayes

Naïve Bayes is a probabilistic method borrowed from statistics, based on the Bayes' theorem, which describes the probability of an event occurring based on previous knowledge of the conditions associated with that particular event.[22] For classification tasks, a model learns the probabilities of an object for belonging to a specific class. The name "naïve" is derived from the fact that a Naïve Bayes model assumes that the occurrence of a particular feature is not dependent on the occurrence of other features.[23] Pakhomov et al. used a Naïve Bayes classifier and another classification algorithm Perceptron, to train models capable of identifying clinical notes of patients with a diagnosis of congestive cardiac failure.[24] The Naïve Bayes classifier yielded better recall on positive samples (95 vs 86%) but had poor accuracy (57 vs 65%).[24] Although the Naïve Bayes classifier is more comfortable to implement and has a lesser training period, its main limitation is the assumption that all features are mutually independent. For example, in the study by Pakhomov et al., the terms "heart" and "failure" were considered to occur independently of each other, rendering the classifier unattractive for text categorisation.[24]

***Artificial Neural Networks** (ANNs)*

Artificial Neural Networks (ANNs) resemble neurons, the functional unit of the central nervous system in humans. Each ANN is characterised by an input layer, a hidden layer and an output layer.[25] The input layer represents features extracted by the model. A feature can be any continuous or categorical variable, such as the left ventricular ejection fraction, weight or gender.

The features in the input layer are used to predict an output, represented by one or more categories. The nodes displayed in the input layer communicate with each node in the hidden layer (**Fig. 5**). Initially, an arbitrary value between zero and one is used to grade the strength of the connection between the input layer and the hidden layer. The weighted sum of the signal from the input layer is then passed through the activation function (**Fig. 7**). The sigmoid function transforms the weighted sum of each input with a negative value to a value close to zero. Positive values are transformed into a value close to one. Most neural networks are now designed with more than one hidden layer, increasing their capacity to classify non-linear data.

1

Kwon et al. studied 2 165 patients with acute heart failure to predict in-hospital mortality, 1-year and 3-year risk of mortality using deep neural networks, random forest, logistic regression, support vector machines and a Bayesian network. Deep neural networks had the largest area under the receiver operating characteristics curve (AUC) of 0.880, 0.782 and 0.813 for predicting in-hospital, 1-year and 3-year risk of mortality, respectively.[26]

Model performance

Models predicting a binary outcome

The AUC is the most common parameter used to evaluate the performance of classification models.[27] Models with good discriminatory abilities have larger AUC, generally greater than 0.70, whereas those with AUC less than 0.50 lack discriminatory abilities.[27] Model performance is also evaluated with a confusion matrix, allowing one to calculate accuracy, precision, sensitivity and specificity. Other performance metrics include the F-measure or F-score, geometric mean and logarithmic loss.[28]

Models predicting a continuous outcome

The overall fit of a linear regression model is evaluated with the r-squared (r^2). Simply put, the r^2 is a measure of how much the prediction error is reduced, relative to how much potential error there is.[29] The error is estimated with indices such as the mean absolute error, mean squared error and the root mean squared error.[30]

Model flexibility

Model overfitting and underfitting are the most common problems encountered when evaluating performance. Overfitting occurs when a model shows high accuracy scores during training and low accuracy scores during validation. Model overfitting is minimised by adding more data to the training set, reducing the number of layers in the neural network and through dropped out nodes, which are nodes that are prevented from contributing in the prediction model.

1

Underfitting occurs when the model fails to classify data or make predictions during the training phase. A low accuracy score and a high loss identify a model that is underfitting.[31]

Limitations of statistical and machine learning predictive models

Despite the existence of numerous models predicting outcomes in patients with heart failure, only a few are easily accessible online as risk score calculators.[11, 32, 33] Although the lack of external validation of the predictive models is the driving factor for the lack of availability, several inherent limitations exist that are responsible for the rather unhurried integration of predictive models into clinical practice.

Some clinicians find risk calculation cumbersome.[34] Also, the majority of risk prediction models were designed before new heart failure drugs that further reduce mortality and heart failure hospitalisations were discovered, in addition to and in some cases replacing prior standard heart failure therapies.[35, 36]

1

Class imbalance refers to the disproportionality between the classes of data used to train the predictive model,[37] a common problem that is not unique to medical data. When the training data with the negative outcome (e.g. dead) has significantly fewer observations compared to the class (e.g. alive) with the majority of observations, the classification algorithm is inclined to favour the majority class. This poses problems as the minority class, which bears the outcome of interest, will have a low accuracy score. Fortunately, it is possible to address the class imbalance problem by manipulating data, algorithm or both.[37]

Conclusion

Both machine learning techniques and statistical methods should be employed when creating predictive models. Prior to clinical application, the best performing model should be externally validated in a cohort of patients not used for model derivation.

Turnitinnarrative.docx

ORIGINALITY REPORT

84%	84%	82%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	link.springer.com Internet Source	68%
2	Dineo Mpanya, Turgay Celik, Eric Klug, Hopewell Ntsinjana. "Machine learning and statistical methods for predicting mortality in heart failure", Heart Failure Reviews, 2020 Publication	16%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

turnitinsystematic-1.docx

by Dineo Mpanya

Submission date: 02-Feb-2023 01:07PM (UTC+0200)
Submission ID: 2004719402
File name: turnitinsystematic.docx (90.14K)
Word count: 2548
Character count: 15173

Predicting mortality and hospitalization in heart failure using machine learning: a systematic literature review

Abstract

Objective: The partnership between humans and machines can enhance clinical decisions' accuracy, leading to improved patient outcomes. Despite this, the application of machine learning techniques in the healthcare sector, particularly in guiding heart failure patients' management, remains unpopular. This systematic review aims to identify factors restricting the integration of machine learning derived risk scores into clinical practice when treating adults with acute and chronic heart failure.

Methods: Four academic research databases and Google Scholar were searched to identify original research studies where heart failure patient data were used to build models predicting all-cause mortality, cardiac death, all-cause and heart failure-related hospitalization.

Results: Thirty studies met the inclusion criteria. The selected studies' sample size ranged between 71 and 716 790 patients, and the median age was 72.1 (interquartile range: 61.1-76.8) years. The minimum and maximum area under the receiver operating characteristic curve (AUC) for models predicting mortality were 0.48 and 0.92, respectively, with the random forest algorithm achieving the highest AUC. Models predicting hospitalization had an AUC of 0.47 to 0.84. Nineteen studies (63%) used logistic regression, 53% random forests, and 37% of studies used

decision trees to build predictive models. None of the models were built or externally validated using data originating from Africa or the Middle-East.

Conclusions: The variation in the aetiologies of heart failure, limited access to structured health data, distrust in machine learning techniques among clinicians and the modest accuracy of existing predictive models are some of the factors precluding the widespread use of machine learning derived risk calculators.

1. Introduction

Predictive analytics is applied across many industries, typically for insurance underwriting, credit risk scoring and fraud detection (Boodhun, 2018, Carcillo, 2018, Moradi, 2019). Both statistical methods and machine learning algorithms are used to create predictive models (Mpanya, 2020). In heart failure, machine learning algorithms create risk scores estimating the likelihood of a heart failure diagnosis and the probability of outcomes such as all-cause mortality, cardiac death and hospitalization (Kwon, 2019, Desai, 2020, Lorenzoni, 2019, Blecker, 2018, Hsich, 2011, Adler, 2020, Chicco, 2020, Golas, 2018, Angraal, 2020).

Clinicians treating heart failure patients may underestimate or overestimate the risk of complications and may battle with dose titration when prescribing oral medication (Gheorghiade, 2012, Hanratty, 2002a). Despite these challenges, risk calculators are still not widely used to guide the management of heart failure patients. Most clinicians find risk calculation time consuming and are not convinced of the value of the information derived from predictive models (Hanratty, 2002b, Eichler, 2007). Moreover, the lack of integration of risk scores predicting heart failure outcomes into management guidelines may diminish clinicians' confidence when using risk calculators. Also, clinicians tend to question the integrity of unsupervised machine learning and deep learning methods since algorithms single-handedly select features (*predictors*) without human input.

Machine learning and its subtype, deep learning, have shown an impressive performance in medical image analysis and interpretation (Lakhani, 2017). Convolutional neural networks (CNN) were trained to classify chest radiographs as

pulmonary tuberculosis (TB) or normal using chest radiographs from 685 patients. The ensemble of CNN's performed well with an area under the receiver operating characteristic curve (AUC) of 0.99 (Lakhani, 2017). The impressive results have resulted in the commercialization of chest x-ray interpretation software-generated using deep learning (Khan, 2020). The availability of such software can play a critical role in remote areas with limited or no access to radiologists, as CNN can potentially identify subtle manifestations of TB on chest radiographs, leading to prompt initiation therapy, curbing further transmission of TB. Amid these capabilities, the uptake of machine learning techniques in the healthcare sector remains limited. The aim of this systematic review is to identify models predicting mortality and hospitalization in heart failure patients and discuss factors that restrict the widespread clinical use of risk scores created with machine learning algorithms.

2. Methods

2.1 Search strategy for identification of relevant studies

A systematic literature search was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Literature searches were conducted in MEDLINE, Google Scholar, Springer Link, Scopus, and Web of Science. The search string contained the following terminology: (Mortality OR Death OR Readmission OR Hospitalization) AND (Machine Learning OR Deep Learning) AND (Heart Failure OR Heart Failure, Diastolic OR Heart Failure, Systolic).

2.2 Review methods and selection criteria

Studies reported in languages other than English were not included. A single reviewer screened titles, abstracts and full-text articles and made decisions regarding potential eligibility. Studies were eligible if they reported models predicting all-cause or cardiac mortality or all-cause or heart failure-related hospitalization in heart failure patients. Models included in the study were created using machine learning algorithms and/or deep learning. We did not include studies using solely logistic regression for a classification task. Logistic regression analysis is a machine learning algorithm borrowed from traditional statistics. When logistic regression is used as a machine learning algorithm, the algorithm is initially trained to identify clinical data patterns using a dataset with labelled classes, a process known as supervised learning. After that, the logistic regression algorithm attempts to classify new data into two or more categories based on "posteriori knowledge."

2.3 Data extraction

The following features were extracted: study region, data collection period, sample size, age, gender, cause of heart failure (ischaemic vs non-ischaemic), predictor variables, handling of missing data, internal and external validation, all-cause mortality and cardiovascular death rate, all-cause hospitalization rate and performance metrics (sensitivity, accuracy, AUC or c-statistics and F-score).

3. Results

3.1 The review process

The initial search yielded 1 835 research papers. After screening titles and abstracts, 1 367 did not meet the inclusion criteria. Excluded papers were predominantly theoretical reviews and conference papers in the field of computer

science. Two hundred and sixty full-text articles were assessed for eligibility. A further 230 studies were excluded, leaving thirty papers legible for analysis (**Fig. 1**). Reasons for excluding 230 studies are provided as **supplementary data**.

3.2 Characteristics of the included studies

The source of data in the majority of the studies were electronic health records (EHR) (*n*=16), followed by claims data (*n*=5), trial data (*n=3),* registry *(n=3)* and data obtained from research cohorts (*n=3).* Data was collected from hospitalized patients in twelve studies. The sample size in the predictive models ranged between 71 and 716 790, with the smallest sample size used to predict survival in patients with advanced heart failure managed with second-generation ventricular assist devices (Kourou, 2016). Within the 30 studies, ten studies created models predicting mortality. Another ten studies predicted hospitalization, and five studies predicted both mortality and hospitalization. The data used to create predictive models was collected between 1993 and 2017 (**Table 1).** Of the 30 included studies, 22 included data originating from North America, seven from Asia and six from Europe. There were no studies conducted in Africa or Middle-East (**Fig. 2).**

3.3 Clinical characteristics of patients with heart failure

The majority of studies reported the patients' age (93%) and gender (87%). The median age was 72.1 (61.1-76.85) years. Between 14.0-83.9% of the extracted studies' participants had ischaemic heart disease (**Table 2**). In total, 30% of studies mentioned Black patients. Between 0.95% to 100% of the individuals were Black, with one study enrolling only African American males with heart failure (McKinley, 2019).

3.4 Machine learning algorithms

Only eight (27%) studies used a single algorithm to build a predictive model. Nineteen studies (63%) used logistic regression, 53% random forests, and 36% of studies used decision trees to create predictive models. The rest of the algorithms are depicted in **Fig.3**.

3.5 Predictors

Twelve (36.4%) studies did not report the number of predictors or features used. The number of predictors in the studies extracted was between 8 and 4 205. Some authors only mentioned the number of predictors and did not list them. Age, sex, diastolic blood pressure, left ventricular ejection fraction (LVEF), estimated glomerular filtration rate, haemoglobin, sodium and blood urea nitrogen were some of the predictors of mortality identified in the extracted studies (Adler, 2020, Angraal, 2020, Chicco, 2020). In contrast, ischaemic cardiomyopathy, age, LVEF, hypotension, haemoglobin, creatinine, and potassium serum levels were features identified as predictors of hospitalization (Lorenzoni, 2019).

3.6 Model development, internal and external validation

When creating a predictive model using machine learning, data is generally partitioned into three or four datasets. In the studies extracted, between 60-80% of the data was used for training models, while the rest was used for testing and/or internally validating the models. Although the data on model validation was scanty, external validation was explicitly mentioned in two studies. None of the models were externally validated using data originating from Africa or the Middle-East.

3.7 Model performance and evaluation metrics

Parameters used to evaluate model performance were the confusion matrix, reporting sensitivity, specificity, positive and negative predictive value, accuracy, and precision. Most studies also reported the f-score, AUC, concordance statistic (C-statistic), and recall. The minimum and maximum AUC for models predicting mortality were 0.477 and 0.917, and models predicting hospitalization had an AUC between 0.469 and 0.836.

4. Discussion

This systematic review highlights several factors that restrict the use of risk scores created with machine learning algorithms in the clinical setting. The existence of clinical information with prognostic significance such as clinical signs, symptoms and functional status decoders such as the New York Heart Association class in the free-text format in EHR systems may result in models with low predictive abilities if such critical data is omitted. Fortunately, newer emerging techniques such as bidirectional long short-term memory with a conditional random fields layer have been introduced to remedy the problem of free-text in EHR (Jagannatha, 2016a, Jagannatha, 2016b).

Risk scores derived from heart failure patients residing in North America or Europe may not be suitable for application in LMIC. In these regions, the predominant cause of heart failure is ischaemic heart disease (IHD), whereas, in sub-Saharan Africa, hypertension is still the leading cause of heart failure (Agbor, 2018). Also, healthcare services' availability and efficiency differ significantly between countries, suggesting

that algorithms trained using data from the high-income countries (HIC) should be retrained using native data before adopting risk calculators.

Despite the endemicity of heart failure in LMIC, risk scores derived from patients residing in LMIC are scanty or non-existent. The lack of EHR systems, registries, and pooled data from multicentre studies is responsible for the absence of risk scores derived from patients in LMIC. If digital structured health data were available in LMIC, models predicting outcomes could be created instead of extrapolating from studies conducted in HIC. As a result of unstructured health data in LMIC, there was an underrepresentation of this population in the training and test datasets.

The AUC was one of the most commonly reported performance metric in the extracted studies. The highest AUC for models predicting mortality was 0.92, achieved by the random forest algorithm in a study by Nakajima et al., where both clinical and physiological imaging data were used to train algorithms (Nakajima, 2020). A model with an AUC equal to or below 0.50 is unable to discriminate between classes. One might as well flip a coin when making predictions. Some of the reasons for the modest performance metrics demonstrated by machine learning algorithms include a training dataset with multiple missing or too few features, absence of ongoing partnership between clinicians and data scientists and class imbalance. In most instances, when handling healthcare data, the negative class tends to outnumber positive classes. The learning environment is rendered unfavourable since there are fewer positive observations or patterns for an algorithm to learn from. For example, when predicting mortality, the class with patients that demised is frequently smaller than the class with alive patients.

Models with perfect precision and recall have an F-measure, also known as the F-Score or F1 Score, equal to one (Hripcsak, 2005). Sensitivity, also known as recall, measures a proportion of positive classes accurately classified as positive (Saito, 2015). Machine learning algorithms in the extracted studies had a sensitivity rate between 7.2 and 91.9%. The low sensitivity, reported by Turgeman and colleague, improved to 43.5% when they used an ensemble method to combine multiple predictive models to produce a single model (Turgeman, 2016).

Although the random forest algorithm appeared to have the highest predictive abilities in most studies, one cannot conclude that it should be the algorithm of choice whenever one attempts to create a predictive model. The random forest algorithm's main advantage is that it is an ensemble-based classifier that takes random samples of data and exposing them to multiple decision tree algorithms. Decision trees are intuitive and interpretable and can immediately suggest why a patient is stratified into a high-risk category, hence guiding subsequent risk reduction interventions, a significant advantage in contrast to deep learning methodologies such as artificial neural networks with a “black box” nature. Once random samples of data have been exposed to multiple decision tree algorithms, the decision trees' ensemble identifies the class with the highest number of votes when making predictions. Random forests also perform well in large datasets with missing data, a common finding when handling healthcare data, and can rank features (*predictors*) in the order of importance, based on predictive powers (Uddin, 2019).

Predictors of mortality identified by machine learning algorithms in the extracted studies were explainable and included features such as the LVEF, hypotension and blood urea nitrogen levels and age. Whether these predictors can be considered significant risk factors for all heart failure, irrespective of genetic makeup, is debatable. The youngest patient in the studies reviewed was 40 years old, but most of the patients included in the predictive models were significantly older, with a median age of 72 years. Risk scores derived from older patients may reduce the applicability of the existing risk calculators in the sub-Saharan African (SSA) context, considering that patients with heart failure in SSA are generally a decade younger (Bloomfield, 2013).

Geographically unique heart failure aetiologies and diverse clinical presentations call for predictive models that incorporate genomic, clinical and imaging data. We recommend that clinicians treating heart failure patients focus on establishing structured EHR systems and comparing outcomes such as mortality and hospitalization in patients managed with and without risk scores. Each centre should also integrate risk calculators with EHR systems without relying on other centres' risk calculators. Clinicians without access to EHR systems should carefully study the cohort used to create risk scores before implementing risk scores in their clinical practice.

5. Limitations

This systematic literature review has several limitations. The systematic literature search was conducted by a single reviewer, predisposing the review to selection bias. We only included original research studies published after 2009. The rationale

for including studies published in the past 11 years was to avoid including studies where rule-based expert systems were used instead of newer machine learning techniques. While the data used to create predictive models was heterogeneous, a meta-analytic component as part of the review would have provided a broader perspective on machine learning algorithms' performance metrics when predicting heart failure patient outcomes.

6. Conclusion

The variation in the aetiologies of heart failure, limited access to structured health data, distrust in machine learning techniques among clinicians and the modest accuracy of predictive models are some of the factors precluding the widespread use of machine learning derived risk calculators.

turnitinsystematic-1.docx

ORIGINALITY REPORT

85%	82%	85%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Dineo Mpanya, Turgay Celik, Eric Klug, Hopewell Ntsinjana. "Predicting mortality and hospitalization in heart failure using machine learning: A systematic literature review", IJC Heart & Vasculature, 2021 Publication	56%
2	www.ncbi.nlm.nih.gov Internet Source	29%
3	doaj.org Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

turnitincluster.docx

by Dineo Mpanya

Submission date: 20-Sep-2022 03:02PM (UTC+0200)
Submission ID: 1904472882
File name: turnitincluster.docx (333.25K)
Word count: 4195
Character count: 23189

The diagnosis and therapy of heart failure are fundamentally determined by a single imaging parameter, the left ventricular ejection fraction (LVEF). The study aims to use unsupervised machine learning algorithms to phenotype heart failure patients into various clusters using multiple clinical parameters instead of LVEF solely. Seven unsupervised machine learning clustering algorithms were used to cluster heart failure patients hospitalised with acute and chronic heart failure. The study population consisted of 636 heart failure patients with a mean age of 55.6 ± 16.9 years and 352 (55.3%) males. The baseline median LVEF was 32% (interquartile range: 25 - 45), and the rate of in-hospital all-cause mortality was 14.5%. Cluster 1(*modest ventricular dysfunction cluster)* comprised of 229 (36.0%) patients with a mean age of 56.2 ± 17.2 years. Patients in Cluster 1 had a relatively higher mean LVEF of 45 ± 17.6, (p<0.001), a smaller mean left ventricular internal diameter end-systole (LVIDs) of 4.1 ± 1.3 cm (p <0.001) and mean left ventricular internal diameter end-diastole (LVIDd) of 5.2 ± 1.1 cm. Cluster 2 (*hypotensive cluster)* comprised of 117 (18.4%) patients with a minimum systolic and diastolic blood pressure of 91 and 60 mmHg, respectively. In cluster 3 (*congestive cluster),* there was a predominance of patients with symptoms of fluid overload, and 93 (64.6%) patients had ascites. Systolic and diastolic blood pressure, age, and the LVEF, had the strongest impact on discriminating between the three clusters**.** Clinicians without access to echocardiography could potentially rely on blood pressure measurements and age to risk stratify heart failure patients. However, larger prospective studies are mandatory for the validation of these clinical parameters.

4.6

Heart failure is a clinical syndrome characterised by a constellation of symptoms and signs of fluid overload mostly in a patient with an underlying cardiovascular disease. The global prevalence and incidence of heart failure continue to rise to the extent that heart failure has been declared a global pandemic (Ponikowski, 2014). Therefore, innovative strategies that may assist in improving the management of heart failure patients are still required.

The current approach to the diagnosis, risk stratification, management, and prognosis of heart failure is fundamentally determined by a single imaging parameter, the left ventricular ejection fraction (LVEF). The LVEF is a crude measurement that is operator-dependant, and not readily available in most hospitals in low-and-middle-income countries (LMIC) not equipped with cardiology services. Unsupervised machine learning algorithms autonomously create clusters by grouping data points with similar attributes, revealing undiscovered patterns without learning from a clearly defined outcome. A recent systematic review identified 81 articles reporting the utility of both supervised and unsupervised machine learning algorithms in heart failure, with 11 (13%) studies reporting classification or clustering of heart failure patients, 25 (31%) on the detection of the onset of heart failure, 18 (21%) on predicting rehospitalisation, 20 (26%) on predicting mortality, and the rest (6%) on predicting outcomes and response to implantable cardiac devices (Jasinska-Piadlo, 2022).

Another systematic review identified 27 studies using unsupervised machine learning for heart failure subtype definition (Banerjee, 2021). This study aims to describe baseline demographic and clinical characteristics of heart failure patients hospitalised in a tertiary-level hospital in Johannesburg, South Africa, and explore the utility of unsupervised machine learning in clustering heart failure patients using a constellation of clinical parameters, instead of a single imaging parameter, the LVEF.

Data were exported from the PMRCardio database, a Microsoft Structured Query Language Server Management Studio (version 15.0.18330.0), that stores clinical data of patients hospitalised in cardiology wards at the Charlotte Maxeke Johannesburg Academic Hospital (CMJAH) in South Africa. The CMJAH is a tertiary-level state-owned institution equipped with a cardiac intensive care unit, a cardiac catheterisation and electrophysiology laboratory and outpatient services. All acutely and chronically ill patients with cardiovascular diseases are hospitalised in dedicated cardiac wards. Their admission data is stored digitally in the PMRCardio database by registrars or residents rotating in the wards. Clinical parameters available in the dataset (features) include demographic data, clinical history and examination findings, laboratory indices, electrocardiogram (ECG), echocardiogram, angiography and in-hospital all-cause mortality status.

The tenth revision of the international classification of diseases (ICD-10) was used to select patients diagnosed with heart failure. The following ICD-10 diagnoses were used to select patients included in the analysis: "heart failure unspecified" or "congestive heart failure or "left ventricular failure", or "dilated cardiomyopathy" or "ischaemic cardiomyopathy." Only consecutive heart failure patients 18 years of age and older with a documented left ventricular ejection fraction (LVEF) hospitalised between October 2009, and December 2018 were included.

Data exported from the database were merged into a single Microsoft Excel Sheet. All retrospective patient data used in the final analysis was further prospectively verified by comparing hard copies of medical records with electronic data. Patients were subsequently stratified into three groups, based on the baseline LVEF: heart failure with reduced ejection fraction – HFrEF (LVEF < 40%), heart failure with midrange ejection fraction – HFmEF (LVEF=40 - 49%), and heart failure with preserved ejection fraction – HFpEF (LVEF ≥ 50%). Admission biochemical results were obtained from the National Health Laboratory Service (NHLS). Approval to conduct the study was received from the University of the Witwatersrand Human Research Ethics Committee (Clearance certificate number: M190515). Permission was also obtained from NHLS senior authorities.

4.7.3 Data pre-processing for clustering of heart failure patients

Packages and libraries required for data analysis were imported into a Jupyter notebook (version 6.0.3). Features with more than 20% missing values, patient identifiers and column data without variation were removed from the dataset. The remaining columns with missing data were imputed using the mean. Categorical features with missing values were ascribed with zero. The data was subsequently scaled and normalised prior to the clustering of patients.

4.7.4 Cluster analysis

Clustering is a form of unsupervised machine learning where data is categorised into groups based on similarities within data points in each cluster. Before clustering the heart failure data, the feature "mortality" that characterises the vitality status of each patient upon discharge (*dead* or *alive*) from the dataset was removed. The rationale for removing the outcome "mortality" lies behind the premise that unsupervised machine learning clusters *unlabelled* data without a clearly defined outcome.

Seven clustering algorithms were used to cluster heart failure patients a partitioning method (K-prototype and K-Means), a hierarchical method (agglomerative and Balanced Iterative Reducing and Clustering hierarchies - BIRCH), density-based (Ordering Points To Identify the Clustering Structure – OPTICS and density-based spatial clustering of applications with noise - DBSCAN), and a model-based clustering method, a Gaussian mixture model.

Distance-based algorithms use k-means, k-modes or both to estimate the distance between data points. Agglomerative clustering also referred to a "bottom-up" approach, initially treats each observation (patient) as a single cluster. The algorithm then successively agglomerates pairs of clusters until a single cluster is formed.

The DBSCAN algorithm groups data points close to each other based on the density of data points in a particular region, marking outliers as points in a low-density region. The Gaussian mixture model uses multiple Gaussian distributions to identify various clusters, with the assumption that each modal distribution represents a particular structure.

Agglomerative clustering was chosen as the best algorithm for clustering heart failure patients based on achieving the highest silhouette score. Using the ward linkage method, the agglomerative clustering algorithm identified three clusters with a silhouette score of 0.7. A silhouette score measures the goodness of fit of clusters by incorporating the distance between points within a cluster and spacing between each cluster. This score ranges between -1 and +1, where the best score, +1 suggests that clusters are far from each other and -1 infers that clusters are poorly describing patterns in the dataset. A dendrogram was also plotted to represent the clusters visually. Shapley additive explanations (SHAP) was used to identify and rank features that played a prominent role in the segmentation of the heart failure cohort into three clusters using the agglomerative clustering algorithm.

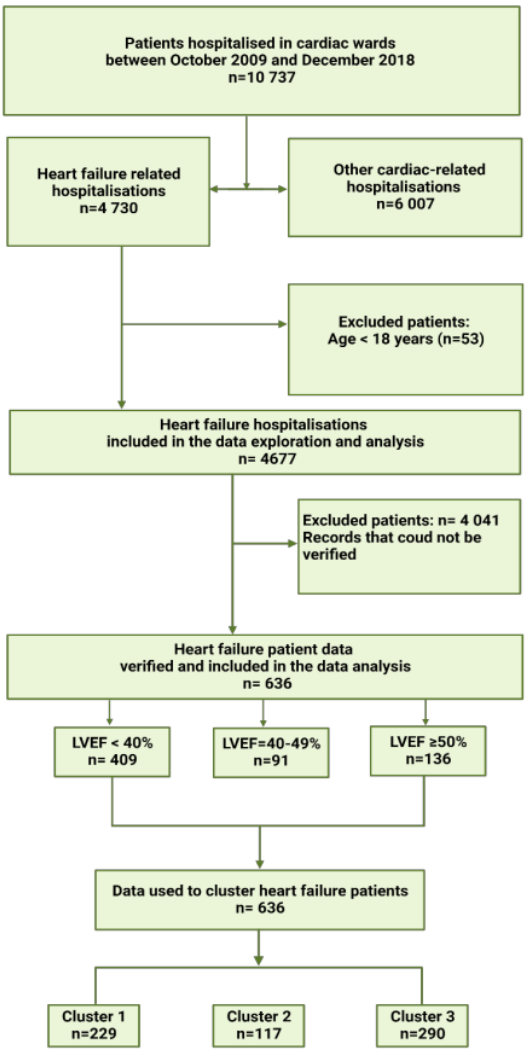


Figure 4.1 Flow chart showing heart failure patient selection for clustering

4.8 Results

Between October 2009 and December 2018, there were 10 737 admissions, of which 4 730 were heart failure related **(Figure 4.1)**. After excluding patients with incomplete data, particularly the LVEF, the final heart failure cohort comprised 636 patients, of which 352 (55.3%) were males. The mean age in the entire cohort was 55.6 ± 16.9 years, and patients with HFmEF were older with a

mean age of 58.0 ± 17.1 years. The baseline median LVEF in the entire study cohort was 32% (25 - 45), and 409 (64.3%) had a LVEF less than 40% (HFrEF), 91 (14.3%) had a LVEF between 40 and 49% (HFmEF), and 136 (21.4%) had a LVEF ≥ 50% (HFpEF). Heart failure patients had a mean systolic and diastolic pressure of 128 ± 30.1 mmHg and 82 ± 19.7 mmHg, respectively. The rest of the baseline clinical characteristics in the study cohort are depicted in **Table 4.1.**

Cluster Analysis

The agglomerative clustering hierarchical algorithm identified three clusters of heart failure patients (**Figure 4.2**). Cluster 1(*modest ventricular dysfunction cluster)* comprised of 229 (36.0%) patients with a mean age of 56.2 ± 17.2 years. Patients in Cluster 1 had a relatively higher mean LVEF of 45 ± 17.6, (p<0.001), a smaller mean left ventricular internal diameter end-systole (LVIDs) of 4.1 ± 1.3 cm (p <0.001) and mean left ventricular internal diameter end-diastole (LVIDd) of 5.2 ± 1.1 cm. The mean urea was 14.5 ± 31.3, while the median urea in the entire study population was 8.5 (6.0 – 12.5) **(Table 4.2)**.

Cluster 2 (*hypotensive cluster)* comprised of 117 (18.4%) patients with a mean age of 54.3 ± 15.9 years. Patients in cluster 2 had a lower mean systolic and diastolic blood pressure of 117 ± 26.5 (p=0.0135) and 76 ± 16.8 mmHg, respectively. The minimum systolic and diastolic blood pressure was 91 and 60 mmHg, respectively. In cluster 3 (*congestive cluster)* there was a predominance of patients with symptoms of fluid overload, and 93 (64.6%) patients had ascites. The post-hoc analysis for all statistically significant values are depicted in **Table 4.3 and 4.4**.

The SHAP values were used to estimate the contribution of each feature in the clustering of heart failure patients. Both low and high values of systolic blood pressure, age, the LVEF, and diastolic blood pressure were the top four features that had an impact on discriminating between the three clusters **(Figure 4.3)**. The mean SHAP value of the systolic blood pressure (SHAP value=0.36) was approximately 2.6 times higher than that of the LVEF (SHAP value=0.14).

Outcomes

Among the 636 patients hospitalised with heart failure, 92 patients died. The rate of in-hospital all-cause mortality was 14.5%. The rate of death was the highest in patients with HFrEF (63%), followed by patients with HFpEF (19.6%) and HFmEF (17.4%). The median length of stay in the hospital in the entire heart failure cohort was 6 (4 -11) days, and it did not differ based on the LVEF category (p=0.4441).

Cluster analysis revealed a higher mortality rate (39.1%) in patients in cluster 3 (*congested cluster)*, followed by cluster 2, the *hypotensive cluster* (38%) and cluster 1 (22.8%), the *modest ventricular dysfunction cluster*. Patients in cluster 3 were hospitalised for 9.0 ± 7.8 days, whereas patients in cluster 1 and 2 were both hospitalised for a mean duration of 9.6 ± 8.9 and 8.6 ± 7.7 days, respectively.

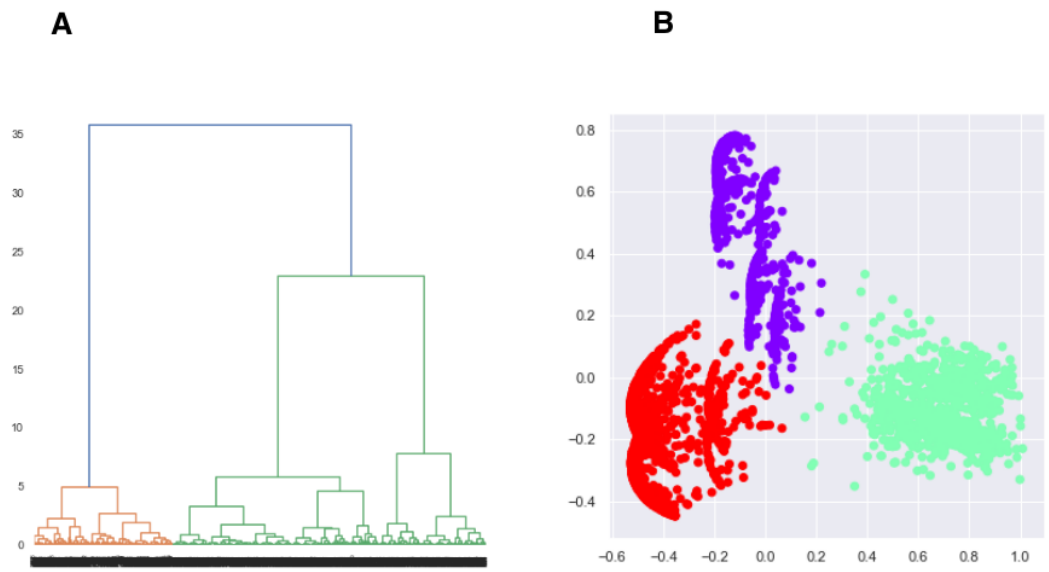


Figure 4.2 Clusters of heart failure patients identified by agglomerative clustering. **A:** Dendrogram showing three clusters. Cluster 1 (orange), cluster 2 and 3 (green). Three clusters were selected based on the silhouette score of 0.7, which is a measure of goodness of fit of patients into each cluster. **B:** A two-dimensional scatterplot of the three clusters of heart failure patients identified by the agglomerative clustering algorithm

Table 4.1 Baseline demographic and clinical characteristics of the study patients stratified according to the left ventricular ejection fraction

Clinical Parameters	**All patients (n=636)**	**HFrEF (n=409)**	**HFmEF (n=91)**	**HFpEF (n=136)**	**p-value**
Age (years), mean (SD)	55.6 ± 16.9	54.6 ± 16.9	58.0 ± 17.1	56.9 ± 16.4	0.1183
Male sex, n (%)	352 (55.3)	218 (61.9)	53 (15.1)	81 (23.0)	0.3720
Ethnicity, n (%)					
African	380 (59.7)	251 (66.1)	49 (12.9)	80(21.1)	0.4038
Caucasian	154 (24.2)	93 (60.4)	31 (20.1)	30 (19.5)	0.0595
Indian	72 (11.3)	46 (63.9)	8 (11.1)	18 (25.0)	0.5831
Mixed Ancestry	29 (4.6)	19 (65.5)	3 (10.3)	7 (24.1)	0.7992
Examination findings:					

Lung crepitations	392 (61.6)	251 (64.0)	60 (15.3)	81 (20.7)	0.6153
Ascites	144 (22.6)	90 (62.5)	20 (13.9)	34 (23.6)	0.7597
Cyanosed	51 (8.0)	32 (62.5)	2 (13.9)	17 (23.6)	0.7597
Hepatomegaly	249 (39.2)	173 (69.5)	36 (14.5)	40 (16.1)	0.0284
Peripheral oedema	436 (68.6)	282 (64.7)	58 (13.3)	96 (22.0)	0.5298
S3	324 (50.9)	227 (70.1)	32 (9.9)	65 (20.1)	0.0015
JVP elevated	493(77.5)	326(66.1)	69(14.0)	98(19.9)	0.1653
Pulse	100 ± 24.3	102 ± 25.1	101 ± 16.8	94 ± 25.1	0.1324
Systolic BP, mmHg	128 ± 30.1	125 ± 28.8	136 ± 35.7	135 ± 27.5	0.0002
Diastolic BP, mmHg	82 ± 19.7	83 ± 20.4	85 ± 20.5	78 ± 16.1	0.0318
Laboratory tests					
Haemoglobin, g/dL	12.9 ± 2.5	13.3 ± 2.4	12.4 ± 2.9	12.5 ± 2.8	0.0016
Sodium, mmol/L	139 ± 5.4	139 ± 5.5	138 ± 5.1	140 ± 5.2	0.1415
Potassium, mmol/L	4.4 ± 0.7	4.4 ± 0.7	4.3 ± 0.9	4.4 ± 0.8	0.5181
Urea, mmol/L	8.5 (6.0 – 12.5)	12.1 ± 16.9	16.5 ± 43.3	13.2 ± 20.8	0.2692
eGFR, mL/min	66.1 (41.2 – 85.0)	66.7 ± 27.8	73.5 ± 35.5	63.7 ± 37.3	0.4604
Echocardiography					
LVIDd, cm	5.6 ± 1.1	6.0 ± 0.9	5.3 ± 1.1	4.5 ± 1.1	<0.001
LVIDs, cm	4.6 ± 1.3	5.2 ± 1.0	4.3 ± 0.8	3.1 ± 1.3	<0.001

Data is presented as mean ± standard deviation (SD) for continuous features with a normal distribution, and median and interquartile ranges (p25-p75) when the distribution is skewed. Categorical data are represented as absolute numbers and percentages. BP= Blood pressure; eGFR= estimated glomerular filtration rate using the modification of diet in renal disease (MDRD) formula; HfrEF=Heart failure with reduced ejection fraction (LVEF < 40%); HfmEF=heart failure with midrange ejection fraction (LVEF=40 - 49%). HfpEF=heart failure with preserved ejection fraction (LVEF ≥ 50%).JVP= jugular venous pressure; LVEF= left ventricular ejection fraction; LVIDd= left ventricular internal diameter end-diastole; LVIDs= left ventricular internal diameter end-systole; S3= third heart sound

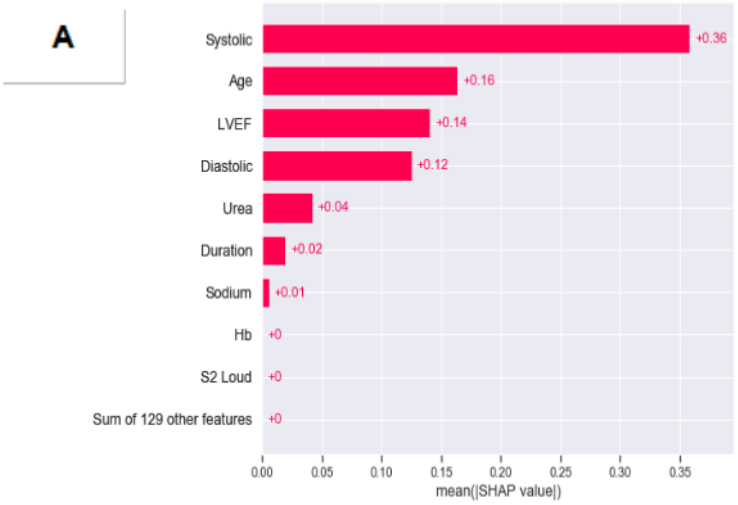
Table 4.2 Baseline demographic and clinical characteristics of the study patients stratified according to clusters.

Data is presented as mean ± standard deviation (SD) for continuous features with a normal distribution,

Clinical Parameters	All patients (n=636)	Cluster 1 (n=229)	Cluster 2 (n=117)	Cluster 3 (n=290)	p-value
Age (years), mean (SD)	55.6 ± 16.9	56.2 ± 17.2	54.3 ± 15.9	55.6 ± 16.9	0.7642
Male sex, n (%)	352 (55.3)	128 (36.4)	66 (18.8)	158 (44.9)	0.9189
Ethnicity, n (%)					
African	380 (59.7)	132 (34.7)	68 (17.9)	180(47.4)	0.5485
Caucasian	154 (24.2)	57 (37.0)	25 (16.2)	72 (46.8)	0.7286
Indian	72 (11.3)	28 (38.9)	17 (23.6)	27 (37.5)	0.2787
Mixed Ancestry	29 (4.6)	12 (41.4)	7 (24.1)	10 (34.5)	0.4468
Examination findings:					
Lung crepitations	392 (61.6)	122 (31.1)	67 (17.1)	203 (51.8)	0.0002
Ascites	144 (22.6)	23 (16.0)	28 (19.4)	93 (64.6)	<0.0001
Cyanosed	51 (8.0)	9 (17.6)	10 (19.6)	32 (62.7)	0.0122
Hepatomegaly	249 (39.2)	65 (26.1)	44 (17.7)	140 (56.2)	<0.0001
Peripheral oedema	436 (68.6)	115 (26.4)	82 (18.8)	239 (54.8)	<0.0001
S3	324 (50.9)	102 (31.5)	66 (20.4)	156 (48.1)	0.0474
JVP elevated	493 (77.5)	148 (30.0)	94 (19.1)	251 (50.9)	<0.001
Systolic BP, mmHg	128 ± 30.1	135 ± 28.6	117 ± 26.5	128 ± 26.2	0.0135
Diastolic BP, mmHg	82 ± 19.7	83 ± 17.3	76 ± 16.8	84 ± 18.8	0.7012
Laboratory tests					
Haemoglobin, g/dL	12.9 ± 2.5	12.7 ± 2.4	13.2 ± 2.3	13.1 ± 2.4	0.0986
Sodium, mmol/L	139 ± 5.4	139 ± 4.7	138 ± 4.9	140 ± 5.8	0.4697
Potassium, mmol/L	4.4 ± 0.7	4.3 ± 0.6	4.5 ± 0.8	4.4 ± 0.8	0.1063
Urea, mmol/L	8.5 (6.0 – 12.5)	14.5 ± 31.3	14.3 ± 16.0	11.2 ± 16.4	0.0917
Echocardiography					
LVEF, %	32 (25 - 45)	45 ± 17.6	31 ±13.8	33 ± 14.7	<0.001
LVIDd, cm	5.6 ± 1.1	5.2 ± 1.1	5.8 ± 1.1	5.9 ± 1.0	<0.001
LVIDs, cm	4.6 ± 1.3	4.1 ± 1.3	5.0 ± 1.0	4.9 ± 1.1	<0.001

and median and interquartile ranges (p25-p75) when the distribution is skewed. Categorical data are

represented as absolute numbers and percentages. BP= Blood pressure; eGFR= estimated glomerular filtration rate using the modification of diet in renal disease (MDRD) formula; JVP= jugular venous pressure; LVEF= left ventricular ejection fraction; LVIDd= left ventricular internal diameter end-diastole; LVIDs= left ventricular internal diameter end-systole; S3= third heart sound. Pulse, pro brain natriuretic peptide and estimated glomerular filtration rate not included in the table since the number of missing values > 20%.



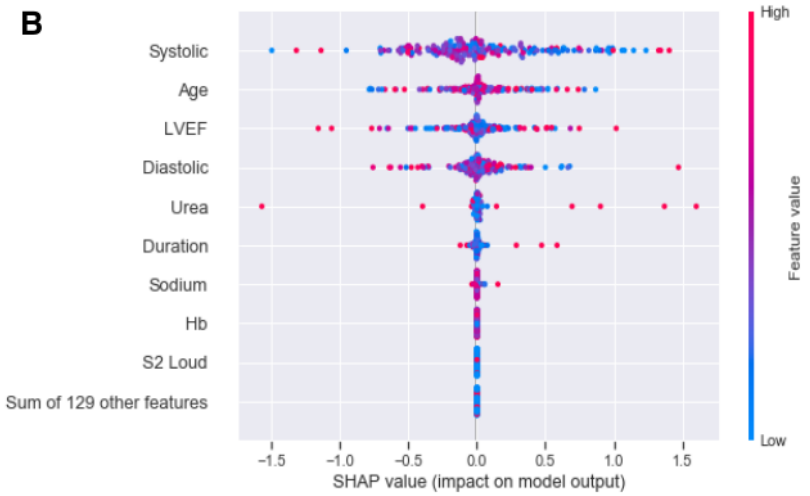


Figure 4.3 Features identified with shapley additive explanations (SHAP) that played a role in the segmentation of the heart failure cohort into three clusters using the agglomerative clustering algorithm. A: Bar plot depicting the importance of each feature. The higher the mean weight, the higher the contribution of the feature B: Beeswarm plot showing the contribution of the top four features. Each dot represents a single patient. Red and blue refers to high and low feature values, respectively. Both high and low values of the systolic and diastolic blood pressure, age and LVEF contributed towards clustering patients. Hb= Haemoglobin; LVEF=left ventricular ejection fraction; S2= second heart sound.

Table 4.3 Post hoc analysis of clinical parameters in the study patients stratified according to clusters

Clinical Parameters	**All patients (n=636)**	**Cluster 1 vs Cluster 2**	**Cluster 1 vs Cluster 3**	**Cluster 2 vs Cluster 3**
Systolic BP, mmHg	128 ± 30.1	0.001	0.0152	0.001
LVEF, %	32 (25 - 45)	0.001	0.001	0.5583
LVIDd, cm	5.6 ± 1.1	0.001	0.001	0.7457
LVIDs, cm	4.6 ± 1.3	0.001	0.001	0.9000

BP= Blood pressure; LVEF= left ventricular ejection fraction; LVIDd= left ventricular internal diameter in diastole; LVIDs= left ventricular internal diameter end-systole

Table 4.4 Post hoc analysis of clinical parameters in the study patients stratified according to the left ventricular ejection fraction

Clinical Parameters	All patients (n=636)	HFrEF vs HFmEF	HFrEF Vs HFpEF	HFmEF Vs HFpEF
Systolic BP, mmHg	128 ± 30.1	0.0049	0.0022	0.9000
Diastolic BP, mmHg	82 ± 19.7	0.6592	0.0626	0.0468
Haemoglobin, g/dL	12.9 ± 2.5	0.0142	0.0156	0.9000
LVIDd, cm	5.6 ± 1.1	0.0010	0.0010	0.0010
LVIDs, cm	4.6 ± 1.3	0.0010	0.0010	0.0010

BP= Blood pressure; LVIDd= left ventricular internal diameter in diastole; LVIDs= left ventricular internal diameter end-systole

4.9 Discussion

In this study, we used retrospective data of consecutive patients hospitalised with new-onset heart failure symptoms or decompensated chronic heart failure to describe demographic and clinical characteristics, and categorise heart failure phenotypes using unsupervised machine algorithms. The mean age in our cohort was 55.6 ± 16.9 years and 14.5% of all patients hospitalised with the diagnosis of heart failure died.

Heart failure is generally associated with poorer outcomes in LMIC, and the onset of heart failure occurs at a younger age of 53 years, affecting economic productivity in this region (Dokainish, 2016, Ajayi, 2020). The most plausible explanations for poorer outcomes include, but are not limited to the unavailability of life-prolonging therapy such as cardiac resynchronization defibrillators and heart transplant services. Furthermore, our hospital is a tertiary-level institution, implying that there might be a delay in the referral of patients and as a result, our patients were sicker. The younger age of onset may be partly related to the dominance of non-ischaemic cardiomyopathy, possibly with a genetic aetiology, considering that in sub-Saharan Africa (SSA), the most common cardiovascular diseases are still hypertensive heart disease, cardiomyopathies, rheumatic heart disease and congenital heart disease (Ajayi, 2020, Dokainish, 2016), unlike in high-income countries, where ischaemic heart disease is the primary cause of heart failure.

After subjecting the heart failure cohort to unsupervised machine learning algorithms, the agglomerative clustering algorithm identified three clusters, the *modest ventricular dysfunction cluster*, with relatively higher LVEF and smaller left ventricular diameter during contraction and relaxation of cardiac muscles, the *hypotensive cluster* with lower systolic and diastolic blood pressure and the *congested cluster* with a high burden of heart failure symptoms and signs of fluid overload. The heart is responsible for maintaining adequate perfusion to all organs. A low systolic and diastolic pressure was a dominant finding in cluster 2, the *hypotensive cluster*. The reduction in myocardial contractility in this cluster likely caused a decrease in cardiac output, systemic vasoconstriction, and hypotension in the *hypotensive cluster* (van Diepen, 2017). This finding resonates with the importance of preventing hypotension in heart failure patients, considering that hypoperfusion may lead to cardiogenic shock and subsequent mortality. The in-hospital all-cause mortality rate in the *hypotensive* cluster was 38% and was similar to the congested cluster, which had the highest mortality rate of 39.1%. It is well established that patients who present with severe symptoms of acute decompensated heart failure experience poor outcomes, suggesting the need to expedite their management, irrespective of the LVEF.

Most studies have confirmed the utility of supervised machine learning algorithms for classification tasks (Kwon, 2019, Segar, 2021, Su, 2017, Ahmad, 2018, Shameer, 2017, Golas, 2018, Miao, 2018, Meng, 2019, Liu, 2020). However, only a few studies have since been published on phenotyping heart failure patients using unsupervised machine learning algorithms (Ahmad, 2014, Ieva, 2017, Moser, 2014, Bose, 2018, Segar, 2020, Nouraei, 2022, Perry, 2021). In a paper summarising 40 studies on heart failure subtype definition using unsupervised machine learning, hierarchical clustering was used in 25 studies and most clusters were derived based on heart failure symptoms only (n=20) (Banerjee, 2021). Harada et al. used K-Means clustering to stratify heart failure patients with a LVEF above 40% and found four distinct groups and concluded that older patients might be at a higher risk of right ventricular afterload mismatch and renal dysfunction (Harada, 2020). Also, in Russia, 25 952 heart failure patients were clustered using K-Means, and 15 distinct heart failure phenotypes were identified. The clusters identified in the study from Russia were characterised and labelled based on comorbidities and causes of heart failure only.

Clustering algorithms have the potential to unravel distinct phenotypes of heart failure that may assist clinicians in triaging or risk profiling patients, hence influencing the management of heart failure patients. Our study is unique since it offers insights into the role of machine learning in better understanding heart failure phenotypes in our region. The systolic blood pressure had a 2.6 times higher impact on differentiating between clusters, followed by the age at the time of admission with a mean SHAP value of 0.16, and the LVEF had a lower SHAP value of 0.14. These findings suggest that blood pressure and age may play a role in the diagnosis, risk stratification, and management of heart failure patients in centres without access to echocardiography.

Each institution should attempt to cluster its heart failure patients and discover new clinical patterns. The clusters will vary across regions, partly due to the variation in the causes of heart failure and the availability of clinical parameters used to create clusters. Future studies should use larger, prospective cohorts of patients, and subsequently, train and test the performance of clustering algorithms using new data that was not used to create original clusters of patients. Most importantly, the prognostic value of the unsupervised machine learning algorithms should be elicited by measuring clinical outcomes and comparing them with outcomes obtained from patients classified solely with the LVEF.

A larger dataset would have provided more perspective into the characteristics of patients in each cluster. The study is also limited by its retrospective nature and was a single-centre study. As such, the findings from this study may not be generalisable to all heart failure treated in LMIC. In addition, pertinent features such as the cause of mortality were not available in the dataset. Most clinical history findings were documented in free text format and were not analysed. Furthermore, a significant number of patients diagnosed with heart failure with missing echocardiogram findings were not included in the analysis. Moreover, discharge oral medications were not captured for all patients, and were not included in the final analysis. Overall, the lack of electronic health records and dedicated data capturers is a major hindrance in LMIC, such that patients may not fully benefit from the capabilities of artificial intelligence.

The next chapter will introduce the topic of predicting outcomes such as mortality and 30-day rehospitalisation using supervised machine learning algorithms. Unlike with unsupervised machine learning algorithms where there is no training or a clearly labelled output/outcome, prior to making predictions, supervised machine learning algorithms are trained using data with a column that specifies whether a patient was dead or alive upon discharge from the hospital.

4.10

An unsupervised machine learning algorithm identified three clusters of heart failure phenotypes, beyond the LVEF. The most important features that had an impact on the formation of clusters were blood pressure, age, and the LVEF. Clinicians without access to echocardiography could potentially rely

on blood pressure measurements and age to risk stratify heart failure patients. However, larger prospective studies are mandatory for the validation of these clinical parameters.

turnitincluster.docx

ORIGINALITY REPORT

13%	11%	10%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ncbi.nlm.nih.gov Internet Source	2%
2	www.perinatalpriorities.co.za Internet Source	1%
3	www.frontiersin.org Internet Source	1%
4	hrcak.srce.hr Internet Source	1%
5	services.rmh.med.sa Internet Source	1%
6	ir.lib.uwo.ca Internet Source	<1%
7	"Monday, 31 August 2009", European Heart Journal, 09/02/2009 Publication	<1%
8	erepository.uonbi.ac.ke:8080 Internet Source	<1%
9	www.smj.org.sg Internet Source	<1%

10	"Current Approach to Heart Failure", Springer Science and Business Media LLC, 2016 Publication	<1 %
11	doi.org Internet Source	<1 %
12	hdl.handle.net Internet Source	<1 %
13	Samantha Sotillo, Jessica L. Ward, Emilie Guillot, Oliver Domenig et al. "Dose-response of benazepril on biomarkers of the classical and alternative pathways of the renin-angiotensin-aldosterone system in dogs", Research Square Platform LLC, 2022 Publication	<1 %
14	synapse.koreamed.org Internet Source	<1 %
15	Chang Liu, Zining Yan, Li Fan, Jun Huang, Dan Shen, Xiangting Song. "Layer-specific speckle tracking analysis of left ventricular systolic function and synchrony in maintenance hemodialysis patients", Research Square Platform LLC, 2019 Publication	<1 %
16	Submitted to University of Southampton Student Paper	<1 %

17	Al-Naher, Ahmed. "The Permit Project: Personalised Renal Function Monitoring Via Information Technology", The University of Liverpool (United Kingdom) Publication	<1 %
18	Submitted to University of Leicester Student Paper	<1 %
19	Tao Yue, Qiting Li, Rongsheng Wang, Zhaoyi Liu, Mengru Guo, Fengmin Bai, Zhanming Zhang, Weifeng Wang, Yahui Cheng, Haijun Wang. "Comparison of Hospital Anxiety and Depression Scale (HADS) and Zung Self-Rating Anxiety/Depression Scale (SAS/SDS) in Evaluating Anxiety and Depression in Patients with Psoriatic Arthritis", Dermatology, 2020 Publication	<1 %
20	Todd R. Vogel, Jamie B. Smith, Robin L. Kruse. "The association of postoperative glycemic control and lower extremity procedure outcomes", Journal of Vascular Surgery, 2017 Publication	<1 %
21	www.jstage.jst.go.jp Internet Source	<1 %
22	www.scielo.br Internet Source	<1 %
23	Repository.Library.NOAA.gov Internet Source	<1 %

24	academic.oup.com Internet Source	<1%
25	cvja.co.za Internet Source	<1%
26	hadaco.com.vn Internet Source	<1%
27	onlinelibrary.wiley.com Internet Source	<1%
28	Golla, Anvesh. "Pattern of Cardiac Changes in Chronic Kidney Disease", Dr. NTR University of Health Sciences (India), 2021 Publication	<1%
29	Jackson, Gabrielle. "Bromocriptine and Cabergoline as Adjunctive Therapy for Peripartum Cardiomyopathy", Weill Medical College of Cornell University, 2022 Publication	<1%
30	Skoglund, Per. "Pulse Pressure as a Predictive Marker for Cardiovascular Events: Relation to Biomarkers and Antihypertensive Treatment", Karolinska Institutet (Sweden), 2021 Publication	<1%
31	app.pulsenotes.com Internet Source	<1%
32	assets.cureus.com Internet Source	<1%

33	e-hir.org Internet Source	<1%
34	journals.viamedica.pl Internet Source	<1%
35	pure.rug.nl Internet Source	<1%
36	www.hatter.uct.ac.za Internet Source	<1%
37	www.igi-global.com Internet Source	<1%
38	repub.eur.nl Internet Source	<1%
39	"Heart Failure: Pharmacologic Management", Wiley, 2006 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

turnitinpredict.docx

by Dineo Mpanya

Submission date: 01-Feb-2023 02:09PM (UTC+0200)
Submission ID: 2004011711
File name: turnitinpredict.docx (704.98K)
Word count: 4303
Character count: 24944

Predicting in-hospital all-cause mortality in heart failure using

Abstract

Background: The age of onset and causes of heart failure differ between high-income and low-and-middle-income countries (LMIC). Heart failure patients in LMIC also experience a higher mortality rate. Innovative ways that can risk heart failure patients in this region are needed. The aim of this study was to demonstrate the utility of machine learning in predicting all-cause mortality in heart failure patients hospitalised in a tertiary academic centre.

Methods: Six supervised classification machine learning algorithms were trained to predict in-hospital all-cause mortality using data from 500 consecutive heart failure patients with a left ventricular ejection fraction (LVEF) less than 50%.

Results: The mean age was 55.2 ± 16.8 years. There were 271 (54.2%) males, and the mean LVEF was 29 ± 9.2%. The median duration of hospitalisation was seven days (interquartile range: 4 – 11), and it did not differ between patients discharged alive and those who died. After a prediction window of 4 years (interquartile range: 2 – 6), 84 (16.8%) patients died before discharge from the hospital. The receiver operating characteristic curve area was 0.82, 0.78, 0.77, 0.76, 0.75, and 0.62 for random forest, logistic regression, support vector machines (SVM), extreme gradient boosting, multilayer perceptron (MLP), and decision trees, and the accuracy during the test phase was 88%, 87%, 86%, 82%, 78%, and 76% for random forest, MLP, SVM, extreme gradient boosting, decision trees and logistic regression. The support vector machines were the best performing algorithm, and furosemide, beta-blockers, spironolactone, early diastolic murmur and a parasternal heave had a positive coefficient with the target feature, whereas coronary artery disease, potassium, oedema grade, ischaemic cardiomyopathy, and right bundle branch block on electrocardiogram had negative coefficients.

Conclusion: Despite a small sample size, supervised machine learning algorithms successfully predicted all-cause mortality with modest accuracy. The SVM model will be externally validated using data from multiple cardiology centres in South Africa before developing a uniquely African risk prediction tool that can potentially transform heart failure management through precision medicine.

Key message

Models predicting outcomes in heart failure are built mainly from data originating in high-income countries, where ischaemic heart disease is predominant. In this study, we used data of in-hospital heart failure patients to train machine learning algorithms to predict all-cause mortality. This is the first sub-Saharan African model demonstrating machine learning algorithms' utility in making predictions. In addition, the study shows that the performance of machine learning algorithms is sensitive to the sample size and credibility of the data used to train algorithms.

1 Introduction

Patients diagnosed with heart failure experience a high mortality rate, with 30-50% of heart failure patients demising within five years from diagnosis [1]. This is despite the introduction of novel pharmacological and intracardiac device therapy. Furthermore, due to the high cost associated with the implantation of device therapy and transplant services, only a select few patients in sub-Saharan Africa (SSA) can access these life-saving interventions. As such, there is a need to risk stratify patients with precision, ensuring that the right patient receives the right therapy at the right time.

Risk calculators are derived from predictive modeling, and they are used to estimate the risk of outcomes such as mortality and rehospitalisation in heart failure. Most readily available risk calculators were created using data predominantly from high-income countries (HIC). For example, the Meta-Analysis Global Group in Chronic Heart Failure (MAGGIC) risk calculator, created from a statistical Poisson model, estimates the risk of one and three-year mortality in heart failure [2]. Another widely available risk calculator predicting in-hospital all-cause mortality was derived from the Get with the Guidelines Heart Failure data of multiple sites in HIC, using a multivariable logistic regression model [3]. To the best of our knowledge, risk calculators derived from SSA do not exist, and existing heart failure predictive models have not been validated using data originating from SSA.

Predictive models are population-specific and are prone to inherent biases such as the availability of therapy, referral pathways, genotype, and varying population data used to create these models. Furthermore, most of the available risk calculators were derived using data from heart failure patients in whom the primary cause of heart failure is ischaemic heart disease, unlike in SSA, where the predominant cause of heart failure is non-ischaemic heart disease. Therefore, there is a need to create uniquely African predictive models, mainly using machine learning techniques, since machine learning algorithms are capable of learning from a labelled dataset prior to making predictions. Also, the accuracy of predictions may improve over time as the algorithms are exposed to a larger dataset. This study aims to use machine learning algorithms to predict all-cause mortality in heart failure patients hospitalised in a tertiary academic centre in Johannesburg, South Africa.

2 Materials and Methods

2.1 Study design and participants

Heart failure patient data were exported from the PMRCardio database, a Microsoft Structured Query Language Server Management Studio (version 15.0.18330.0) that stores clinical data of patients hospitalised in cardiology wards. The Charlotte Maxeke Johannesburg Academic Hospital (CMJAH) is a tertiary-level state-owned institution equipped with general cardiac wards, a cardiac intensive unit, a catheterisation and electrophysiological laboratory, and outpatient clinics.

All acutely and chronically ill patients with cardiovascular diseases are hospitalised in dedicated cardiac wards. Their admission data is stored digitally in the PMRCardio database by trainee physicians rotating in the cardiology wards. Clinical parameters available in the dataset

(features) include demographic data, clinical history and examination findings, laboratory indices, electrocardiogram (ECG) data, echocardiography data, angiography data, the status of the patient upon discharge (alive or dead), the date of subsequent admission and discharge oral medication (**Supplementary file 1**). In our institution, heart failure patients are routinely prescribed guideline-directed medical therapy.

After exporting data from the PMRCardio database, data were merged into a single Microsoft Excel Sheet and imported into Python software, version 3.10.0. The international classification of diseases, tenth revision (ICD-10), was used to select patients diagnosed with heart failure due to any aetiology. The following ICD-10 diagnoses were used to select patients included in the analysis: "heart failure unspecified", "congestive heart failure ", left ventricular failure", "dilated cardiomyopathy", or "ischaemic cardiomyopathy." Only heart failure patients 18 years of age and older with a left ventricular ejection fraction (LVEF) < 50% hospitalised between 2009 and 2018 were included in the dataset used to build predictive models. The rationale for only including patients with a LVEF < 50% is that in our institution, patients with a LVEF between 40% and 49% (*heart failure with a mid-range ejection fraction)* are managed similarly to those with LVEF<40%.

All retrospective patient data used in the final analysis was further prospectively verified by comparing hard copies of medical records with electronic data. Patients without a documented LVEF were excluded from the analysis. Biochemical data captured at the time of admission was obtained from the National Health Laboratory Service (NHLS). Approval to conduct the study was received from the University of the Witwatersrand Human Research Ethics Committee (Clearance certificate number: M190515). Permission to conduct the study was also obtained from NHLS senior authorities. Informed consent was not obtained from the patients as this was a retrospective chart review. The study protocol conformed to the ethical guidelines of the 1975 Declaration of Helsinki as reflected in a priori approval by the institution's human research committee.

2.2 Data Cleaning and Pre-processing

Packages and libraries required for data analysis were imported into Python software, version 3.10.0. The following software, packages and libraries were installed: Jupyter notebook version 6.0.3, scikit-learn version 1.02, statsmodels version 0.13.2, seaborn version 0.11.2, numpy version 1.21.6, scipystats version 1.5.4, matplotlib version 3.1.3 and pandas version 1.3.5. Features with more than 20% missing values, patient identifiers, and column data without variation were removed from the dataset. The remaining continuous features with missing data were imputed using the mean value. Categorical features were coded as 0 for no and 1 for yes, while blank or missing values were ascribed with zero. Data used to train machine learning algorithms was transformed and scaled between 0 and 1. The rationale for scaling the data is to standardise the weight of each feature. For example, sodium levels equal to 120 will have more weight than a potassium level of 5.0.

2.3 Exploratory Data Analysis and Hyperparameter Optimisation

Exploratory data analyses were carried out visually and quantitatively. Data were partitioned into training and test datasets, with 70% of the data used for training. Grid searching was implemented for all classification models, where parameters associated with the best

performance metrics are discovered and selected during model hyper-tuning. Since the data is unbalanced with more patients in the survivors' class, cost-sensitive learning was applied at a ratio of 0.14: 0.86, where a higher weight was allocated to the minority class (*dead=1*).

All demographic and clinical parameters captured during arrival at the hospital were used as features. A filter method was then used to identify features that are positively or negatively correlated with all-cause mortality. Filtering was done using correlation coefficients. To ascertain feature-to-target correlation, a correlation coefficient was estimated for each feature. The threshold was set at 0.5. The lowest and highest values were furosemide (r=:-0.277395, $p<0.001$) and potassium (r= 0.180126, $p<0.001$). Since all features were weakly correlated with the target value, none of the features were removed.

To build predictive models, we only used features captured during the first hospitalisation. As such, the model with the best performance metrics will be used to predict the risk of in-hospital mortality using patient data obtained at the time of arrival at the cardiology wards at the CMJAH.

In this study, the prediction window is 4 years (interquartile range: 2 -6 years). The prediction window starts from the time of hospitalisation (index date), where the risk of death is assessed, and ends at the time when a patient dies while hospitalised. The EHR system only stores a single entry of data. As such, there is no time window. For example, although clinical parameters are measured several times while the patient is hospitalised, only the first measurement (*baseline)* was captured in the EHR system. The parameters presented in the paper were collected from the baseline index admission for each of the 500 patients. The following six algorithms were used to build a model predicting all-cause in-hospital mortality:

2.3.1 Decision trees

Decision trees implement a sequential decision process based on whether the conditions set are true or false. The decision tree algorithm was trained, and the tree was plotted based on grid search results by placing the criterion as Gini, maximum depth at five, and the minimum sample split at eight.

2.3.2 Random Forest

A random forest classifier uses multiple decision trees, and each of the decision trees outputs a prediction (dead vs. alive). To estimate the final output, the random forest algorithm then counts the number of votes for each class predicted by each decision tree. Grid search identified the following parameters to be responsible for best model performance: the number of estimators=1 757, minimum samples split=5, minimum samples leaf=2, maximum depth=150, and criterion=entropy.

2.3.3 Support vector machines

Support vector machines classify data by creating a line or hyperplane between the classes. The maximum distance between the classes (support vectors) and the hyperplane is then chosen. After applying grid search, the best parameters for creating a model were a linear kernel, gamma=0.001, and C=10. The "C" is a parameter that controls the effect of support vectors.

2.3.4 Logistic regression

For the machine learning logistic regression model, the coefficient of each feature was obtained after training and testing the model's performance. A conventional statistical univariate and multivariate logistic regression model was also built to compare predictors of mortality identified by machine learning algorithms with those extracted from the statistical model. Clinical parameters with a p-value less than 0.05 were selected and included in the final multivariate logistic regression model to identify independent predictors of all-cause mortality. Confidence interval percentiles were set between 0.025 – 0.975.

2.3.5 Extreme gradient boosting (XGBoost)

Extreme gradient boosting is an ensemble of weak algorithms; usually, decision trees added to the model sequentially. Feature importance is calculated based on the number of times the feature appears in isolated trees.

2.3.6 Multilayer perceptron (MLP)

The MLP network consists of one input layer, hidden layers and an output layer. Grid search was used to find the best model parameters. The search for optimal parameters was performed with the following limits: 0 – 0.4 dropout rate, batch size between 10 and 100, and epochs between 1 and 150, and the activation function was chosen to be either to be the rectified linear unit (ReLU) or the sigmoid activation function. The learning rate was set at 0.01. The hyperparameters resulting in the best performing networks were as follows: a multi-layer network with 124 nodes using the ReLU activation function, batch size of 60 and 20 epochs. Shapley additive explanations (SHAP) values were used to identify features influencing the construction of the model.

Six supervised machine learning algorithms were trained to predict all-cause in-hospital mortality in heart failure patients with a LVEF less than 50% using 123 features. The performance metrics of each algorithm were summarised using the confusion matrix to calculate the accuracy, recall, precision, and F1 score, except for the MLP, where only the accuracy was reported since the rest of the performance metrics are not routinely available after training MLP.

Features identified by the algorithms were ranked, weighed and compared with predictors identified by logistic regression using the conventional statistical method. An area under the receiver operating characteristic (ROC) curve was used to evaluate the ability of each model to discriminate between the negative (*alive=0*) and positive (*dead=1*) classes.

2.4 Biostatistics

Categorical data are reported as absolute numbers and percentages, while the mean with standard deviation (±) and median with 25 - 75th percentile interquartile ranges were used to summarise numerical data with a normal and non-normal distribution, respectively. Normality was assessed visually with histograms and quantitatively with the Shapiro-Wilk and skewness tests. The Pearson chi-square and unpaired t-tests were used to compare categorical and normally distributed numerical data, respectively. In contrast, the Wilcoxon signed-rank test was used to compare numerical data with a non-normal distribution between patients that survived and died. A p-value less than 0.05 represented a statistically significant difference in the distribution of the features between survivors and non-survivors.

3 Results

3.1 Demographic and clinical findings

Between January 2009 and December 2018, there were 4 730 acute and chronic heart failure-related hospitalisations (**Fig. 1**). After excluding patient records with missing LVEF and data that could not be verified using original hand-written patient records and biochemical data captured in the NHLS website, a total of 500 heart failure patients with a LVEF < 50% were included in the descriptive cohort analysis. The mean age was 55.2 ± 16.8 years, and 271 (54.2%) of these patients were males. The mean LVEF was 29 ± 9.2%. In this cohort, there were 300 (60.0%), 124 (24.8%), 54 (10.8%), and 22 (4.4%) Black, White, Indian and Mixed ancestry patients, respectively. The rest of the baseline clinical characteristics of the study cohort are depicted in **Table 1**.

There were 159 (31.8%) patients with hypertension and 268 (53.6%) patients with dilated cardiomyopathy. A total of 209 (41.8%) patients had hepatomegaly, 110 (22.0%) had ascites, 350 (55.0%) had mitral regurgitation, 340 (68.0%) had bipedal oedema, and 395 (79.0%) had an elevated jugular venous pressure on clinical examination. The mean resting

pulse rate was 102 ± 23 beats per minute. Upon discharge from the hospital, there were 299 (59.8%) patients on oral beta-blocker therapy, and angiotensin-converting enzyme (ACE) inhibitors, furosemide and spironolactone were taken by 201 (40.2%), 300 (60.0%), 362 (72.4%) patients, respectively.

3.2 Outcomes

In the selected 500 patients hospitalised with heart failure, 84 (16.8%) patients died before discharge from the hospital. The median duration of hospitalisation was seven days (interquartile range: 4 – 11), and it did not differ between patients discharged alive and those who died. Among the 416 patients discharged alive, 29 (7.0%) patients were readmitted within 30 days, 42 (10.1%) beyond 30 days, and 340 were not readmitted.

3.3 Performance Metrics

Multilayer perceptron had an accuracy of 87% during training and testing. The AUROC was 0.82, 0.78, 0.77, 0.76, 0.75, and 0.62 for random forest, logistic regression (ML), SVM, XGBoost, MLP, and decision trees, respectively **(Fig 2**). Support vector machines achieved an accuracy of 86% during the test phase, followed by XGBoost, decision trees and logistic regression at 82%, 78% and 76%, respectively (**Table 2**). The recall rate was above 50% in all algorithms, except for SVM, with a recall rate of 44% and 43%, respectively during training and testing using data of patients that died while hospitalised. Logistic regression had the lowest precision of 39% and 29% during training and testing.

3.4 Feature Ranking

3.4.1 Multilayer perceptron

The top nine features identified by the MLP model were peripheral oedema grade, furosemide, enalapril, hepatomegaly, LVIDd, aspirin, spironolactone, serum potassium and the LVEF **(Fig 3).**

3.4.2 Logistic regression (ML)

Peripheral oedema grade had a coefficient of 0.688. Furosemide, angiotensin receptor blockers, a fourth heart sound on auscultation and aspirin, negatively correlated with all-cause mortality **(Fig 4)**.

3.4.3 Support vector machines

The following top five features had a positive coefficient with the target feature: furosemide, beta-blockers, spironolactone, early diastolic murmur and a parasternal heave, whereas coronary artery disease, potassium, oedema grade, ischaemic cardiomyopathy, and right bundle branch block on electrocardiogram had negative coefficients. The rest of the features are depicted in **Fig 4**.

3.4.4 Extreme gradient boosting

Enalapril was the most common feature used to split decision trees and was used 20 times, followed by oral beta-blocker therapy, elevated jugular venous pressure, furosemide and CAD **(Fig 4)**.

3.4.5 Decision trees

Decision trees identified beta blockers as the root node (Gini =0.5). In addition, the following features were placed as leaf nodes: systolic and diastolic blood pressure, urea, duration, the number of hospitalisations, and CAD.

3.4.6 Random Forest

Urea, potassium, duration of admission, LVIDd, serum sodium levels, age, LVIDs, diastolic blood pressure, haemoglobin, and the LVEF, were identified as the top ten features discriminating the two classes (Alive vs. Dead) **(Fig 5)**.

3.4.7 The conventional statistical logistic regression model

In the multivariate logistic regression model, the following variables were independent predictors of all-cause in-hospital mortality: diastolic blood pressure, coronary artery disease, enalapril, spironolactone, furosemide, sodium and potassium **(Table 3)**.

4 Discussion

This study used data from 500 patients with a LVEF less than 50% on the index or baseline hospitalisation to train six supervised machine learning algorithms to predict all-cause mortality. Although the random forest algorithm achieved a higher AUC, the striking difference in training and testing performance metrics suggests that the model is overfitting data. Support vector machines had an AUC of 0.77, and an accuracy of 91% and 86% during training and testing, making it an ideal model for validation. The most plausible explanation for the SVM showing a desirable performance lies in the ability of the algorithm to process and model complex and non-linear data by creating a decision boundary that separates classes.

The all-cause mortality and rehospitalisation rate in 500 patients admitted with heart failure were 16.8 % and 17.1%, respectively, while the median duration of hospitalisation was seven (interquartile range: 4 - 11) days. In another tertiary-level hospital in South Africa, the in-hospital mortality rate among patients admitted with acute heart failure was 8.4% and the mean duration of hospitalisation was nine ± 12 days [1]. Similarly, in a multicentre study conducted across Africa involving 169 heart failure patients recruited in South Africa, the one-year all-cause mortality was 11.8% [2]. In a meta-analysis involving 67 255 heart failure patients hospitalised in Australia, the pooled 30-day all-cause mortality was 8% [3]. Similarly, in another systematic review of research studies conducted in (low-and-middle-income countries (LMIC), the in-hospital all-cause mortality was 8%, with a confidence interval of 6 - 10% [4]. The most plausible explanation for our cohort's higher in-hospital all-cause mortality is the delay in presentation to the hospital. As a tertiary academic centre, most patients referred to our centre are initially managed in primary and secondary level institutions and eventually referred to our centre in advanced stages of heart failure.

Despite the variation in the aetiology of heart failure between patients residing in LMIC and HIC, and limited access to life-prolonging intracardiac device therapy, particularly in patients attending state-owned treatment facilities, we found similar predictors of all-cause mortality using machine learning algorithms and a conventional statistical logistic regression model. After training the SVM, furosemide, beta-blockers, spironolactone, early diastolic murmur and a parasternal heave, yielded positive coefficients, whereas coronary artery disease, potassium, oedema grade, ischaemic cardiomyopathy, and right bundle branch block on electrocardiogram had negative coefficients. The support vector machines and the machine learning logistic regression algorithm identified similar predictors as the MLP model. Some of the predictors of all-cause mortality in heart failure reported in the literature include age, sex, diastolic blood pressure, LVEF, serum sodium and creatinine, estimated glomerular filtration rate, haemoglobin, haematocrit, and blood urea nitrogen [5; 6; 7]. Predictors of mortality specifically in patients with a LVEF above 50% include blood urea nitrogen and the body mass index [6].

Healthcare databases are notoriously famous for housing imbalanced datasets. Class imbalance refers to the disproportionality between the data classes used to train the predictive model [8], a common problem that is not unique to medical data. For example, when the training data with the negative outcome (*dead*) has significantly fewer observations than the majority class, the classification algorithm favours the majority class, by focusing the learning process on the majority class. Our dataset had fewer entries attributed to the negative outcome and was imbalanced. This accounts for the suboptimal performance shown by the algorithms when learning and making predictions specifically for the minority class. However, instead of under-sampling the majority

class or generating new synthetic values to increase the number of samples in the minority class, we used class weights to excessively penalise false negative and false positive results from the minority class (*dead*).

Six supervised machine learning algorithms were successfully trained to predict mortality despite an imbalanced and relatively small dataset. The success implies that data quality is also essential to consider before building predictive models. The project's next phase is to increase the sample size by collecting data from multiple cardiac centres around sub-Saharan Africa. Adding new data that was not previously used to train or test the algorithms tends to improve the models' performance metrics. Once the performance of the classifier has been validated, a risk calculator will be developed and implemented for use in Africa.

The performance of machine learning algorithms is commonly attributed to the size of the dataset. Indeed, the higher the number of observations the better the performance metrics. Some authors have attempted to derive an equation for calculating a suitable sample size for machine learning algorithms by calculating the error rate of prediction [9; 10]. In our opinion, the performance of machine learning algorithms is also influenced by the number of features (clinical parameters) used, the ratio of observations within the classes, and the credibility of the data. In our study, 500 patients with 123 features were used to train algorithms. Consider a model created with over a million patients, but with only 5 features. Its performance might be suboptimal. An ideal sample size calculation should incorporate the number of observations and features and a measure of data credibility. In general, the error rate should decrease as the size of the training dataset increases. The AUC for each of the classification algorithms can be improved by increasing the size of the training dataset. Other techniques for improving the AUC such as feature normalization and scaling, setting class weights, and grid search were applied in this study.

Machine learning algorithms cannot process data with empty fields. As such innovative ways of handling missing data have been introduced. These include deleting incomplete records, filling incomplete fields with the mean, median or mode, and predicting missing values using machine learning algorithms and the last observation carried forward method [11]. In this study, we removed clinical parameters with more than 20% missing values and imputed missing values using the mean value of each feature. The rationale for using the mean is that overall the mean value (Gaussian distribution) of each feature does not change. However, this method does not factor for the covariance between features.

This was a single-centre study conducted in a tertiary centre since the CMJAH division of cardiology is the only department equipped with an electronic health record system. As such, the findings from this study may not be generalisable to all heart failure patients treated in SSA. In addition, most pertinent clinical parameters such as comorbidities, causes of heart failure, angiographic, and ECG findings were available in the database but underreported, as some of the data were captured in the free text format and could not be extracted for analysis and incorporation into the predictive models. Also, we excluded a significant amount of patient data due to the unavailability of handwritten patient records or objective clinical evidence supporting the heart failure diagnosis. This was a major limitation, since handwritten patient records are routinely removed from the CMJAH after five years from the time of initial admission, due to the unavailability of storage facilities. Furthermore, pertinent clinical data, such as the presence of cardiogenic shock, and the administration of ionotropes and vasopressors, were documented in the EHR system in free-text format and could not

be extracted. Nevertheless, the study provides proof of concept for developing a heart failure model derived from LMIC predicting all-cause mortality.

5 Conclusion

Despite a small sample size, machine learning algorithms successfully predicted all-cause mortality with modest accuracy. The SVM model will be externally validated using data from multiple cardiology centres in South Africa before developing a uniquely African risk prediction tool.

turnitinpredict.docx

ORIGINALITY REPORT

90%	90%	90%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.frontiersin.org Internet Source	90%
2	www.ncbi.nlm.nih.gov Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		