

https://dspace.org/

Using Supervised Machine Learning Techniques To Classify Respiratory Tract Infections

Madimabe, Metsekae Richard

2025
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/49763
Madimabe, Metsekae Richard. (2025). Using Supervised Machine Learning Techniques To Classify Respiratory Tract Infections [Master's dissertation, University of the Witwatersrand, Johannesburg]. WIReDspace. https://hdl.handle.net/10539/49763

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

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

Using Supervised Machine Learning Techniques To Classify Respiratory Tract Infections

Student:
Mr. Metsekae Richard Madimabe

Supervisors:
Prof. Nicola Page, Mr. Samuel Chikasha

A Research Report submitted to the Faculty of Health Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science in Epidemiology - Public Health Informatics

Johannesburg, 29 September 2025

DECLARATION

I, Metsekae Richard Madimabe declare that this Research Report is my own, unaided work. It is being submitted for the Degree of Master of Science in Epidemiology at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



(Signature of candidate)

29 day of September 2025 in Wits Education Campus

DEDICATION

In loving memory of my parents. . . . The greatest honor goes to you.

No one will ever know the strength of your love but me, with these few words below, just know how much I feel.

Love, a word that comes and goes
But few people really know what it means to really love somebody
Love, though the tears may fade away, I'm so glad your love will stay
Cause I love you and you show me, what it really means to love . . .

ACKNOWLEDGMENTS

I would firstly like to take this opportunity to thank my main supervisor, Prof. Nicola Page, for her guidance and expertise in the development of this work. Thank you so much for facilitating and cultivating such a positive learning environment, the work you have created for me, sharing and imparting your knowledge so that i can reach my academic goals. Thank you Mr. Samuel Chikasha for your expertise in ensuring my research study is on track to meet faculty deadline.

ABSTRACT

OBJECTIVES: Acute respiratory tract infections (RTIs) are a leading cause of morbidity and mortality among children under five years of age, and the fourth leading cause of mortality in adults globally. Misclassification of RTIs remains a significant challenge due to inconsistent diagnostic practices, particularly in low- and middle-income countries (LMICs), especially within sub-Saharan Africa. This study aimed to develop a clinical decision support system using a machine learning approach to assist healthcare providers in the accurate diagnosis and classification of RTIs..

METHODS: This was a secondary data analysis research study of data collected from a retrospective observational cohort study of respiratory tract infections (RTI) and non-RTI cases, namely acute febrile disease of unknown cause (AFDUC) and gastrointestinal tract infection (GTI), surveyed between July 2018 and December 2020. The primary objectives of the study were to develop, fine-tune, and evaluate a predictive model capable of classifying RTI cases using training datasets that combined data from RTI and AFDUC cases, as well as RTI and GTI cases. The machine learning workflow involved two key phases: data pre-processing and feature engineering, both of which were essential in preparing the data for model development, testing, validation, and final evaluation aimed at achieving accurate diagnostic classification. Classification and prediction models were built using three machine learning algorithms: artificial neural networks (ANN), random forest (RF), and support vector machines (SVM). K-fold cross-validation was employed to optimize and fine-tune the models for improved performance. Model evaluation was conducted using performance metrics derived from confusion matrices, which facilitated the assessment of the receiver operating characteristic (ROC) curves and area under the curve (AUC). Furthermore, feature importance analysis was carried out to identify the most influential variables contributing to robust model performance, while variable correlation analysis was used to explore interactions among the independent variables.

RESULTS: Amongst the syndromes, RTI had highest overall numbers of cases enrolled, largely from age group 1 year and below with 42.0% of cases when compared to AFDUC and GTI. For correlation measurements, few variables correlated positively with value between 0.55 and 0.71, particularly age group which correlated with chest pain, headache and HIV status in all models. The evaluation of each model and classifier reached performance metric scoring of above 0.8 for accuracy, precision, specificity and F1-score, which was desirable. The classified RTI ROC curve shape was robust with the expected shape, which was able to measure the AUC above 0.92 for three classes and AUC above 0.95 for two classes.

CONCLUSIONS: In this study, we differentiated RTIs from other infections that may present with similar signs and symptoms—such as cough, fever, and elevated temperature—observed during patient admission to the hospital. The predictive models developed and presented in this

study provide accurate estimations for distinguishing RTI cases from AFDUC and GTI cases. These findings offer important evidence that RTIs can be effectively and efficiently screened using machine learning approaches, without the limitations typically associated with time, manpower, and cost constraints.

Contents

List of Figures **vii**

List of Tables **viii**

1 INTRODUCTION AND LITERATURE REVIEW **1**

1.1 Introduction 1
1.2 Problem Statement 2
1.3 Justification 2
1.4 Aims and Objectives 3
1.5 Literature Review 3
1.5.1 Aetiology of acute respiratory tract infections 3
1.5.2 Epidemiology of respiratory tract infections 4
1.5.3 Clinical practice applied to respiratory tract infections 5
1.5.3.1 Standard clinical healthcare practice 5
1.5.3.2 Clinical practice and healthcare challenges 6
1.5.4 Evidence-based clinical practice and decision support 7
1.5.4.1 Evidence-based diagnosis of respiratory tract infections . . . 7
1.5.4.2 Artificial intelligence and decision-rule in clinical practice . . 8
1.5.4.3 Machine learning techniques in clinical practice 9

2 CLASSIFICATION THEORY **11**

2.1 Research Methodology Framework 11
2.1.1 Data pre-processing and feature engineering 11
2.1.1.1 Data pre-processing 12
2.1.1.2 Feature engineering 13
2.1.2 Machine learning classifiers 15
2.1.2.1 Support Vector Machine 15
2.1.2.2 Random Forest 16
2.1.2.3 Artificial Neural Network 17
2.1.3 Model optimisation 18
2.1.4 Model evaluation 18

3 MATERIALS AND METHODS **21**

3.1 Primary Study Design and Setting 21
3.1.1 Primary Study Population and Sample Size 21
3.1.2 Primary Study Inclusion and Exclusion Criteria 21
3.2 Secondary Study Design and Population 22
3.3 Data Collection and Management 23
3.4 Collected Study Variables 23

3.5 Data Pre-processing and Model Design 24
3.6 Experimental Environment 27

4 RESULTS AND FINDINGS 28
4.1 Patient Characteristics and Datasets 28
4.2 Pre-processed Data and Analysis 28
4.3 Model Development Outcome and Evaluation 32

5 DISCUSSION AND CONCLUSION 37
5.1 Discussion 37
5.1.1 Main findings 37
5.1.2 Clinical implications of prediction model 38
5.1.3 Strength and limitation of this study 39
5.2 Conclusion 39

REFERENCES 41

A APPENDICES 49

List of Figures

1 Support Vector Machine 15
2 Random Forest 16
3 Artificial Neural Network 17
4 Confusion Matrix 19
5 ROC Curve and Area Under Curve 20
6 Finalised Classification Task 1, 2 and 3 Datasets 23
7 Classification Model Design 24
8 Missing Values per Variable 25
9 Finalised Classification Task 1, 2 and 3 Datasets 28
10 Classification Task 1 Predictors, Correlation Coefficient Matrix 30
11 Classification Task 2 Predictors, Correlation Coefficient Matrix 31
12 Classification Task 3 Predictors, Correlation Coefficient Matrix 32
13 Performance Comparison of Classification Tasks 33
14 Random Forest Area under ROC Curve for all Classification Tasks 34
15 Neural Network MLP Area under ROC Curve for all Classification Tasks . . . 34
16 Support Vector Machine Area under ROC Curve for all Classification Tasks . . 35
17 Gini and Permutation Feature Importance for Classification Task 1 35
18 Gini and Permutation Feature Importance for Classification Task 2 36
19 Gini and Permutation Feature Importance for Classification Task 3 36
20 Random Forest Confusion Matrix for all Classification Tasks 49
21 Random Forest Confusion Matrix for all Classification Tasks 49

22 Random Forest Confusion Matrix for all Classification Tasks 49

List of Tables

1 Python Libraries for ML 12
2 Comparison of patient characteristics for all syndromes 29
3 Classification Tasks Performance Comparison Measured by 4 Importance Scores 33

List of Abbreviations

AI	Artificial Intelligence
ANDEMIA	African Network for Improved Diagnostics, Epidemiology and Management of Common Infectious Agents
ANN	Artificial Neural Networks
AUC	Area Under the Curve
BOW	Bag of Words
CAP	Community-acquired Pneumonia
DT	Decision Tree
FN	False Negatives
FP	False Positives
GNB	Gaussian Naive Bayes
HIV	Human Immunodeficiency Virus
hMPV	Human Metapneumovirus
hRV	Human Rhinovirus
IQR	interquartile range
KNN	K-Nearest Neighbour
LMIC	Low- and Middle-income Countries
LRTI	Lower Respiratory Tract Infection
MDI	Mean Decrease in Impurity
MICE	Multivariate Imputation by Chained Equations

MLP	Multilayer Perceptron
NICD	National Institute for Communicable Disease
NLP	Natural Language Processing
PLHIV	People living with HIV
RF	Random Forest
ROC	Receiver Operating Characteristics
RSV	Respiratory Syncytial Virus
RTI	Respiratory Tract Infection
StatsSA	Statistics South Africa
SVM	Support Vector Machine
TF-IDF	Term Frequency and Inverse Document Frequency
TN	True Negatives
TP	True Positives
URTI	Upper Respiratory Tract Infection
WHO	World Health Organisation

CHAPTER 1

1 INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

Acute respiratory tract infections (RTIs) are a leading cause of morbidity and mortality in children 5 years and below (Seidu et al., 2019), and fourth leading cause of mortality in adults globally in 2019 (WHO, 2020). According to the GBD 2021 Upper Respiratory Tract Infection (URTI) and Otitis Media Collaborators (2024), RTIs accounted for 292,908.6 per 100,000 new incidents rates in children 2–4 years of age and 108,190.2 per 100,000 new incidents rates in adults 70–74 years (Sirota et al., 2025). Additionally, the GBD 2021 Lower Respiratory Tract Infection (LRTI) Collaborators reported that 502,000 deaths occurred in children below the age of 5 years and 1,110,000 deaths occurred in adults 70 years and older globally (Bender et al., 2024).

In sub-Sahara countries, upper RTIs accounted for 71% of global deaths (14,000 deaths) in all ages, where deaths in children between the ages of 2–4 years was 8% (1,680) and 1.4% (275) death in older adults between ages of 70–74 of global deaths (Sirota et al., 2025). Furthermore, lower RTIs accounted for 261,000 deaths in children below the age of 5 years and 50% of deaths occurred in sub-Sahara (Bender et al., 2024).

In South Africa, Statistics South Africa (StatsSA) reported that in 2020, respiratory tract infections were the fifth leading cause of death in all ages, and second leading cause of death in children below the age of 5 years, accounting for 5.1% (1,299) of total children's deaths (StatsSA, 2024). In adults aged 65 and above, RTIs accounted for 4.5% (9,111) of total deaths, ranking it the number six leading cause of death in the elderly (StatsSA, 2024). The age groups that are at highest risk of mortality due to RTIs are the group below age of 5 years and the elderly (StatsSA, 2024).

According to the GBD 2021 Lower Respiratory Infections and Antimicrobial Resistance Collaborators (Bender et al., 2024), the highest global mortality estimates proportions were due to *S. pneumoniae*, and second highest by other viral aetiologies.

For the past several decades, artificial intelligence (AI) research for various application has been ongoing in all industries especially the commerce sector (Bohr and Memarzadeh, 2020). However, in the healthcare sector, two studies of interest used retrospective data obtained from hospital records, for patients diagnosed with lower RTI and pneumonia. Both studies used clinical symptoms, vital signs, and biological test results as features of the machine learning prediction model that classified respiratory infections and pneumonia severity (Naydenova et al., 2016; Olayemi et al., 2019). The aim of both studies was to design a clinical tool to aid in the diagnosis of respiratory tract infections using clinical features that are measurable and prevalent.

To classify RTI, AFDUC and GTI were chosen as non-RT infection comparators, since they share one common clinical feature, which is fever and classified as high temperature 38C and above. When comparing RTI to AFDUC, many patients present with fever only without focal symptoms, while respiratory signs develop later in the infection, however, RTI requires identification respiratory symptoms to be classified. When AFDUC symptoms develop later, similar systematic signs such as Myalgia and headache would be identified amongst patients of both syndromes. Although patients with GTI would also present with fever, which makes it the only common feature to RTI patients. Another uncommon feature is the localisation of the infections, as for RTI it is localised in the respiratory tract, while GTI is localised in the gut, specifically the stomach or intestines (Gutierrez et al., 2024; Minodier et al., 2017).

1.2 Problem Statement

The diagnosis and classification of RTI cases depends on the accurate assessment and diagnosis of clinical signs and symptoms by medical specialists and practitioners. To date, medical practitioners have been the gold standard in the diagnosis and classification of RTI cases. Nevertheless, the assessment and diagnosis process done by medical practitioners' exhibit drawbacks such as cost, time, inconsistency amongst many. Misclassification of RTI cases can lead to the unnecessary treatment costs of antibiotics administered in self-limiting infections, delayed treatment due to prolonged time for differential diagnoses, and inconsistent diagnosis of RTI cases with similar clinical features resulting in poor outcomes.

1.3 Justification

Machine learning approaches can be used as an alternative to improve the clinical decision-making process and overall diagnosis and classification of RTI cases. Significant costs were

associated with antibiotic treatment in patients with self-limited respiratory infections. Machine learning can identify patients requiring antibiotics treatment and improve appropriate prescription of antibiotics (Oonsivilai et al., 2018). In cancer research for non-communicable diseases, machine learning has not only performed better than the physicians, but also predict patients' symptoms with high accuracy, specificity, and sensitivity in real-time, limiting time constrains. Additionally, in a clinical cardiovascular disease research project, machine learning was used to predict the survival of patients with heart failure, which became a supporting tool for clinicians (Chicco and Jurman, 2020; Chicco and Rovelli, 2019). Findings from these areas of research, revealed the significance of machine learning prediction models in clinical decision support of disease diagnosis and survival prediction for chronic patients, limiting inconsistencies. The research output generated from these studies, have encouraged our research in the use of machine learning approaches to improve the clinical decision support for diagnosis and classification of RTI cases as well as highlight the type of data points that could assist clinicians for inclusion in future studies.

1.4 Aims and Objectives

The research project aim was to apply supervised machine learning techniques to build prediction models that would classify RTI cases and non-RTI cases, and evaluate optimum model to serves as powerful clinical tool to aid practitioner's diagnosis. The objectives to achieve the aim were:

1. To describe and compare the clinical, demographic, and socioeconomic characteristics of RTI cases and non-RTI cases.

2. To apply supervised machine learning to classify RTI and non-RTI cases and identify key clinical, demographic, and socio-economic predictors of RTI.

3. To evaluate the performance of different supervised machine learning models and identify the optimal model for classifying RTI and non-RTI cases.

1.5 Literature Review

1.5.1 Aetiology of acute respiratory tract infections

Bacterial pathogens were thought to be the leading cause of RTIs in both children and adults (Suarez et al., 2015). However, increasing evidence suggests that viral respiratory infections

play an important role in RTI hospitalisation, which could progress to acute respiratory distress syndrome, and ultimately mortality (Cohen et al., 2015b,c; Suarez et al., 2015). One of the common and severe respiratory infections is community-acquired pneumonia (CAP), which account for 40% cases of hospitalisation, of which 20% of these cases are immunocompromised patients (Azoulay et al., 2020; Niederman and Torres, 2022). For a long time, bacteria have been isolated from CAP patients, particularly from young children and older adults. Findings from a study with a purpose of analysing the aetiology, incidence and outcome of CAP, reported a viral aetiology in 23.4% of hospital patients admitted and presenting with fever (Tao et al., 2018; To et al., 2019).

Amongst the infectious pathogens detected in RTI cases, influenza, human metapneumovirus (hMPV) and respiratory syncytial virus (RSV) were the major viruses detected in children below 5 years of age along with *Streptococcus pneumoniae* as co-infection (Cohen et al., 2015a). Amongst the adults who were hospitalised, viruses including influenza A and B, RSV and hMPV together with bacteria (*S. pneumoniae*, *Staphylococcus aureus* and *Moraxella catarrhalis*) as coinfections were detected (Suarez et al., 2015). A study conducted in South Africa, detected 37% human rhinovirus (hRV), 26% RSV, 26% adenovirus and 10% enterovirus in children RTI cases (Cohen et al., 2015c). Respiratory viruses cause of morbidity and mortality in immunocompromised RTI cases, including patients with acquired immunodeficiency syndrome (Venter et al., 2011).

1.5.2 Epidemiology of respiratory tract infections

The diagnosis of acute respiratory tract infections can be challenging in identifying the cause, because the infections can affect both the lower and the upper respiratory tract (Wu et al., 2024). The upper respiratory tract is represented by the sinuses, nasopharynx, larynx, peritonsillar structures and epiglottis, whereas the lower respiratory tract is represented by pulmonary parenchyma and bronchial tree. Acute respiratory infections may exhibit rhinitis, sinusitis, otitis, pharyngitis, epiglottitis, tracheitis as upper tract disease, bronchitis, bronchiolitis and pneumonia as lower tract disease, or both. Acute upper respiratory tract infections are largely viral in nature, with majority of infections mild and self-limiting, and some cases may advance to severe infection requiring additional examination and further tests. Acute lower respiratory tract infections are mild and short-lived episodes that affect the airways and cause bronchitis, which could advance to severe life-threatening pneumonia through pulmonary parenchy-

mal inflammation. Virus and bacteria are responsible for LRTI infections; however, bacteria are responsible for the majority of community-acquired pneumonia (Rogan, 2016; Saleri and Edward, 2019).

Symptoms and signs that present clinically, may sometimes overlap depending on the type of infection and respiratory tract location, especially in cases of co-infection by both virus and bacteria (Sumitomo and Kawabata, 2024). It is globally accepted that cough, particularly chronic cough that last 10 or more days is an indicator of clinical representation of acute respiratory infections. Together with fever (38°C), symptoms such as nasal congestion and sometimes nasal discharge and localised pain are experienced by patients. Cough is associated with the mechanisms that cause inflammation of bronchoalveolar and airway, which ultimately induce sputum (Cho et al., 2019; Jaume et al., 2020). However, clinical representation of RTIs is non-specific in some clinical cases (Gentilotti et al., 2022), which maybe attributable to the geographic area and microorganism pathogenesis.

Risk factors associated with RTI and hospitalisation may not be constant across all ages, as age itself is a major risk factor (e.g. children below 1 years and adults 65 years and older) (Le Roux et al., 2019; Storms et al., 2017). A study conducted by Le Roux et al. (2019), reported that in South Africa, low birth weight, delayed or missed vaccinations, and maternal smoking were risk factors for RTI in children below the age of 2 years. However, infants in their first 6 months of life were at higher risk of hospitalisation with RTI, especially if born to human immunodeficiency virus (HIV)-infected mothers. In addition to age and smoking, HIV is a known risk factor in adult RTI hospitalisations. Another risk factor in adults aged 65 years and older who were hospitalised with pneumonia was more than three co-morbidities (Abadom et al., 2016; Storms et al., 2017).

1.5.3 Clinical practice applied to respiratory tract infections

1.5.3.1 Standard clinical healthcare practice

Standard clinical practice has a set of guidelines that are vital components of the healthcare system, which provide cost-effective medical management of common and uncommon health conditions. At the centre of clinical practice guideline, there is a principle of classifying disease or infections in the form of diagnosis which is regarded as gold-standard of clinical decision making (Croft et al., 2015).

Clinical guidelines have been designed to assist healthcare providers in the diagnosis and treatment of acute respiratory tract infections, however, guidelines may fall short due to level of healthcare facilities and service rendered. To provide the appropriate treatment that would ensure patient recovery, accurate diagnosis and cause of illness must be properly determined. Furthermore, the guidelines can be used by healthcare providers to inform patient on the treatment they would likely to receive when diagnosis and cause of illness has been confirmed (De São José et al., 2014; McCracken, 2001).

Clinical practice guidelines are vigorous, and systematic in their approach, and should be able to address the respiratory tract infections and pathogens involved. Each guideline for the diagnosis of respiratory tract infections has been based on either viral and /or bacterial infections, or those without pathogens aetiology (De São José et al., 2014; McCracken, 2001).

1.5.3.2 Clinical practice and healthcare challenges

There are many challenges and problems that affect the employment of comprehensive, effective and reputable healthcare services to the communities including where the facilities are located. The internal challenges are systemic in nature, ranging from healthcare management, performance measurement, control systems, legislation and guidelines (Ippolito et al., 2023). All internal challenges can be managed to improve healthcare and impact the clinical management, by introducing artificial intelligence, in particular machine learning (Kelly et al., 2019).

The external healthcare management challenge often include the negative attitude towards clinical guidelines, coming from healthcare facilities managers, healthcare providers, and patients themselves. Many external factors can affect the ability of the healthcare provider to execute the appropriate guideline recommendations, while patient expectations of some treatments can hinder healthcare provider's adherence to the guidelines, especially its recommendations (McCracken, 2001).

The main driver of the issues and problems encountered in the management and healthcare service delivery is the cost implications involved in the running of healthcare facilities. The budget provided to the facilities may not be enough to run all essential services, rendering clinical administration ineffective as costs escalate (Ostadi et al., 2023).

Zhang et al. (2016) reported that the cost implications for treating RTIs in children is measured in millions of dollars annually, severely affecting low- and middle-income countries (LMICs). This can be attributed to low gross national income. In high-income countries, treatment for acute respiratory tract infections per case of outpatients was reported to be US$ 334.6 in USA, US$ 251.1 in Germany and US$ 254.9 in Australia (Zhang et al., 2016). In LMICs, treatment for hospitalised acute respiratory tract infection per case cost US$ 249.7 in Zambia and US$ 162.1 in Kenya (Zhang et al., 2016). In South Africa, treatment cost for outpatients with acute respiratory tract infection was US$ 263.1 per case, and cost for treatment in a tertiary hospital for acute to severe respiratory tract infection was US$ 491.4 per case (Zhang et al., 2016).

1.5.4 Evidence-based clinical practice and decision support

Evidence-based medicine is achieved when clinical prediction tools, clinical decision rules and/or clinical decision support tools are used as prediction algorithm and risk models for prognosis and diagnosis of infections. Decision rules aim to directly encourage healthcare providers' conduct, while prediction rules provide support to healthcare providers in predicting risk with the premise that precise predictions would lead to improved decisions (Cowley et al., 2019; Sutton et al., 2020).

Evidence-based clinical practice is a framework that utilises the power of statistics and probability to determine the cause of disease and infections. Infections and diseases have been diagnosed with accuracy and precision when statistical tests are applied to measure past cases and disease likelihoods are predicted, which create further knowledge in the prognosis (Newman and Kohn, 2020; Phillips, 2010).

To complement the clinical practice guidelines, evidence-based decision could be powerful tools to improve clinical decision making on the foundation of grouped clinical signs and symptoms. However, most published evidence-based clinical decision support tools have been developed for serious infections primarily at hospital emergency departments, which lack external validation for primary healthcare practice (Van Ierland et al., 2015).

The purpose of evidence-based clinical decision tools is to estimate the likelihood of a clinical condition by taking into account a small amount of highly valid predictors and variables. Their

aim is to assist healthcare providers in making decisions under unfavourable conditions and improve diagnostic, and decision-making, with the primary purpose of improving the quality of patient care (Sutton et al., 2020). Some evidence-based clinical decision tools aim to 'rule out' conditions by classifying patients who are very unlikely to have an illness, ultimately reducing unnecessary testing without compromising patient care (Cowley et al., 2019). Kwan et al. (2020) reported that, when an evidence-based clinical decision support system is employed, patients receiving desired care increased proportionally by 5.8% (4.0% - 7.6%).

1.5.4.1 Evidence-based diagnosis of respiratory tract infections

The diagnosis of respiratory tract infection using evidence-based clinical guidelines to reduce variability has been previously adopted to support clinical decisions. A study evaluated an evidence-base clinical guidelines and decision support tool, developed to decrease the prescription rate of antibiotics in RTI cases presenting at the private hospital in Quito, Ecuador. The clinical decision support tool included guidelines to diagnose and identify self-limiting respiratory tract infections. Researchers in the study also analysed the short- and long-term effects of the clinical guideline tool implementation. A three phase implementation process for the tool was initiated, the first one in 2010, followed by the post-implementation in 2011 and the last implementation in 2018.The study showed the reduction in antibiotics prescription by healthcare providers, resulting in preventive overuse of medication for self-limiting respiratory tract infections (Sánchez et al., 2021). Another evidence-based clinical decision system was developed to aid healthcare providers to ruling out pneumonia, which also aided in reducing rate of unneeded antibiotic prescriptions (Steurer et al., 2011).

A review of prediction rule tools developed in the 128 centres in eight European countries (Toll et al., 2008), aimed to measure different validation methods to assist with the strengthening of evidence-based clinical decision and prediction rule tools. In this study, researchers indicate that a major limitation to all clinical decision systems was the validation of such systems, as testing the generalisation of evidence predictions rules across a gold standard of healthcare providers, was difficult (Toll et al., 2008).

1.5.4.2 Artificial intelligence and decision-rule in clinical practice

Machine learning is a branch of AI that use specific statistical and mathematical traits to identify patterns from data to solve problems and improve with experience. Machine learning is a

computation of intelligent computer programs and written algorithm that learn from provided data and predict the outcome (Kersting, 2018). Machine learning was first used in the 1952, two years after companies started developing machines with ability to make decision like humans (Kaul et al., 2020). Electronics industries adopted AI computational modelling in their robotics engineering and innovation; however, medical science was slow to adopt AI. In the early 1960s, medical science was introduced to AI when biomedical web-based knowledge database systems were developed and implemented (Kaul et al., 2020).

Data captured in electronic health systems from patients seen at healthcare facilities and research sites, which forms big data and infrastructure for data warehousing, are in use for purposes of guiding evidence-based clinical practice in recent times. Artificial intelligence approaches, specifically Gaussian Naive Bayes (GNB), deep and reinforcement learning, are remarkably well-matched to deal with the amount of patient data collected, type and format of data and impending questions in primary healthcare. Artificial intelligence can assistant healthcare providers in simplifying difficult task of stratifying patients at risk for interventions, identifying those most at risk of severe infection and disease (Giordano et al., 2021).

Since 2010, there has been an exponential rise in electronic data and breakthroughs in computing power which contributed to significant advancements in AI, including machine learning, for clinical decision support for diagnosis, treatment, and prognosis. Although there have been favourable outcomes in the use of AI and machine learning, there also been challenges in interpreting machine learning results and translating these into clinical terms (Hernandez-Boussard et al., 2020; Yang et al., 2023). Most of the challenges experienced relate to data sample lacking representation and missing values, results interpretation, and model setting such as hyperparameter tuning for accurate prediction and classification (Ostojic et al., 2024; Stam et al., 2023).

1.5.4.3 Machine learning techniques in clinical practice

The use of machine learning for development of clinical prediction models was not common until the early 90s, when artificial neural networks (ANNs) and decision tree (DT) algorithms were used for cancer diagnosis and detection (Cruz and Wishart, 2006). In a published review on the use of machine learning techniques; ANN, GNB, support vector machine (SVM) and DT were the four main techniques used for development of cancer prediction models. The main finding for cancer prediction models was improved inference in cancer progression and future

feature selection, especially in modelling predictions for cancer susceptibility, recurrence and survival (Kourou et al., 2015).

There are many studies which report respiratory tract infection morbidity, mortality and the associated risk factors. Most of these studies make use of conventional machine learning prediction models (Troeger et al., 2018). The published reports on machine learning applications in the RTI domain are very limited, especially those that assist in RTI diagnosis and classification using prediction models. Available reports used retrospective data to model machine learning classification of respiratory tract infections caused by viruses (Mai and Krauthammer, 2017), predicting past influenza outbreaks (Zhang and Nawata, 2017) and the risk of mortality of non-severe pneumonia cases (Tuti et al., 2017). Findings for a viral RTI classification model attained a positive predictive value ranging from 50.1 % to 64.7% and a negative predictive value ranging from 52.1% to 100% (Mai and Krauthammer, 2017). One study on influenza outbreaks, reports an accuracy of 95% (Zhang and Nawata, 2017). Similar work using risk of mortality of non-severe pneumonia cases, attained an area under the curve (AUC) above 0.7 (Tuti et al., 2017).

The finding of these studies proved that machine learning models can classify respiratory tract infection and pneumonia, thus improving clinical decision making. Both studies recommend using larger sample sizes with more clinical features than what is currently monitored during diagnosis (Naydenova et al., 2016; Olayemi et al., 2019). Both studies achieved over 80% precision, accuracy, sensitivity and AUC. There is little research that reports the application of machine learning in the classification of RTI cases. Therefore, the current research seeks to close the gap in the existing body of knowledge by applying novel machine learning approaches to determine and classify RTI and non-RTI cases.

CHAPTER 2

2 CLASSIFICATION THEORY

2.1 Research Methodology Framework

Supervised machine learning is defined as processes that use historical data, learn noticeable relationships, and use that knowledge to make prediction on new data. This process of learning was made possible by statistical algorithms employed to read raw data, select suitable features, and build a suitable prediction model for classification of target outcome on new data (Biswas et al., 2021). A framework overview described below and in figure 7 in materials and methods section, outlined in graphics, the processes applied and employed to classify possible RTI cases from other classes, following a classification model design and evaluation.

The initial step was to collect the dataset from where it was securely stored and prepare data for analysis. The main dataset, which contain collated RTI, gastrointestinal tract infections (GTI) cases and acute febrile disease of unknown cause (AFDUC), was cleaned, configured, feature scaled and put through many pre-processing procedures that identified features suitable for model design and development. After pre-processing, statistical analysis was conducted, and then the dataset was imputed and normalised for model design. Additionally, post statistical analysis, data imputation and normalisation, the main dataset was split into three subsets of classification classes, RTI, AFDUC and GTI as model 1 data subset, RTI and AFDUC as model 2 data subset, and lastly RTI and GTI as model 3 data subset. Each data subset was split into train and test datasets, which were used to train all models designed and to test each model through series of evaluation metrics listed below.

2.1.1 Data pre-processing and feature engineering

Data pre-processing and feature engineering were the two phases of machine learning that prepared the data for model development, testing, validation and evaluation for classification of a target diagnosis outcome. The processes of pre-processing, feature engineering, data splitting and classification of target outcome were done using Python libraries listed in table 1 below.

Table 1: Python Libraries for ML

Library	Function
NumPy	Indexing, sorting, and reshaping array of real numbers in multi-dimension.
Pandas	Processing data sets into pre-defined Data Frames supported by NumPy multi-dimensional arrays.
Matplotlib Seaborn	Graph plotting and visualization of multivariate data points.
ScyPy	For statistical analysis (ANOVA package, chi-squared contingency packages).
Scikit-learn	Pre-processing packages for data cleaning and manipulation, Classification packages for model design and evaluation (SVM, RF, ANN, test and train, confusion matrix and evaluation metrics.

2.1.1.1 Data pre-processing

Pre-processing mainly involved data cleaning and normalisation, which also included transformation and integration of data, outliers and noise identification (García et al., 2016).Data cleaning included the removal of noise and inconsistent data, while data transformation and integration was responsible for conversion and consolidation of data. Data normalisation was performed in both pre-processing and feature selection phases, which included data and variable values recoded. Missing values were dealt with by either removing the whole record for critical variables and imputing a value in replacement, while stop words were removed.

Missing Value Imputation

Analyses based on incomplete cases were shown to be biased when the data was missing at random. However, removing the whole record with missing variable data points did not bias the results statistically but inferences impacted model fitting during design. Such biases were overcome by performing missing value imputations. Data imputation included replacing missing values with the mean of the observed values for that variable and using information from related observations within the collected variables.

For the imputation of both categorical and numerical (ordinal and continuous) values, two machine learning techniques were employed and the best one chosen and used. The first technique employed was K-Nearest Neighbour (KNN) imputation, where imputed values for cases with missing values corresponded to the value of similar cases. Imputation tests employed by KNN was the distance calculation of variable values in the presence of missing values to measure the imputation value. The second imputation technique used was Multivariate Imputation by Chained Equations (MICE), which operates under the assumption that values or observations

are missing at random. Imputation tests employed by MICE included both linear and logistic regression models suitable for both continuous and categorical variables (Memon et al., 2023).

Remove Stop Words

Stop words are commonly used words such as ('i', 'saw', 'we', 'don't', 'know') that are naturally filtered out during a natural language processing (NLP) task to improve text analysis (Kashina et al., 2020). Stop word removal was essential to improve machine learning model's priority on words which would translate in achievement of accurate predictions. Stop word removal also helped in the reduction of dimensionality in the data set. NLTK Python library has no less than 179 stop words available for English language package. Examples of English stop words of letter "a" are: (a, about, above, after, again, against, all, am, an, and, any, are, aren't, as, at).

Data Normalisation

The collected clinical and non-clinical data, was pre-processed using a normalisation technique essential in the process of assembling the collected data. Consistency across features was varying data ranges maintained, moreover, it was feasible to modify numerical variables in a given dataset. Amongst a variety of normalisation techniques, the min–max normalising technique was ranked the best and has been regularly deployed for linear transformation and data normalisation. The data was altered to settle within a pre-established range in this occurrence. The maximum value for each characteristic was one, and the minimum value was zero; and the values that remained were transformed to decimal numbers ranging from 0 to 1.

2.1.1.2 Feature engineering

Feature engineering was a process that improved the performance of a model, using algorithms that can apply: feature scaling and selection, outliers and data encoding, dimension reduction and data balancing. The importance of selecting suitable variables as predictors and features in a prediction model to classify target class, as preferred outcome, rests on the understanding what the algorithm was supposed to learn.

Feature scaling was a process that involved feature integration within the same magnitude and a common range (García et al., 2016). Features scaling included operations on text language processing namely, tokenisation, stemming and lemmatisation. The tokenisation process

segmented text into single words, clauses, and sentences by separating words and removing punctuation. Stemming reduced related words to a common stem (Kashina et al., 2020), while lemmatisation removed the inflectional endings of words and returned the dictionary form of the word.

Tokenisation

In plain terms, tokenisation is a procedure of turning a series of words in a sentences into tokens. Tokenisation was the most essential technique in natural language processing pipeline. Tokenisation turned a significant portion of text in a sentence into a string character, called token. There were many tokenisation algorithms to choose from, however, we chose to use word piece tokeniser over many available such as: Penn-tree, Byte pair encoding and Sentence piece tokenisers just to name a few.

Stemming

Stemming is a procedure of converting and changing the form of words into their base form. Stemming technique helped to produce morphological variations of a base word to express grammatical attribute. Stemming is a process that bind a portion of a word with suffixes or prefixes such as (-ed, -ize, -s, -de, mis). So, stemming resulted in text that are not genuine words. Stemming algorithms are extensively used in pre-processing, before machine learning modelling design.

Lemmatisation

Lemmatisation is a natural language processing algorithm that process text on the morphological analysis of each word. Lemmatisation also uses wordnet, a morphological function, built on a grammatical knowledge base platform. Lemmatisation was conducted following stemming, and it was carried out on the tokenised words. Following stemming, lemmatisation is the highly essential pipeline of natural language processing technique. In this research study, stemming and lemmatisation were used extensively, e.g., ‘convulsions’ became ‘convulsion’, ‘dehydrations’ became ‘dehydration’ and ‘infections’ became infection.

Feature Selection

Feature selection has a variety of manual and data-driven techniques that falls into three classes, namely the filter, wrapper, and embedded methods. The filter techniques included methods that had computed the correlation of each feature to the outcome variable class. The wrapper tech-

nique allowed for measure of subset of features that were uninformative and proposed a score for them. Lastly, the embedded technique performed feature selection by regularisation which encouraged models to use few features as opposed to many features (Ou et al., 2020; Zheng and Casari, 2018).

Lastly, the techniques used for feature selection to complete the data processing phase included, the Term Frequency and Inverse Document Frequency (TF-IDF) and the Bag of Words (BOW) technique. TF-IDF had shown to be the best technique when compared to the BOW during feature engineering process, which also conserved some semantic nature of the sequence. The phrase "term frequency" expressed the number of times a word occurs in a text sequence, such as the likelihood of a word to appear in a text documentation. The formular of TF-IDF consisted of two piece, the initial piece computed the Term Frequency and the second piece estimated the Inverse Document Frequency which lay between 0 and 1.

2.1.2 Machine learning classifiers

2.1.2.1 Support Vector Machine

The support vector machine uses a minimum of two data points belonging to two classes separated by a hyperplane line, and the techniques' goal was to classify target outcome in a clean and unused dataset and indicate which class they belong to.

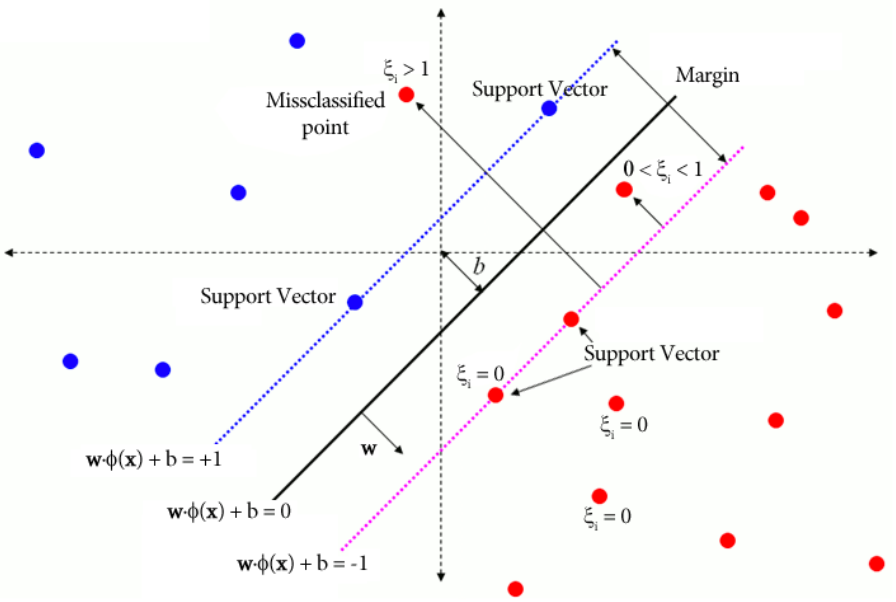


Figure 1: Support Vector Machine

The SVM figure 1 above, classified the point above or on the hyperplane as class +1 (X), and the point below the hyperplane as class -1 (Y), and w as weight, x as independent variable, and

b bias, as per the equation 1 below.

$$f(x_i) = \begin{pmatrix} +1 & if\ w \cdot x + b \geq 0 \\ -1 & if\ w \cdot x + b < 0 \end{pmatrix} \tag{1}$$

When class X and class Y are separated by hyperplane, points from each class that are closer to the line will be marked as support vectors. The further apart the support vectors are from each class, the bigger the margins between them increased. The bigger the margins, the higher our confidence that the algorithm can classify RTI case. The technique was chosen because overfitting has minimum effect on predictors during modelling.

2.1.2.2 Random Forest

The random forest (RF) consisted of multiple combinations of decision trees, a technique of combining many draft models to achieve one optimal final model, illustrated in figure 2 below. Using a decision tree technique for target outcome classification, depends on the chosen feature and the condition applied to the technique's computational function.

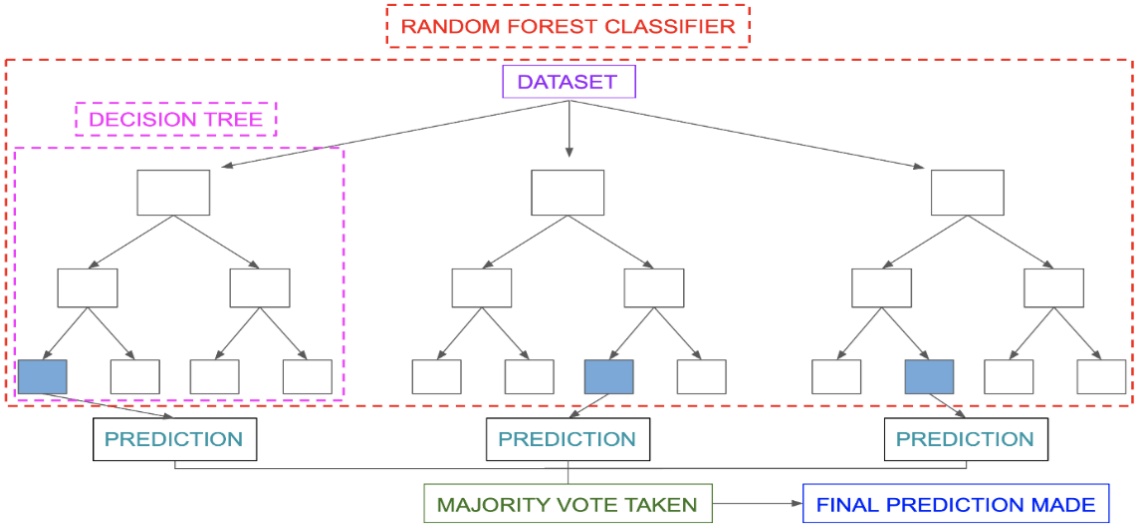


Figure 2: Random Forest

Since we were conducting a prediction model for RTI classification, a feature (e.g., sex or HIV status) represents a condition that was split into two branches (notes), male and female branches for sex and HIV- and HIV+ branches for HIV status. Then each branch was further split into two more branches (yes/no) based on another condition (e.g., age <1) presented by feature 'age'. Until branch splitting was not required anymore, the technique was stopped and a decision was given.

$$I_{Gini}(X_k) = \sum_{i=1}^{C} p(i|k)(1 - p(i|k)) \tag{2}$$

The random forest equation 2 above measures the Gini impurity (outcome), and was based on the probability of misclassification of a data point categorised using a label chosen randomly from the branch. To explain each parameter, $p(i|k)$ was the frequency of the label i at a branch and C was the number of unique labels. This algorithm was chosen because the same features were used more than once with different conditions each time to maximize on an expected decision of RTI outcome classification.

2.1.2.3 Artificial Neural Network

Multilayer perceptron (MLP), is an artificial neural network made up of multilayer neurons that works in similar ways to the brain processing an input signal. In this study, multilayer neuron network technique modelled a complex patterns of neurons that were able to predict a target outcome as illustrated in figure 3. All features were included as predictors on the input layer (1st layer) of the algorithm code, then the technique processed the input as 'brain neuron' and passed the processed information through to the middle layer called hidden layer. The hidden layer then processed the information as inputs and the output information was transmitted to the output layer as an expected prediction outcome (Bre et al., 2018; Zhang et al., 2018). This technique was chosen for its high accuracy in prediction models which were closer to the expected and actual value.

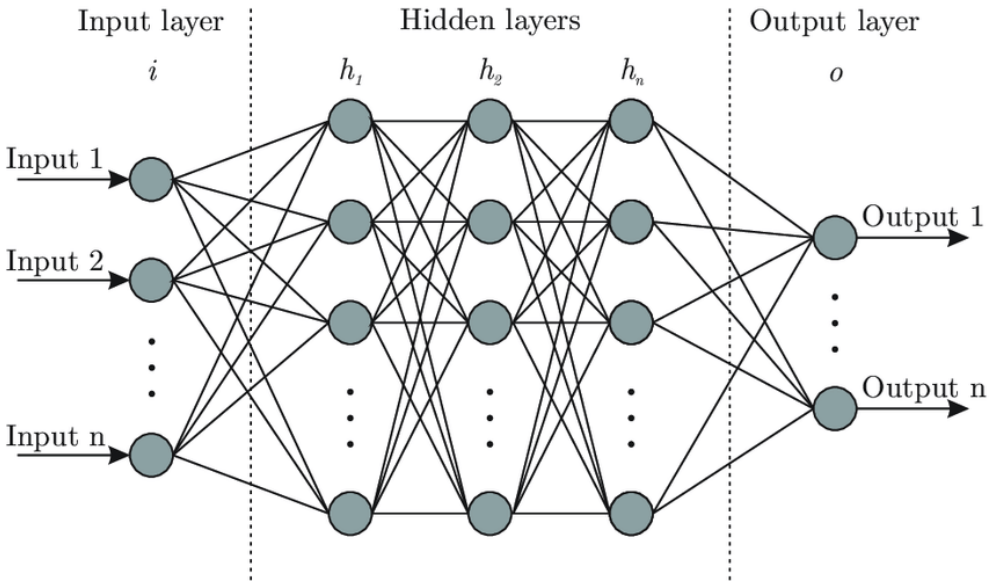


Figure 3: Artificial Neural Network

The two ANN equations below measure the function outcome probability (Y_0/Y_1), based on a

of linear regression model. The equation 3, is a linear format of the classification and equation 4, the sigmodal format classification. The parameters w^{-T} represents the weights applied on each independent variable $\bar{x}$, and b represent the bias in input layer.

$$f(x) = z = w^{-T}(\bar{x} + b) \tag{3}$$

$$f(x) = z = \sigma(w^{-T}(\bar{x} + b)) \tag{4}$$

The interpretation of the artificial neural network prediction outcome is a horizontal tree diagram. The tree diagram was composed of black and grey lines connecting all added features and/or variables in input layer to all nodes in the hidden layer for statistical computation. The outcome information in the hidden layer was then connected to the output layer classification outcome Y_1 (RTI cases) and Y_0 (non-RTI cases) with black and grey lines. The black lines indicated the positive weights and grey lines indicated the negative weights. Line thickness was in proportion to the relative magnitude of each weight (Zhang et al., 2018).

2.1.3 Model optimisation

During the model training, two optimisation techniques, namely K-fold and GridSearch cross-validation, were applied to strengthen and optimise the best classification model. GridSearch cross-validation was the process of performing automated hyperparameter tuning, to determine the optimal values for a trained model. On the other hand, k-fold cross-validation was a method of data resampling, used to choose and fine tune hyper-parameters of a model by maximizing the performance of prediction on a limited data (Ou et al., 2020). Firstly, the data was randomly split into 10 subsets (10-fold iterations), then the model was trained ten times with 9 subsets as training sets and 1 subset as testing set. In these 10 iteration runs, each data point (case) ended up in the 1 testing subset exactly once.

2.1.4 Model evaluation

Evaluation of the machine learning model was a crucial and important step that determined how effective the model was, based on performance metrics using test dataset. For the purpose of model evaluation, six performance metrics were used, which were adapted metrics from the confusion matrix in figure 4.

A confusion matrix is made up of the actual values and predicted values, measured as true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). Each metrics

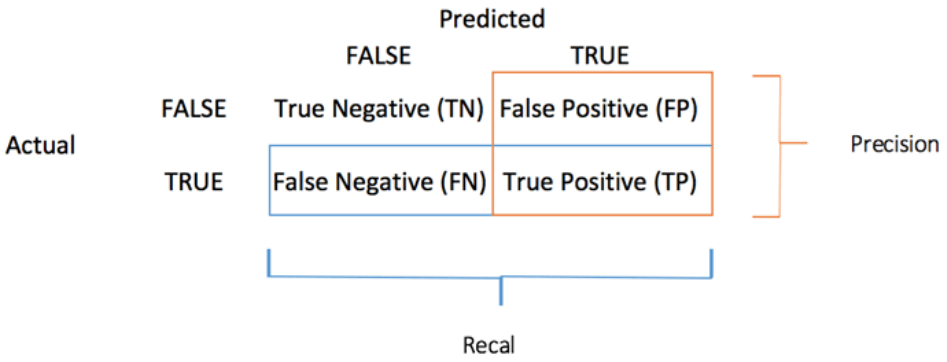


Figure 4: Confusion Matrix

is briefly explained below:

1. **Accuracy**: Measure of total number of true predicted cases over the total number of all predicted cases.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{5}$$

2. **Precision**: Measure of all true positive predicted cases over total number of positive predicted cases. Fraction of true predicted cases out of total true cases.

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

3. **Specificity**: Measure of total number of true negatives cases over total number of all negative cases.

$$Specificity = \frac{TN}{FP+TN} \tag{7}$$

4. The **F1-score** was used to measure accuracy of predictions by the model. This was the weighted harmonic mean of the models' precision and recall.

$$F1_{score} = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{8}$$

5. **Receiver operating characteristics** (ROC) curve, a classification performance metric, figure 5, which plotted a probability curve of true positives (sensitivity) against false positive (specificity) rates. Sensitivity referred to the proportion of positive RTI cases that were correctly classified as positives (true positive rate), while specificity referred to the proportion of negative non-RTI cases that were correctly classified as negatives (true negative rate). Area under

the curve was an overall performance measure of classifier to distinguish between two or more classes, which was maximized by ROC curve hugging the top left corner of the graph (Jebb et al., 2017; Sparks et al., 2008; Uddin et al., 2019). Area under the ROC curve is a graphical schematic that assesses the ability of each classifier and model, to measure the predicted and actual values beyond the threshold of 0.5 AUC.

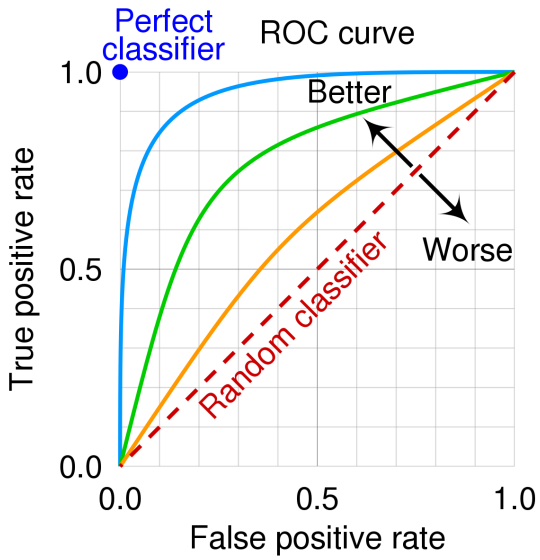


Figure 5: ROC Curve and Area Under Curve

The ROC curve graph x-axis is 1 – specificity and y-axis sensitivity. On the graph, a diagonal line is drawn from the bottom left corner to the top right corner of the graph, representing 50% of both sensitivity and specificity of test model and area under the ROC curve of 0.5. On the graph, we put points for cut-off parameter value of the test model, plotting the true positive rate and false positive rate of the test at that value and connecting these points that created a ROC curve. The desired outcome performance of the model was a ROC curve far above the diagonal line with the measured AUC between 0.8 and 1, to accept the model.

CHAPTER 3

3 MATERIALS AND METHODS

3.1 Primary Study Design and Setting

The African Network for Improved Diagnostics, Epidemiology and Management of Common Infectious Agents (ANDEMIA) (Schubert et al., 2021), was a syndromic hospital-based sentinel surveillance program investigating the aetiology and clinical characteristics of acute respiratory tract infections (RTI), acute gastrointestinal tract infections (GTI), and acute febrile disease of unknown cause (AFDUC) in sub-Saharan Africa. The study included 12 rural and urban healthcare facilities in four African countries: the Republic of South Africa, Burkina Faso, Democratic Republic of the Congo and Côte d'Ivoire). In South African setting, the study was conducted at Kalafong Hospital, an urban site, and Mapulaneng/Matikwana hospitals, the rural sites, enrolling informed consented patients presenting with either RTI, GTI and AFDUC syndromes, conducted between 2018 and 2022.

3.1.1 Primary Study Population and Sample Size

The study population included patients of all ages who meet the RTI, GTI or AFDUC case definitions and a sample size of 1600. ANDEMIA aimed to identify the aetiology of common infections in selected African settings, assess associated risk factors, and monitor antimicrobial resistance. Additionally, the program aimed to expand the understanding of the causes and management of common infectious diseases in sub-Saharan Africa, particularly among children, and to improve the detection and management of emerging pathogens. To achieve its aim, a comprehensive diagnostic tool/protocol was employed, which included classic bacteriology, antimicrobial resistance screening, multiplex real-time PCR, and high throughput sequencing. Following the COVID-19 outbreak, PCR testing for SARS-CoV-2 and full genomic analysis information was initiated. Furthermore, standardised questionnaires were employed to collect clinical data, demographic, behavioural, and socioeconomic data for epidemiologic analyses.

3.1.2 Primary Study Inclusion and Exclusion Criteria

RTI cases were enrolled when the following criteria were met:

- Fever $\geq$ 38°C or history of fever (or self-reported) with onset within the last 10 days.
- $\geq$ 1 respiratory signs/symptoms in the last 10 days: cough, expectoration, dyspnoea (shortness of breath), pulmonary consolidation, and chest pain.
- Signs of severity necessitating hospitalisation.

RTI cases were not enrolled when the criteria listed above were met and the patient was not hospitalised or admitted for less than 24 hours.

AFDUC cases were enrolled when the following criteria were met:

- Fever $\geq$ 38°C or history of fever (or self-reported) with onset within the last 10 days,
- Absence of evident cause other than acute neurological disease,
- Signs of severity necessitating hospitalisation.

AFDUC cases were not enrolled when the criteria listed above were not met and the patient was not hospitalised or admitted for less than 24 hours.

GTI cases were enrolled when the following criteria were met:

- 3 loose or liquid stools in the last 24 hours,
- Diarrhoea $\leq$ 4 weeks,
- Signs of severity necessitating hospitalisation.

GTI cases were not enrolled when the criteria listed above were not met and the patient was not hospitalised or admitted for less than 24 hours.

3.2 Secondary Study Design and Population

This was a secondary data analysis research study of data collected from a retrospective observational cohort study of respiratory tract infections: RTI, and two non-RTI cases: AFDUC and GTI, enrolled in ANDEMIA surveillance study. All patients enrolled with each syndromes were included and there was no additional exclusion criteria applied.

3.3 Data Collection and Management

ANDEMIA case information was collected on paper forms at each site, and transcribed into electronic form in Voozanoo™, a secure online database capturing system. Voozanoo™, is a copyright user-oriented web-based platform for development of databases for research in clinical studies and other fields of science. The data entered and stored in Voozanoo™, had no personal identification of the enrolled patients, but rather a designated and anonymised identification number. The final dataset size before pre-rocessing was 1204 rows for RTI and 1748 rows for non-RTI (982 AFDUC and 766 GTI). The data analysis was carried out at the National Institute for Communicable Disease (NICD), Johannesburg, South Africa.

3.4 Collected Study Variables

In this study, all structured variables collected were used as features (predictors) for machine learning model design and evaluation, with only three being continuous numerical and the remainder categorical—the majority of which were nominal, primarily binary in nature, while the rest were ordinal, based on ranked categories. Additionally, a few cases diagnosed with both RTI and GTI were excluded from both statistical analysis and model design as the sample size was small (n = 169 rows). See below the finalised classifications tasks datasets Figure 6.

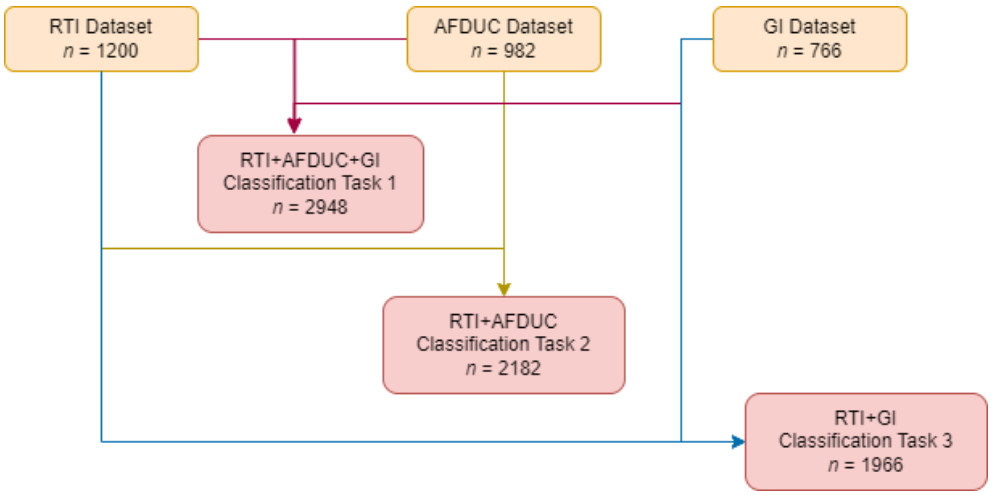


Figure 6: Finalised Classification Task 1, 2 and 3 Datasets

Amongst these predictors, history of fever (Wright and Auwaerter, 2020) and cough (Morice et al., 2020), were the main signs and symptoms associated with clinical respiratory infections. Other clinical signs and symptoms collected were patient's hospital admission temperature readings, shortness of breath, expectoration, chest pain, pulmonary consolidation and

number of days since symptoms onset, which were primary risk-factors likely associated with RTI outcome. Additional clinical and non-clinical features collected included basic patient demographic and socio-economic characteristics, such as level of education, employment status, geo-location, drinking and cooking water source. Additional, unstructured variable derived through natural language processing—such as free-text other respiratory symptoms descriptions—were obtained during secondary data collection. However, they were excluded from the final modeling and statistical analyses due to several limitations: a high proportion of missing values and most overlap with existing structured variable: namely cough, and insufficient data density or variability. These variables remain accessible for future exploratory work or alternative modeling strategies. However, the cough value was extracted with a logic condition to supplement the cough variable that was coded missing or no, when cough was indicated in the other respiratory symptom variable as additional respiratory sign.

3.5 Data Pre-processing and Model Design

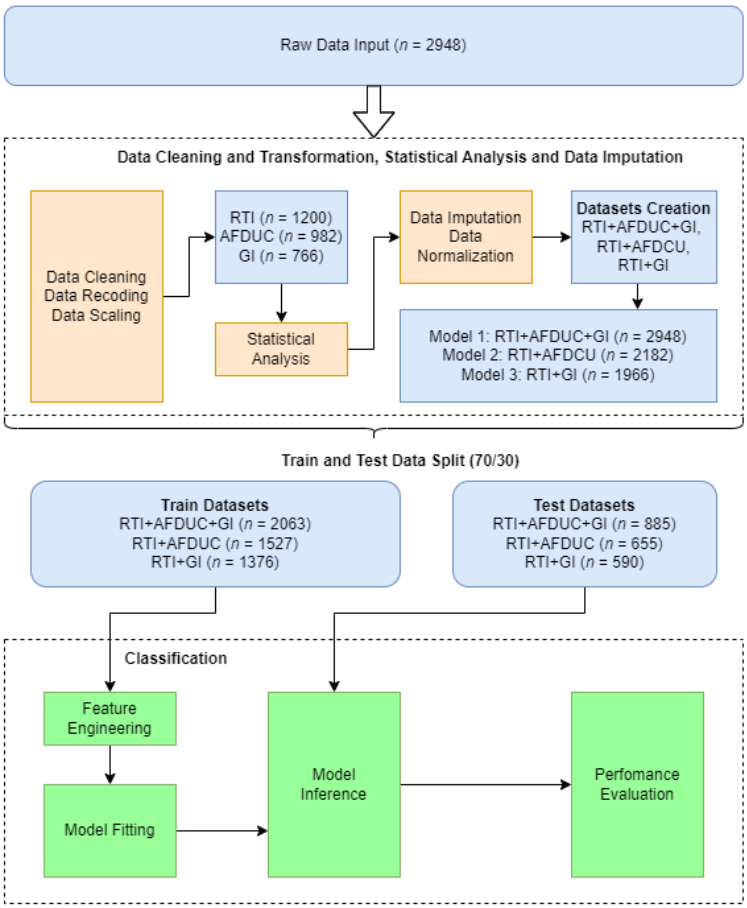


Figure 7: Classification Model Design

After downloading the main dataset, incomplete records with only records ids were re-

moved, data cleaned and outlier deleted final and a final cleaned dataset containing a total sample size of 2948 data-points (rows) created. Following data cleaning, a single final dataset was used for analysis with sample size per syndrome of 1200 rows for RTI cases, 982 rows for AFDUC cases and 766 rows for GTI cases was created, with 30 predictors for all syndromes (Figure 7). The missingness test was computed on this single dataset, to observe the pattern of missing values. The majority of variables had missingness levels below 20%, and a few had levels between 40% and 50%, only one variable had missingness level of 53%. Figure 8 show the levels of missing values, and the most important variables had fewer missing values except one core variable missing 50% of the values. The missingness percentages were computed with simple NumPy® mean/sum times 100 and the diagram was computed with missingno® python library package.

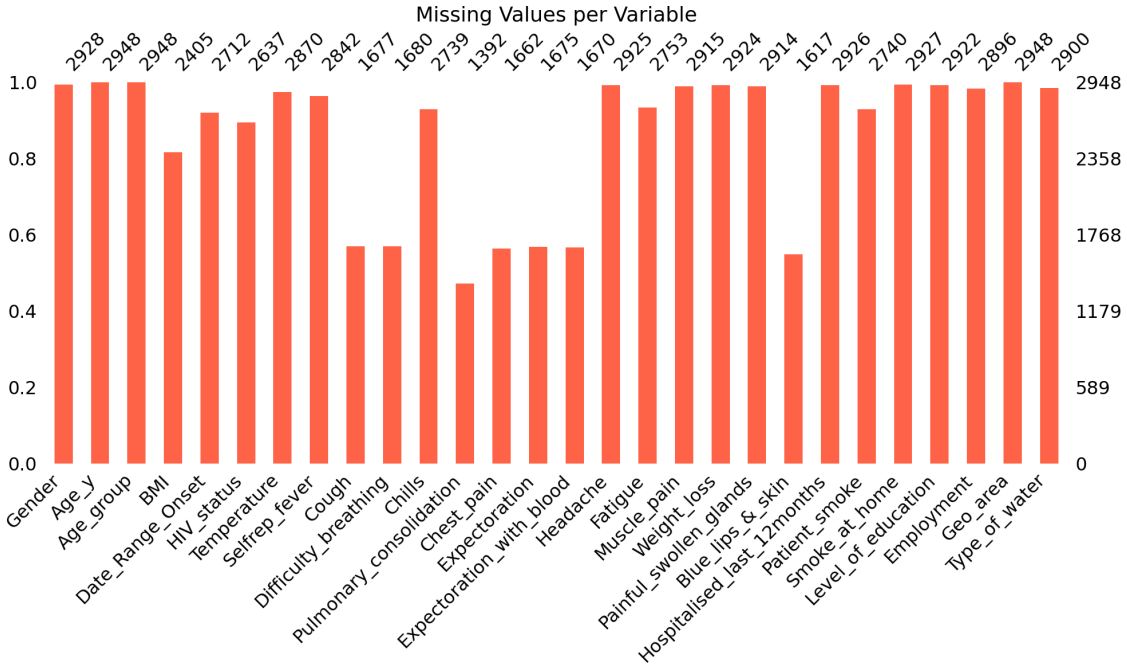


Figure 8: Missing Values per Variable

Before statistical analysis test were computed following data cleaning, the final and cleaned dataset categorical variables were recoded to numerical variables, that match the format as either ranked or Boelean. Recoding was done with Numpy® library for the non-missing values and the three syndromes or classes were compared to each other using two statistical tests: one-way ANOVA for continuous variables and X^2 chi-square test for categorical variables. The reason to use ANOVA of the non-normally distributed continuous variables and X^2 chi-square on the categorical variables was because the sample size was high enough for the computation. The statistical test p-values and significance was reported in Table 2 and interpreted in the dis-

cussion.

Following statistical analysis, the variables with missing values were then imputed using sklearn impute library, specifically using Iterative imputer and KNN imputer. Reasons using these two techniques was that over 60% of the variables had less than 10 percent of missing values, which made it possible to use these techniques. Furthermore, the variables were normalised using Min–Max normalising technique. Before data imputation and normalisation, the final dataset was separated into classes, based on the syndrome (RTI, AFDU and GTI), and computation was performed on each class separated into individual datasets, see figure 7. Following imputation and normalisation, three classification tasks' datasets were created that was modelled independently with three machine learning techniques each. Classification task 1 had all three classes merged together (RTI, AFDUC and GTI datasets), while classification task 2 was only RTI and AFDUC datasets merged together and lastly, classification 3 with RTI and GTI datasets merged together.

Additionally, correlation tests were computed on continuous variables and also on recoded categorical variable, to measure two variables association and possible multicollinearity. The correlation method was performed on the variables that had values imputed for each classification task dataset. For positive correlation, values ranged from zero (negligible correlation) to one (strong correlation (0 < correlation <= 1) and negative correlation values ranged from zero (negligible correlation) minus to one (strong correction) (-1 <= correlation < 0). For the association of two variables, when positive correlation is observed, means one variable increase, so does the other, and the negative correlation is observed, means when one variable increase, the other decrease.

Each classification task datasets were modelled with three classification algorithms, model design and development, and lastly evaluation. The three classification task datasets were split into data for training and fitting the model, and data for validation and testing of the model. Initially, cross validation (CV) technique was used to fine tune the hyper-parameters for better model development, after the data was split first into train and test data. The sklearn package with model selection packages was used, particularly GridSearchCV was computed. A normal model training, testing and evaluation procedure was performed parallel to the CV procedure. We later decided to drop the CV procedure, as it did not yield a better model, when evaluation procedures were performed. Furthermore, the three models were developed, using three algo-

rithms namely: random forest, neural network MLP and support vector machine. Each model was trained, tested and validated and lastly, evaluated using 3 classification tasks datasets, see Figure 9, which adhere to standard ML workflow.

The data was split into 70% for model training, 30% for model testing. The features that were selected to design and build the prediction model were also measured to determine their significance in building a robust model. The two techniques were employed to measure and compare each feature's significant outcome between the two methods. Mean Decrease in Impurity (MDI) and Permutation Importance were the techniques used to measure the feature's importance during model development. MDI measured and calculated the Gini impurity reduction score, while permutation measured the importance of a feature following a random shuffle.

3.6 Experimental Environment

Cleaning of data, feature design, statistical analysis, model design and development, and model evaluation were done on Python scripting language (version 3.12.4; Python Software Foundation, Delaware, USA) 64-bit based on the SPYDER (version 6.0.3 (standalone)) Integrated Development Environment (IDE), installed on Microsoft Windows 11 professional 64-bit operating system, an 11th Gen Intel Core i5-1155G7 @2.50GHz processor with 8Gb random access memory. The following Python libraries were used: Pandas (version 2.2.3) in high performance data structure manipulation and transformation, while NumPy (version 1.26.4) facilitated the manipulation of numerical data processing and Matplotlib (version 3.10.0) and Seaborn (version 0.13.2) mainly for plotting graphs. Scipy (version 1.14.1) library was used mainly for statistical analysis and Scikit-learn (version 1.6.1) library, which contained a wide range of packages such as pre-processing, model_selection, modelling classifiers and metrics packages, was used for model training and evaluation.

CHAPTER 4

4 RESULTS AND FINDINGS

4.1 Patient Characteristics and Datasets

AFDUC and GTI datasets were fully merged with RTI dataset, resulting in a design classification 1 dataset. Furthermore, AFDUC and GTI were merged again with RTI separately to make classification task 2 dataset with RTI and AFDUC, and classification task 3 dataset with RTI and GTI as illustrated by Figure 9.

Model Classifier	Classiffication Tasks
Random Forest	Task 1: (RTI + AFDUC + GI) Task 2: (RTI + AFDUC) Task 3: (RTI + GI)
Artificial Neural Network	Task 1: (RTI + AFDUC + GI) Task 2: (RTI + AFDUC) Task 3: (RTI + GI)
Support Vector Machine	Task 1: (RTI + AFDUC + GI) Task 2: (RTI + AFDUC) Task 3: (RTI + GI)

Figure 9: Finalised Classification Task 1, 2 and 3 Datasets

4.2 Pre-processed Data and Analysis

Statistical analysis was done on the cleaned, recoded and un-imputed ANDEMIA dataset containing all three classes/syndromes as listed in Table 2. Amongst the three syndromes, 21.9% of cases were of people living with HIV (PLHIV) and 34.9% of cases were of children under one year of age. Amongst the syndromes, RTI had highest overall numbers of cases enrolled, largely from age group one year and below with 42% of cases when compared to AFDUC and GTI.

However, cases diagnosed with AFDUC, were higher in age groups 1 – 4 years with 40.2% (395/982), 5 – 24 years with 14.1% and 25 – 49 years with 18.9% (of cases observed. GTI syndrome had the lowest overall numbers of cases, accounting for the second highest in age group one year and below with 38.9% cases and highest numbers of cases in the 50 years and above group with 8.9%. The outcome of the age group comparison between the three symptoms

showed significance ($p < 0.000$), indicating that differences were observed between the classes in terms of age.

Table 2: Comparison of patient characteristics for all syndromes

Sample Characteristics	**Total (*n*=2948)**	**RTI (*n*=1200)**	**AFDUC (*n*=982)**	**GTI (*n*=766)**	***p*-value**
Male	1519 (51.9%)	591 (49.8%)	532 (54.4%)	396 (51.8%)	0.1067
Age (years)-median(IQR)	1 (0.6-23)	1 (0.3-27)	2 (1-21.8)	1 (0.7-18)	0.1061
Age group (years)					
<1 year	1030 (35%)	504 (42%)	228 (23.2%)	298 (38.9%)	<0.001
1-4 years	925 (31.4%)	293 (24.4%)	395 (40.2%)	237 (30.9%)	
5-24 years	281 (9.5%)	85 (7.1%)	138 (14.1%)	58 (7.6%)	
25-49 years	511 (17.3%)	220 (18.3%)	186 (18.9%)	105 (13.7%)	
50> years	201 (6.8%)	98 (8.2%)	35 (3.6%)	68 (8.9%)	
Clinical signs and symptoms					
PLHIV	578 (21.9%)	256 (23.7%)	172 (19.9%)	150 (21.7%)	0.1339
BMI	16.3 (14.1-19.4)	16.6 (14.4-20.3)	16.3 (14.1-19.3)	15.7 (13.6-18.4)	0.4336
Temperature(°C)-median(IQR)	37.5 (37-38.3)	37.5 (37-38)	38 (37.3-38.6)	37.4 (36.8-38)	<0.001
Self-reported fever	2532 (89.1%)	1023 (90.9%)	935 (97%)	574 (76.2%)	<0.001
Cough	1497 (91%)	1106 (96.8%)	227 (80.2%)	164 (74.2%)	<0.001
Shortness of breath	1045 (62.2%)	934 (79.3%)	42 (14.8%)	69 (31.5%)	<0.001
Pulmonary consolidation	62 (4.5%)	56 (5.8%)	2 (0.7%)	4 (2.5%)	0.0009
Chest pain	354 (21.3%)	308 (26.5%)	17 (6%)	29 (13.3%)	<0.001
Expectoration	464 (27.7%)	409 (34.8%)	20 (7.1%)	35 (16%)	<0.001
Expectoration with blood	63 (3.8%)	58 (5%)	2 (0.7%)	3 (1.4%)	0.0005
Weight loss	1177 (40.3%)	448 (37.5%)	262 (27%)	467 (61.6%)	<0.001
Chills	522 (19.1%)	186 (18.1%)	231 (24%)	105 (14.1%)	<0.001
Fatique	1805 (65.6%)	584 (56.8%)	669 (68.9%)	552 (73.3%)	<0.001
Headache	611 (20.9%)	204 (17.1%)	308 (31.7%)	99 (13%)	<0.001
Muscle pain	126 (4.3%)	47 (4%)	51 (5.2%)	28 (3.7%)	0.2126
Painful swollen glands	30 (1%)	7 (0.6%)	15 (1.5%)	8 (1.1%)	0.0911
Blue lips & skin	39 (2.4%)	33(2.9%)	0 (0)	6 (2.8%)	0.0151
Hospitalised in the last 12 months	323 (11%)	125 (10.6%)	114 (11.7%)	84 (11%)	0.7131
Socio-economic characteristics					
Patient smoke	94 (3.4%)	37 (3.4%)	38 (4.1%)	19 (2.7%)	0.3197
Smoke at home	678 (23.2%)	260 (22%)	235 (24%)	183 (24%)	0.4449
Level of education					
None	62 (2.1%)	27 (2.3%)	16 (1.6%)	19 (2.5%)	<0.001
<= 6 years	214 (7.3%)	102 (8.6%)	40 (4.1%)	72 (9.5%)	
7-10 years	768 (26.3%)	309 (26%)	239 (24.5%)	220 (29%)	
>10 years	360 (12.3%)	154 (13%)	132 (13.5%)	74 (9.7%)	
Employment, Self report					
Unemployed	560 (70.8%)	240 (71.2%)	162 (62.5%)	151 (77.4%)	0.1324
Full time	137 (17.3%)	51 (15.1%)	59 (22.8%)	27 (13.8%)	
Part time	51 (6.4%)	22 (6.5%)	19 (7.3%)	10 (5.1%)	
Self-employed	30 (3.8%)	12 (3.6%)	11 (4.2%)	7 (3.5%)	
Employment, Parent/Guardian					
Unemployed	1450 (69.2%)	600 (73.5%)	477 (66.8%)	373 (66.3%)	0.0185
Full time	471 (22.5%)	162 (19.9%)	174 (24.4%)	135 (24%)	
Part time	101 (4.8%)	35 (4.3%)	35 (4.9%)	31 (5.5%)	
Self-employed	37 (1.8%)	7 (0.9%)	18 (2.5%)	12 (2.1%)	
Geographic Area					
Urban	1790 (60.7%)	666 (55.5%)	688 (70.1%)	436 (56.9%)	<0.001
Village	1158 (39.3%)	534 (44.5%)	294 (29.9%)	330 (43.1%)	
Water Source					
Piped Water	2715 (93.6%)	1108 (93.7%)	916 (94.4%)	691 (92.4%)	0.22
Rain/Spring/River	185 (6.4%)	74 (6.3%)	54 (5.6%)	57 (7.6%)	

The overall median age of all patients was one year (IQR of 0.6-23 years), with most cases in infants, and the comparison between the classes was not significant ($p = 0.1061$) indicating limited difference in age.

The majority of patients enrolled were male accounting for 51.9% of all cases, with a high proportion coming from the urban area (60.7%) and a significant difference observed when compared to the rural area ($p < 0.001$). Most patients (93.6%), had access to piped and improved water sources, and a difference was observed ($p = 0.22$) when compared to cases that collected water from springs and rivers. A high proportion of unemployment was observed in this study, with 70.8% in self-reporting patients and 69.2% in parents and legal guardians. In patients who self-reported employment status, 17.3% were employed full time, while 6.4% had part time employment and 3.8% were self-employed. There were no significant differences in the self-reported employment status in the various syndromes ($p = 0.1324$). Parents and legal guardians reported 22.4% full-time employment, 4.8% part time employment and 1.8% self-employment. A difference in employment status was noted between syndromes with slightly less people employed in the RTI class compared to the AFDUC or GTI class ($p = 0.0185$). The overall majority of patients 18 years and older (12.3%) reported to have more than ten years of education, and of those, 13% were RTI cases, 13.5% AFDUC cases and 9.7% GTI cases. When we look at the rest of education group, 26.3% had 7 to 10 years of education, while 7.3% had education for only 6 years or less and 2.1% reported no education at all.

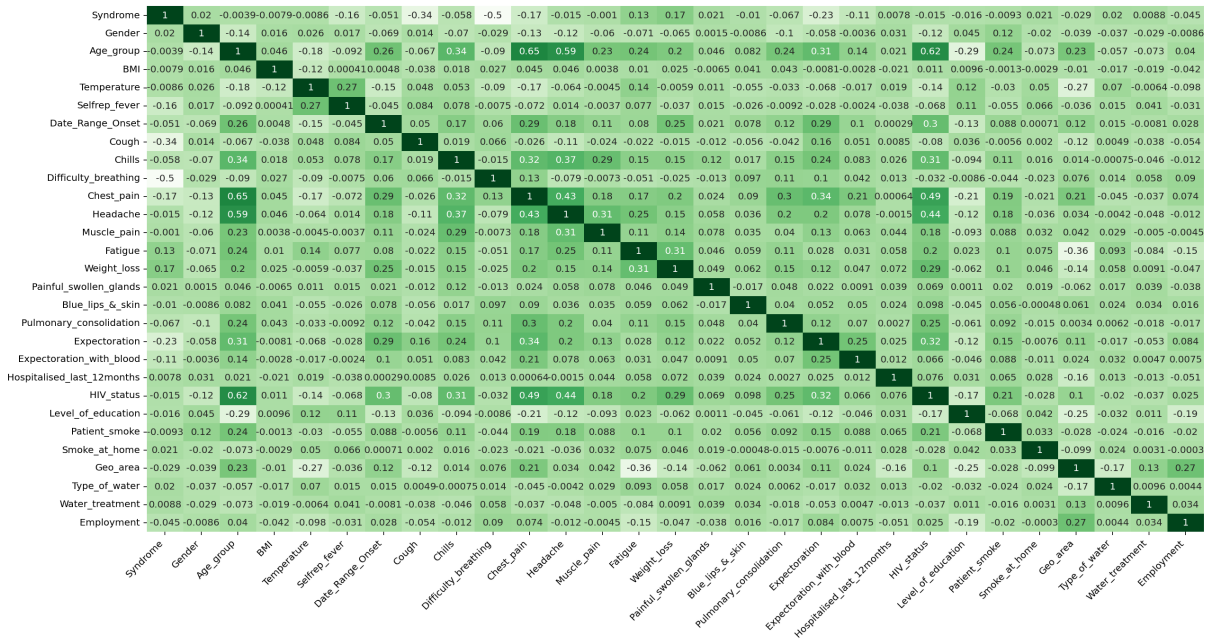


Figure 10: Classification Task 1 Predictors, Correlation Coefficient Matrix

When comparing the clinical signs and symptoms amongst cases between the three syndromes, there was no major differences in symptoms such cough ($p < 0.001$), self-reported fever ($p < 0.001$), shortness of breath ($p < 0.001$), expectoration ($p < 0.001$) and pulmonary consolidation

(p = 0.0028). The clinical signs that showed no difference across the three syndromes' symptoms were muscle pain (p = 0.3254), painful swollen glands (p = 0.1574) and hospitalisation in the last 12 months (p = 0.6298). Even the patient smoke risk factor selected as a feature was not statistically significant (p = 0.5044) when datasets were compared, possibly due to missing values across the datasets.

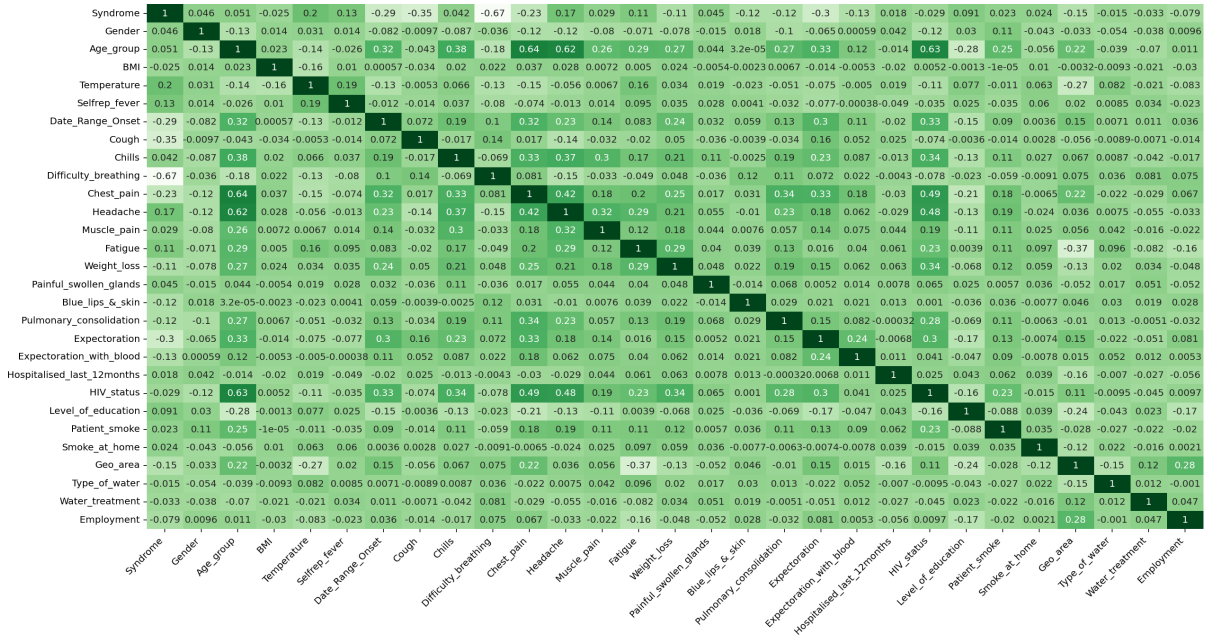


Figure 11: Classification Task 2 Predictors, Correlation Coefficient Matrix

A correlation graphic output outlining the association between two sets of variable data was created as shown in figures 10, 11 and 12 for all the models. We had a very few variables that correlated positively, even fewer variables that correlated negatively. When you look at variables that had positive relationship, their correlation values were the highest, scoring above 0.5 and reaching 0.6 and above for each model. For classification task 1 (Figure 10), age group variable correlated positively with chest pain variable, reaching a value of 0.65, correlated with headache at value of 0.59, and HIV status variable reaching a value of 0.62.

Difficulty breathing was the only variable that had a negative relationship with syndrome variable at value -0.50 for classification task 1. However, difficulty breathing variable that had a negative correlation to syndrome variable, when one increase, the other will decrease.

The variable interaction pattern observed in classification task 1, was also observed in classification task 2 and 3 variables correlations outcomes. For classification task 2 (Figure 11), age

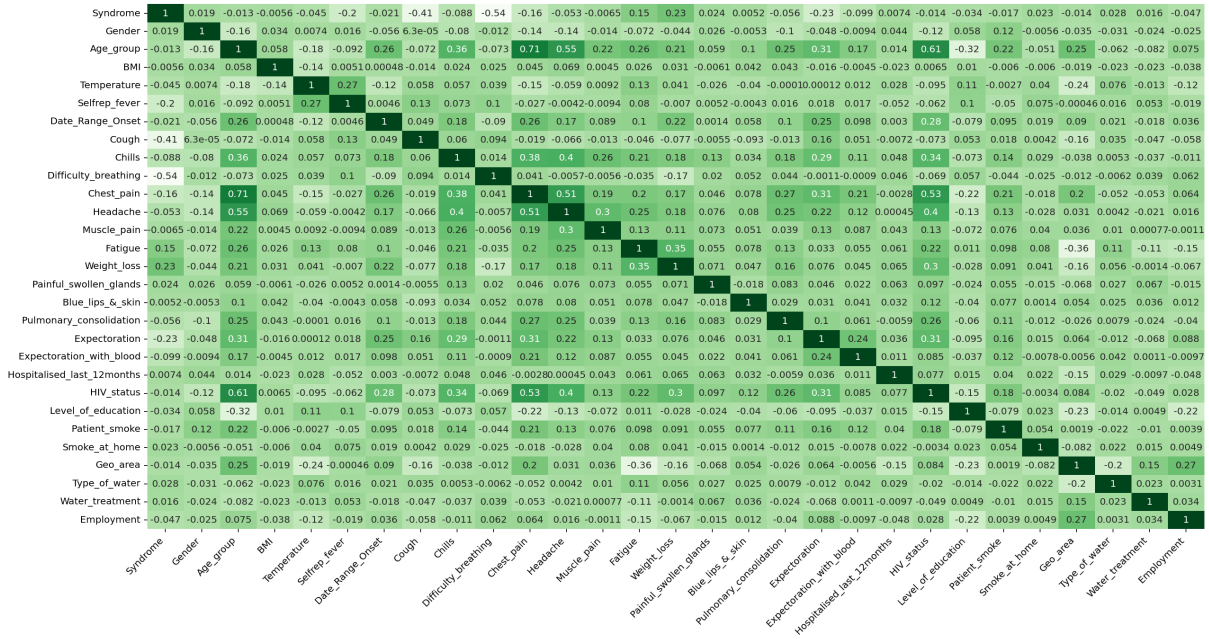


Figure 12: Classification Task 3 Predictors, Correlation Coefficient Matrix

group variables had a positive coefficient with moderate correlation to chest pain, reaching a value of 0.64, while headache variable had positive coefficient of 0.62, and HIV status reaching a positive coefficient of 0.63. The variable, difficulty breathing had a negative coefficient of -0.67, moderately correlating with syndrome variable.

For classification task 3 (Figure 12), age group correlated from moderate to strong with most variables, reaching positive coefficient of 0.55 to 0.71 range. The age group variable had a strong correlation, with positively coefficient of 0.71 with chest pain, positively coefficient of 0.55 with headache, and lastly, positively coefficient of 0.61 with HIV status. HIV status correlated moderately with chest pain variable, which had a positive coefficient of 0.53, while difficulty breathing correlated moderately with RTI, GTI and AFDUC syndrome variables, with a negative coefficient value of -0.54.

4.3 Model Development Outcome and Evaluation

Table 3 present the performance metrics (accuracy, precision, specificity and F1-score) for the top three chosen classifiers and all of the metrics are presented for each classification task. In Table 3, we observed that the classifier, random forest, scored higher on classification task 2 and out-performed other classifiers on all three classification tasks. However, classification task 2 presented the highest performance score when compared to other classification tasks across the classifiers, and even within the same classifier.

Table 3: Classification Tasks Performance Comparison Measured by 4 Importance Scores

Classifiers and Tasks		Accuracy	Precision	Specificity	F1-Score
Random Forest	Task 1	0.86	0.86	0.93	0.86
	Task 2	0.95	0.95	0.94	0.95
	Task 3	0.92	0.92	0.89	0.92
Neural Network MLP	Task 1	0.83	0.83	0.92	0.83
	Task 2	0.92	0.92	0.91	0.92
	Task 3	0.90	0.90	0.88	0.90
Support Vector Machine	Task 1	0.81	0.82	0.92	0.81
	Task 2	0.91	0.92	0.92	0.91
	Task 3	0.89	0.90	0.89	0.89

Each classifier and classification task performance metrics were also plotted as illustrated by figure 13. The graph helped visualise the classifier and model performance, specifically demonstrating the best performing overall classifier as random forest, which was followed by support vector machine. Both classifiers performed very well on classification task 2 and 3 datasets, while all classifiers performed poorly on the classification task 1 dataset, specifically support vector machine classifier.

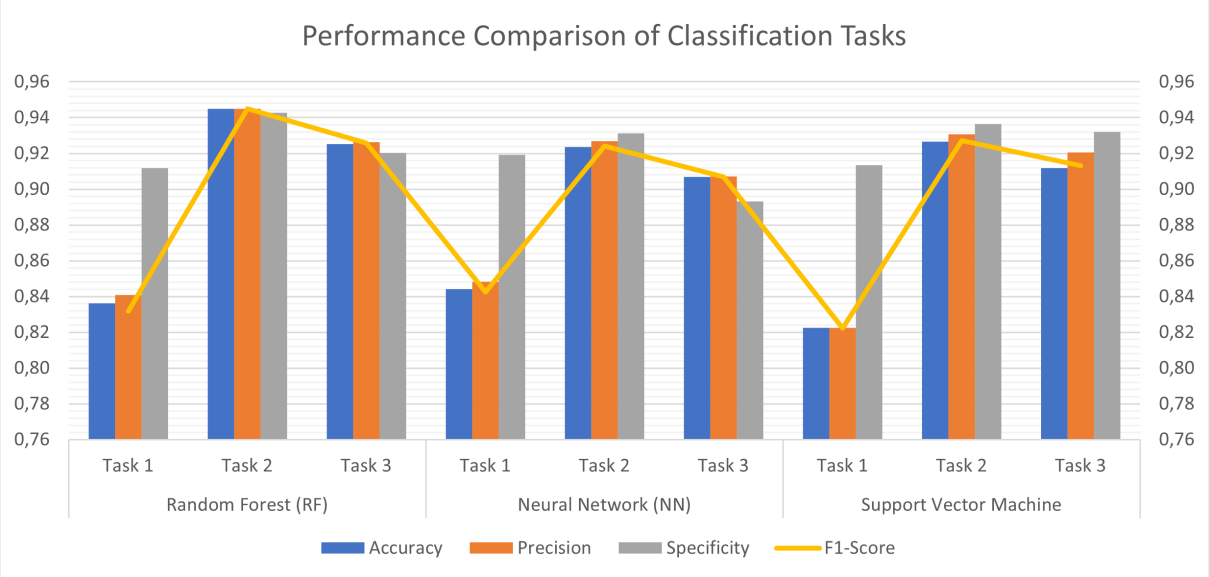
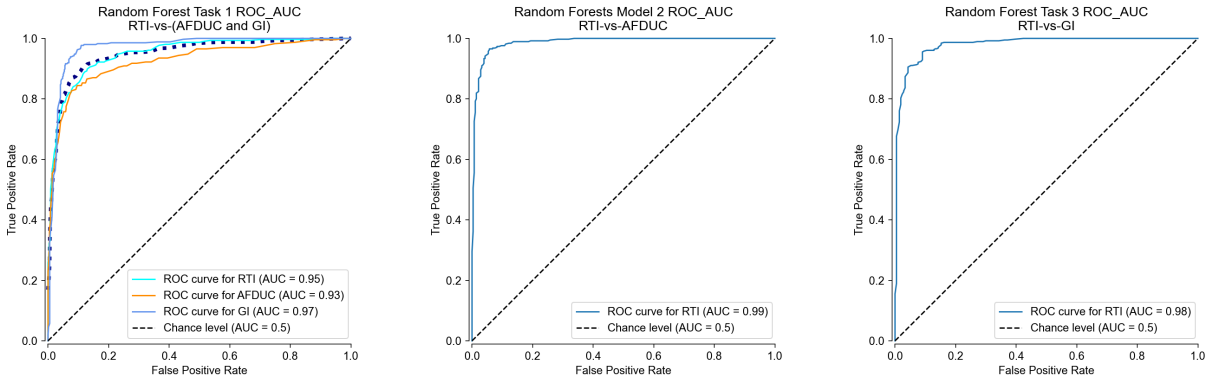


Figure 13: Performance Comparison of Classification Tasks

In this study, Area Under the ROC curve (AUROC) analysis was the methods performed to demonstrate the relationship between true and false positive rates of the predicted and actual values, as illustrated in figures 14, 15 and 16. Classification task 2 had the highest accuracy, precision, specificity and F1 scores in all classifiers, random forest scoring highest, followed by support vector machine and neural network scoring last. Classification task 3 performed well, slightly inferior to classification task 2 performance evaluation metric scoring on all classifiers.

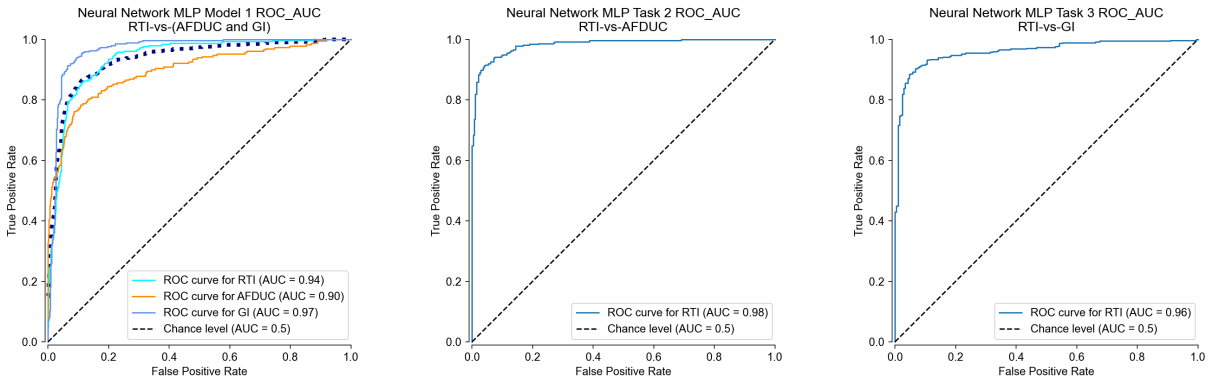
Only classification task 2 and 3 achieved scoring above 0.91 for all performance evaluation matrices. Each of the designed models performed well under each classifier, by achieving AUC above 0.8. Although, the performance was the lowest when compared to classification task 2 and 3, classification task 1 was able to achieve AUC above 0.90, for all classifiers, even though 3 classes were compared and measured.



(a) Task 1 AUROC Curve (b) Task 2 AUROC Curve (c) Task 3 AUROC Curve

Figure 14: Random Forest Area under ROC Curve for all Classification Tasks

Using all three classes as illustrated in classification task 1 for each classifier, we have observed that AUC is reduced (Figure 14a). However, classification task 2 (Figure 14b) and 3 (Figure 14c) only have two classes, which generated higher AUC values (0.99) for random forest classifier.

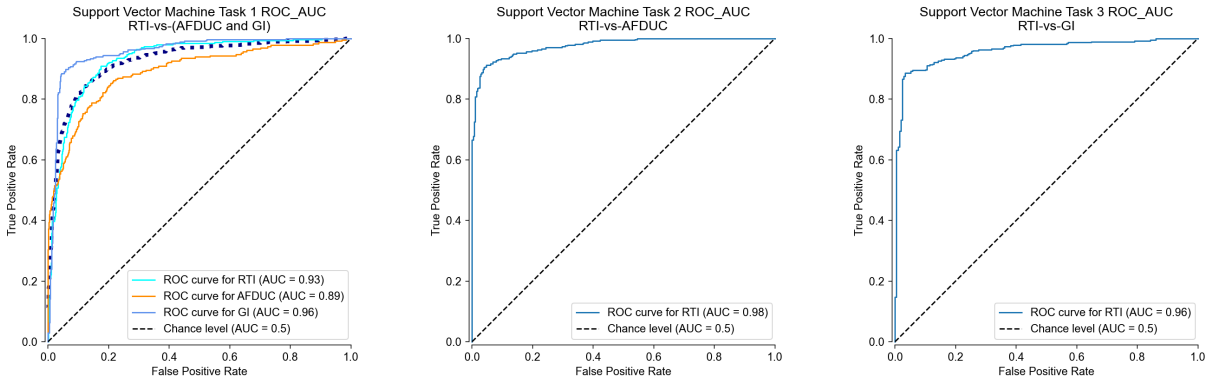


(a) Task 1 AUROC Curve (b) Task 2 AUROC Curve (c) Task 3 AUROC Curve

Figure 15: Neural Network MLP Area under ROC Curve for all Classification Tasks

Comparison of classifiers based on the ROC and AUC value showed that random forest had the best outcome as the overall classifier, and that the classification task 2 dataset achieved the highest AUC and best shaped ROC. Support vector machine classifier had the poorest ROCs in

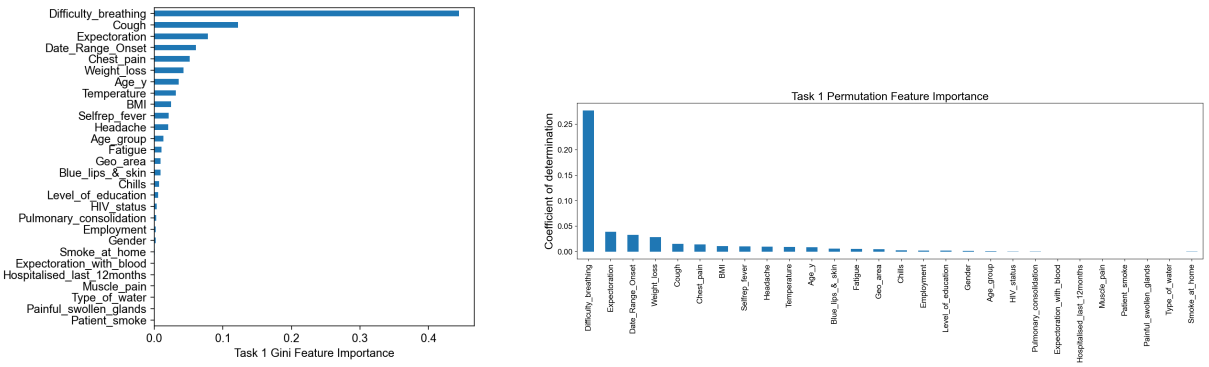
all the classification tasks, especially in classification task 3. Overall, classification task 3 ROC shape was slightly unusual and less convex at the top, when compared to classification task 1 and 2, for neural network MLP and support vector machine classifiers.



(a) Task 1 AUROC Curve (b) Task 2 AUROC Curve (c) Task 3 AUROC Curve

Figure 16: Support Vector Machine Area under ROC Curve for all Classification Tasks

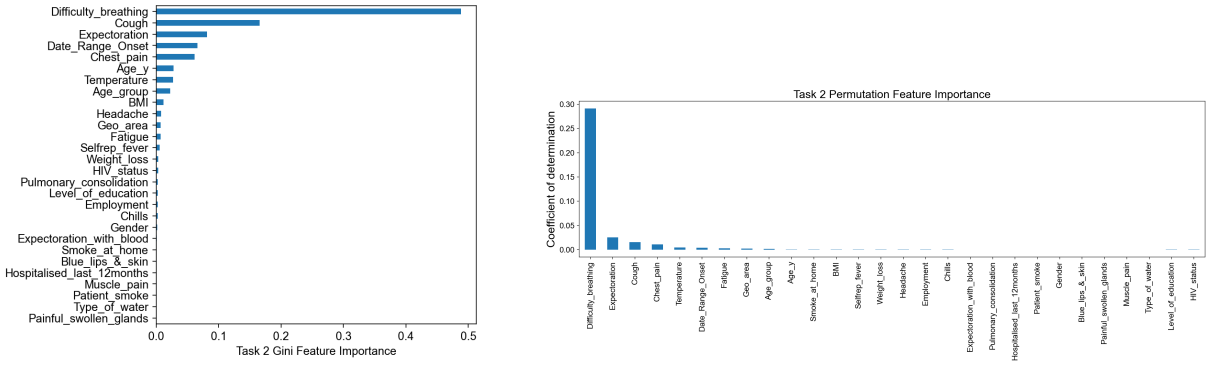
MDI, Gini impurity and permutation importance scores were measured across a set of features attributed to each model. The MDI computation was derived from Random Forest Classifier model, as it was the best model from the three that were measured. The average mean value across all models datasets were calculated and result, which are presented in Figures 17, 18 and 19. The feature with the high Gini impurity score, was placed higher in the important features' scale and the feature with low Gini impurity score, was placed lower in the importance feature scale for each model. Both permutation and MDI importance showed a similar ability in distinguishing highly important features.



(a) Task 1 Gini Feature Importance (b) Task 1 Permutation Feature Importance

Figure 17: Gini and Permutation Feature Importance for Classification Task 1

It appears that the variables difficulty breathing, cough and expectoration, were the most important features in all three models by Gini score feature importance. Furthermore, similar

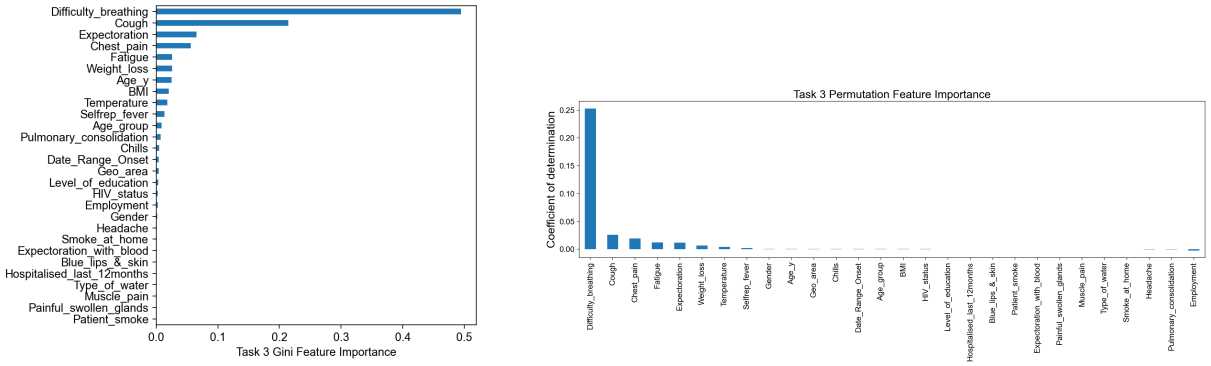


(a) Task 2 Gini Feature Importance

(b) Task 2 Permutation Feature Importance

Figure 18: Gini and Permutation Feature Importance for Classification Task 2

variables also scored higher by permutation feature importance. At most, only a few features were regarded as important by both techniques. However, after the top three features, the remaining variable's importance varied in scoring, while others scored higher by Gini impurity, and scored lower by permutation importance.



(a) Task 3 Gini Feature Importance

(b) Task 3 Permutation Feature Importance

Figure 19: Gini and Permutation Feature Importance for Classification Task 3

However, two features in particular demonstrated altered importance. Feature date range onset scored within the top ten features in classification task 1 and 2, then became the worst performing feature in classification task 3 scoring outside top 10 for both Gini impurity and permutation importance scores. Variable weight loss, performed well in classification task 1 and 3, scoring within the top 10, and performed worse in classification task 2 scoring lower outside top 10.

CHAPTER 5

5 DISCUSSION AND CONCLUSION

5.1 Discussion

5.1.1 Main findings

In this study, the predictive models for the diagnosis of RTI were developed by applying various machine learning methodologies. The applicability of the model was tested, analyzed and evaluated to fulfil the study aim and objectives. We initially used cross-validation data splitting technique to fine tune the design models during development phase, and identified that RTI classification outcome was not different from classified RTI outcome from data split with standard procedure. However, the data processing requirement for system transactions was very high with the cross-validation splitting, with the duration of the process taking longer than the standard splitting procedure.

Correlation of independent features was measured, to determine the multicollinearity that was likely to emerge between features. The machine learning prediction model ranked the age group variable high amongst important features, indicating that age is an important risk factor for pneumonia (Sun et al., 2022). Furthermore, the age group feature, correlated with other features from moderate to high degree, specifically with HIV status, headache and chest pain features, in all three models. The rest of the features had low degree of correlation or close to no correlation, with mostly negative values close to zero. Feature importance for all three models scored age in the top ten important features for RTI classification. Moreover, a study reported that age was a risk factor for a recurring lower RTI (Kara et al., 2024), and another study associated age as potential risk factor for acute RTI (Nshimiyimana and Zhou, 2022). Most features that were measured and ranked within the top ten, were features that had low correlation amongst each other, and the age group feature.

In all models, classification of RTI was predicted with high accuracy, precision and specificity. The score above 80% was mostly desirable for each model performance metric, which included the AUC and ROC curve scoring of above 0.8. Although model 1 had the lowest prediction accuracy, precision and specificity scores in all three classifiers, the likelihood reason was the

prediction of RTI class amongst two more classes. We achieved the best prediction results using random forest models, able to classify RTI within one and two classes, with high performance scoring of above 80% with all scoring metrics. The classified RTI ROC curve shape was robust with the expected shape, which was able to measure the AUC above 0.92 for three classes and AUC above 0.95 for two classes.

5.1.2 Clinical implications of prediction model

The prediction model outcome of this study suggests that the minimum number of features used to design and build a machine learning classification model, achieved satisfactory levels of prediction performance. Most critical signs and symptoms presented and used to diagnose RTI during medical examination, were amongst the most important feature that built all three prediction models. Each model was able to predict RTI outcome in a test data, amongst AFDUC and GTI syndromes, from previously diagnosed cases by medical professionals. Similar findings were reported for other machine learning prediction models, outperforming and matching medical doctor's clinical judgment when assessing clinical risk mortality score in patients with sepsis (van Doorn et al., 2021).

Looking at each model, even though the performed was different, there was insight that was observed. The classification tasks 2 dataset, which was a subset of RTI and AFDUC datasets, share similarity in that both contain data from respiratory infection patients. Classification task 2 outperformed other tasks in all the models (algorithms). The when you compare the models , with classification task 2, random forest performed better than both SVM and ANN, due classification or class boundaries which makes SVM and ANN harder to classify.

The model can be used as the primary screening for RTIs, to differentiate self-limiting illness requiring a few days of recovery, to acute illness that requires a longer recovery period. In this study, we observed that, the date_range_onset variable was a top 10 important features that build a robust model to classified RTI, with cases experiencing acute infection in the 10 days prior to hospital admission. The longer the duration of clinical signs and symptoms, the longer the respiratory infection caused by viral infection, particularly in children (Paalanne et al., 2023). Machine learning has been widely used in healthcare facilities, apart from being used to predict health determinants of patients, it is also used in the collection of patient history and improve patient experience in healthcare services (Javaid et al., 2022).

5.1.3 Strength and limitation of this study

Since the study had up to 50% of children below the age of one year, we had a limitation of not using the childhood vaccination records to determine the association of not receiving vaccine and pneumococcal and respiratory tract infection. The main limitation of the study was the sample size, which was quite small for machine learning computational power. The advantage was that, we did not have to worry about overfitting issues, which are inherent in large sample size. Risk of bias is one of the metrics we did not measure in this study, especially since we have used a small sample size and moderate number of missing data values.

5.2 Conclusion

In this study, we differentiated the RTI from other infections that may have similar signs and symptoms, such as cough, fever, and temperature presented by patient during hospital admission. Although socioeconomic features were used to design the model, the permutation and MDI scoring was low, which made them less important when model was built. Compared to the standard and conventional statistical applications, the approach of applying supervised machine learning techniques showed significant performance in prediction of RTI outcome. The results of this work demonstrate the superiority of random forest models' applicability in machine learning classification of RTI. In conclusion, the predictive models presented in this study presents an accurate estimated of RTI classification from AFDUC and GTI, which provides important evidence that RTIs can be effectively screened without time constraints, manpower and cost implications.

Although the performance of the RTI model exceeded our expectations, the median age of the cases included in the model datasets was one year, which indicated the majority of cases were infants and children. The management of illness is different for adults and children, especially infants in the first two months of life and children two months and up to five years. The general model we designed, only looked at the overall symptoms and vital signs, which may not be unique to infants and children, and mostly common in adults. Some of the symptoms not included which are prevalent in infants include chest indrawing, nasal flaring and granting which are associated with respiratory infections in the very young. A different approach is needed that will look at modelling RTIs classification in different age groups, as shown in our findings which suggest age as a risk factor. This will allow the use of specific symptoms, likely to be diagnosed in infants and children as model predictors, for age groups in children.

The other issue that affects the classification of RTIs is the misdiagnosis of pleuropulmonary and nasal and throat infections, especially the community acquired pneumoniae and other respiratory infections (Eshaghi et al., 2025). Misdiagnosis leads to delays in the identification of aetiology, which extends the period of infection and may drive disease progress from acute to severe infection. If we can minimize the time to diagnosis, RTIs can be more readily classified and the specific treatment administered. Our observations suggest that infection date onset is critical in the identification of RTI, and possibly the severity of the infection.

We observed limitations in this study, the minor one being the sample size, which affects prediction models with overfitting, resulting in less desirable classification outcomes. One good feature of machine learning is that, it does not suffer from bias through statistical power determination which other modelling are dependent on. Rather, having a large sample size, which is critical in machine learning prediction models, is the only option that would guarantee model accuracy of above 80% (Rajput et al., 2023). Sample size should be at a certain threshold to make meaningful prediction observations, and when that threshold is reached, there may not be an additional benefit to the classification outcome. Another minor limitation was the exclusion of some socio-economic variables, in particular the breastfeeding variable, in which the lack of breastfeeding reported a risk factor in one study (Kassaw et al., 2024).

A major limitation we have observed in this study was the missing data of vital clinical signs and symptoms. The assumption is that, data was missing at random, which is common in every healthcare facility, suggesting inadequate handling of patient medical history and socio-economic information. Although, the missingness of data was addressed with techniques that imputed values were the data was missing, significant bias was avoided. Although imputation of data works, the issue is how significant and unique is the missing value to each patient. As reported elsewhere, imputation purpose in classification models is to reduce bias of covariates on prediction outcome over reduction of bias of other covariates (Blythe et al., 2023). Imputation should be the last resort to missing values, care when collecting and transcribing data should be done with care to limit missingness.

REFERENCES

Abadom, T. R., Smith, A. D., Tempia, S., and et al. (2016). Risk factors associated with hospitalisation for influenza-associated severe acute respiratory illness in south africa: A case-population study. *Vaccine*, 34:5649–5655.

Azoulay, E., Russell, L., Van de Louw, A., and et al. (2020). Diagnosis of severe respiratory infections in immunocompromised patients.

Bender, R. G., Sirota, S. B., Swetschinski, L. R., and et al. (2024). Global, regional, and national incidence and mortality burden of non-covid-19 lower respiratory infections and aetiologies, 1990–2021: a systematic analysis from the global burden of disease study 2021. *The Lancet Infectious Diseases*.

Biswas, A., Saran, I., and Wilson, F. P. (2021). Introduction to supervised machine learning. *Kidney360*, 2:878–880.

Blythe, R., Parsons, R., Barnett, A. G., and et al. (2023). Vital signs-based deterioration prediction model assumptions can lead to losses in prediction performance. *Journal of Clinical Epidemiology*, 159:106–115.

Bohr, A. and Memarzadeh, K. (2020). *The rise of artificial intelligence in healthcare applications*.

Bre, F., Gimenez, J. M., and Fachinotti, V. D. (2018). Prediction of wind pressure coefficients on building surfaces using artificial neural networks. *Energy and Buildings*, 158:1429–1441.

Chicco, D. and 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.

Chicco, D. and Rovelli, C. (2019). Computational prediction of diagnosis and feature selection on mesothelioma patient health records. *PLoS ONE*, 14.

Cho, P. S., Birring, S. S., Fletcher, H. V., and et al. (2019). Methods of cough assessment.

Cohen, C., Moyes, J., Tempia, S., and et al. (2015a). Mortality amongst patients with influenza-associated severe acute respiratory illness, south africa, 2009-2013. *PLoS ONE*, 10:2009–2013.

Cohen, C., Walaza, S., Moyes, J., and et al. (2015b). Epidemiology of severe acute respiratory illness (sari) among adults and children aged 5 years in a high hiv-prevalence setting, 2009-2012. *PLoS ONE*, 10:2009–2012.

Cohen, C., Walaza, S., Moyes, J., and et al. (2015c). Epidemiology of viral-associated acute lower respiratory tract infection among children <5 years of age in a high hiv prevalence setting, south africa, 2009-2012. *Pediatric Infectious Disease Journal*, 34:66–72.

Cowley, L. E., Farewell, D. M., Maguire, S., and et al. (2019). Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. *Diagnostic and Prognostic Research*, 3.

Croft, P., Altman, D. G., Deeks, J. J., and et al. (2015). The science of clinical practice: Disease diagnosis or patient prognosis? evidence about "what is likely to happen" should shape clinical practice. *BMC Medicine*, 13:1–8.

Cruz, J. A. and Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, 2:59–77.

De São José, B. P., Camargos, P. A. M., Da Cruz Filho, A. S., and et al. (2014). Diagnostic accuracy of respiratory diseases in primary health units. *Revista da Associacao Medica Brasileira*, 60:599–612.

Eshaghi, S., Sheybani, F., Hedjazi, A., and et al. (2025). Infectious causes of death: An autopsy-based study of 546 cases. *Open Forum Infectious Diseases*, 12.

García, S., Ramírez-Gallego, S., Luengo, J., and et al. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*, 1.

Gentilotti, E., Nardo, P. D., Cremonini, E., and et al. (2022). Diagnostic accuracy of point-of-care tests in acute community-acquired lower respiratory tract infections. a systematic review and meta-analysis.

Giordano, C., Brennan, M., Mohamed, B., and et al. (2021). Accessing artificial intelligence for clinical decision-making.

Gutierrez, R., Landa, M., Sambou, M., and et al. (2024). Aetiology of non-malaria acute febrile illness fever in children in rural guinea-bissau: a prospective cross-sectional investigation. *Frontiers in Epidemiology*, 4.

Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J. P., and et al. (2020). Minimar (minimum information for medical ai reporting): Developing reporting standards for artificial intelligence in health care.

Ippolito, A., Sorrentino, M., Capalbo, F., and et al. (2023). How technological innovations inerformance measurement systems overcome management challenges in healthcare. *International Journal of Productivity and Performance Management*, 72:2584–2604.

Jaume, F., Valls-Mateus, M., and Mullol, J. (2020). Common cold and acute rhinosinusitis: Up-to-date management in 2020.

Javaid, M., Haleem, A., Singh, R. P., and et al. (2022). Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, 3:58–73.

Jebb, A. T., Parrigon, S., and Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, 27:265–276.

Kara, S., Peker, N. A., Arslan, Y. K., and et al. (2024). Risk factors for lower respiratory tract infection and recurrent lower respiratory tract infection in children. *Van Medical Journal Van Med J*, 31:53–59.

Kashina, M., Lenivtceva, I. D., and Kopanitsa, G. D. (2020). Preprocessing of unstructured medical data: The impact of each preprocessing stage on classification. In *Procedia Computer Science*, volume 178, pages 284–290. Elsevier B.V.

Kassaw, A. K., Bekele, G., Kassaw, A. K., and et al. (2024). Prediction of acute respiratory infections using machine learning techniques in amhara region, ethiopia. *Scientific Reports*, 14.

Kaul, V., Enslin, S., and Gross, S. A. (2020). History of artificial intelligence in medicine. *Gastrointestinal Endoscopy*, 92:807–812.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., and et al. (2019). Key challenges for delivering clinical impact with artificial intelligence.

Kersting, K. (2018). Machine learning and artificial intelligence: Two fellow travelers on the quest for intelligent behavior in machines. *Frontiers in Big Data*, 1:1–4.

Kourou, K., Exarchos, T. P., Exarchos, K. P., and et al. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17.

Kwan, J. L., Lo, L., Ferguson, J., and et al. (2020). Computerised clinical decision support systems and absolute improvements in care: Meta-analysis of controlled clinical trials. *The BMJ*, 370.

Le Roux, D. M., Nicol, M. P., Myer, L., and et al. (2019). Lower respiratory tract infections in children in a well-vaccinated south african birth cohort: Spectrum of disease and risk factors. *Clinical Infectious Diseases*, 69:1588–1596.

Mai, M. V. and Krauthammer, M. (2017). Controlling testing volume for respiratory viruses using machine learning and text mining. *AMIA Annual Symposium Proceedings*, 2016:1910–1919.

McCracken, G. H. (2001). Clinical practice guidelines for the diagnosis and treatment of respiratory tract infections. *SUP. THE AMERICAN JOURNAL OF MANAGED CARE*, 7.

Memon, S. M., Wamala, R., and Kabano, I. H. (2023). A comparison of imputation methods for categorical data. *Informatics in Medicine Unlocked*, 42:101382.

Minodier, L., Masse, S., Capai, L., and et al. (2017). Clinical and virological factors associated with gastrointestinal symptoms in patients with acute respiratory infection: A two-year prospective study in general practice medicine. *BMC Infectious Diseases*, 17.

Morice, A. H., Millqvist, E., Bieksiene, K., and et al. (2020). Ers guidelines on the diagnosis and treatment of chronic cough in adults and children. *European Respiratory Journal*, 55.

Naydenova, E., Tsanas, A., Howie, S., and et al. (2016). The power of data mining in diagnosis of childhood pneumonia. *Journal of the Royal Society Interface*, 13.

Newman, T. B. and Kohn, M. A. (2020). *Evidence-based diagnosis: an introduction to clinical epidemiology*. Cambridge University Press.

Niederman, M. S. and Torres, A. (2022). Severe community-acquired pneumonia. *European Respiratory Review*, 31.

Nshimiyimana, Y. and Zhou, Y. (2022). Analysis of risk factors associated with acute respiratory infections among under-five children in uganda. *BMC Public Health*, 22.

Olayemi, C. O., Sunday, O. A., Olasehinde, O. O., and et al. (2019). Application of machine learning to the diagnosis of lower respiratory tract infections in paediatric patients. *i-manager's Journal on Pattern Recognition*, 5:2018.

Oonsivilai, M., Mo, Y., Luangasanatip, N., and et al. (2018). Using machine learning to guide targeted and locally-tailored empiric antibiotic prescribing in a children's hospital in cambodia. *Wellcome Open Research*, 3.

Ostadi, B., Daloie, R. M., and Sepehri, M. M. (2023). A new scenario-based simulation model for cost management of healthcare services through improving the efficiency of the health centers. *Journal of Healthcare Engineering*, 2023.

Ostojic, D., Lalousis, P. A., Donohoe, G., and et al. (2024). The challenges of using machine learning models in psychiatric research and clinical practice.

Ou, X., Huang, J., Chang, Q., and et al. (2020). First time quality diagnostics and improvement through data analysis: A study of a crankshaft line. *Procedia Manufacturing*, 49:2–8.

Paalanne, N., Honkila, M., Mattila, S., and et al. (2023). Duration of clinical symptoms in children with acute respiratory infection. *Acta Paediatrica, International Journal of Paediatrics*, 112:813–819.

Phillips, B. (2010). Clinical decision rules: how to build them. *Archives of Disease in Childhood. Education and Practice Edition*, 95:83.

Rajput, D., Wang, W. J., and Chen, C. C. (2023). Evaluation of a decided sample size in machine learning applications. *BMC Bioinformatics*, 24.

Rogan, M. (2016). *Respiratory Infections, Acute*, pages 332–336. Elsevier Inc.

Saleri, N. and Edward, T. R. (2019). Respiratory infections. *Travel Medicine*, pages 527–527.

Sánchez, X., Orrico, M., Morillo, T., and et al. (2021). Reducing unnecessary antibiotic prescription through implementation of a clinical guideline on self-limiting respiratory tract infections. *PLoS ONE*, 16.

Schubert, G., Achi, V., Ahuka, S., and et al. (2021). The african network for improved diagnostics, epidemiology and management of common infectious agents. *BMC Infectious Diseases*, 21.

Seidu, A. A., Dickson, K. S., Ahinkorah, B. O., and et al. (2019). Prevalence and determinants of acute lower respiratory infections among children under-five years in sub–saharan africa: Evidence from demographic and health surveys. *SSM - Population Health*, 8:100443.

Sirota, S. B., Doxey, M. C., Dominguez, R.-M. V., and et al. (2025). Global, regional, and national burden of upper respiratory infections and otitis media, 1990–2021: a systematic analysis from the global burden of disease study 2021. *The Lancet Infectious Diseases*, 25:36–51.

Sparks, R., Carter, C., Donnelly, J. B., and et al. (2008). Remote access methods for exploratory data analysis and statistical modelling: Privacy-preserving analytics *Computer Methods and Programs in Biomedicine*, 91:208–222.

Stam, W. T., Ingwersen, E. W., Ali, M., and et al. (2023). Machine learning models in clinical practice for the prediction of postoperative complications after major abdominal surgery. *Surgery Today*, 53:1209–1215.

StatsSA (2024). Mortality and causes of death in south africa: Findings from death notification 2020.

Steurer, J., Held, U., Spaar, A., and et al. (2011). A decision aid to rule out pneumonia and reduce unnecessary prescriptions of antibiotics in primary care patients with cough and fever. *BMC Medicine*, 9.

Storms, A. D., Chen, J., Jackson, L. A., and et al. (2017). Rates and risk factors associated with hospitalization for pneumonia with icu admission among adults. *BMC Pulmonary Medicine*, 17:1–9.

Suarez, N. M., Bunsow, E., Falsey, A. R., and et al. (2015). Superiority of transcriptional profiling over procalcitonin for distinguishing bacterial from viral lower respiratory tract infections in hospitalized adults. *Journal of Infectious Diseases*, 212:213–222.

Sumitomo, T. and Kawabata, S. (2024). Respiratory tract barrier dysfunction in viral-bacterial co-infection cases.

Sun, X., Douiri, A., and Gulliford, M. (2022). Applying machine learning algorithms to electronic health records to predict pneumonia after respiratory tract infection. *Journal of Clinical Epidemiology*, 145:154–163.

Sutton, R. T., Pincock, D., Baumgart, D. C., and et al. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success.

Tao, R. J., Luo, X. L., Xu, W., and et al. (2018). Viral infection in community acquired pneumonia patients with fever: A prospective observational study. *Journal of Thoracic Disease*, 10:4387–4395.

To, K. K., Chan, K. H., Ho, J., and et al. (2019). Respiratory virus infection among hospitalized adult patients with or without clinically apparent respiratory infection: a prospective cohort study. *Clinical Microbiology and Infection*, 25:1539–1545.

Toll, D. B., Janssen, K. J., Vergouwe, Y., and et al. (2008). Validation, updating and impact of clinical prediction rules: A review.

Troeger, C., Blacker, B., Khalil, I. A., and et al. (2018). Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet Infectious Diseases*, 18:1191–1210.

Tuti, T., Agweyu, A., Mwaniki, P., and et al. (2017). An exploration of mortality risk factors in non-severe pneumonia in children using clinical data from kenya. *BMC Medicine*, 15:1–12.

Uddin, S., Khan, A., Hossain, M. E., and et al. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, 19:1–16.

van Doorn, W. P., Stassen, P. M., Borggreve, H. F., and et al. (2021). A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis. *PLoS ONE*, 16.

Van Ierland, Y., Elshout, G., Berger, M. Y., and et al. (2015). Translation of clinical prediction rules for febrile children to primary care practice: An observational cohort study. *British Journal of General Practice*, 65:e224–e233.

Venter, M., Lassauniere, R., Kresfelder, T. L., and et al. (2011). Contribution of common and recently described respiratory viruses to annual hospitalization in children in south africa. *Journal of Medical Virology*, 83:1458–1468.

Wright, W. F. and Auwaerter, P. G. (2020). Fever and fever of unknown origin: Review, recent advances, and lingering dogma. *Open Forum Infectious Diseases*, 7.

Wu, Q., Pan, X., Han, D., Ma, Z., and Zhang, H. (2024). New insights into the epidemiological characteristics of mycoplasma pneumoniae infection before and after the covid-19 pandemic.

Yang, H. S., Rhoads, D. D., Sepulveda, J., and et al. (2023). Challenges and considerations of developing and implementing machine learning tools for clinical laboratory medicine practice.

Zhang, J. and Nawata, K. (2017). A comparative study on predicting influenza outbreaks. *BioScience Trends*, 11:533–541.

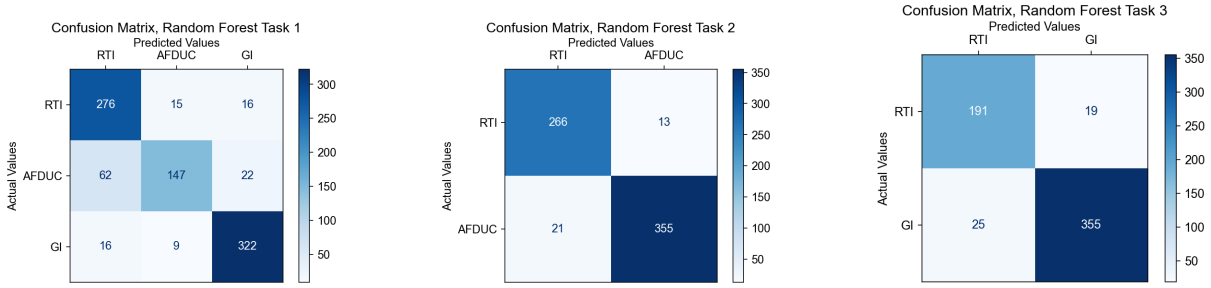
Zhang, S., Sammon, P. M., King, I., and et al. (2016). Cost of management of severe pneumonia in young children: Systematic analysis. *Journal of Global Health*, 6.

Zhang, Z., Beck, M. W., Winkler, D. A., and et al. (2018). Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Annals of Translational Medicine*, 6:216–216.

Zheng, A. and Casari, A. (2018). *Feature engineering for machine learning : principles and techniques for data scientists*. O'Reilly Media.

A APPENDICES

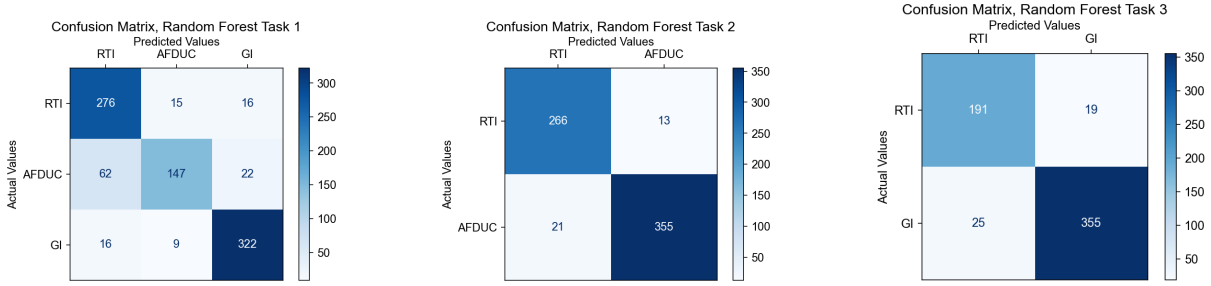
Appendix 1: Random Forest Confusion Matrix



(a) Model 1 Confusion Matrix (b) Model 2 Confusion Matrix (c) Model 3 Confusion Matrix

Figure 20: Random Forest Confusion Matrix for all Classification Tasks

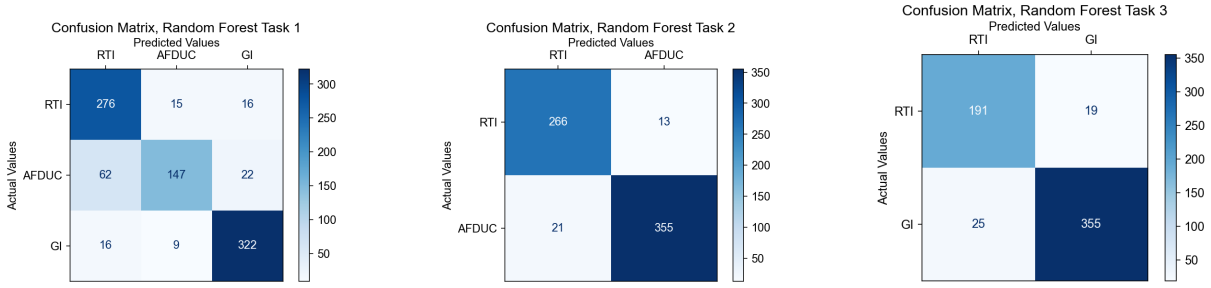
Appendix 2: Neural Network MPL Confusion Matrix



(a) Model 1 Confusion Matrix (b) Model 2 Confusion Matrix (c) Model 3 Confusion Matrix

Figure 21: Random Forest Confusion Matrix for all Classification Tasks

Appendix 3: Support Vector Machine Confusion Matrix



(a) Model 1 Confusion Matrix (b) Model 2 Confusion Matrix (c) Model 3 Confusion Matrix

Figure 22: Random Forest Confusion Matrix for all Classification Tasks

Appendix 4: Ethics Clearance Certificate

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Mr Metsekae Richard Madimabe

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M240709

NAME: (Principal Investigator)	Mr Metsekae Richard Madimabe
DEGREE:	MSc in Epidemiology - Public Health Informatics
DEPARTMENT:	Biostatistics and Epidemiology
STUDY SITE(S):	N/A
TITLE:	Using Supervised Machine Learning Techniques to Classify Respiratory Tract Infections
DATE CONSIDERED:	26/07/2024
DECISION:	Approved unconditionally
CONDITIONS:	Secondary data analysis of previously approved study
NOTE:	Original Application Number MED24-06-009. Subsequent Online Application Number M241003-A-0001.
SUPERVISOR:	Prof N. Page and Mr S. Chikasha
APPROVED BY:	Paul Ruff Paul Ruff (Oct 3, 2024 15:08 GMT+2) Professor P. Ruff, Chairperson, HREC (Medical)

DATE OF APPROVAL: 03/10/2024 **EXPIRY DATE:** 02/10/2029

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATOR(S)

I/we fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report** in the month date of convened meeting each year until study is closed. Failure to submit annual report will result in ethics clearance certificate suspension. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. Unreported changes to the application may invalidate the clearance given by the HREC (Medical). Email a signed copy of ethics clearance certificate to HREC-Medical.ResearchOffice@wits.ac.za.

Principal Investigator Signature Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Appendix 5: Plagiarism Declaration Report

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Metsekae Richard Madimabe (Student number: 1253595) am a student registered for the degree of Master of Science in Epidemiology in the academic year 2025.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature:  Date: 27 February 2025

Appendix 6: Similarity Index Report

MSc Research Report_Final_st#1253595-3.pdf

ORIGINALITY REPORT

12%	9%	9%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	orca.cardiff.ac.uk Internet Source	<1%
2	www.mdpi.com Internet Source	<1%
3	academic.oup.com Internet Source	<1%
4	Sujata Dash, Subhendu Kumar Pani, Joel J. P. C. Rodrigues, Babita Majhi. "Deep Learning, Machine Learning and IoT in Biomedical and Health Informatics - Techniques and Applications", CRC Press, 2022 Publication	<1%
5	d.docksci.com Internet Source	<1%
6	journals.plos.org Internet Source	<1%
7	edoc.rki.de Internet Source	<1%
8	Abdulkadir Karaci, Osman Ozkaraca, Ethem Acar, Ahmet Demir. "Prediction of traumatic pathology by classifying thorax trauma using a hybrid method for emergency services", IET Signal Processing, 2020 Publication	<1%
9	www.researchgate.net Internet Source	<1%
10	www.medrxiv.org	