# FORECASTING TUBERCULOSIS INFECTIONS USING ARIMA AND HYBRID NEURAL NETWORK MODELS AMONG CHILDREN BELOW 15 YEARS IN HOMA BAY AND TURKANA COUNTIES, KENYA

**BY**

**SIAMBA STEPHEN NYONGESA**

**A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN BIOSTATISTICS IN THE SCHOOL OF SCIENCE**
**UNIVERSITY OF ELDORET, KENYA**

**NOVEMBER, 2022**

# DECLARATION

This thesis is my original work and has not been submitted for any academic award in any institution; and shall not be reproduced in part or full, or in any format without prior written permission from the author and/or University of Eldoret.

**Declaration by the candidate**

**Signature: …………………………….. Date: ……………………………**

**Siamba Stephen Nyongesa**

**SC/PGM/052/11**

**Declaration by supervisors**

This thesis has been submitted with our approval as University Supervisors.

**Signature: …………………………….. Date: ……………………………**

**Dr. Argwings Otieno**

Head of Department

Mathematics Department

University of Eldoret, Kenya

**Signature: …………………………….. Date: ……………………………**

**Dr. Julius Koech**

Department of Mathematics

University of Eldoret, Kenya

# DEDICATION

To my father (late) Livingstone, mother Rebecca, my siblings Eric, Esther, and Kithim, words cannot cap it all, but a heartfelt thank you, and to my dear wife (Eunice) and children (Becky, Arlene, and Sebastian), I do all this for you.

# ABSTRACT

Tuberculosis (TB) among children under the age of 15 is a significant public health problem, particularly in resource-constrained settings and is among top ten most dangerous causes of death worldwide, and ranks among the top five most lethal infectious agents in Kenya. However, the real burden of tuberculosis among children in Kenya is unclear. In modelling infectious diseases, Autoregressive Integrated Moving Average (ARIMA) and hybrid ARIMA models have been widely used. However, few studies in Kenya have utilized ARIMA or hybrid ARIMA models to model infectious diseases. This study sought to forecast TB infections in children under the age of 15 Homa Bay and Turkana Counties in Kenya using ARIMA and hybrid neural network models and specifically sought to compare the; performance of the models in predicting TB notification cases, accuracy produced by the models, and the forecasted temporal trends of TB notification cases among children below 15 years. The study hypothesized that the hybrid ARIMA-ANN model yields more accurate predictions and forecasts. The study used monthly TB confirmed cases reported for Homa Bay and Turkana Counties between 2012 and 2021. The ARIMA model was chosen using the Akaike Information and Bayesian Information Criteria. The ANN model was developed using the Multi-Layer Perceptrons (MLPs) three-layer feed-forward architecture. The hybrid ARIMA model was developed by combining the fitted cases using the ARIMA model and the residuals from the ANN. The hybrid ARIMA model (ARIMA-ANN) outperformed the single ARIMA(0,0,1,1,0,1,12) and ANN (1,1,2)[12] models in terms of predictive and forecast accuracy. The hybrid ARIMA model outperformed the ANN (1,1,2)[12] and ARIMA (0,0,1,1,0,1,12) models in terms of prediction accuracy, $p<0.001$. In Homa Bay and Turkana Counties, the 12-month predicted TB incidence of 175 to 198 infections per 100,000 children in 2022. The hybrid ARIMA model provides superior prediction accuracy and forecast performance. The findings of this study suggest that TB cases in children are underreported, and that the incidence of TB in children may be greater than previously assumed. Tuberculosis monitoring data needs to be re-evaluated in order to comprehend current inadequacies. To get the TB battle back on track, it is critical to reallocate critical resources to the National TB program.

## TABLE OF CONTENTS

**DECLARATION** ..... **ii**

**DEDICATION** ..... **iii**

**ABSTRACT** ..... **iv**

**TABLE OF CONTENTS** ..... **v**

**LIST OF TABLES** ..... **ix**

**LIST OF FIGURES** ..... **x**

**LIST OF ABBREVIATIONS** ..... **xi**

**OPERATIONAL DEFINITION OF TERMS** ..... **xiii**

**ACKNOWLEDGEMENTS** ..... **xiv**

**CHAPTER ONE** ..... **1**

**INTRODUCTION** ..... **1**

1.1 Background to the study ..... 1

1.2 Problem statement ..... 4

1.3 Objectives of the study ..... 6

  1.3.1 Main Objective ..... 6

  1.3.2 Specific Objectives ..... 6

  1.3.3 Research Questions ..... 6

  1.3.4 Research Hypothesis ..... 6

1.4 Significance of the study ..... 6

1.5 Limitations of the study ..... 8

**CHAPTER TWO** ..... **9**

**LITERATURE REVIEW** ..... **9**

2.1 Modelling TB disease burden ..... 9

2.2 The concept of Time Series ..... 15

- 2.2.1 Autoregressive (AR) and Moving Average (MA) processes ..........................15
- 2.2.2 Autoregressive Moving Average (ARMA) models .......................................17
- 2.2.3 Autoregressive Integrated Moving Average models (ARIMA) models .........17
- 2.2.4 Seasonal Autoregressive Integrated Moving Average (SARIMA) models ....18

2.3 Artificial Neural Networks (ANNs).......................................................................19

- 2.3.1 ANN models...........................................................................................19
- 2.3.2 Architecture of the ANNs ....................................................................19
- 2.3.3 Hybrid models and training.................................................................22

2.4 Stationarity..................................................................................................23

2.5 ARIMA Model identification and specification ....................................................25

- 2.5.1 Model Identification.............................................................................25
- 2.5.2 Autocorrelation (ACF) and partial autocorrelation (PACF) functions ...........26

2.6 Predictive and Forecasting accuracy measures.....................................................27

**CHAPTER THREE .........................................................................................30**

**RESEARCH METHODOLOGY ..............................................................................30**

3.1 Study Area ...................................................................................................30

3.2 Materials and methods.....................................................................................31

- 3.2.1 Study design.........................................................................................31
- 3.2.2 Sample size and population....................................................................32
- 3.2.3 The Data ..............................................................................................32
- 3.2.3 Methodology proposed...........................................................................32

3.3 Ethical Considerations ....................................................................................36

**CHAPTER FOUR...............................................................................................38**

**RESULTS .........................................................................................................38**

4.1 Overview........................................................................................................38

4.2 Exploratory data analysis ........................................................................38

4.2.1 Data transformation ........................................................................39

4.2.2 Splitting data into training and testing data ....................................................42

4.3 Model comparison in predicting TB cases ...........................................................43

4.3.1 Model estimation ........................................................................43

4.3.1.1 Residual diagnostics ........................................................................45

4.3.1.2 Performance of the ARIMA model ........................................................46

4.3.1.3 Accuracy assessment of the ARIMA model ............................................47

4.3.2 Artificial Neural Network (ANN) Model Fitting ...........................................49

4.3.3 Hybrid Model Fitting ........................................................................50

4.4 Accuracy comparison based on different parameter specifications .......................53

4.5 Performance comparison of temporal forecast of TB trends ................................55

**CHAPTER FIVE ........................................................................59**

**DISCUSSION ........................................................................59**

5.1 Introduction ........................................................................59

5.2 Performance comparison of predictive accuracy of models .................................59

5.3 Accuracy comparison produced by different parameter specifications .................60

5.4 Model performance comparison in temporal forecast of TB trends ......................61

**CHAPTER SIX ........................................................................64**

**CONCLUSION AND RECOMMENDATIONS ....................................................64**

6.1 Conclusion ........................................................................64

6.2 Recommendations ........................................................................64

6.3 Limitations ........................................................................66

**REFERENCES ........................................................................68**

**APPENDICES ........................................................................................................81**

APPENDIX I: University Clearance to undertake research ........................................81

APPENDIX II: NACOSTI Research permit.................................................................82

APPENDIX III: Approval from Elizabeth Glaser Pediatric AIDS Foundation ..........83

APPENDIX IV: Similarity Report..................................................................................84

## LIST OF TABLES

Table 2.1: Theoretical ACF and PACF of AR(p), MA(q)and ARMA(p, q) models...27

Table 4.1: Model comparison ........................................................................44

Table 4.2: Model parameters .........................................................................45

Table 4.3: Comparison of ARIMA (0,0,1,1,10,1,12) forecasts and actual test data....47

Table 4.4: ARIMA model accuracy.............................................................49

Table 4.5: Accuracy comparison of the NNAR model...............................................50

Table 4.6: Hybrid model accuracy................................................................52

Table 4.7: Model accuracy comparison........................................................54

Table 4.8: Predictive accuracy comparison ..................................................55

Table 4.9: Comparison of TB case forecasts for the next 12 months..........................57

## LIST OF FIGURES

Figure 2.1: A three-layer feed-forward ANN architecture (Zhang, 2003) ..................20

Figure 3.1: Map of Kenya with Homa Bay and Turkana Counties marked out (Source: https://d-maps.com) ..........................................................................................31

Figure 3.2: Proposed methodology ...............................................................35

Figure 4.1: Monthly trend of TB cases between 2012 and 2021 ................................38

Figure 4.2: Decomposition of additive series ..............................................39

Figure 4.3: TB case monthly cycle plot .......................................................40

Figure 4.4: Monthly TB cases ACF plot.......................................................41

Figure 4.5: Monthly TB cases PACF plot ....................................................41

Figure 4.6: Training and testing set plot .....................................................42

Figure 4.7: Residual diagnostic plots..........................................................46

Figure 4.8: Plot of ARIMA (0,0,1,1,0,1,12) forecasts compared to actual test data ...48

Figure 4.9: Plot of ARIMA (0,0,1,1,0,1,12) model fitted TB and actual TB cases....48

Figure 4.10: Plot of NNAR (1, 1, 2)[12] fitted values against training data ...............50

Figure 4.11: NNAR model 24 month predicted TB cases ..........................................50

Figure 4.12: Plot of Hybrid model forecasted TB cases against actual testing data....52

Figure 4.13: Plot of Hybrid model fitted TB cases against actual training data.........53

Figure 4.14: Plot of ARIMA, ANN, and ARIMA-ANN model fitted TB against actual TB cases..........................................................................................54

Figure 4.15: ARIMA (0,0,1,1,0,1,12) 12 month forecast of TB cases ........................56

Figure 4.16: The NNAR (1,1,2)[12] model 12-month TB case prediction .................56

Figure 4.17: ARIMA-ANN 12 month forecast of TB cases.......................................57

Figure 4.18: Point forecast comparison .......................................................58

## LIST OF ABBREVIATIONS

| | |
|---|---|
| ACF | Autocorrelation Function |
| ACVF | Autocovariance Function |
| ADF | Augmented Dickey-Fuller |
| AIC | Akaike Information Criterion |
| AIDS | Acquired Immune-Deficiency Virus |
| ANN | Artificial Neural Networks |
| AR | Autoregressive |
| ARMA | Autoregressive Moving Average |
| ARIMA | Autoregressive Integrated Moving Average |
| BIC | Bayesian Information Criterion |
| COVID-19 | Corona Virus Disease 2019 |
| DM | Diebold-Mariano |
| FNN | Feedforward Neural Network |
| HIV | Human Immunodeficiency Virus |
| KNBS | Kenya National Bureau of Statistics |
| MA | Moving Average |
| MAE | Mean Absolute Error |
| MAPE | Mean Absolute Percent Error |
| MCAR | Missing Completely at Random |
| MLE | Maximum Likelihood Estimation |
| MLPs | Multi-Layer Perceptrons |
| MoH | Ministry of Health |
| NACOSTI | National Commission of Science, Technology and Innovation |

| | |
|---|---|
| NN | Neural Networks |
| NNAR | Neural Network Auto-Regressive |
| NNETAR | Neural Network Autoregression |
| NTLLDP | National Tuberculosis, Leprosy and Lung Disease Program |
| NTP | National Tuberculosis Program |
| PACF | Partial Autocorrelation Function |
| PP | Phillip-Perron |
| RMSE | Root Mean Square Error |
| SARIMA | Seasonal Autoregressive Integrated Moving Average |
| SSA | Sub-Saharan Africa |
| TB | Tuberculosis |
| TIBU | Treatment Information from Basic Unit |
| WHO | World Health Organization |
| MCAR | Missing Completely at Random |
| UN | United Nations |
| WN | White Noise |

# OPERATIONAL DEFINITION OF TERMS

| | |
|---|---|
| **Hybrid** | A combination of fitted values from the ARIMA model and ANN residual fitted values in order to account for linear and non-linear properties existing in the data resulting in better model performance compared to single models |
| **Forecast** | Forecasting involves taking models fit on historical data and using them to predict future observations |
| **Noise** | A non-systematic component that is nor Trend/Seasonality within the data |
| **Package** | In the context of the R statistical software, packages are collections of functions and data sets developed by the community that improve existing base R functionalities, or by adding new ones |
| **Seasonality** | Also a component of a time series. Seasonality is a general systematic linear or (most often) non-linear component that changes over time and does repeat |
| **Signal** | Information contained in a series |
| **Stationarity** | A common assumption in many time series techniques is that the data are stationary. A stationary process has the property that the mean, variance and autocorrelation structure do not change over time. |
| **Time series** | A series of data points indexed in time order. |
| **Trend** | One of the four components of a time series. Trend shows the general tendency of the data to increase or decrease during a long period of time and can be linear or non-linear. A trend is a smooth, general, long-term, average tendency. |

## ACKNOWLEDGEMENTS

I thank the almighty God for without him I could not have reached this far. My gratitude also goes to the School of Science and Department of Mathematics and Computer Science of University of Eldoret for granting me the opportunity to present this work at the University of Eldoret 2nd Postgraduate students conference.

I also extent my heartfelt appreciation to the Elizabeth Pediatric AIDS Foundation for granting me the opportunity to utilize the TB data as contained within the approval of their Patient and Program Outcomes Protocol.

# CHAPTER ONE

# INTRODUCTION

## 1.1 Background to the study

Tuberculosis (TB) is a highly infectious infection caused by the bacillus *Mycobacterium tuberculosis* and is among the top 10 most dangerous infectious agent-related causes of death (above HIV/AIDS), claiming nearly 4,000 lives each day. According to Lin and Liao (2013), around 33% of the worldwide population has been infected with tuberculosis, mainly in underdeveloped nations wherein TB is a major source of illness and mortality. The most common type of TB is pulmonary TB and spreads from one infected person to another through the air when the infected person coughs, speaks, or sings and nearby persons can inhale the TB bacteria and get infected as well.

The cause of TB was somewhat unknown until 1882 when the bacillus responsible for TB was discovered by Dr. Robert Koch and was eventually named *Mycobacterium tuberculosis* (Sakula, 1982). Generally, approximately 100 to 200 million people will develop active TB disease in their lifetime (WHO, 2018). People with risk factors such as malnutrition, diabetes, cigarette use, alcohol use, and HIV infection are more likely to acquire active TB illness (WHO, 2017, Shimeles *et al.*, 2019).

Globally, children accounted for about 7% of over 10 million new TB cases (WHO, 2017) in 2017. However, WHO (2018), reported that 55% of 1 million new TB cases among children were unreported while this was 65% among children under 5 years.

In 2019, about 1.4 million of people died from TB (WHO, 2020). Adolescent and pediatric TB is usually overlooked amid challenges faced regarding diagnosis (WHO, 2020). The challenge with sub-optimal treatment and under-reporting of TB cases among children stems from the fact that clinically and epidemiologically, and this makes care and prevention of TB a challenge (Cowger *et al.,* 2019). Furthermore, TB diagnosis in children is difficult (Krauss *et al*., 2015) especially among younger children who are not able to expectorate sputum and most often TB symptoms present similar to flu symptoms. Developing countries account for the largest proportion of new TB. In 2015, Asian countries contributed 61% of global new TB cases while Africa contributed over a quarter of the new cases (WHO, 2018). Globally, about 7 countries contributed about two-thirds of all new TB cases in 2016 (WHO, 2018) while in 2019, 87% of all new TB infections were accounted by 30 high TB burdened countries while 8 countries contributed 67% of new TB cases in 2019 (Floyd *et al.,* 2018, WHO, 2019).

When compared to other regions throughout the world, Sub-Saharan Africa (SSA) has a substantially higher TB burden thus imposing a great burden a massive load on already overburdened health-care systems (Zumla *et al.,* 2015). About 1.5 million people died of TB in 2013 with SSA accounting for over a quarter of the deaths (Zumla *et al.,* 2015). Despite having just 12% of the world population, SSA accounted for roughly 29% of the global total 9 million TB infections in 2014, with 254,000 deaths owing to TB (WHO, 2014).

In 2016, African contributed a quarter of all global reported TB cases despite accounting for only 10% of the global population (WHO, 2020). Among children, TB is often an overlooked cause of mortality because it is only accurately diagnosed in about 45% of children (Dodd *et al*., 2017, WHO, 2018). In 2017, approximately 1.2

million TB cases among children and adolescents aged 14 years or younger and about 205,000 deaths attributed to TB infection were reported (WHO, 2018).

In Kenya, tuberculosis is among the five leading causes of death and has a significant impact on the lives of the people. In addition, Kenya is among the top 30 TB burdened countries (WHO, 2016) and is among 14 countries faced with a triple burden of TB, TB_HIV and Multi-Drug Resistant TB (Kimani *et al.,* 2021, WHO, 2019). In 2015, the TB incidence was 233 per 100,000 people, with a fatality rate of 20 per 100,000 from all kinds of TB in Kenya. However, Kenya, identifies only 72% of bacteriologically confirmed TB infections and 80% of all cases (WHO, 2016).

Between 1990 and 2007, the number of TB cases reported in Kenya grew from 11,000 to 116,723 cases (Kipruto *et al.,* 2015, WHO, 2016) majorly because of the HIV pandemic and increased case detection as a result of greater diagnostic capability in the health system, as well as improved access to care as a result of health facility decentralization. Furthermore, the country has made significant progress in tuberculosis diagnosis through provision of resources that enable diagnosis and management of TB (Kimani *et al.,* 2021).

Forecasting may be accomplished using a variety of methodologies, ARIMA, and Neural Network models (Hyndman, 2018). The ARIMA models have increasingly been used in public health globally because of its advantages to effectively model the behavior of health outcomes where random variation is common. In the context of seasonal time series modeling, a variant of ARIMA known as SARIMA is used (Hamzacebi, 2008). The ARIMA model and its several modifications are based on Box-Jenkins principle (Box and Jenkins, 1976.

Furthermore, ARIMA models are popular because of their flexibility to represent time series with simplicity. The most significant constraint of these models is the linear assumption which is not possible in many circumstances. In order to overcome this disadvantage, various non-linear stochastic models have been proposed (Zhang, 2003, 2007). However, implementation of these non-linear stochastic models is not always straight-forward compared to linear models. However, many methods have been used in forecasting of infectious diseases (Ren *et al.*, 2013).

Although the ARIMA model has been popular among these models in its application to modelling infectious and non-infectious diseases, it has been limited by its inability to detect non-linear patterns in the data. Generally, rarely does data present as solely linear or non-linear and in most cases. As a result, there is a need to investigate robust ARIMA-based models, such as hybrid models, that can evaluate and represent both linear and non-linear patterns in the data. Because real-world applications mostly contain non-linear patterns, ANN models have been proposed resulting in significant improvements in prediction accuracy. However, ANN models cannot address both linear and non-linear patterns; hence, hybrid models have been used more recently to address this gap (Yolcu *et al.*, 2013; Khashei and Bijari, 2012).

## 1.2 Problem statement

Since 2000, the yearly TB incidence rate has decreased by 1.5% on average worldwide (Aryee *et al.*, 2018). However, to achieve the End TB first milestones, this rate needs to decrease by 4-5% on average annually. The Ministry of Health in Kenya has put in measures and interventions aimed at curtailing the spread of TB including integration of TB services in health care facilities. As a result, Kenya is among the 6 out of 15 high burden countries that were identified to be on target (Cha *et al*., 2020, WHO, 2015) to achieve the End TB first milestones.

Despite significant gains and global interventions to eradicate TB, the disease is still responsible for significant global morbidity and mortality with children most at risk (Marais *et al*., 2004). Marais (2011), noted that accurate identification of TB cases in children and the lack of good surveillance data make it difficult to accurately quantify TB burden among children. Furthermore, TB disease symptoms such as cough, fever, night sweats, weight loss and other symptoms can be mild for many months when a person develops active TB disease and this is even complicated among children (WHO, 2018).

Insufficient assumptions utilized in the modeling of health surveillance data can negatively impact the results, consequently having adverse effects when such results are used to inform decision making. More often, real world disease surveillance data rarely depict a purely linear association. On the other hand, nonlinear models have the disadvantage of lacking good theoretical underpinnings to explain its functioning although they are able to forecast with great accuracy when compared to established linear ARIMA models. To attain higher forecasting accuracy, use of hybrid models can lead to better predictive performance (Khashei and Bijari, 2012, Taskaya and Ahmad, 2005). This study proposes application of a real-world data-driven hybrid model to comprehend the dynamics of TB infection data among children below 15 years.

## 1.3 Objectives of the study

### 1.3.1 Main Objective

To forecast tuberculosis infections using ARIMA and hybrid neural network models among children below 15 years in Homa Bay and Turkana Counties in Kenya.

### 1.3.2 Specific Objectives

The specific objectives were to compare:

i. The performance of hybrid ARIMA-ANN model and ARIMA model in predicting TB notification cases
ii. The accuracy produced by different parameter specifications of the models used in modelling TB notification cases
iii. The forecasted temporal trends of TB notification cases

### 1.3.3 Research Questions

i. How does the models compare in forecasting TB notification cases?
ii. What is the degree of accuracy provided by various parameter specifications of the models used in modeling TB notification cases?
iii. How do the models compare in terms of forecasting temporal trends of TB notification cases?

### 1.3.4 Research Hypothesis

$H_{0:}$ The hybrid ARIMA model outperforms the single ARIMA and ANN models in terms of predictive and forecast accuracy.

## 1.4 Significance of the study

Tuberculosis identification of children aged 0-14 years had improved with the number of children diagnosed and initiated on TB treatment increasing from 4,483 in 2015 to 7,714 in 2017 translating to an increase of up to 72% (National Tuberculosis, Leprosy, and Lung Disease, 2019)). Furthermore, children identified bacteriologically with TB increased from 10% to 18% in the same period. However, identification of TB cases among children has continually been proven to be a challenge as about 66% of cases in SSA remain undiagnosed or un-reported (Jenkins, 2016).

According to Brent (2012), one of the major issues impeding worldwide efforts to eliminate tuberculosis is the inability to reliably detect and diagnose TB patients, which has remained sub-optimal and this affects the availability of high-quality surveillance data. In addition, while the seasonality of TB case notification has been extensively explored in the general population, this has not been the case among children in Kenya.

Brent (2012), further mentioned that the clinical presentation of tuberculosis in children differs from adults due to a combination of immunological, morphological, and epidemiological characteristics that make diagnosis challenging (Newton *et al.*, 2008). Furthermore, due to extended interaction with TB infected adults, the risk of children developing active TB is higher (Marais *et al.*, 2004, Schaaf *et al.*, 2003). As a result, reducing TB associated mortality among children requires a thorough understanding of the challenges posed by TB case notification to treatment and their outcomes (Mwangwa *et al*., 2017). It is also important to provide answers to whether the temporal trend from time series data of TB cases can be used in gaining clear information about future trends of TB cases among children especially in TB endemic counties in Kenya.

## 1.5 Limitations of the study

The research had no control on the accuracy of the data reported in the TIBU system. Regardless, because TB data recorded in the TIBU system is used at the nationally to report TB detection and management, the researcher assumed the data had been submitted to stringent data quality standards prior to reporting.

The study utilized data aggregated by month between 2012 and 2021 and the data obtained did not contain any missing data. However, in the event that the data could have contained missing values, this would have been treated as missing completely at random (MCAR) and data imputation techniques would have been employed to fill in the missing data.

In the year 2019 through to 2021, health care seeking behavior, transmission of TB, diagnosis, treatment and prevention and control efforts were greatly impacted by the COVID-19 pandemic (Alene *et al.*, 2020, Cilloni *et al.*, 2020).

# CHAPTER TWO

# LITERATURE REVIEW

## 2.1 Modelling TB disease burden

Mathematical models have been used in simulation of epidemic dynamics. While acknowledging that existing TB control interventions have been partially successful, Houben *et al*. (2014), asserted that, in the context of limited resources, these models can result in better practices that would amplify better health and economic benefits.

Garnet *et al.* (2011), noted that mathematical models are useful in projection of the potential public health and economic effect of interventions. Suyama *et al.* (2003), noted that epidemiologists have used applied mathematics to analyze data in public health and disease surveillance areas. However, owing to the diagnostic challenges of TB in children, not much is known about TB infection trends among children and from other potential associated factors hence the greater need for utilization of mathematical models to determine the temporal TB trend. In addition, since TB infections among children presents as a unique challenge that is compounded by paucity of data, it presents an area that is rarely explored.

Wang *et al.* (2018), employed time-series analysis, specifically, seasonal ARIMA and hybrid models to characterize the monthly TB notification rate in China between 2005 and 2017. The seasonal trend of tuberculosis incidence was investigated in these models.

Cao *et al.* (2013), conducted a study predict TB epidemics and analyzed its seasonality in China using the SARIMA, hybrid seasonal ARIMA, and Generalized Neural Networks (GNN) models to fit the data from 2005 to 2010, and noted better performance of the hybrid modes. The study's seasonal tendency projected a lower monthly incidence in January and February and a greater incidence from March through June.

Wah *et al.* (2014), used ARIMA model to assess the relationship between population characteristics and yearly TB cases and found out that TB risk among non-residents was significantly linearly decreasing compared to Singapore residents, but with no clear seasonal trend in TB cases.

Xiao *et al.* (2018), investigated the impact of climatic conditions on tuberculosis incidence in Southwest China from 2006 to 2015. (DLNM). After adjusting for autocorrelation, they discovered that variations in climatic parameters such as temperature, humidity, wind, and sunlight were strongly correlated with TB incidence. The co-dynamics of tuberculosis with climatic conditions were investigated in this study.

Zeming and Yanning (2020), conducted a study to predict HIV-AIDS incidences in China in 2017 using monthly HIV-AIDS data using ARIMA, back propagation and a hybrid model. They found that the hybrid model offered better predictive power.

Zhou *et al.* (2016) conducted a study to forecast the prevalence of schistosomiasis in Qianjiang, China using an ARIMA-NARNN (Non-linear Autoregressive Neural Network) model on yearly schistosomiasis data from 1956 to 2012 and found that the hybrid model produced high quality prediction accuracy. They recommended using the hybrid model to identify schistosomiasis prevalence in other schistosomiasis endemic areas, including other infectious illnesses.

Yu *et al.* (2014) conducted a study to forecast occurrence of hand, foot, and mouth disease using monthly incidence case data from 2008 to 2012 in Shenzhen, China. They used a hybrid seasonal ARIMA and NARNN model and found out that the best-fitting model was the hybrid seasonal ARIMA-NARNN model with forecasts indicating a clear increase in hand, foot, and mouth disease occurrences in Shenzhen.

Li *et al.* (2019), conducted a research to predict TB cases in China using ARIMA and ARIMA-generalized regression neural network (GRNN) hybrid models on monthly TB incidence data in Lianyungang from January 2007 to December 2016 and found that the hybrid model offered better performance compared to the single ARIMA model.

The use of innovative machine learning algorithms in modeling illness incidence is widely established in Africa. In various variations, these models have been used to simulate and anticipate the short- and long-term patterns of non-infectious diseases including cancer and malaria (Anokye *et al.,* 2018, Ebhuoma *et al.,* 2018).In these studies; as much as the ARIMA model offered a way of predicting cases, it did not guarantee perfect forecasts especially over a longer forecast horizon, can best be applied on data that is stable over time with minimum outliers (Anokye *et al.,* 2018) and would not be suitable if there is no clear strategy of dealing with outliers and suffer from lack of enough data which can result in either under-fitting or over-fitting of the model (Ebhuoma *et al.,* 2018).

ARIMA and seasonal ARIMA models have recently been used to predict and forecast COVID-19 cases in Sub-Saharan Africa. While recognizing that time series models have been widely used to estimate the prevalence or spread of infectious diseases, Takele (2020), utilized the ARIMA model to predict Covid-19 prevalence Ethiopia, Djibouti, Sudan, and Somalia in East Africa. They emphasized that the nature of COVID-19 distribution may alter future predictions of COVID-19 cases, notably in the context of the four nations studied. Furthermore, the study may have considered the influence of seasonality, such as the days of the week when COVID-19 infections were highest or lowest.

Furthermore, Umunna and Olanrewaju (2020), modelled HIV prevalence in Minna in Niger state in Nigeria using ARIMA and SARIMA models using monthly HIV data from 2007 to 2018 and found out that the SARIMA model was the best for forecasting monthly HIV prevalence. Of interest in their findings was that the average fitted value from January 2007 was half of the actual value reported which in essence would indicate under-fitting and might have been better addressed by considering a more robust approach for model evaluation during model development. In addition, outliers which might have accounted for extraneous variation might have been present within the data basing on the 95% prediction intervals including negative values. Furthermore, the optimal SARIMA model might have been impacted by the existing non-linearities within the data which were not effectively accounted for by the model.

Ade *et al.* (2016), conducted a study to model TB incidences using data between 2000 and 2014 in Benin using ARIMA model and found out that the TB cases exhibited seasonal changes. In addition, Aryee *et al.* (2018), carried out a study to forecast TB incidences at Korle-Bu Teaching Hospital's chest clinic using data between 2008 and 2017 and applied the ARIMA model. Though they found no indication of a growing or decreasing trend in the incidence of tuberculosis, they did observe that the best model does not necessarily yield the best results. As a result, the study might have used a more accurate model and technique for better accuracy.

In Africa, Azeez *et al.* (2016), carried out a study to model TB incidence in South Africa using SARIMA and hybrid SARIMA-NNAR models and found out that the SARIMA-NNAR had a better goodness-of-fit compared to the SARIMA model and recommended that strong action is required to minimize infectious disease spread.

Hybrid ARIMA models have been utilized to model short-term and long-term infectious disease incidence in other parts of the world but with little application of the same in Africa with majority of the applications confined to single ARIMA models (Anokye *et al.,* 2018; Azeez *et al.,* 2016; Manikandan *et al.,* 2016, Achieng *et al.,* 2020). Application of hybrid ARIMA models have been applied majorly in other health sectors such as agriculture and commerce with very little application in public health.

Gashu *et al.* (2018), conducted a study on TB cases in Ethiopia by fitting a time series model on TB case data between 2010 and 2016 and discovered that the TB cases exhibited a seasonal pattern and differences between the Amhara and Oromia regions. The findings of this study, in terms of seasonality of TB cases were similar to those by Azeez *et al.* (2016), Cao *et al.* (2013), and Wah *et al.* (2018).

Frah and Alkhalifa (2016), were able to fit a Box-Jenkins model in time series analysis of TB cases in Sudan and showed a fairly downward trend pattern. However, the study did not explore the seasonality of TB cases and at granular level such as age in order to establish presence of any trends. In Ghana, Aryee *et al.* (2018), estimated the incidence of TB cases reported at a tertiary hospital in Ghana using a Box-Jenkins approach by utilizing monthly data reported between 2008 and 2017 and found no evidence of obvious fluctuation in the trend of TB cases. The study concluded that the best model does not necessarily result in the best accuracy due to the lack of clear evidence of trend, implying the necessity to investigate improved models.

In Uganda, Jaganath *et al.* (2019), carried out a study aimed at establishing the seasonality of childhood TB cases in Kampala between the periods 2010 to 2015 respectively. In their study, they explored the role of meteorological factors and influenza cases on TB diagnoses. Monthly mean plots were compared in their analysis, and Poisson regression was used to analyze the connection between TB diagnoses and meteorological parameters. They discovered a clear association with TB diagnoses, notably an overlap with pulmonary TB cases.

While the Box-Jenkins time series models and hybrid models based on the Box-Jenkins model have been used in exploring seasonality of TB cases with meteorological and population-based factors particularly in other SSA countries, not much has been done in Kenya. In Kenya, very little work in relation to comparison of effective models to assess TB case trend particularly among the most vulnerable population of children has been done. One of the main reasons of the increased risk of pediatric TB infection is due to seasonality (Tedijanto *et al.*, 2018, Padberg *et al.*, 2015, Wubuli *et al.*, 2017). Children aged 4 years or younger had about 4 times the seasonality of TB risk compared to older children (Willis *et al.,* 2012) while in China, children aged 0-14 years exhibited the highest seasonal variation (Wubuli *et al.*, 2017).

Pediatric TB in equatorial regions, especially in the context of seasonality, have not been explored by studies in the past covering the entire four distinct seasons but rather variations between wet and dry seasons. Furthermore, very few studies have focused on pediatric TB cases and its seasonality while exploring the different generic models based on the Box-Jenkins model. As a result, it is critical to conduct research aimed at evaluating and identifying TB cases among children under the age of 15 years in order to give recommendations on targeted resource mobilization and allocation, particularly in tropical countries such as Kenya.

## 2.2 The concept of Time Series

A time series is a sequence of data points measured over time, and it is technically described as a collection of vectors Y(t) where t=0, 1, 2, ... and indicates the elapsed time. The vector Y(t) is often a random variable with values recorded and ordered chronologically. A time series is often made up of four components: trend, cyclical, seasonal, and irregular. As a result, the goal of time series analysis is to separate time series variance into trend, periodic, and stochastic components (Sarpong, 2013).

### 2.2.1 Autoregressive (AR) and Moving Average (MA) processes

An AR model is a representation of a type of random process and is used to describe certain time-varying processes within the time series data (Bakar and Rosbi, 2017). The fundamental principle behind these models is that the current value of a series $Y_t$ is a function of p previous values, that is, $Y_{t-1}, Y_{t-2}+, \dots, + Y_{t-p}$. As a result, an AR process of order p is expressed as follows;

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \varepsilon_t \qquad (1)$$

Where $\{\varepsilon_t\} \sim WN(0, \sigma^2)$ and is uncorrelated with Ys for each s < t

For simplicity, equation 1 assumes that the mean of $Y_t$ is zero.

On the other hand, when the mean, $E(Y_t) = \mu \neq 0, Y_t$ is written as $Y_t - \mu$, to obtain;

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + \ldots + \phi_p(Y_{t-p} - \mu) + \varepsilon_t,$$

Which can be written as;

$$Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t, \tag{2}$$

Where; $\alpha = \mu(1 - \phi_1 - \ldots - \phi_p)$

Furthermore, equation 1 may be written as;

$$Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \ldots - \phi_p Y_{t-p} = \varepsilon_t,$$

Using the backshift operator, however, $BY_t = Y_{t-1}$, we write;

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)Y_t = \varepsilon_t$$

Alternatively, we may use simple notation and write;

$$\phi(B)Y_t = \varepsilon_t, \tag{3}$$

Where the autoregressive (AR) operator is denoted by $\phi(B)$;

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \tag{4}$$

Generally, $\phi(B)$ can be defined as the characteristic polynomial of the process and its roots determine whether the process is stationary or non-stationary.

Therefore, the AR (p) can be viewed as a solution to the equation;

$$Y_t = \frac{1}{\phi(B)} \varepsilon_t \tag{5}$$

A MA model is one that makes use of the association between actual or observed values and the error term from a MA model applied to lagged values and this implies that the output variable is linearly related to the present and previous values of an error term (Bakar and Rosbi, 2017).

As a result, $\{Y_t\}$ is a MA process of order q if;

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}, \tag{6}$$

Where; $\varepsilon_t \sim \text{White Noise (WN)}\ (0, \sigma^2)$, and the constants are $\theta_1, \ldots, \theta_q$

A MA(q) process, on the other hand, may be stated as follows:

$$Y_t = \theta(B)\varepsilon_t, \tag{7}$$

Where the MA operator;

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q = 1 + \sum_{j=1}^{q} \theta_j B^j \tag{8}$$

is a linear combination of values in the shift operator;

$$B^k \varepsilon_t = \varepsilon_{t-k}$$

### 2.2.2 Autoregressive Moving Average (ARMA) models

An ARMA (p, q) model is a combination of AR(p) and MA(q) models and is a type of stochastic process where the auto-covariance functions are determined by a finite number of unknown parameters. In general, an ARMA process of orders p and q may be expressed as (Lee, lecture 4);

$$Y_t = \sum_{j=1}^{p} \phi_j Y_{t-j} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \ \forall\, t \in \mathbb{Z} \tag{9}$$

The ARMA (p, q) process is written in lag operator notation as;

$$\phi(B) Y_t = \theta(B) \varepsilon_t \ \forall\, t \in \mathbb{Z} \tag{10}$$

### 2.2.3 Autoregressive Integrated Moving Average models (ARIMA) models

The ARIMA model was developed by Box and Jenkins in 1960 and like other forecasting methods, it requires historical data on the variable under consideration. They are represented in the forecast equation as ARIMA (p, d, q), where p is the number of AR terms, d is the number of non-seasonal variations, and q is the number of lagged errors (Shrivastav and Ekata, 2012; Pfaff, 2008).

The ARIMA model assumes that the time series is stationary with the following assumptions;

i. Residuals are normally distributed and independent, having a zero mean and homogenous variance: this is expressed mathematically as $\varepsilon_t \sim N(\mu, \sigma^2)$. This implies that for model adequacy, the correlation in the observations has been eliminated.

ii. Variance homogeneity and residual zero mean: plots of standardized residuals vs predicted values are used to examine variance homogeneity over time; nevertheless, if the variance is heterogeneous, logarithmic processing can be applied to achieve homogeneity.

iii. Residuals are independent: ACFs and PACFs should demonstrate that the residuals are a white noise process.. Mathematically, this assumption can be presented as $\varepsilon_t \sim \mathrm{iid}N(0, \sigma^2)$.

### 2.2.4 Seasonal Autoregressive Integrated Moving Average (SARIMA) models

The Seasonal ARIMA model is a multiplicative model composed of non-seasonal and seasonal components. ARIMA (p, d, q) (P, D, Q)$^S$ is the representation for a SARIMA model.

A backshift operator is defined as $BY_t = Y_{t-1}$.

Thus, a non-seasonal AR process can be written as;

$$\phi(B) = 1 - \phi_1(B) - \cdots - \phi_p(B^p) \tag{11}$$

A non-seasonal MA process can be written as;

$$\theta(B) = 1 + \theta_1(B) + \cdots + \theta_q(B^q) \tag{12}$$

A seasonal AR process can be written as;

$$\vartheta(B^S) = 1 - \vartheta_1(B^S) - \cdots - \vartheta_P(B^{PS}) \quad (13)$$

A seasonal MA process can be written as;

$$\Theta(B^S) = 1 + \Theta_1(B^S) + \cdots + \Theta_Q(B^{QS}) \quad (14)$$

A SARIMA model may be expressed explicitly without differencing as;

$$\vartheta(B^S)\phi(B)(Y_t - \mu) = \theta(B)\Theta(B^S)\,\varepsilon_t, \forall\, t \in \mathbb{Z} \quad (15)$$

The left of equation 15 comprises the seasonal and non-seasonal AR process and on the right of the equation we have the seasonal and non-seasonal MA processes. Also, because monthly TB case data are used in this study, S=12.

## 2.3 Artificial Neural Networks (ANNs)

### 2.3.1 ANN models

These models have been proposed as alternative and superior modeling tools for forecasting (Khashei and Bijari, 2010). The goal of ANNs is to develop a model for duplicating human brain cognition into a computer (Kihoro *et al.*, 2006). As a result, ANNs are biologically driven (Larie and Cockrell, 2021), and are also data driven and self-adaptive (Zhang *et al.*, 1998). Consequently, there is no need to define a particular model or make any assumptions about the data distribution.

### 2.3.2 Architecture of the ANNs

The most widely used ANNs in forecasting are Multi-Layer Perceptrons (MLPs), which use a single hidden layer feed-forward network (FNN) (Cinar, 2020). The model is composed of three layers: the input layer, the hidden layer, and the output layer, that are connected by acyclic linkages. According to Darji *et al.* (2015), a neural network can be composed of a single or multi-layered network of neurons produced when one neuron interacts with another. Figure 2.1 depicts a three-layer feed-forward design of ANNs.

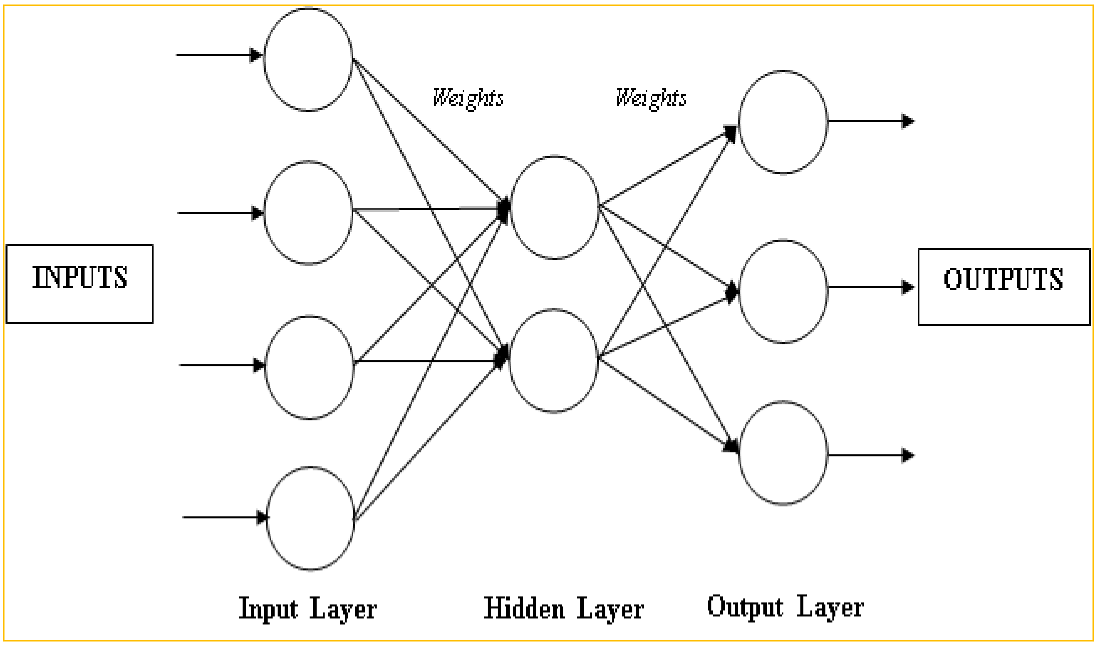


**Figure 2.1: A three-layer feed-forward ANN architecture (Zhang, 2003)**

According to Zhang (2007), the model output can be computed as:

$$Y_t = \gamma_0 + \sum_{j=1}^{q} \gamma_j h\left(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} Y_{t-i}\right) + \varepsilon_t, \forall\, t \tag{16}$$

Where $Y_{t-i}$ (i=1, 2, …, p) are the p inputs and $Y_t$ is the output; p and q are the number of input and hidden nodes respectively, $\gamma_j$ (j=0, 1, 2, …, q) and $\beta_{ij}$ (i=0, 1, 2, …, p; j=0, 1, 2, …, q) are the connection weights and $\varepsilon_t$ is the random shock; bias terms are $\gamma_0$ and $\beta_{0j}$. There is no scientific rule for deciding on *q*. The logistic function h(.) is commonly used as the non-linear activation function, where:

$$h(.) = \frac{1}{1 + e^{-x}} \tag{17}$$

In reality, the model in equation 16 conducts a non-linear functional mapping from past time series data to future value. That is to say:

$$Y_t = f\left(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}, v\right) + \varepsilon_t \tag{18}$$

In this case, *v* is a vector comprising all parameters and f(.) is a function. According to Zhang (2003), ANNs have the capacity to represent non-linearity between input variables and output variables, and universal approximators that can estimate a huge class of functions with an exceptional degree of precision.

Generally, for any input layer of a neural network, there is always a single layer. The number of neurons in the input layer is determined by the shape of the data. The number of neurons is equal to the number of features of the data. In this case, the number of neurons in the input layer is the number of optimal lags (*p*) from the time series based on the lag selection of the neural network function. However, in general, some neural network configurations add one additional node to account for the bias term.

With regard to the output layer, every neural network has exactly one output layer. However, determining the number of neurons in the output layer is completely determined by the selected model configuration. In the case of this study, the output layer is $Y_t$.

As a result, if the number of hidden layers is set to 1, the number of neurons/nodes in the hidden layer is the mean of the neurons in the input and output layers, that is, in this example, the mean of the number of neurons in the input layer, p, and the number of neurons in the output layer, 1.

As a result, the input layer's number of neurons is;

$$\frac{p+1}{2} \tag{19}$$

Where p is the number of optimal lags and can be determined from the optimal ARIMA model.

In general, to prevent over-fitting, the number of neurons should be kept below;

$$N_h = \frac{N_s}{(\alpha * (N_i + N_o))} \tag{20}$$

The number of input neurons is $N_i$, $N_o$ is the number of output neurons (1 in the case of this study), $N_s$ is the number of samples in the training set (96 months of TB cases in the case of this study) and $\alpha$ is an arbitrary scaling factor between 2 and 10. In this study, the automated selection was be able to prevent potential overfitting.

### 2.3.3 Hybrid models and training

In addition to selecting the suitable number of hidden nodes, selecting the appropriate number of lagged observations p is vital since this is the most critical parameter in an ANN model because it plays a significant influence in shaping the series' non-linear autocorrelation structure (Stokes *et al.*, 2020).

On the contrary, because there is no formal theory to aid in the determination of the optimal value of *p*, experiments are conducted to determine the suitable or optimal p as well as q. This is one of the most significant gaps that limit the application of ANNs.

As indicated in equation 21, a time series can be presented as having linear and nonlinear components in an additive model.

$$\mathrm{Y_t = l_t + n_t} \tag{21}$$

Where $l_t$ and $n_t$ are the linear and non-linear components respectively estimated separately where the linear component is developed first followed by the nonlinear component (fitted using the ANN structure). The residual term, $e_t$ is obtained by subtracting the predicted value $\widehat{\mathrm{Y}}_t$ from the actual value $\mathrm{Y_t}$ at time t as shown below;

$$\mathrm{e_t = Y_t - \widehat{Y}_t} \tag{22}$$

In the event that there are linear correlations in the residuals, the linear (ARIMA) model is not sufficient in predicting the data. Existence of significant non-linear patterns in the residuals would imply the limitation of the ARIMA models and by modelling residuals using ANNs, the non-linear relationships are discovered. The MAPE, MAE and RMSE are used in evaluating the forecasting accuracy. The best model was used to predict short-term tuberculosis cases in 2022.

## 2.4 Stationarity

An AR process is considered to be stable or stationary if the parameters exist within a specified range, for instance, if there is only one AR parameter, then it must fall in the $-1 < \phi < 1$ range. Conversely, past effects would accumulate and successive $\mathrm{Y_t}'\mathrm{s}$ would move towards infinity rendering the series non-stationary. There exists duality between the MA and AR process (Box and Jenkins, 1976). Regardless of the MA parameters, an MA(q) process is always stationary (Adhikari and Agrawal, 2013).

In order to assess time series stationarity, the Augmented Dickey-Fuller (ADF) and Phillips-Perron (PP) unit root tests can be used on the non-differenced and differenced time series. Furthermore, the ACF and PACF plots can be examined to assess stationarity (Arltová and Fedorová, 2016). The ADF test is an improved version of the Dickey-Fuller (DF) test (Dickey and Fuller, 1981). The null hypothesis of the ADF test is that there is a unit-root and the alternative hypothesis is that there is no unit-root where the test statistic $\mathrm{DF}_k = \frac{\hat{\gamma}}{\mathrm{SE}(\hat{\gamma})}$ is calculated and compared with the critical value of the Dickey-Fuller test.

Furthermore, this study used the Phillips-Perron (PP) unit root test (Phillips and Perron, 1988). The PP test is more complicated than the ADF unit root test, although it has the same null hypothesis and employs the same critical values. The PP test is a non-parametric adjustment to the t-statistic that uses the Dickey-Fuller equation;

$$\Delta Y_t = \mu Y_{t-1} + v + \lambda_t + \varepsilon_t \tag{23}$$

Since $\varepsilon_t \sim \mathrm{iidN}(0, \sigma^2)$ and can be heteroscedastic, the test estimates the equation;

$$Y_t = Y_{t-1} + v + \lambda_t + \varepsilon_t \tag{24}$$

The PP test estimates the non-augmented DF test equation and adjusts the coefficient's t-ratio so that serial correlation has no effect on the test statistic's asymptotic distribution. The PP test is statistically written as;

$$\overline{t_\mu} = t_\mu \sqrt{\frac{\gamma_0}{f_0}} - \frac{T(f_0 - \gamma_0)[se(\mu)]}{2(\sqrt{f_0})s} \tag{25}$$

Where $\gamma_0$ and $f_0$ are estimates of $\sigma_\varepsilon$ and $\sigma$

In incidences where the time series is non-stationary, differencing would be employed and testing for stationarity with each differencing until stationarity is achieved.

## 2.5 ARIMA Model identification and specification

### 2.5.1 Model Identification

This study used the correlogram and partial correlogram (autocorrelation and partial autocorrelation plots). The mechanical technique was also used in the study to estimate model parameters by testing each model at different values of p, d, and q. Furthermore, table 2.1 displays theoretical models but does not provide the ultimate ideal model to be examined. As a result, numerous potential models may be constructed, and the penalty function statistics such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) can be used to pick the optimal model (Chakrabarti and Ghosh, 2011). The AIC and BIC are measures of an estimated statistical model's quality of fit. The AIC and BIC measurements are:

$$\mathrm{AIC}(p) = n\ln(\hat{\sigma}_e^2/n) + 2p \tag{26}$$

$$\mathrm{BIC}(p) = n\ln(\hat{\sigma}_e^2/n) + p + p\ln(n) \tag{27}$$

Where *n* is the number of observations used to fit the model, p is the number of lag parameters and the sum of sample square residuals are presented as $\hat{\sigma}_e^2$. The optimal model is selected by the number of model parameters that minimizes the AIC or BIC. In this study, the aim was to select the model parameters that minimize both the AIC and BIC.

In order to determine if an observable series is linearly independent, the Ljung-Box statistic (Ljung and Box, 1978) is utilized and is defined as follows:

$$Q(h) = T(T+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{T-k} \tag{28}$$

The sample autocorrelation at lag k is $\widehat{\rho_k}$, the sample size is T and the number of lags accounted for in the test is h. When the null hypothesis is satisfied, the statistic is asymptotically $\chi^2$ distributed with h degrees of freedom.

### 2.5.2 Autocorrelation (ACF) and partial autocorrelation (PACF) functions

It is critical to analyze the ACF and PACF as part of univariate analysis to establish the appropriate model for a particular time series since these approaches indicate how the observed observations in a time series are associated. Plotting the ACF and PACF versus sequential time delays is critical for modeling and forecasting, and these plots assist establish the order of AR and MA components, respectively. Their mathematical definitions are;

For a timeseries $Y_t \ \forall\, t \in \mathbb{Z}$, the autocovariance at lag k can be defined as:

$$\gamma_k = \mathrm{Cov}(Y_t, Y_{t+k}) = E[(Y_t - \mu)(Y_{t+k} - \mu)] \tag{29}$$

The autocorrelation coefficient (Cochrane, 1997) at lag k can be defined as:

$$\rho_k = \frac{\gamma_k}{\gamma_0} \tag{30}$$

Where $\mu = E[Y_t]$ is the time series mean, the autocovariance at lag zero, $\gamma_0$, is the time series variance while $\rho_k$ (autocorrelation coefficient) is dimensionless and independent of the scale of measurement and $-1 \leq \rho_k \leq 1$. In essence, the autocovariance $\gamma_k$ is the theoretical autocovariance function (ACVF) and $\rho_k$ is the theoretical autocovariance function (ACF) at lag k (Box and Jenkins, 1976).

The PACF, on the other hand, is employed after correcting for data at intermediate lags to evaluate the correlation between an observation k lags ago and the present observation., that is, at lags <k.

Generally, the theoretical framework for AR, MA and ARMA is stated in table 2.1.

**Table 2.1: Theoretical ACF and PACF of AR(p), MA(q)and ARMA(p, q) models**

| **Model** | **ACF** | **PACF** |
|---|---|---|
| AR(p) | Dies down (exponential decay) | Cuts off after lag p after spikes between lags 1 to p |
| MA(q) | Cuts off after lag q after spikes at lags 1 to q | Dies down (exponential decay) |
| ARMA(p ,q) | Dies down (exponential decay) | Dies down (exponential decay) |

## 2.6 Predictive and Forecasting accuracy measures

Because time series forecasting is crucial, especially in the trend and forecast of infectious illnesses such as tuberculosis, adequate attention should be made while picking a given model. As a result, multiple performance metrics (Hamzacebi, 2008; Zhang, 2007) have been developed to quantify forecast accuracy and compare different models.

The forecast accuracy for the ARIMA and ARIMA-ANN suggested models was measured using the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percent Error (MAPE) in this study. The RMSE is the square root of the calculated Mean Square Error (MSE) with each property of MSE holding for RMSE. The MAPE represents the percentage of average absolute error and is independent of the scale of measurement but affect by data transformation. The MAE measures the average absolute deviation of predicted values from actual values.

These 3 measures can be defined as;

$$\mathrm{RMSE} = \sqrt{\frac{\sum e_t^2}{n}} \tag{31}$$

$$\mathrm{MAPE} = \left[\frac{\sum |e_t|/Y_t}{n}\right] \times 100 \tag{32}$$

$$\mathrm{MAE} = \frac{\sum|e_t|}{n} \tag{33}$$

Where $t = 1, 2, 3, \dots, n$, *e* is the forecast error in the period t, $Y_t$ is the real value of the time period t, *n* is the observations in the period; $e_t = Y_t - \widehat{Y}_t$ where $\widehat{Y}_t$ is the predicted value.

Furthermore, according to Lewis (1982), the MAPE values for model selection as $\mathrm{MAPE} \leq 10\%$ for high forecast accuracy, $10\% < \mathrm{MAPE} \leq 20\%$ for good forecasting accuracy, $20\% < \mathrm{MAPE} \leq 50\%$ for reasonable forecasting accuracy.

The Diebold-Mariano (DM) test is used to test the predictive accuracy of any two models under comparison (Diebold and Mariano, 1995) and test the hypothesis that the two forecasts are equally accurate, whereas the alternative hypothesis states that the two forecasts are not equally accurate. The DM test takes into consideration a sample path of loss differentials $\{d_t\}_{t=1}^{T}$. In case of a squared loss function, then we have, $d_t = \varepsilon_t^2 - \breve{\varepsilon}_t^2$ where $\varepsilon$ and $\breve{\varepsilon}$ are the losses or forecast errors from two forecast models. Assuming that the loss differential is a stationary covariance series, the sample average, $\bar{d}$, will asymptotically converge to a normal distribution.

$$\sqrt{T}\bar{d} \overset{d}{\rightarrow} N(\mu, 2\pi f_d(0)) \tag{34}$$

Diebold and Mariano proposed testing the null hypothesis, which states that forecast errors from two different forecasts would result in roughly the same loss, $E[\varepsilon_t^2 - \breve{\varepsilon}_t^2] \cong 0$, when compared to a two-sided alternative. If the null hypothesis is true, in a new experiment, the generated p-values show the likelihood of attaining the realized forecast error differential or a more severe one. The DM test statistic is as follows;

$$\mathrm{DM} = \frac{\bar{d}}{\sqrt{\frac{2\pi\hat{f}_d(0)}{T}}} \tag{35}$$

Where, $2\pi\hat{f}_d(0)$ is a consistent estimate equal to $\sum_{\tau=-(T-1)}^{(T-1)} \omega_\tau \gamma_d(\tau)$ where;

$$\gamma_d(\tau) = \frac{1}{T}\sum_{t=|\tau|+1}^{T}(d_t - \bar{d})(d_{t-|\tau|} - \bar{d}) \tag{36}$$

# CHAPTER THREE

# RESEARCH METHODOLOGY

## 3.1 Study Area

Homa Bay County is one of the former districts of Nyanza province in Kenya (Figure 3.1). This county is situated on the shores of Lake Victoria, which provides a significant source of income. Homa Bay County is approximately 3,155 square kilometers and lies approximately 0.6221° S, 34.3310° E and has a population of about 1,131,950 (Kenya National Bureau of Statistics, 2019). Homa Bay town is the largest town as well as the headquarters of the Homa Bay county which is made up of 8 Sub-Counties. Homa Bay county has a HIV prevalence that is 4.5 times higher than the national prevalence (Otieno and Okuku, 2017). Rono and Migwambo (2018), identified TB/HIV co-infection as a potential risk factor for the highest numbers of TB-related deaths in Homa Bay county.

Turkana County is located 3.3122° N, 35.5658° E within the former Rift Valley province of Kenya (Figure 3.1). It is by far the largest county in Kenya by land area and occupies approximately 68,680 km2. It is bordered by Uganda, South Sudan and Ethiopia and is largely an arid area with approximately 926,976 population (Kenya National Bureau of Statistics, 2019). Lodwar is the largest town and is the capital of Turkana County. The population in Turkana is majorly nomadic and is considered a hardship area prone to drought (Ojakaa *et al.*, 2014).

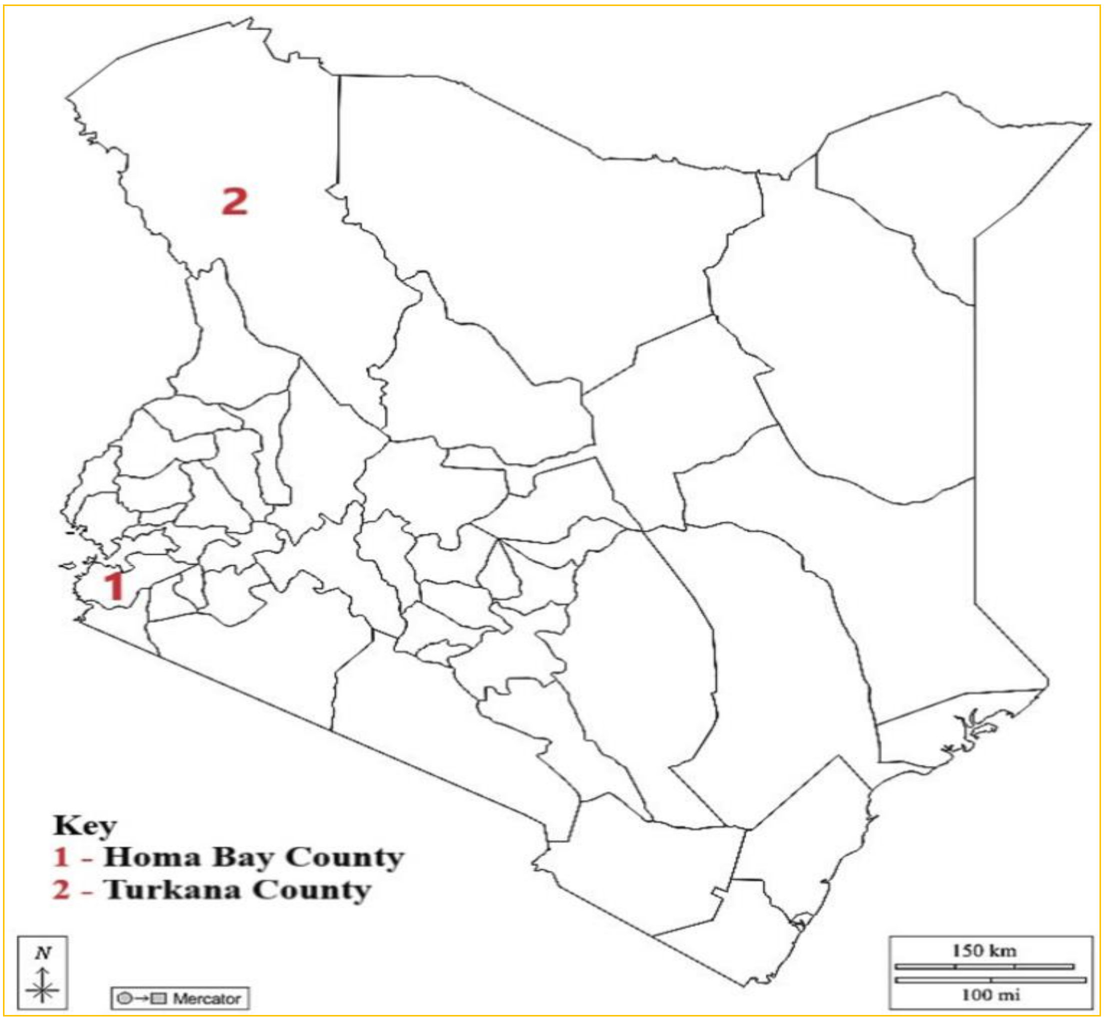


**Figure 3.1: Map of Kenya with Homa Bay and Turkana Counties marked out (Source: https://d-maps.com)**

## 3.2 Materials and methods

### 3.2.1 Study design

This was a retrospective cohort study that used secondary routinely collected data on children under the age of 15 in Homa Bay and Turkana Counties who had a confirmed TB diagnosis.

### 3.2.2 Sample size and population

The population of this study included all children under the age of 15 years who were screened and tested for TB in health institutions in Kenya's Homa Bay and Turkana Counties. Within this population, the study examined all the TB data reported in the TIBU system between 2012 and 2021 by health facilities in Homa Bay and Turkana Counties for children under the age of 15 years.

### 3.2.3 The Data

This study analyzed monthly tuberculosis (TB) case notification data from the Treatment Information from Basic Unit (TIBU) system. The data covered the period from January 2012 to December 2021. The Division of Leprosy, Tuberculosis and Lung Disease and Kenya's Ministry of Health (MoH) implemented the transition from a paper-based recording and reporting system to TIBU, an electronic system, in 2012 (MoH, 2010).

The TIBU system is a national case-based surveillance system that stores individual tuberculosis cases reported to the national TB program and has had nationwide coverage since then (MoH, 2012). The National Commission for Science, Technology, and Innovation (NACOSTI) granted permission to carry out the research through a research permit and a letter of authorization was received from the Elizabeth Glaser Pediatric AIDS Foundation to use the data within the Patient and Program Outcomes Protocol.

### 3.2.3 Methodology proposed

The proposed methodology for this study is depicted in Figure 3.2, which is based on the Box-Jenkins methodology for the ARIMA and incorporates the ANN and hybrid

ARIMA models. The TIBU system data was transformed to time series data, which includes TB notification instances recorded between 2012 and 2021. The time series data was then divided in an 80:20 ratio, with 80% used as training data for model construction and 20% used for model validation. The PP and ADF tests were used to determine stationarity in the training data. The ACF and PACF plots were also examined to determine the order of the ARIMA model with the best p and q parameters. To test the hypothesis of residual independence, constant variance, and zero mean, the Ljung-Box Q statistic was utilized.

After ensuring constant variance and zero mean residuals, several ARIMA models were created using the training data, and the most parsimonious model with the lowest AIC and BIC was chosen. The predicted TB cases from 2012 to 2019 were fitted using the ARIMA model and analyzed for accuracy using the RMSE, MAE, and MAPE accuracy metrics. The best model was also subjected to a sliding window cross-validation procedure to ensure that the best of the best ARIMA model is selected with the best accuracy parameters. In this case, 11-month training data was modelled using the best model and the 12-month TB cases were forecasted. This procedure was repeated for proceeding months until data for December 2019 had been forecasted. In each case, the RMSE, MAE, and MAPE parameters were obtained and averaged across within the cross-validation process. This ensured that the best model produced had the best accuracy. The cross-validation procedure was also used to select the best forecast horizon without compromising the accuracy of the model. The best model was then used to forecast data for 24 months. Following that, the best ARIMA model was utilized to anticipate TB notification instances for 2022.

In the instance of the ANN model, the NNETAR function was used to fit the training data to the ANN model. The NNETAR function used a MLP with a single hidden layer neural network to build a multilayer perceptron. In general, the NNETAR function fits a NNAR(p, P, k)m model, with p and P values chosen automatically when not supplied. This model considered seasonality in the time series. As a consequence, P=1 was chosen as the default value, and p was picked from the best ARIMA (linear) model structure that suited the seasonally adjusted data. If k is not supplied, it defaults to k=(p+P+1)/2, rounded to the nearest integer. The ANN model was pre-set with a decay parameter of 0.001 and a maximum iteration of 200. Setting the decay parameter to 0.001 prevented the weights from becoming too large, Setting the maximum iteration to 200 guaranteed that the model could test many models until the ideal model with the lowest RMSE was developed. Setting this value to 200 is neither large, 1000 or above, nor small, below 100, to cause overfitting or underfitting. The fitted values were assessed for accuracy by compared with the actual values from the training data and once the best ANN model was obtained, TB cases for the next 24 months were forecasted and accuracy measures obtained by comparing with the test data. Thereafter, the TB cases for 2022 were forecasted.

The residuals from the best ARIMA model were retrieved and fitted using the ANN model using the NNETAR function in the ARIMA-ANN model. This was to ensure that any remaining signal, which is the non-linear data not adequately modelled by the ARIMA model, was accounted for. The ARIMA point prediction from the best model and the residual point forecast from the residual ANN model were merged to generate the ARIMA-ANN, hybrid, model. Over a 24-month period, the hybrid model was utilized to estimate TB notification instances, and the projected values were compared to actual test results. Measures of accuracy were obtained. Following that, the hybrid model was used to forecast TB cases for 2022.

The forecast and predictive accuracy of the three models were compared using RMSE, MAE, and MAPE for the former and the Diebold-Mariano test for the latter, and the best model was chosen and proposed. In addition, the forecasts from the 3 models were compared and the incidence calculated.

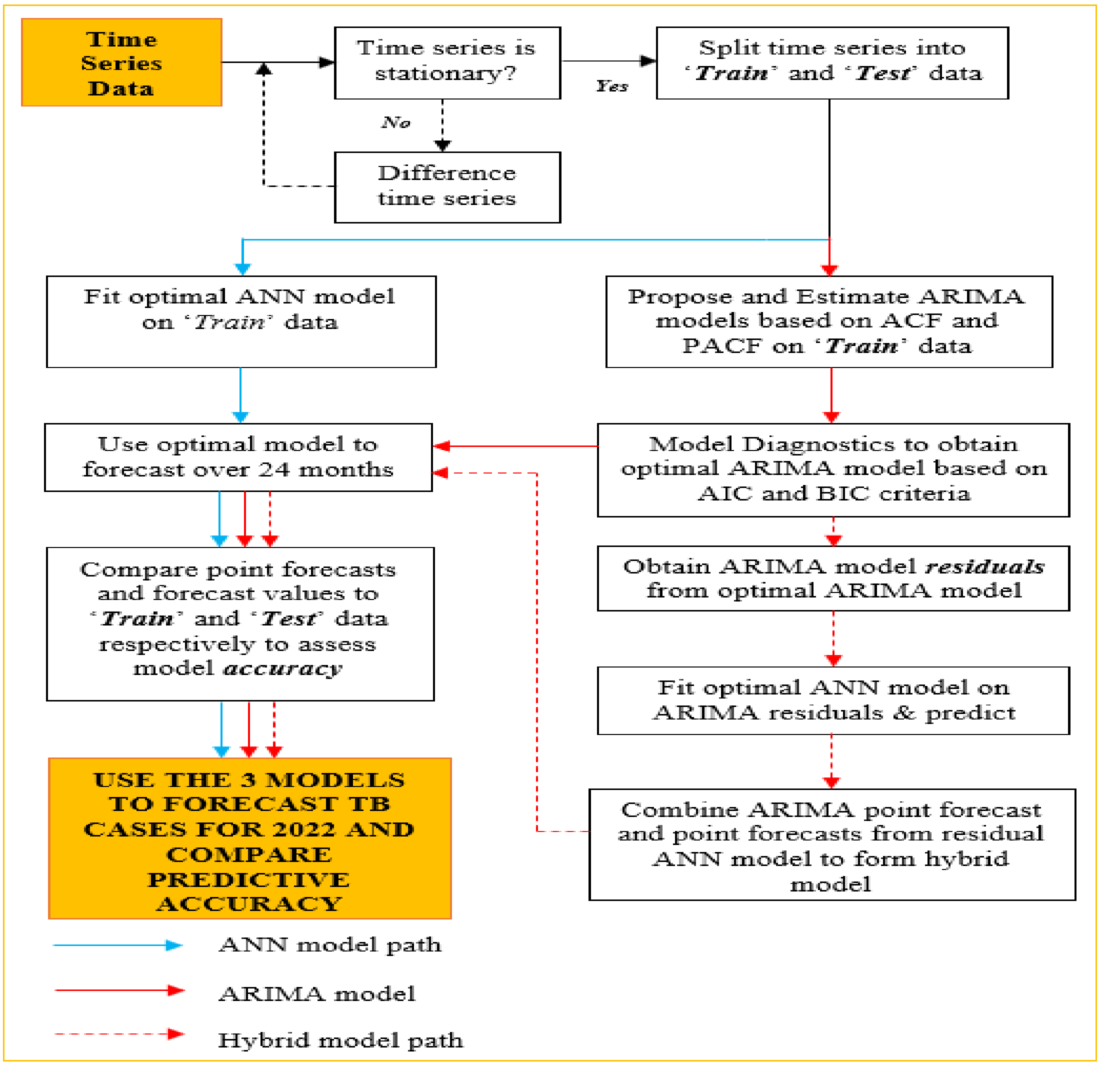

Time Series Data
Time series is stationary?
Yes
Split time series into 'Train' and 'Test' data
No
Difference time series
Fit optimal ANN model on 'Train' data
Propose and Estimate ARIMA models based on ACF and PACF on 'Train' data
Use optimal model to forecast over 24 months
Model Diagnostics to obtain optimal ARIMA model based on AIC and BIC criteria
Compare point forecasts and forecast values to 'Train' and 'Test' data respectively to assess model accuracy
Obtain ARIMA model residuals from optimal ARIMA model
Fit optimal ANN model on ARIMA residuals & predict
USE THE 3 MODELS TO FORECAST TB CASES FOR 2022 AND COMPARE PREDICTIVE ACCURACY
Combine ARIMA point forecast and point forecasts from residual ANN model to form hybrid model
ANN model path
ARIMA model
Hybrid model path

**Figure 3.2: Proposed methodology**

## 3.3 Ethical Considerations

Human subject protection and human rights principles must be observed/applied in all research activities involving human subjects, whether experimental or descriptive. This also includes any relevant information or data. The following aspects of research ethics were considered in this study:

a) **Permission to conduct the research**: Permission was sought from NACOSTI who issued a research permit for license number: NACOSTI/P/22/19567 (Appendix II), the Elizabeth Glaser Pediatric AIDS Foundation on data use within the approved Patient and Program Outcomes Protocol (Appendix III) and research authorization and introduction letter from the Board of Post-Graduate Studies of University of Eldoret (Appendix I).

b) **Informed consent:** This study used retrospective source of data which was accessed from an existing health system database on TB. Aggregated data (devoid of patient identifying information) was collected from the TIBU system, thus, there was no need for informed consent or assent or waiver of the same. In addition, within the confines of the Patient and Program Outcomes Protocol, through which this study utilized the TB data, waiver of consent was granted

c) **Confidentiality and Anonymity:** Patient data whether patient level data or aggregate level data is sensitive information with potential to identify patients or groups of patients. The researcher has the responsibility of ensuring that that patient data is handled, stored and shared in a confidential manner and with authorities given permission to access or handle such information. In this study, patient confidentiality was not compromised as patient names of other identifiers were not collected. The study only collected aggregate data from the

TIBU system. Passwords were only shared by authorized County health personnel to facilitate access to the TIBU system to carry out the study and were kept secret.

d) **Publication in journal(s):** As a requirement, the findings, recommendations and conclusions of the study were published within an academic journal. However, a disclaimer about the findings, conclusions and recommendations of the publication was stated.

# CHAPTER FOUR

# RESULTS

## 4.1 Overview

This section outlines the statistical output obtained based on the time series methodologies used so as to achieve each of the study objectives.

## 4.2 Exploratory data analysis

This study used monthly TB notification case data aggregated from the TIBU system from 2012 to 2021, respectively. There were 120 data points in total. Figure 4.1 depicts the trend of tuberculosis cases reported among children under the age of 15 in Homa Bay and Turkana Counties. The figure shows a slight increase in between 2018 and 2021. (Figure 4.1).

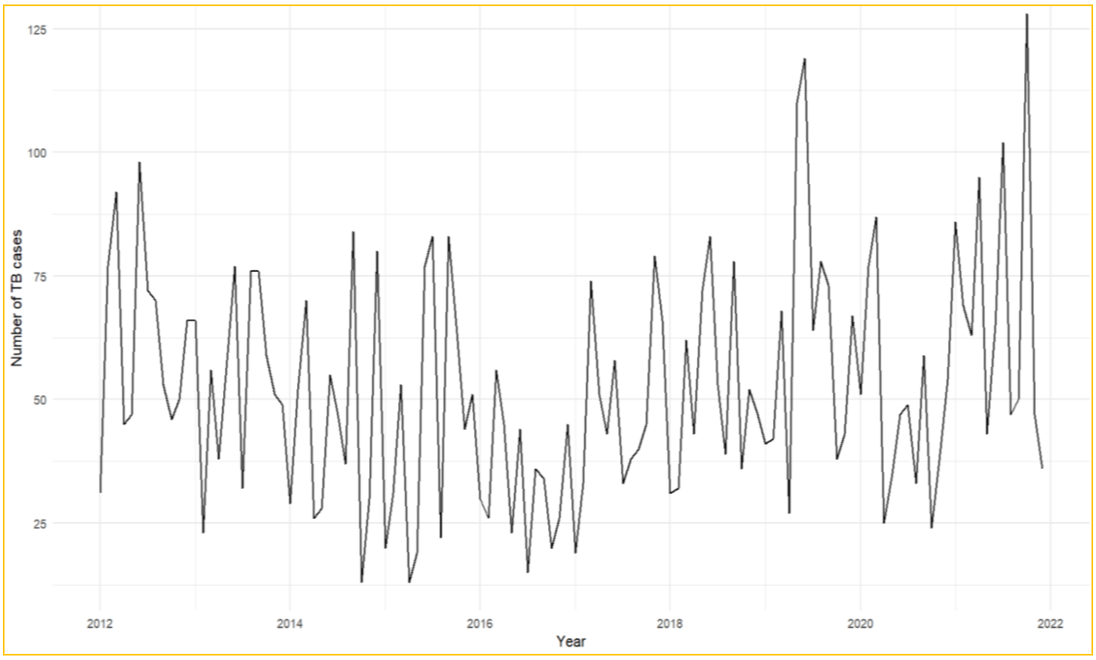


**Figure 4.1: Monthly trend of TB cases between 2012 and 2021**

### 4.2.1 Data transformation

In the development of the hybrid model, the ARIMA model is first developed and this requires that the data is stationary. The plot in figure 4.2 shows an unclear trend in the series and this would require inspection to determine if the stationary or not and the ADF and PP tests were used in this case. In the event of non-stationarity, the series would then be differenced and tests applied with each cycle of differencing until stationarity is achieved.

Figure 4.2 presents the decomposed time series of TB cases composed of the trend, seasonal and random components respectively. The decomposed series clearly shows an increasing trend from mid-2018 to late-2019 and a slight decrease thereafter. The seasonal component shows quarterly seasonality with the random component having a mean of approximately 0. Because reported TB cases vary seasonally, the results above suggest that the ARIMA model should account for seasonality.

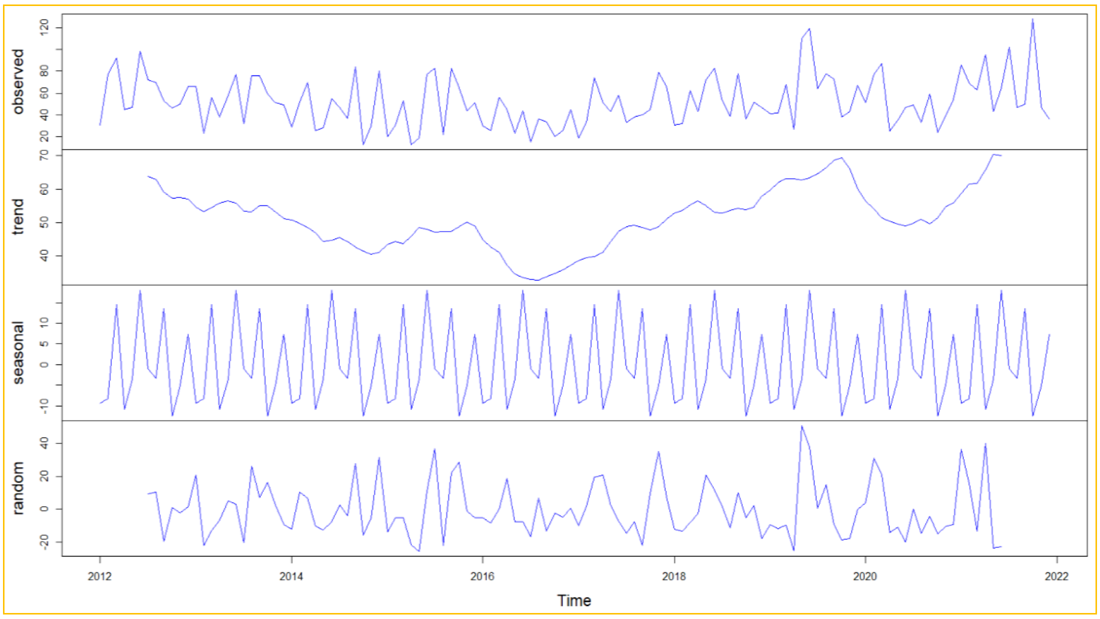


**Figure 4.2: Decomposition of additive series**

Figure 4.3 demonstrates a distinct seasonality in TB cases, with TB cases on average being greater in March, June, September, and December throughout the different years. This suggests that seasonality must be accounted for in the ARIMA model.

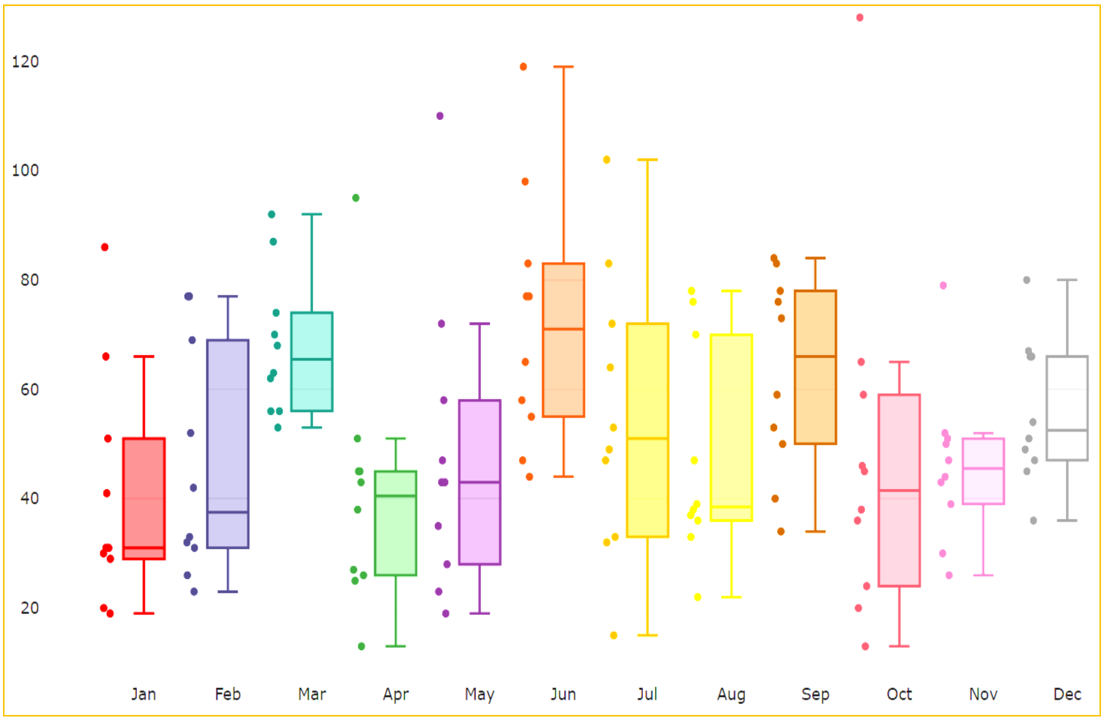


**Figure 4.3: TB case monthly cycle plot**

Figures 4.4 and 4.5 depict the ACF and PACF plots that were used to determine optimal p and q values. The ACF plot in figure 4.4 exhibits non-decay sequence showing stationarity of the series. However, there are possible significant autocorrelations at lag 1. This implies that the series potentially exhibits a MA(0) or MA (1) process. Furthermore, the p-values for the ADF and PP tests p-values were <0.001, providing strong evidence against failing to reject the null hypothesis and the conclusion is that the series is stationary, thus no difference is required, implying that the value of d is 0 at a 95% confidence interval, which is greater than 0.05.

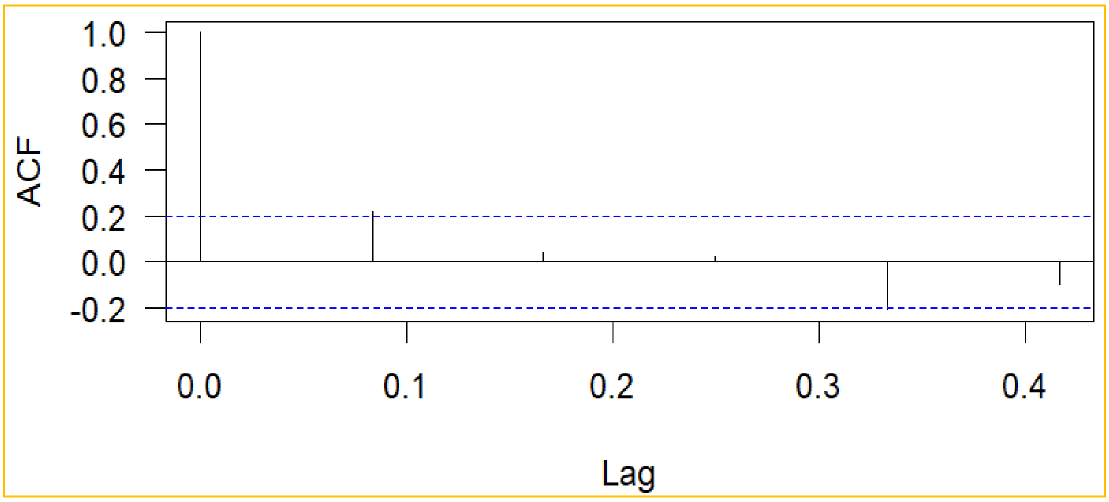


**Figure 4.4: Monthly TB cases ACF plot**

Figure 4.5 shows and confirms a non-decaying series, hence stationarity. The PACF plot also shows a potential decay and a significant spike at lag 4 which shows seasonality, quarterly, confirming the need to account for it in the possible ARIMA models. The models to be experimented on would be potentially at the first lag. This means that the series is most likely a AR(0) or AR(1) process.

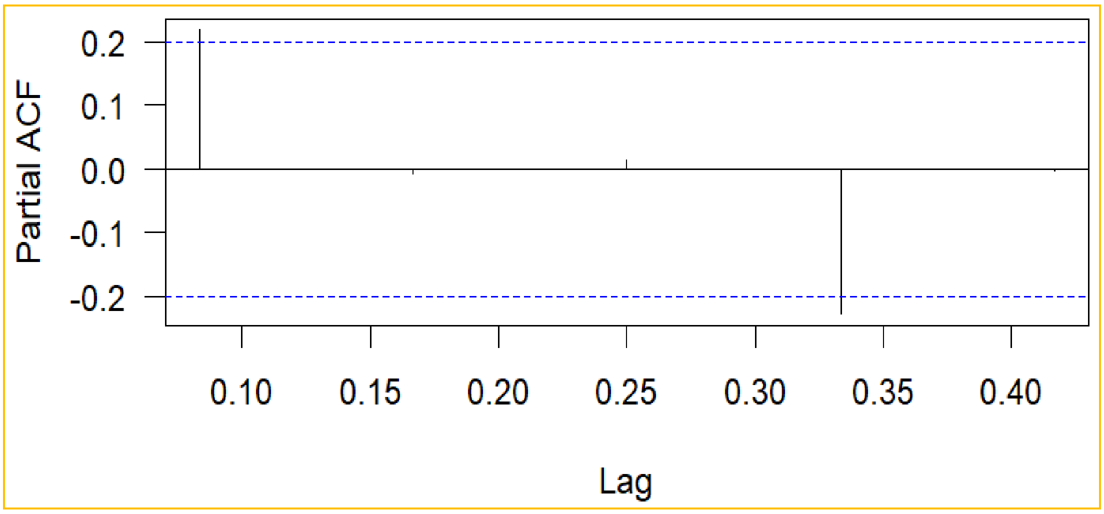


**Figure 4.5: Monthly TB cases PACF plot**

### 4.2.2 Splitting data into training and testing data

The goal of separating data into a training and testing set is to prevent the model from overfitting. The training set was used to develop the model and the test set for validation. The survey data ranged from January 2012 to December 2019, with 80% from January 2020 to December 2021. Based on the available chronological data points, the 80:20 strategy of partitioning the data into a training and testing set has been proved to yield test error rate estimates with minimal bias and variance (James *et al.*, 2013). When the RMSE, MAE, and MAPE are at their ideal minimum values, the model learns successfully and can predict values that are significantly closer to the true values (Medar, Rajpurohit, & Rashmi, 2017). Figure 4.6 is a plot of the training set ('train') and the testing set ('test') comprising 96 and 24 records respectively.

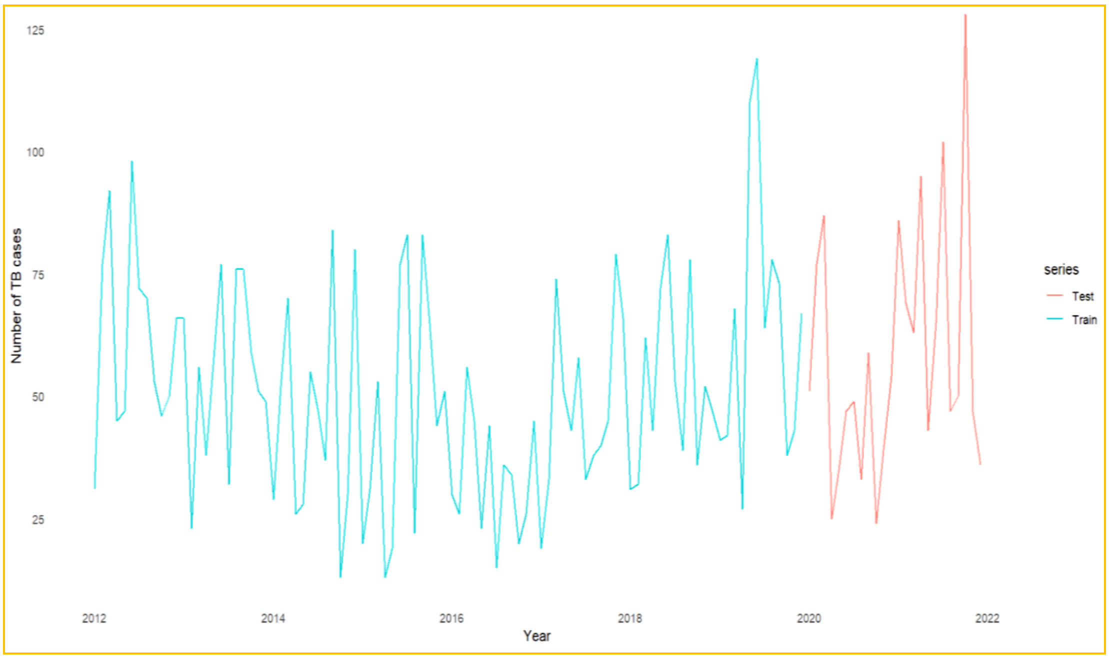


**Figure 4.6: Training and testing set plot**

### 4.3 Model comparison in predicting TB cases

#### 4.3.1 Model estimation

To choose the best parsimonious model with the lowest estimated AIC or BIC values, AIC and the BIC were utilized. In order to select the best model, 14 ARIMA models were considered and tested, each with seasonality accounted for. The Ljung-Box Q statistic was used to assess model residual independence. The null hypothesis of the Ljung-Box Q test is that the residuals are independently distributed. Table 4.1 displays the best comparative potential ARIMA models based on the ACF and PACF plot findings.

ARIMA (0,0,1,1,0,1,12) is the best model with lowest AIC and BIC values of 858.9 and 871.7 respectively and the estimated model parameters were p=0, d=0, and q=1 and P=1, D=0, and Q=1. Accounting for seasonality in the model improved accuracy of the model. The best model's Ljung-Box Q test yielded a p-value of 0.971, indicating that the ARIMA (0,0,1,1,0,1,12) model residuals were not serially linked.

Exploring the model residual diagnostics showed significant auto-correlations and partial auto-correlations at lag 3 respectively. Though this does not significantly impact on the distribution of the residuals, it shows that there is potentially some signal remaining in the residuals that has not been captured by the ARIMA model. Furthermore, the best model selected presented with the least RMSE=18.74, MAE=14.39 and MAPE=39.00 when compared to other models under consideration.

**Table 4.1: Model comparison**

| **Model** | **AIC** | **BIC** |
|---|---|---|
| ARIMA (0,0,0,0,0,1,12) | 865.8 | 873.5 |
| ARIMA (0,0,0,1,0,0,12) | 863.6 | 871.2 |
| ARIMA (0,0,0,1,0,1,12) | 864.5 | 874.7 |
| ARIMA (0,0,1,0,0,1,12) | 863.5 | 873.7 |
| ARIMA (0,0,1,1,0,0,12) | 861.1 | 871.3 |
| **ARIMA (0,0,1,1,0,1,12)** | **858.9** | **871.7** |
| ARIMA (1,0,0,0,0,0,12) | 871.2 | 878.9 |
| ARIMA (1,0,0,1,0,0,12) | 862.1 | 872.3 |
| ARIMA (1,0,0,0,0,1,12) | 864.7 | 874.9 |
| ARIMA (1,0,0,1,0,1,12) | 859.1 | 871.9 |
| ARIMA (1,0,1,0,0,0,12) | 867.5 | 877.8 |
| ARIMA (1,0,1,0,0,1,12) | 861.9 | 874.7 |
| ARIMA (1,0,1,1,0,0,12) | 860.6 | 873.5 |
| ARIMA (1,0,1,1,0,1,12) | 960.5 | 875.9 |

The best (see Table 4.1) ARIMA model comprised a non-seasonal AR(1), a seasonal MA(1) and non-seasonal MA(1) polynomials. These three polynomials are presented as:

Let the backshift operator be presented as $BY_t = Y_{t-1}$,

A non-seasonal AR(1) polynomial can be written as;

$$\phi(B^{12}) = 1 - \phi_1 B^{12} \quad (37)$$

A Seasonal MA(1) polynomial can be written as;

$$\Theta(B^{12}) = 1 + \Theta_1 B^{12} \quad (38)$$

A Non-seasonal MA(1) polynomial can be written as;

$$\theta(B) = 1 + \theta_1 B \quad (39)$$

As a result, the model equation is;

follows: $(1 - \phi_1 B^{12})(1 + \Theta_1 B^{12})(1 + \theta_1 B)(Y_t - \mu) = \varepsilon_t, \forall\, t \in \mathbb{Z}$ (40)

Where $\{\varepsilon_t\} \sim \mathrm{WN}(0, \sigma^2)$, and $\sigma^2 = 351.1$

However, when $\mathrm{E}(\mathrm{Y_t}) = \mu \neq 0$, $Y_t$ is replaced by $Y_t - \mu$.

The estimated coefficients as (see Table 4.2);

ma1 = $\theta_1 = 0.291$

sar1 = $\phi_1 = 0.997$, and

sma1 = $\Theta_1 = -0.953$

$\mu = Intercept = 50.902$

Inserting these estimated model parameters in the model, we have the model equation as;

$$(1 - 0.997B^{12})(1 - 0.953B^{12})(1 + 0.291B)(Y_t - 50.902) = \varepsilon_t \tag{41}$$

**Table 4.2: Model parameters**

| | Estimate | Std. Error | Z-value | Pr(>\|z\|) |
|---|---|---|---|---|
| ma1 | 0.291 | 0.108 | 2.701 | 0.007* |
| sar1 | 0.997 | 0.015 | 68.167 | <0.001** |
| sma1 | -0.953 | 0.127 | -7.512 | <0.001** |
| Intercept | 50.902 | 5.064 | 10.052 | <0.001** |

#### 4.3.1.1 Residual diagnostics

Following model fitting, the model should be verified for fit using the usual model diagnostic checking method, namely residual analysis. Four charts were utilized in model diagnostic testing to examine the underlying assumptions. Figure 4.7 depicts four plots: an ACF plot, a PACF plot, a white noise probability plot at various lags, and a quantile-quantile (Q-Q) plot of the theoretical quintiles to test for residual normality. The Q-Q plot showed that the residuals were normally distributed. At lag 3, inspection of the ACF and PACF plots to assess residual randomness and find patterns or extreme values found high auto-correlations, suggesting possible existing signal in the residuals that has been not adequately modelled.

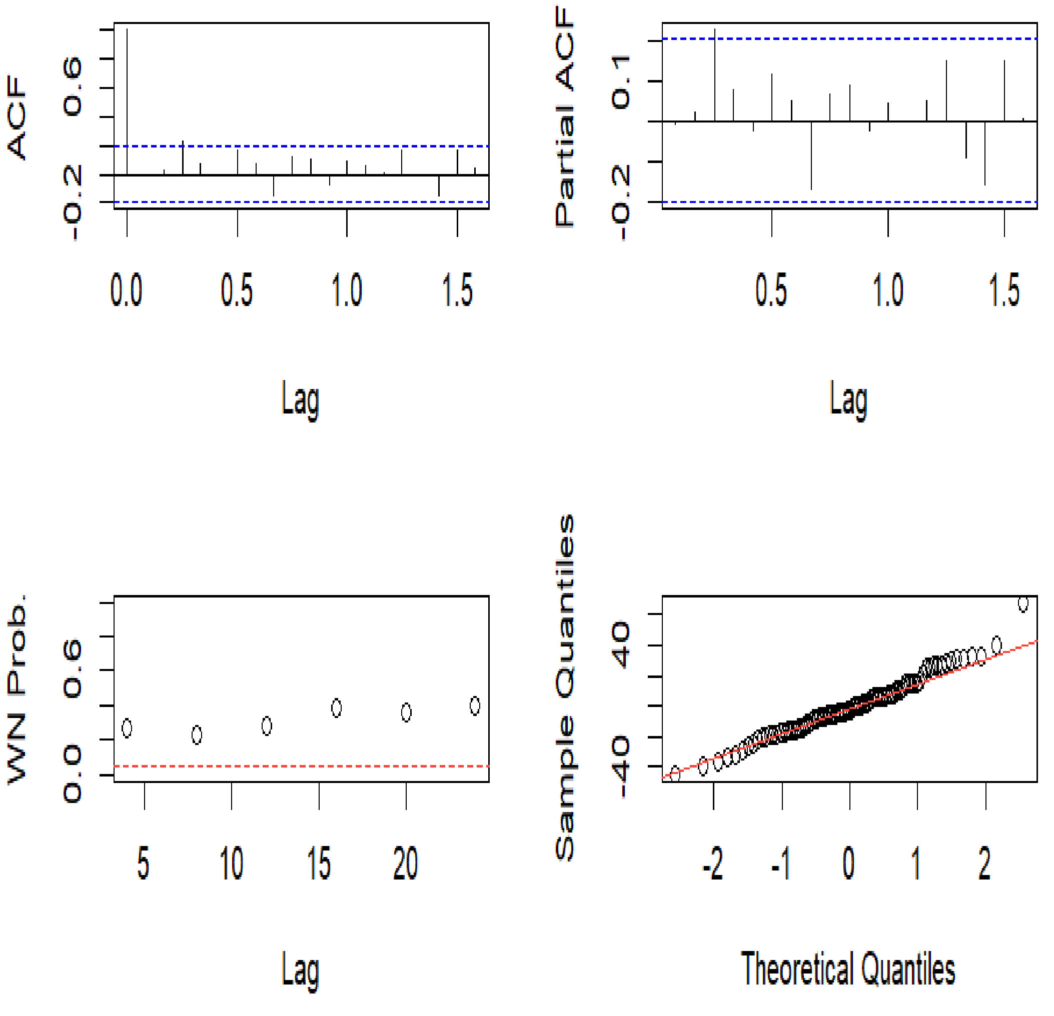


**Figure 4.7: Residual diagnostic plots**

### 4.3.1.2 Performance of the ARIMA model

Evaluation of the forecast accuracy of the ARIMA model was carried out by comparing the forecasted values against the actual test data. Figure 4.8 presents the forecast/fitted plot for 24 months with 80% and 95% prediction intervals. Furthermore, figure 4.9 presents the rolling regression of the actual (observed) training data compared with the ARIMA (0,0,1,1,0,1,12) fitted data while Table 4.3 compares the ARIMA (0,0,1,1,0,1,12) 24 month forecast data to the actual test data. The ARIMA model (0,0,1,1,0,1,12).

Table 4.3: Comparison of ARIMA (0,0,1,1,10,1,12) forecasts and actual test data

| Month | Forecast data | Actual test data |
|---|---|---|
| Jan-20 | 41 | 51 |
| Feb-20 | 42 | 77 |
| Mar-20 | 63 | 87 |
| Apr-20 | 39 | 25 |
| May-20 | 51 | 35 |
| Jun-20 | 71 | 47 |
| Jul-20 | 50 | 49 |
| Aug-20 | 50 | 33 |
| Sep-20 | 62 | 59 |
| Oct-20 | 43 | 24 |
| Nov-20 | 48 | 39 |
| Dec-20 | 57 | 54 |
| Jan-21 | 38 | 86 |
| Feb-21 | 42 | 69 |
| Mar-21 | 63 | 63 |
| Apr-21 | 39 | 95 |
| May-21 | 51 | 43 |
| Jun-21 | 71 | 65 |
| Jul-21 | 50 | 102 |
| Aug-21 | 50 | 47 |
| Sep-21 | 62 | 50 |
| Oct-21 | 43 | 128 |
| Nov-21 | 48 | 47 |
| Dec-21 | 57 | 36 |

#### 4.3.1.3 Accuracy assessment of the ARIMA model

The assessment of the ARIMA model accuracy is presented in Table 4.4 and shows the comparison of the accuracy measures between the training and testing data.

Comparison of the 24-month forecasts from the ARIMA (0,0,1,1,0,1,12) model against the actual test data for 2020 to 2021 shows a mean of 54 cases forecasted against an actual mean of 59 cases. The rolling regression plot in figure 4.9 of the actual training data against the ARIMA (0,0,1,1,0,1,12) fitted values shows that data from the fitted values represent a mean of 51 TB cases compared to a mean of 51 TB cases from the actual training data. This implies that while the model under fits on the test data, it produces better results that are comparable to the actual training data.

Table 4.4 demonstrates that the model performs slightly worse on testing data, RMSE=18.74 against RMSE=29.17 on testing data. When a fitted model fails to effectively account for important information within the data and would need to be accounted for using a more robust method or model.

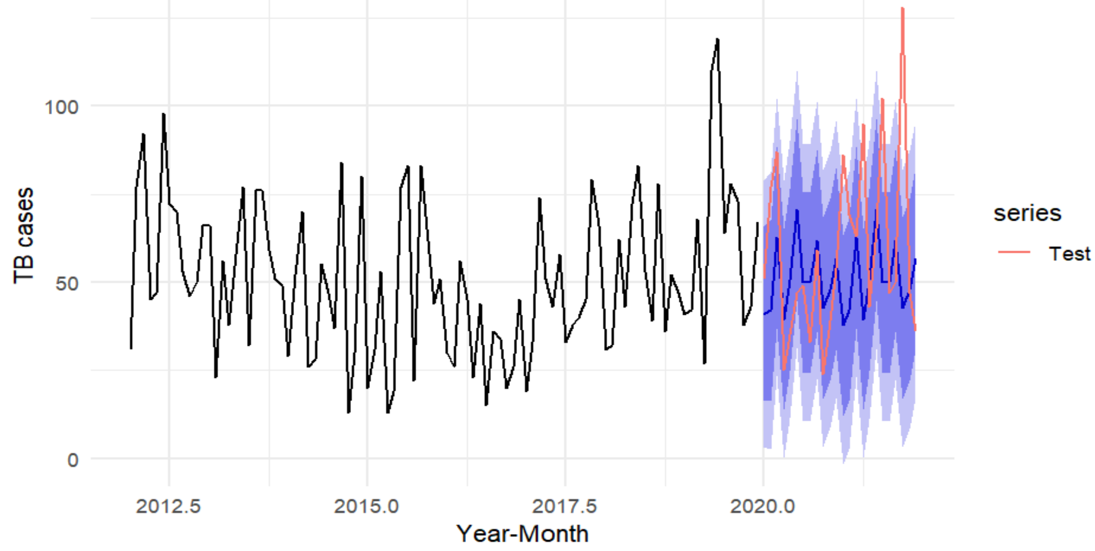


**Figure 4.8: Plot of ARIMA (0,0,1,1,0,1,12) forecasts compared to actual test data**

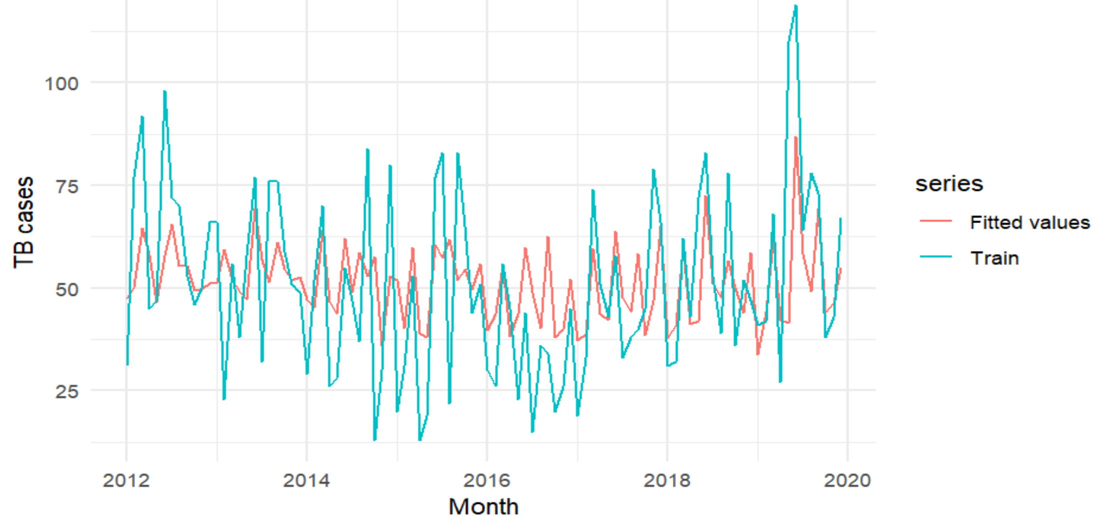


**Figure 4.9: Plot of ARIMA (0,0,1,1,0,1,12) model fitted TB and actual TB cases**

Table 4.4: ARIMA model accuracy

| Data | RMSE | MAE | MAPE |
|---|---|---|---|
| Training | 18.74 | 14.39 | 39.00 |
| Testing | 29.17 | 20.47 | 33.15 |

### 4.3.2 Artificial Neural Network (ANN) Model Fitting

The Neural Network Auto-Regressive (NNAR) function of R's 'nnetar' package was used to fit the training data using within the ANN model structure. The 'nnetar' package was used to provide fully automated model definition, which allows for the automatic and optimal selection of the lag parameter (p) and the number of nodes (k) within the hidden layer. NNAR (1, 1, 2)[12] was the best NNAR model, producing an average of 20 networks each of them being a 2-2-1 network with 9 weights. Figure 4.10 depicts comparison of the forecasted values from the NNAR model and the actual training data. The RMSE was 18.56 and 28.65 on the training and testing data respectively (Table 4.4). Furthermore, the NNAR (1, 1, 2)[12] mean number of TB cases was 50 from 2012 to 2019, compared to the actual mean number of 51 from the training data. Further, comparison between the 24 month forecasted TB cases showed a mean of 57 TB cases over the period 2020 to 2021 compared to 59 TB cases from the actual testing data over the same period. Generally, the NNAR (1, 1, 2)[12] model produces predictions and forecast values that are almost similar to the actual values.

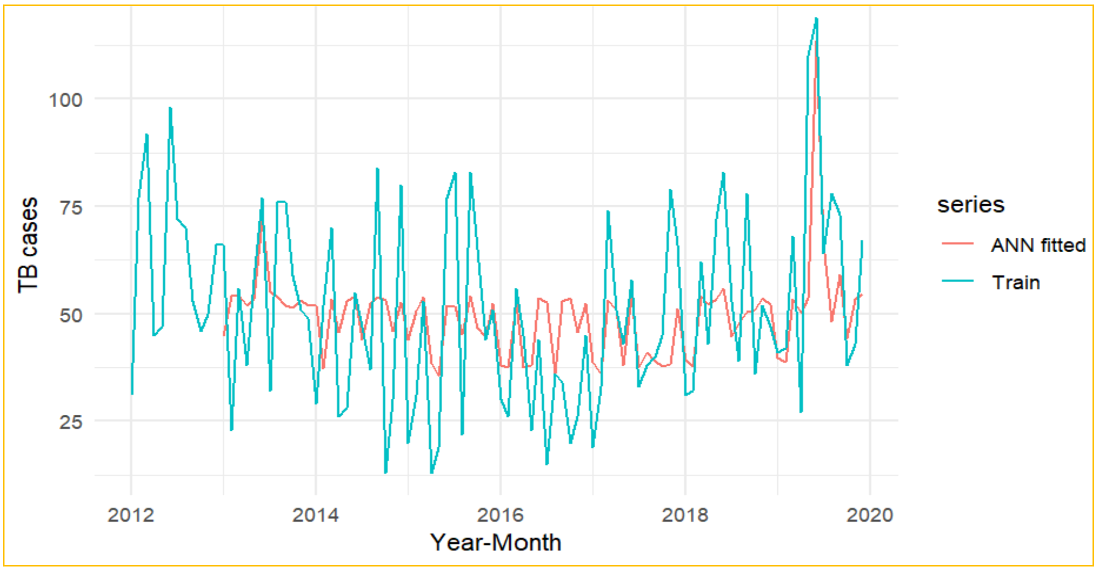


**Figure 4.10: Plot of NNAR (1, 1, 2)[12] fitted values against training data**

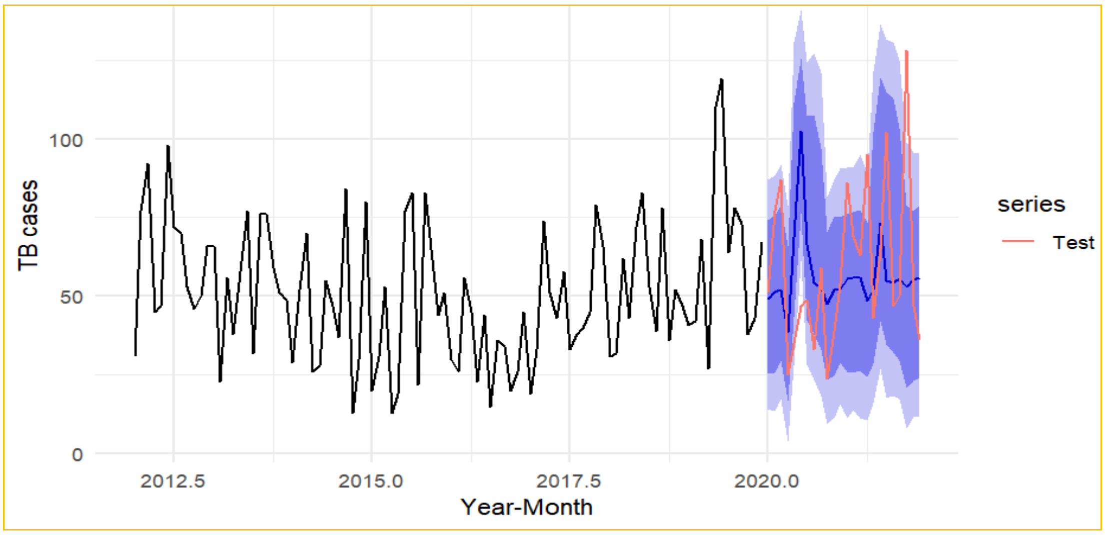


**Figure 4.11: NNAR model 24 month predicted TB cases**

**Table 4.5: Accuracy comparison of the NNAR model**

| Data | RMSE | MAE | MAPE |
|---|---|---|---|
| Training | 18.56 | 14.58 | 29.89 |
| Testing | 28.65 | 21.95 | 38.86 |

### 4.3.3 Hybrid Model Fitting

The ARIMA model has been shown to fail to account for linearities in the data while the ANN model fails to adequately model linearities existing in the data in addition to there not being a formal way of establishing model parameters. As such, most often, the residuals of ARIMA model contain remaining signal and white noise even when such a model fits well. In such a case, there is need to account for such signal from the residuals by modelling them separately in order to extract signal contained to an extent that the resulting residuals would be entirely white noise. In any case that the residuals of the ARIMA model contain both signal and white noise, there is need to model and extract the signal using the ANN model. The residuals were modelled independently in order to allow selection of an appropriate/optimal residual ANN model which again would capture the signal and retain the noise. This is the essence of ARIMA model hybridization, and the expectation is that the resulting hybrid model will improve the accuracy of the forecasts.

The best ARIMA model residuals were fitted using an ANN model. As indicated in 4.3.2, the ANN model on the ARIMA (0,0,1,1,0,1,12) was defined. The neural network autoregressive ('nnetar'()) function was applied to the residuals in this investigation. The residual diagnostics are shown in Figure 4.7, and while the residual mean is close to zero, there are worries regarding the auto-correlations at lag 3.

The RMSE was 19.08 and 27.61 on the training and testing data respectively (Table 4.6). Furthermore, the mean number of TB cases from the ARIMA-ANN model was 51 cases compared to a mean of 51 TB cases from the actual training data. In addition, comparison of the testing data against the forecasted TB cases from the ARIMA-ANN model show a mean of 52 cases from the ARIMA-ANN model against 59 cases from the actual testing data. Figure 4.12 and figure 4.13 present these comparisons.

**Table 4.6: Hybrid model accuracy**

| Data | RMSE | MAE | MAPE |
|---|---|---|---|
| Training | 19.08 | 15.32 | 42.42 |
| Testing | 27.61 | 19.69 | 32.89 |

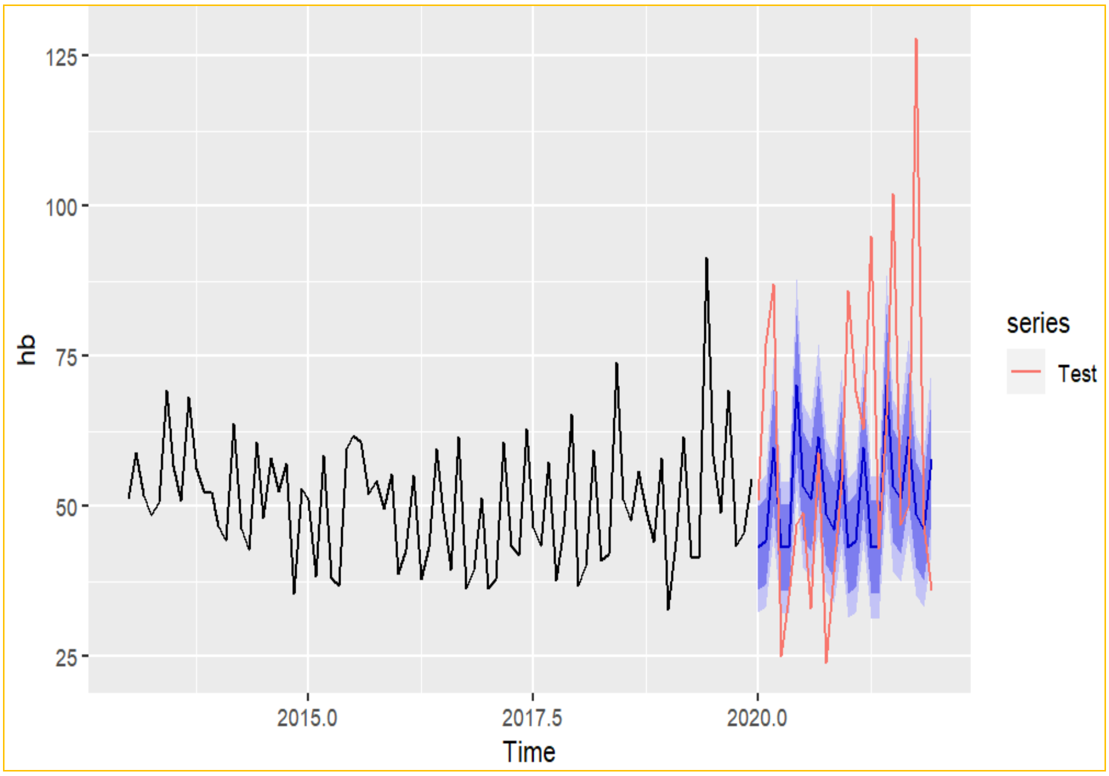


**Figure 4.12: Plot of Hybrid model forecasted TB cases against actual testing data**

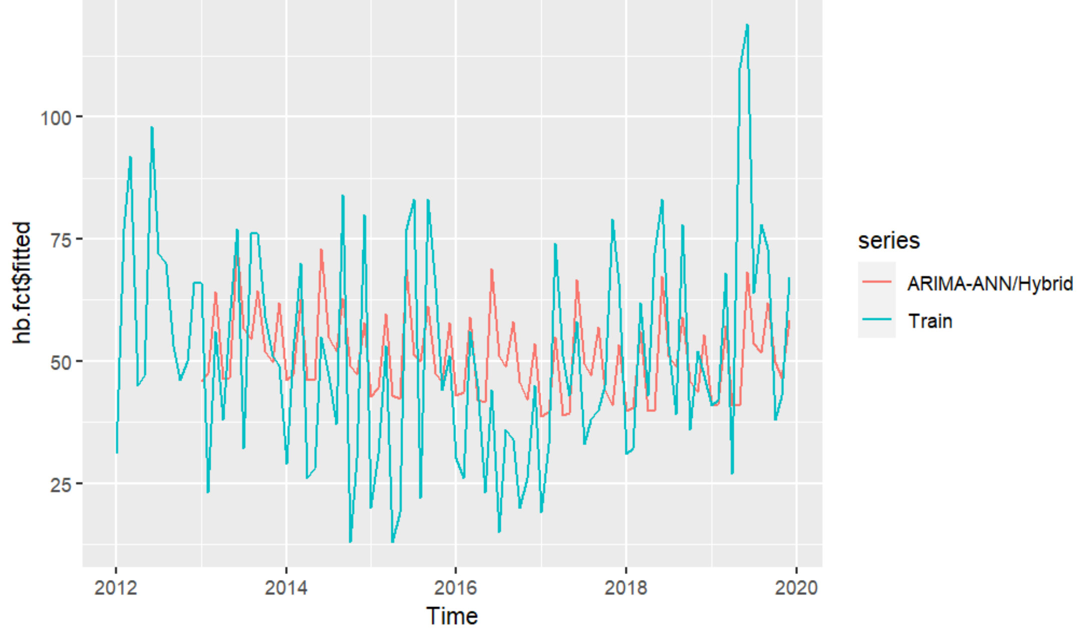


**Figure 4.13: Plot of Hybrid model  fitted TB cases against actual training data**

## 4.4 Accuracy comparison based on different parameter specifications

Table 4.7 presents the model accuracy measures compared across the different models on the training and testing data. This enables the identification of the best performing model among the three models. As a visual evaluation of model performance, Figure 4.14 shows the comparison of the predicted and actual values.

Table 4.7 shows that, while the three models perform nearly identically on training data, the ARIMA, NNAR and ARIMA-ANN models presented RMSE values of 18.74, 18.56, and 19.08, respectively, the ARIMA-ANN model performs better than the other two models on the test data with the lowers RMSE of 27.61. Moreover, the ARIMA-ANN model presents the lowest MAPE of 32.89 and the lowest MAE of 19.69 on the testing data. On the other hand, while the NNAR (1,1,2)[12] performs better on the training dataset compared to the other models, it performs worse than the ARIMA-ANN model on the testing. Because the neural network model is utilized to simulate the ARIMA residuals in order to extract signal, the hybrid model performs better on the testing dataset.

A visual inspection of model performance on actual training data, figure 4.14 confirms the results in table 4.7 where the hybrid plot is close to the actual training data with an almost similar trend compared to the other two models. On the training dataset, the NNAR and hybrid models outperform the ARIMA model, but on the testing dataset, the hybrid model beats the other models.

**Table 4.7: Model accuracy comparison**

| Data | Accuracy measure | ARIMA | NNAR | Hybrid |
|---|---|---|---|---|
| **Training** | RMSE | 18.74 | 18.56 | 19.08 |
| | MAE | 14.39 | 14.58 | 15.32 |
| | MAPE | 39.00 | 29.89 | 42.42 |
| **Testing** | RMSE | 29.17 | 28.65 | 27.61 |
| | MAE | 20.47 | 21.95 | 19.69 |
| | MAPE | 33.15 | 38.86 | 32.89 |

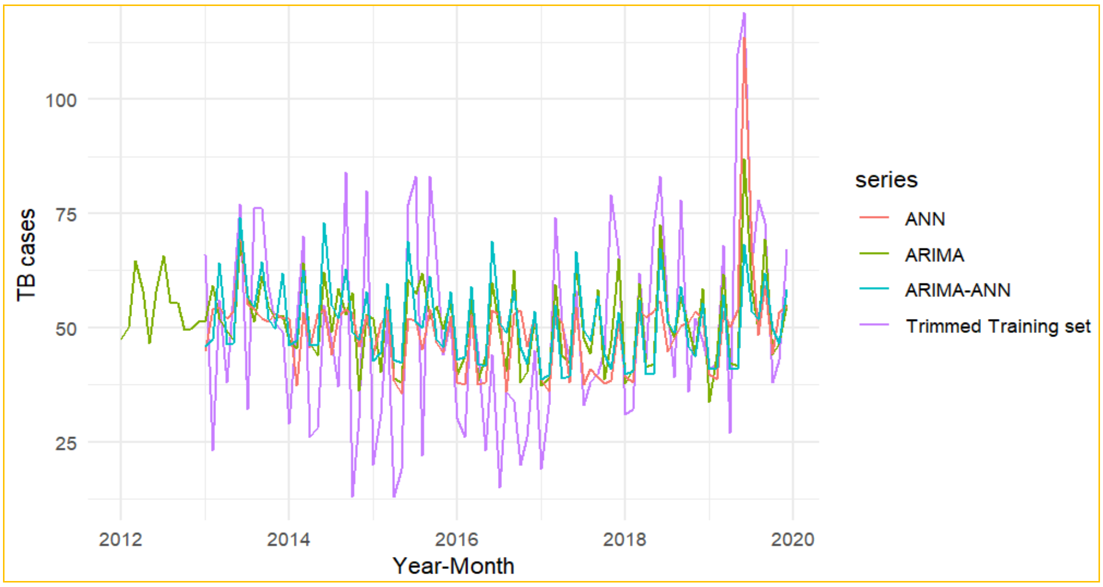


**Figure 4.14: Plot of ARIMA, ANN, and ARIMA-ANN model fitted TB against actual TB cases**

The models' predictive performance was also examined using the DM test. Table 4.8 shows that the prediction accuracies between the NNAR and ARIMA models were not statistically different, p=0.466. On the other hand, the findings show that the NNAR and ARIMA models each present with significantly different prediction accuracies compared to the hybrid model, p<0.001 respectively. These findings confirm that the hybrid ARIMA model is superior in terms of producing better prediction and forecast accuracies compared to the single ARIMA and NNAR models respectively.

**Table 4.8: Predictive accuracy comparison**

| **Model** | **DM statistic** | **Loss Function Power** | **P-value** |
|---|---|---|---|
| ARIMA Vs NNAR | 0.732 | 2 | 0.466 |
| NNAR Vs Hybrid | 6.260 | 2 | <0.001 |
| ARIMA Vs Hybrid | 8.732 | 2 | <0.001 |

### 4.5 Performance comparison of temporal forecast of TB trends

ARIMA (0,0,1,1,0,1,12), NNAR (1,1,2)[12], and ARIMA-ANN models were developed and utilized to predict TB cases for the following 12 months. The predicted TB cases are shown in figures 4.15, 4.16, and 4.17, as well as table 4.9. For the year 2022, the ARIMA (0,0,1,1,0,1,12), ANN (1,1,2)[12], and ARIMA-ANN models predicted 55, 59, and 52 TB cases per month in Turkana and Homa Bay Counties, respectively. The overall number of TB cases anticipated for Turkana and Homa Bay Counties is 657, 706 and 629 based on the ARIMA (0,0,1,1,0,1,12), ANN (1,1,2)[12], and ARIMA-ANN models, respectively. The monthly forecasts compare with actual 12-month TB cases reported for 2021 which total 664 with an approximate mean of 56 TB cases reported per month. Figure 4.18 shows the comparison of the 12-month forecasts from the 3 models.

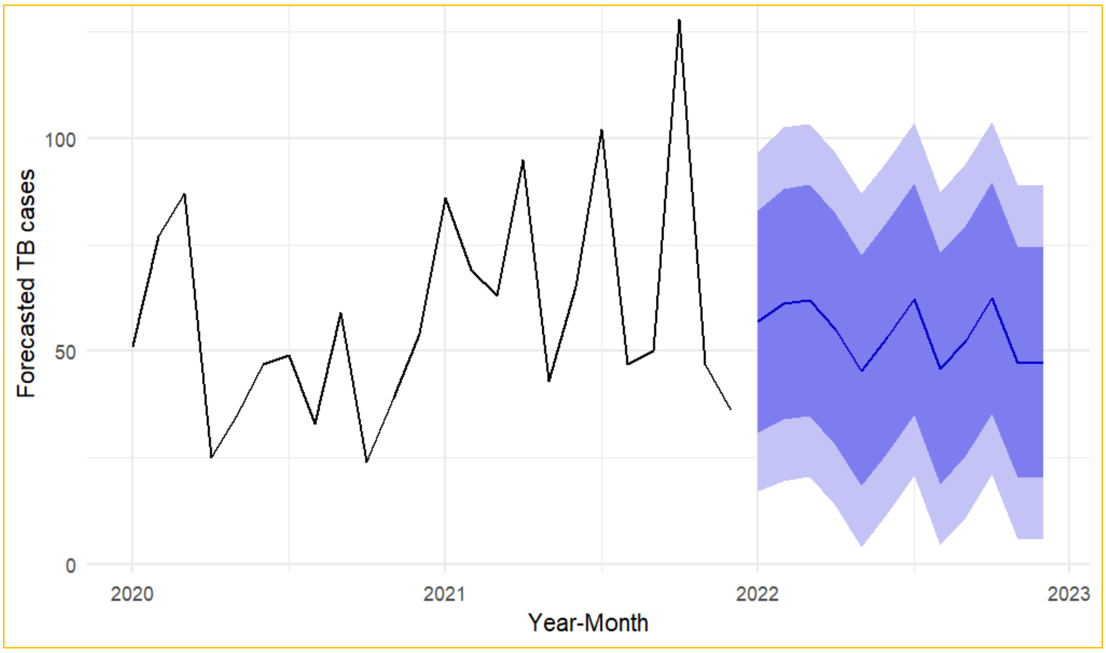


Figure 4.15: ARIMA (0,0,1,1,0,1,12) 12 month forecast of TB cases

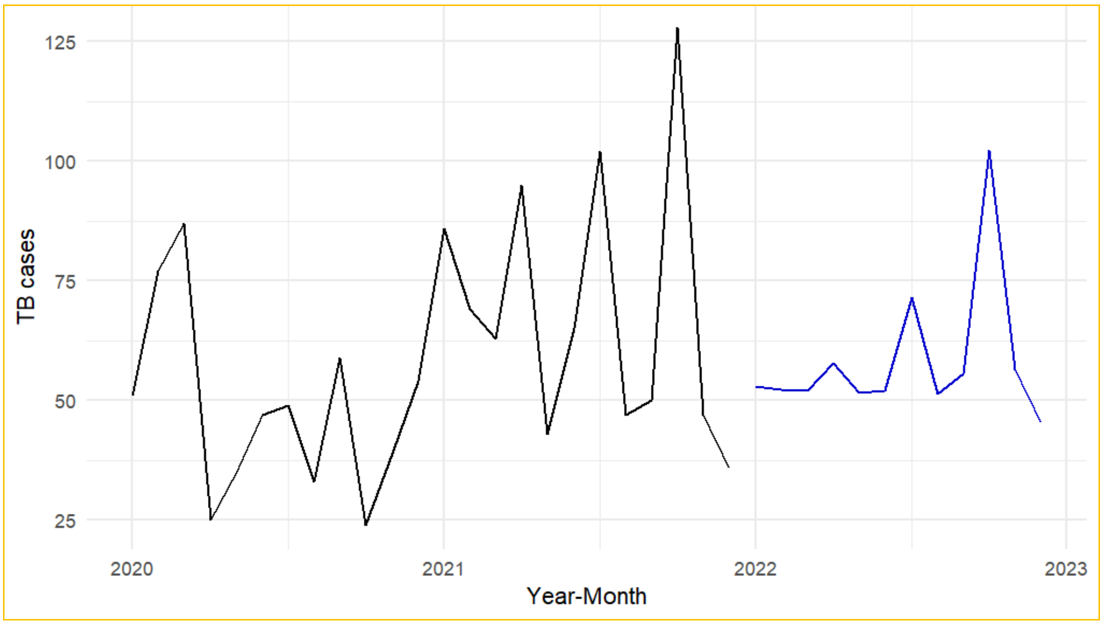


Figure 4.16: The NNAR (1,1,2)[12] model 12-month TB case prediction

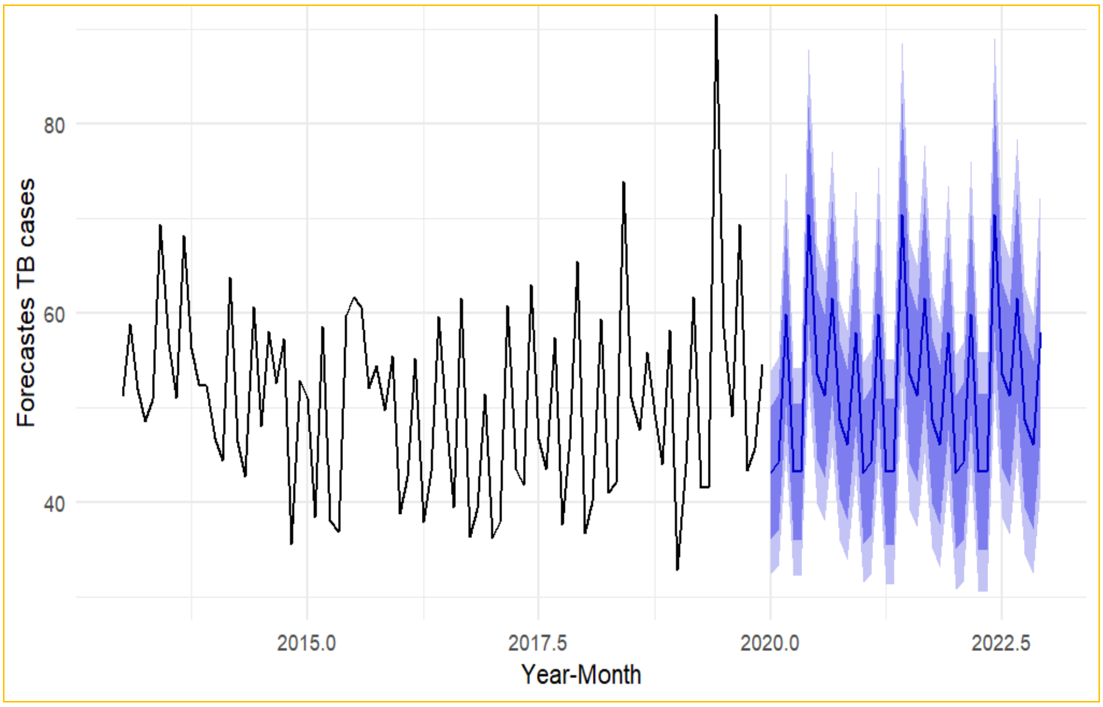


**Figure 4.17: ARIMA-ANN 12 month forecast of TB cases**

**Table 4.9: Comparison of TB case forecasts for the next 12 months**

| Month | ARIMA (0,0,1,1,0,1,12) | NNAR (1,1,2)[12] | ARIMA-ANN |
|---|---|---|---|
| Jan-22 | 57 | 53 | 44 |
| Feb-22 | 61 | 53 | 45 |
| Mar-22 | 62 | 52 | 60 |
| Apr-22 | 56 | 58 | 43 |
| May-22 | 46 | 52 | 44 |
| Jun-22 | 54 | 52 | 71 |
| Jul-22 | 63 | 72 | 54 |
| Aug-22 | 46 | 52 | 52 |
| Sep-22 | 53 | 56 | 62 |
| Oct-22 | 63 | 103 | 49 |
| Nov-22 | 48 | 57 | 46 |
| Dec-22 | 48 | 46 | 59 |
| **Mean** | **55** | **59** | **52** |
| **Total** | **657** | **706** | **629** |

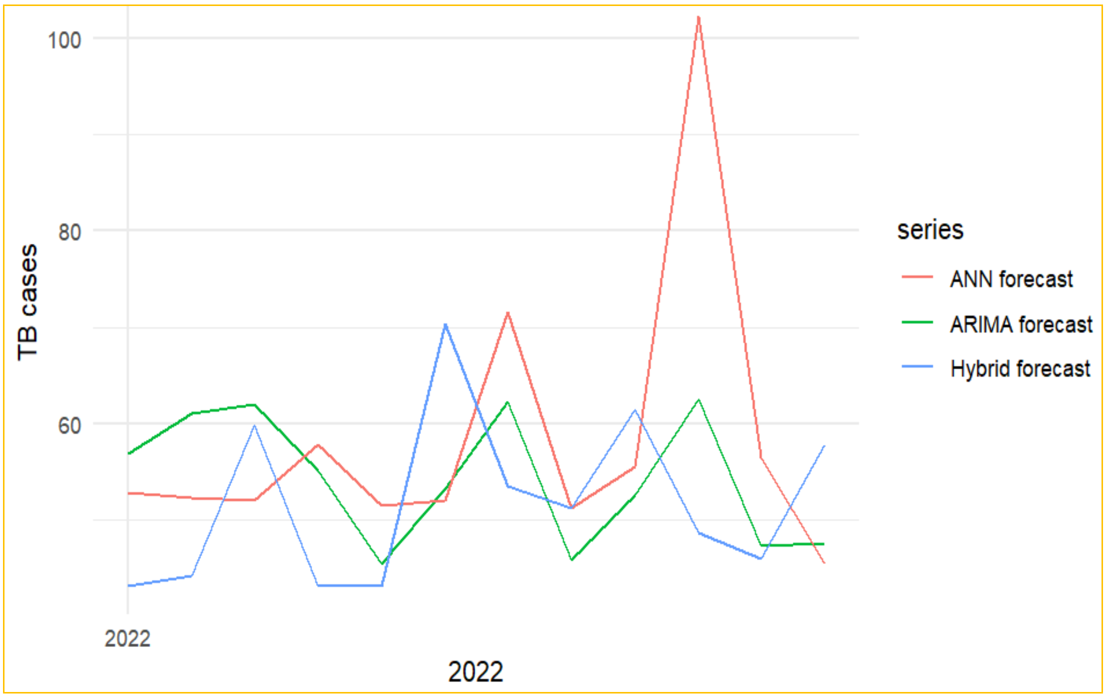


Figure 4.18: Point forecast comparison

# CHAPTER FIVE

# DISCUSSION

## 5.1 Introduction

The findings of the study are outline in the section. The findings of this study are compared against other findings in published literature especially in the tuberculosis as well as infectious disease domain. Each of the findings are discussed based on the stated specific objectives.

## 5.2 Performance comparison of predictive accuracy of models

There are very few studies in Sub-Saharan Africa that have explored or applied ANN or ARIMA-ANN models in predicting and forecasting TB cases in the general populations or sub-populations. As such, the findings from this study would aid in enforcing the need to explore these models towards understanding TB infections and other infectious diseases in general.

The results of this study show that, while all three models were capable of predicting TB cases among children under the age of 15, the hybrid model outperformed the single ARIMA and ANN models in terms of predictive and forecast accuracy. These findings are consistent with those of Azeez *et al.* (2016), who examined the prediction capacities of pure SARIMA and hybrid SARIMA models of TB incidence and discovered that the hybrid model performed better. More recently, Nyoni and Nyoni (2021) used a multilayer perceptron neural network to model and forecast tuberculosis occurrences in Bolivia, concluding that the model was trustworthy in projecting tuberculosis incidences in Bolivia with a predicted decreasing trend.

The results of this work on the prediction capability of hybrid ARIMA-ANN models are consistent with those by Zhou *et al.* (2016), who employed a hybrid ARIMA-ANN model to estimate schistosomiasis prevalence in people. When compared to the ARIMA and ANN models, they achieved fewer modeling and testing errors from the ARIMA-ANN model and proposed that the hybrid model can be used to surveillance data for early warning systems aimed at controlling and eliminating schistosomiasis illness.

Within other contexts outside infectious diseases, the hybrid model has been shown to produce better forecast accuracies when compared with single models. Zhou *et al.* (2018), who utilized the hybrid ARIMA-ANN model to examine the trends of new admission inpatients in order to provide a methodological foundation for minimizing congestion in health institutions. They concluded that although the hybrid model did not necessarily outperform the single ARIMA and ANN model performances, it was worth exploring using different set of data.

## 5.3 Accuracy comparison produced by different parameter specifications

Model accuracy measures were compared across the different models on the training and testing data to establish the model with the best predictive accuracy. The findings in this study revealed that although the three models provided almost the same predictive accuracy on the training data, the ARIMA-ANN model performed better on the testing data compared to the other two models. When the RMSE from the ARIMA-ANN was compared against the single ARIMA and NNAR models on the testing data, accuracy improved by 5.34% and 3.6% respectively.

The results of this study demonstrate that, while the three models perform nearly identically on training data, the hybrid model performs better on the test data. These findings are in line with those by Khashei and Bijari (2012), who indicated that the hybrid model yield better forecasts of infectious disease data.

Infectious disease data most often presents with both linear and nonlinear characteristics and single models would not adequately suffice in modelling such data. In line with the findings of this study, the use of hybrid models would be more effective in modelling such complex autocorrelation structures (Chakraborty *et al.*, 2021).

### 5.4 Model performance comparison in temporal forecast of TB trends

According to the conclusions of this study, for the year 2022, the study forecasts a mean of 52 to 59 TB cases per month among children under the age of 15 in Homa Bay and Turkana Counties compared to the mean of 49 cases per month in 2020 and 70 cases per month in 2021. This ideally confirms that TB cases among children below 15 years is under-reported. According to this analysis, there would be 624 to 708 TB cases recorded in 2022.

In 2019, the estimated population in Kenya was 52 million (KNBS, 2019) of whom about 43% (~22,360,000) (UN, 2017) were children below 15 years. Approximately 140,000 people were TB infected with TB of whom about 85,000 (61%) were reported to the NTP in 2019 (WHO, 2020). Consequently, basing on the estimated TB infections in 2019, the estimated TB incidence was about 269 TB cases per 100,000. Of the TB cases diagnosed and notified to the NTP, about 10-20% are children under 15 years (Dangisso, Datiko and Lindtjørn, 2015) and this represents about 14,000 to

28,000 children infected with TB in 2019. This translates to a TB case incidence of 63 to 125 cases per 100,000 children.

In Kenya, the forecasted TB incidence for 2020 was 259 cases per 100,000 population (WHO, 2020), translating to roughly 134,680 TB cases in 2020 of whom approximately 20% (26,936) are children (Okwara *et al*., 2017), equating to around 121 TB cases per 100,000 children in 2020. The findings of this study show a forecasted mean of 624 to 708 TB cases in 2022, and given that up to two-thirds of TB cases among children are not reported annually, (WHO, 2018), this translates to 1,782 to 2,023 forecasted TB cases in 2022 from this study.

In 2022, the estimated population of children under the age of 15 in Homa Bay and Turkana Counties is 1,020,795 people (U.S. Census Bureau, 2019 release). As a result, this study predicts 175 to 198 TB infections per 100,000 population among children in the two counties. These findings show and confirm that TB cases among children below 15 years are grossly under-reported.

This study's findings indicate that TB case notifications will most likely be greater in 2022 than in 2021. These are bleak findings that are consistent with the WHO (2021) newsletter, especially because of the COVID-19 pandemic. According to the newsletter, the COVID-19 pandemic has resulted in disruption in access and provision of TB services. Kenya recorded a 28% drop in TB diagnosis in 2020 (Mbithi *et al*., 2021), despite reporting that programmatic interventions during the COVID-19 period resulted in better case detection in the second six-months of the COVID-19 pandemic. The negative impact of the COVID-19 pandemic in terms of health care seeking behavior, and resource availability on TB detection and diagnosis has been documented in other Sub-Saharan Countries as well (Thekkur *et al.*, 2021) and this majorly attributed to the measures put in place by the country governments to curtail the spread of the COVID-19 disease including movement restrictions, conversion of health facilities to COVID-19 management units, re-allocation of resources from most in need public health areas, and even closure or restriction of key government services. In line with the findings of this study, Omondi *et al.* (2017), conducted a study in Kisii county among children under the age of 15 years and found that notification rates had decreased but it was unclear whether this was due to a decrease in TB cases or improved diagnostics.

While the global annual TB incidence rate reduced by 1.5% on average since 2000, the achievement of the End TB strategy demands a reduction by an average of 5% annually. The findings in this study show that there is a significant dent in the achievement of the End TB strategy owing partly to the COVID-19 pandemic and under-diagnosis of children in Kenya.

In addition, this study showed seasonal variations of reported cases among children with highest cases reported in the months of March, June, September and December on average. These findings support those of Azeez *et al.* (2016), Cao *et al.* (2013), Wah *et al.* (2014), and Gashu *et al.* (2018), who were able to demonstrate seasonal fluctuations in TB case notification but with regional variances owing to differences in weather patterns.

# CHAPTER SIX

# CONCLUSION AND RECOMMENDATIONS

## 6.1 Conclusion

Based on the results in chapter four and the discussion in chapter five, this study concludes that;

i. The hybrid model outperforms the standalone ARIMA and ANN models in terms of predictive and forecasting accuracy.

ii. More precise forecasts are produced by the hybrid model as measured by the RMSE on short-term 12 months forecasts.

iii. The forecasts from the hybrid model over a short-term 12-month period shows no increase or drop in the TB cases recorded and averaging between 52 and 59 TB cases notified per month in 2022.

## 6.2 Recommendations

This study recommends that;

The hybrid ARIMA models can be used in prediction and short-term forecast of TB cases reported among children below 15 years since it produces better predictive and forecast accuracy.

To develop more accurate models and forecasts, a substantial amount of data should be used to allow better learning of the neural network, particularly when ANN structure is used to model the data.

The findings in this study show that existing challenges in TB case notification among children below 15 years due to the mentioned potential factors. However, from the findings, if the forecasted trend is sustained, reaching the 50% TB infection reduction within the End TB strategy by 2025 will be challenging. This will be exacerbated more by the current COVID-19 pandemic. Eventually reaching the 2035 milestone of End TB strategy will face headwinds and achieving an annual 5% reduction in TB infections is bound to be challenging. Consequently, there is need to re-examine TB monitoring data in order to comprehend current gaps. In order to get the TB battle back on track, critical financial and non-financial resources must be reallocated to the TB program.

This study also offers interesting and promising recommendations for future researchers. There is need to explore the proposed hybrid model using sufficiently large amount of data to allow better learning and offer higher order autocorrelations in order to produce more accurate forecasts. In addition, while this study only explored models on a univariate dataset of TB case notification, there is need to explore inclusion of exogenous variables such as HIV infection numbers in order to further incorporate the relationship between TB and HIV infections which more often go hand-in-hand.

This study also showed and confirmed the presence of seasonal variations in pediatric TB cases reported. As such, interventions can be placed in to optimize TB screening and identification within the high and low seasons in order to increase diagnosis.

Finally, this study utilized the NNETAR structure of ANN. With a larger dataset, future research can utilize other structures of ANN such as multi-layer perceptron and recurrent neural networks utilizing either feed-forward mechanism or back-propagation mechanism in modelling while exploring inclusion of exogenous variables as well.

## 6.3 Limitations

Since the TIBU system was used to gather and report the data for this study, the study had no control over the data's correctness and quality.

Data from 2012 to 2021, divided into training and testing data, were used in this study. In order for deep learning algorithms like ANNs to effectively learn, a lot of data is typically required. In this study, 96 records out of the total 120 records were used as training data. Even though this accounted for 80% of the records, it could not have been enough. To partly overcome this, the study allowed the algorithm to automatically pick the lag order, set a defined decay parameter and a maximum iteration value. However, these actions would not have addressed all of the learning gaps. Furthermore, the testing data consisted of just 24 records, which was evidently insufficient to allow the algorithm to learn better. As such, use of more data as it becomes available can improve the model further as well as including data from more Counties that are TB endemic.

The data for Turkana and Homa Bay County were merged for this study. However, when it comes to pediatric TB and diagnosis processes, these two counties are quite distinct. As such, the number of TB cases reported might present differently when each county is considered separately.

In addition, the year 2019 to 2021 was compounded by the COVID-19 pandemic and this could have had an impact on TB diagnosis as well as management. However, this study could not quantify the COVID-19 impact on TB cases reported in the TIBU system among children below 15 years as this was beyond the scope of this study. A possible recommendation to such a scenario is to utilize models such as interrupted time series to measure possible impact of COVID-19 on TB detection, diagnosis and management.

# REFERENCES

Achieng E, Otieno V and Mung'atu J. (2020) Modeling the trend of reported malaria cases in Kisumu County, Kenya [version 1; peer review: 1 approved with reservations, 1 not approved]. *F1000Research* 2020, 9:600.

Ade, S., Békou, W., Adjobimey, M., Adjibode, O., Ade, G., Harries, A. D., & Anagonou, S. (2016). Tuberculosis case finding in Benin, 2000–2014 and beyond: a retrospective cohort and time series study. *Tuberculosis research and treatment*, *2016*.

Adhikari, R., & Agrawal, R. K. (2013). An introductory study on time series modeling and forecasting. *arXiv preprint arXiv:1302.6613*.

Aidoo, E. (2011). *Forecast performance between sarima and setar models: An application to Ghana inflation rate*. mathesis, Uppsala University.

Alene, K. A., Wangdi, K., & Clements, A. C. (2020). Impact of the COVID-19 pandemic on tuberculosis control: an overview. *Tropical medicine and infectious disease*, *5*(3), 123.

Anokye, R., Acheampong, E., Owusu, I., & Isaac Obeng, E. (2018). Time series analysis of malaria in Kumasi: Using ARIMA models to forecast future incidence. *Cogent social sciences*, *4*(1), 1461544.

Arltová, M., & Fedorová, D. (2016). Selection of unit root test on the basis of length of the time series and value of AR (1) parameter. *Statistika-Statistics and Economy Journal*, *96*(3), 47-64.

Aryee, G., Kwarteng, E., Essuman, R., Nkansa Agyei, A., Kudzawu, S., Djagbletey, R., ... & Forson, A. (2018). Estimating the incidence of tuberculosis cases reported at a tertiary hospital in Ghana: a time series model approach. *BMC Public Health*, *18*(1), 1-8.

Azeez, A., Obaromi, D., Odeyemi, A., Ndege, J., & Muntabayi, R. (2016). Seasonality and trend forecasting of tuberculosis prevalence data in Eastern Cape, South Africa, using a hybrid model. *International journal of environmental research and public health*, *13*(8), 757.

Bakar, N. A., & Rosbi, S. (2017). Data Clustering using Autoregressive Integrated Moving Average (ARIMA) model for Islamic Country Currency: An Econometrics method for Islamic Financial Engineering. *The International Journal of Engineering and Science* (IJES), 6(6), 22-31.

Box GE, Jenkins GM (1976). *Time Series Analysis, Forecasting and control*. Revised. Holden-Fay, San Francisco.

Brent, A. J. (2012). Childhood TB surveillance: bridging the knowledge gap to inform policy. *Journal of tropical medicine*, 2012.

Cao, S., Wang, F., Tam, W., Tse, L. A., Kim, J. H., Liu, J., & Lu, Z. (2013). A hybrid seasonal prediction model for tuberculosis incidence in China. *BMC medical informatics and decision making*, *13*(1), 1-7.

Cha J., Thwaites G.E., Ashton P.M. (2020). An Evaluation of Progress Towards the 2035 WHO End TB Targets in 40 High Burden Countries. *medRxiv*.10.02.20175307.

Chakrabarti, A., & Ghosh, J. K. (2011). AIC, BIC and recent advances in model selection. *Philosophy of statistics*, 583-605.

Chakraborty, T., Chakraborty, A. K., Biswas, M., Banerjee, S., & Bhattacharya, S. (2021). Unemployment rate forecasting: A hybrid approach. *Computational Economics*, *57*(1), 183-201.

Chatfield, C. (2000). *Time-series forecasting*. Chapman and Hall/CRC, USA.

Cilloni, L., Fu, H., Vesga, J. F., Dowdy, D., Pretorius, C., Ahmedov, S., ... & Arinaminpathy, N. (2020). The potential impact of the COVID-19 pandemic on the tuberculosis epidemic a modelling analysis. *EClinicalMedicine*, *28*, 100603.

Cinar, A. C. (2020). Training feed-forward multi-layer perceptron artificial neural networks with a tree-seed algorithm. *Arabian Journal for Science and Engineering*, *45*(12), 10915-10938.

Cochrane, John. (1997). *Time Series for Macroeconomics and Finance. Graduate School of Business*, University of Chicago, and spring.

Cowger, T. L., Wortham, J. M., & Burton, D. C. (2019). Epidemiology of tuberculosis among children and adolescents in the USA, 2007–17: an analysis of national surveillance data. *The Lancet Public Health*, 4(10), e506-e516.

Dangisso, M. H., Datiko, D. G., & Lindtjørn, B. (2015). Spatio-temporal analysis of smear-positive tuberculosis in the Sidama Zone, southern Ethiopia. *PloS one*, *10*(6), e0126369.

Darji, M. P., Dabhi, V. K., & Prajapati, H. B. (2015, March). Rainfall forecasting using neural network: A survey. In 2015 international conference on advances in computer engineering and applications (pp. 706-713). IEEE.

Dickey, D.A., Fuller, W.A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* 49 1057-1072.

Diebold, F.X. and R.S. Mariano. (1995). Comparing Predictive Accuracy. *Journal of Business and Economic Statistics*, 13: 253-63.

Dodd, P. J., Yuen, C. M., Sismanidis, C., Seddon, J. A., & Jenkins, H. E. (2017). The global burden of tuberculosis mortality in children: a mathematical modelling study. *The Lancet Global Health*, 5(9), e898-e906.

Ebhuoma, O., Gebreslasie, M., & Magubane, L. (2018). A seasonal autoregressive integrated moving average (SARIMA) forecasting model to predict monthly malaria cases in KwaZulu-Natal, *South Africa. South African medical journal*, *108*(7).

Floyd, K., Glaziou, P., Zumla, A., & Raviglione, M. (2018). The global tuberculosis epidemic and progress in care, prevention, and research: an overview in year 3 of the End TB era. *The Lancet Respiratory Medicine*, *6*(4), 299-314.

Frah, Ehab & Alkhalifa, Abdalla. (2016). *Tuberculosis Cases in Sudan; Forecasting Incidents 2014-2023 using Box & Jenkins ARIMA Model*. 2016. 108-114. 10.5923/j.ajms.20160603.04.

Garnett G., Cousens S., Hallett T., Steketee R. and Walker N. (2011). Mathematical models in the evaluation of health programmes, *Lancet* 378: 515–525.

Gashu, Z., Jerene, D., Datiko, D. G., Hiruy, N., Negash, S., Melkieneh, K., ... & Hadgu, A. (2018). Seasonal patterns of tuberculosis case notification in the tropics of Africa: a six-year trend analysis in Ethiopia. *PLoS One*, *13*(11), e0207552.

Hamzacebi C. (2008). “Improving artificial neural networks’ performance in seasonal time series forecasting”, *Information Sciences* 178, pages: 4550-4559.

Houben, R. M., Dowdy, D. W., Vassall, A., Cohen, T., Nicol, M. P., Granich, R. M., Shea, J. E., Eckhoff, P., Dye, C., Kimerling, M. E., White, R. G., & TB MAC TB-HIV meeting participants (2014). How can mathematical models advance tuberculosis control in high HIV prevalence settings? *The international journal of tuberculosis and lung disease: the official journal of the International Union against Tuberculosis and Lung Disease*, *18*(5), 509–514.

https://d-maps.com

Jaganath D, Wobudeya E, Sekadde MP, Nsangi B, Haq H, Cattamanchi A (2019). Seasonality of childhood tuberculosis cases in Kampala, Uganda, 2010-2015. *PLoS ONE* 14(4): e0214555.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning,*112 Springer.

Jenkins, H.E. (2016). Global burden of childhood tuberculosis. *Pneumonia* 8, 24.

Kenya National Bureau of Statistics. (2019). Kenya Population and Housing Census: Volume III.

Khashei, M., & Bijari, M. (2010). An artificial neural network (p, d, q) model for timeseries forecasting. *Expert Systems with applications*, *37*(1), 479-489.

Khashei, M., & Bijari, M. (2012). A new class of hybrid models for time series forecasting. *Expert Systems with Applications*, *39*(4), 4344-4357.

Kihoro, J., Otieno, R., & Wafula, C. (2006). Seasonal Time Series Forecasting: A Comparative Study of Arima and ANN Models. *African Journal of Science and Technology, 5*.

Kimani, E., Muhula, S., Kiptai, T., Orwa, J., Odero, T., & Gachuno, O. (2021). Factors influencing TB treatment interruption and treatment outcomes among patients in Kiambu County, 2016-2019. *Plos one*, *16*(4), e0248820.

Kipruto, H., Mung'atu, J., Ogila, K., Adem, A., Mwalili, S., Masini, E., & Kibuchi, E. (2015). The epidemiology of tuberculosis in Kenya, a high TB/HIV burden country (2000-2013). *International Journal of Public Health and Epidemiology Research*, *1*(1), 2-13.

Krauss M.R., Harris D.R., Abreu T., Ferreira F.G., Ruz N.P., Worrell C., *and Hazra R*. (2015). Tuberculosis in HIV-infected infants, children, and adolescents in Latin America. *Brazilian J Infect Dis*., 19 (1), pp. 23-29.

Larie, D., An, G., & Cockrell, R. C. (2021). The Use of Artificial Neural Networks to Forecast the Behavior of Agent-Based Models of Pathophysiology: An Example Utilizing an Agent-Based Model of Sepsis. *Frontiers in Physiology*, *12*.

Lee, J. (2018). Univariate time series modeling and forecasting (Box-Jenkins method), Econ 413, Lecture 4. *Department of Economics, University of Illinois*. Lewis, C. (1982). *International and Business Forecasting Methods*. London: Butterworths.

Li, Z., Wang, Z., Song, H., Liu, Q., He, B., Shi, P., Ji, Y., Xu, D., & Wang, J. (2019). Application of a hybrid model in predicting the incidence of tuberculosis in a Chinese population. *Infection and drug resistance*, *12*, 1011–1020.

Li, Z., Li, Y. (2020) A comparative study on the prediction of the BP artificial neural network model and the ARIMA model in the incidence of AIDS. *BMC Med Inform Decis Mak* 20, 143.

Lin Y-J, Liao C-M. (2013). Seasonal dynamics of tuberculosis epidemics and implications for multidrug-resistant infection risk assessment. *Epidemiol Infect*. 142:358–70.

Ljung, G.M. and Box, G.E.P. (1978). On a Measure of Lack of Fit in Time Series Models. *Biometrika* 65: 297–303.

Manikandan M, Velavan A, Singh Z, Purty AJ, Bazroy J, Kannan S. (2016) Forecasting the trend in cases of Ebola virus disease in West African countries using auto regressive integrated moving average models. Int J Community Med Public Health; 3:615-8.

Marais, B. J., Gie, R. P., Schaaf, H. S., Hesseling, A. C., Obihara, C. C., Nelson, L. J., Enarson, D. A., Donald, P. R. and Beyers, N. (2004). The clinical epidemiology of childhood pulmonary tuberculosis: a critical review of literature from the pre-chemotherapy era [state of the art]. *International Union Against Tuberculosis and Lung* Disease. 8(3):278–85.

Marais, B. J. (2011). Childhood tuberculosis: epidemiology and natural history of disease. *The Indian Journal of Pediatrics, 78*(3), 321-327.

Mbithi, I., Thekkur, P., Chakaya, J. M., Onyango, E., Owiti, P., Njeri, N. C., ... & Harries, A. D. (2021). Assessing the real-time impact of COVID-19 on TB and HIV services: the experience and response from selected health facilities in Nairobi, Kenya. *Tropical Medicine and Infectious Disease, 6*(2), 74.

Medar, R., Rajpurohit, V. S., & Rashmi, B. (2017). Impact of training and testing Data splits on accuracy of time series forecasting in Machine Learning. In *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)* (pp. 1-6). IEEE.

Ministry of Health, Government of Kenya, Division of Leprosy, Tuberculosis, and Lung Disease (2010). *National Monitoring and Evaluation Plan.* Government of Kenya

Ministry of Health, Government of Kenya, Division of Leprosy, Tuberculosis, and Lung Disease (2012). *Annual Report*.

Mwangwa F, Chamie G, Kwarisiima D, Ayieko J, Owaraganise A, Ruel T.D, Plenty A, Tram KH, Clark T.D, Cohen C.R, Bukusi EA, Petersen M, Kamya M.R, Charlebois E.D, Havlir D.V and Marquez C. (2017). Gaps in the Child Tuberculosis Care Cascade in 32 Rural Communities in Uganda and Kenya. *Journal of clinical tuberculosis and other mycobacterial diseases*.9:24–9. Epub 2018/01/02. pmid:29291251; PubMed Central PMCID: PMC5743212.

Newton, S. M., Brent, A. J., Anderson, S., Whittaker, E., and Kampmann, B. (2008). Paediatric tuberculosis. *Lancet Infect. Dis.* 8, 498–510. doi: 10.1016/S1473-3099(08)70182-70188

Nyoni, S. P., & Nyoni, T. (2021). Modeling and Forecasting TB Incidence in Bolivia Using the Multilayer Perceptron Neural Network. *International Research Journal of Innovations in Engineering and Technology*, *5*(3), 301.

Ojakaa, D., Olango, S., & Jarvis, J. (2014). Factors affecting motivation and retention of primary health care workers in three disparate regions in Kenya. *Human resources for health, 12*(1), 1-13.

Okwara, F. N., Oyore, J. P., Were, F. N., & Gwer, S. (2017). Correlates of isoniazid preventive therapy failure in child household contacts with infectious tuberculosis in high burden settings in Nairobi, Kenya–a cohort study. *BMC infectious diseases, 17*(1), 1-11.

Omondi, J., Kathure, E., Gachara, D., Kosgei, R. J., Wegunda, P., Magomere, R., ... & Omesa, E. (2017). Diagnostic methods and treatment outcomes for TB in children under 15 years in Kisii County, 2012-2016. *East African Medical Journal, 94*(10), S77-S89.

Otieno, E. J., & Okuku, M. T. (2017). Socio-Cultural Factors Contributing to the Spread of HIV and AIDs in Homa Bay County, Kenya. *Catholic University of Eastern Africa and Center for Democracy Research and Development*.

Padberg I., Batzing-Feigenbaum J., Sagebiel D. (2015). Association of extra-pulmonary tuberculosis with age, sex and season differs depending on the affected organ. *Int J Tuberc Lung Dis*.19(6):723–8. pmid:25946367

Pfaff, B. (2008). *Analysis of Integrated and Cointegrated Time Series with R*. 2Ed, Springer Science +Business Media, LLC, NY, USA.

Phillips, P. and Perron, P. (1988), "Testing for a unit root in time series regression", *Biometrika*, Vol. 75 No. 2, pp. 335-346.

Ren, H., Li, J., Yuan, Z. A., Hu, J. Y., Yu, Y., & Lu, Y. H. (2013). The development of a combined mathematical model to forecast the incidence of hepatitis E in Shanghai, China. *BMC infectious diseases*, *13*(1), 1-6.

Rono, V. K., & Migwambo, C. O. (2018). Socio-economic and demographic correlates of tuberculosis-related mortality in Homa Bay County, Kenya. *East African Medical Journal*, *95*(9), 1918-1926.

Sakamoto, Y., Ishiguro, M., and Kitagawa G. (1986). *Akaike Information Criterion Statistics*. D. Reidel Publishing Company.

Sakula, A. (1982). Robert Koch: centenary of the discovery of the tubercle bacillus, 1882. Thorax, 37(4), 246-251.

Sarpong, S. A. (2013). Modeling and forecasting maternal mortality; an application of ARIMA models. *International Journal of Applied Science and Technology*, 3(1), 19-28.

Schaaf, H. S., Michaelis, I. A., Richardson, M., Booysen, C. N., Gie, R. P., Warren, R., ... & Beyers, N. (2003). Adult-to-child transmission of tuberculosis: household or community contact? *The International Journal of Tuberculosis and Lung Disease*, *7*(5), 426-431.

Shimeles, E., Tilahun, M., Hailu, T., Enquselassie, F., Aseffa, A., Mekonnen, A., & Wondimagegn, G. (2019). Time interval for diagnosis of tuberculosis and related expenditure in selected health centers in Addis Ababa, Ethiopia. *Advances in Public Health*, *2019*.

Shrivastav, A. K., and Ekata. D. (2012). Applicability of Box Jenkins ARIMA Model in Crime Forecasting: A case study of counterfeiting in Gujarat State. *International Journal of Advanced Research in Computer Engineering and Technology*, 1(4), 2278, 1323.

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., ... & Collins, J. J. (2020). A deep learning approach to antibiotic discovery. *Cell*, *180*(4), 688-702.

Suyama, J., Sztajnkrycer, M., Lindsell, C., Otten, E. J., Daniels, J. M., & Kressel, A. B. (2003). Surveillance of infectious disease occurrences in the community: an analysis of symptom presentation in the emergency department. *Academic emergency medicine, 10*(7), 753-763.

Takele, R. (2020). Stochastic modelling for predicting COVID-19 prevalence in East Africa Countries. *Infectious Disease Modelling*, *5,* 598-607.

Taskaya-Temizel, T., & Ahmad, K. (2005). Are ARIMA neural network hybrids better than single models? In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* (Vol. 5, pp. 3192-3197). IEEE.

Tedijanto C, Hermans S, Cobelens F, Wood R, Andrews JR. (2018). Drivers of Seasonal Variation in Tuberculosis Incidence: Insights from a Systematic Review and Mathematical Model. *Epidemiology*. 2018;29(6):857–66. pmid:29870427

Theil, H. (1958). *Economic forecasts and policy*. Amsterdam: North-Holland Pub. Co.

Thekkur, P., Tweya, H., Phiri, S., Mpunga, J., Kalua, T., Kumar, A. M., ... & Harries, A. D. (2021). Assessing the impact of COVID-19 on TB and HIV programme services in selected health facilities in Lilongwe, Malawi: operational research in real time. *Tropical Medicine and Infectious Disease, 6*(2), 81.

https://www.census.gov/datatools/demo/idb/#/table?COUNTRY_YEAR=2022&COUNTRY_YR_ANIM=2022

Umunna, N. C., & Olanrewaju, S. O. (2020). Forecasting the Monthly Reported Cases of Human Immunodeficiency Virus (HIV) at Minna Niger State, Nigeria. *Open Journal of Statistics, 10*(3), 494-515.

United Nations, Department of Economic and Social Affairs, Population Division. (2017). World Population Prospects, the 2017 Revision, United Nations, New York.

Wah, W., Das, S., Earnest, A., Lim, L. K. Y., Chee, C. B. E., Cook, A. R., ... & Hsu, L. Y. (2014). Time series analysis of demographic and temporal trends of tuberculosis in Singapore. *BMC Public Health, 14*(1), 1-10.

Wang H, Tian CW, Wang WM, Luo XM (2018). Time-series analysis of tuberculosis from 2005 to 2017 in China. *Epidemiol Infect*.146:935–9.

Willis MD, Winston CA, Heilig CM, Cain KP, Walter ND, Mac Kenzie WR. (2012). Seasonality of tuberculosis in the United States, 1993 – 2008. *Clin Infect Dis*.54(11):1553–60.

World Health Organization (2014). Guidance for national tuberculosis programmes on the management of tuberculosis in children. *World Health Organization, Geneva, Switzerland.*

World Health Organization. (2015). Global tuberculosis report 2015, 20th ed. World Health Organization. https://apps.who.int/iris/handle/10665/191102

World Health Organization. (2016). Global tuberculosis report 2016. World Health Organization. https://apps.who.int/iris/handle/10665/250441

World Health Organization. (2017). Global tuberculosis report 2017. World Health Organization. https://apps.who.int/iris/handle/10665/259366.

World Health Organization. (2018). Global tuberculosis report 2018. World Health Organization. https://apps.who.int/iris/handle/10665/274453.

WHO, G. (2020). Global tuberculosis report 2020. *Glob. Tuberc. Rep*.

World Health Organisation (WHO). *Tuberculosis fact sheet*. (2018). Available on http://www.who.int/mediacentre/factsheets/fs104/en/.

Wubuli A, Li Y, Xue F, Yao X, Upur H, Wushouer Q. Seasonality of active tuberculosis notification from 2005 to 2014 in Xinjiang, China. *PLoS ONE*. 2017;12(7): e0180226. pmid:28678873

Xiao Y., He L., Chen Y., Wang Q., Meng Q., Chang W., Xiong L., and Yu Z. (2018). The influence of meteorological factors on tuberculosis incidence in Southwest China from 2006 to 2015. *Sci Rep* 8, 10053.

Yolcu, U., Egrioglu, E., & Aladag, C. H. (2013). A new linear & nonlinear artificial neural network model for time series forecasting. *Decision support systems*, *54*(3), 1340-1347.

Yu, L., Zhou, L., Tan, L., Jiang, H., Wang, Y., Wei, S., & Nie, S. (2014). Application of a new hybrid model with seasonal auto-regressive integrated moving average (ARIMA) and nonlinear auto-regressive neural network (NARNN) in forecasting incidence cases of HFMD in Shenzhen, China. *PloS one*, *9*(6), e98241.

Zhang G.P. (2007). "A neural network ensemble method with jittered training data for time series forecasting", *Information Sciences* 177, pages: 5329–5346.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* 50, 159–175.

Zhang, G. P., Patuwo, B. E., and Hu, M. Y. (1998). Forecasting with artificial neural networks: the state of the art. *Int. J. Forecast.* 14, 36–62.

Zhou, L., Xia, J., Yu, L., Wang, Y., Shi, Y., Cai, S., & Nie, S. (2016). Using a hybrid model to forecast the prevalence of schistosomiasis in humans. *International journal of environmental research and public health*, *13*(4), 355.

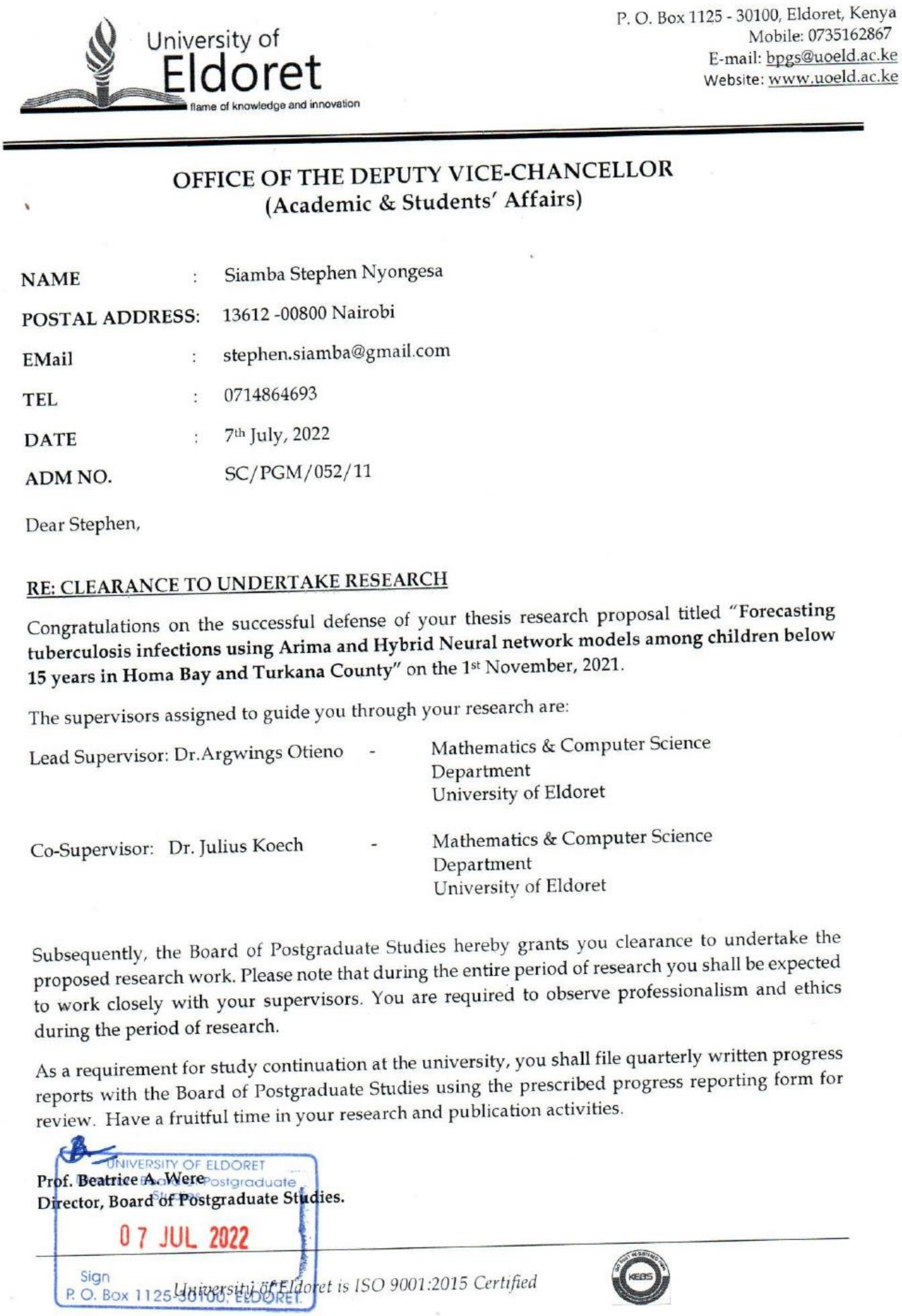
Zhou, L., Zhao, P., Wu, D., Cheng, C., & Huang, H. (2018). Time series model for forecasting the number of new admission inpatients. *BMC medical informatics and decision making*, *18*(1), 1-11.

Zeming L., Li, Yanning A. (2020) comparative study on the prediction of the BP artificial neural network model and the ARIMA model in the incidence of AIDS. *BMC Med Inform Decis Mak* 20**,** 143.

Zumla A, Petersen E, Nyirenda T, Chakaya J. (2015). Tackling the Tuberculosis Epidemic in sub-Saharan Africa – unique opportunities arising from the second European Developing Countries Clinical Trials Partnership (EDCTP) programme 2015-2024. *International Journal of Infectious Diseases*, Volume 32, 2015, Pages 46-49.

# APPENDICES

## APPENDIX I: University Clearance to undertake research



University of
Eldoret
flame of knowledge and innovation

P. O. Box 1125 - 30100, Eldoret, Kenya
Mobile: 0735162867
E-mail: bpgs@uoeld.ac.ke
Website: www.uoeld.ac.ke

**OFFICE OF THE DEPUTY VICE-CHANCELLOR**
**(Academic & Students' Affairs)**

**NAME** : Siamba Stephen Nyongesa
**POSTAL ADDRESS:** 13612 -00800 Nairobi
**EMail** : stephen.siamba@gmail.com
**TEL** : 0714864693
**DATE** : 7th July, 2022
**ADM NO.** SC/PGM/052/11

Dear Stephen,

**RE: CLEARANCE TO UNDERTAKE RESEARCH**

Congratulations on the successful defense of your thesis research proposal titled "**Forecasting tuberculosis infections using Arima and Hybrid Neural network models among children below 15 years in Homa Bay and Turkana County**" on the 1st November, 2021.

The supervisors assigned to guide you through your research are:

Lead Supervisor: Dr.Argwings Otieno - Mathematics & Computer Science Department
University of Eldoret

Co-Supervisor: Dr. Julius Koech - Mathematics & Computer Science Department
University of Eldoret

Subsequently, the Board of Postgraduate Studies hereby grants you clearance to undertake the proposed research work. Please note that during the entire period of research you shall be expected to work closely with your supervisors. You are required to observe professionalism and ethics during the period of research.

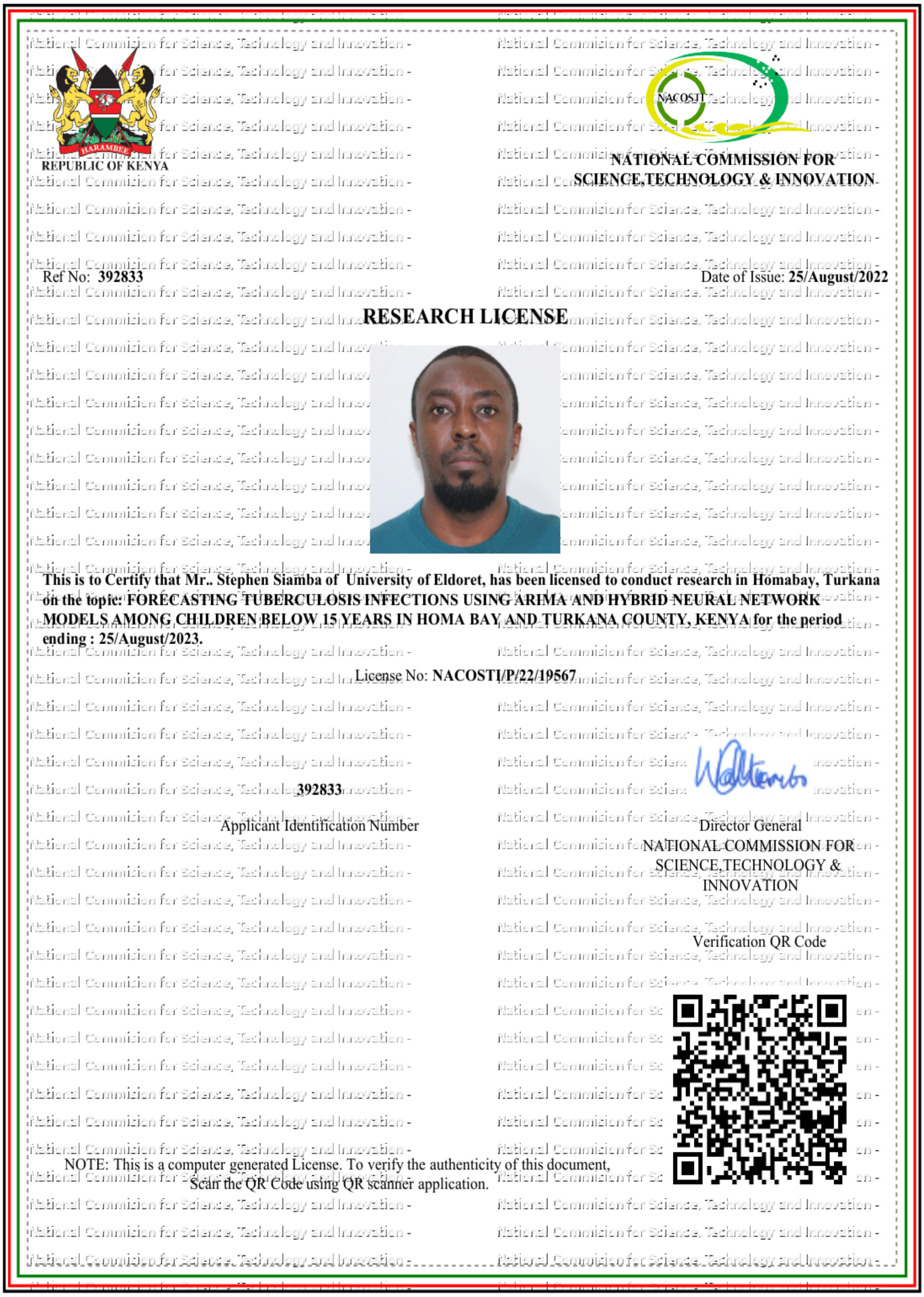
As a requirement for study continuation at the university, you shall file quarterly written progress reports with the Board of Postgraduate Studies using the prescribed progress reporting form for review. Have a fruitful time in your research and publication activities.

UNIVERSITY OF ELDORET
Board of Postgraduate Studies
07 JUL 2022
Sign
P. O. Box 1125-30100, ELDORET.

**Prof. Beatrice A. Were**
**Director, Board of Postgraduate Studies.**

*University of Eldoret is ISO 9001:2015 Certified*

## APPENDIX II: NACOSTI Research permit



REPUBLIC OF KENYA

NATIONAL COMMISSION FOR
SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **392833**

Date of Issue: **25/August/2022**

### RESEARCH LICENSE

**This is to Certify that Mr.. Stephen Siamba of University of Eldoret, has been licensed to conduct research in Homabay, Turkana on the topic: FORECASTING TUBERCULOSIS INFECTIONS USING ARIMA AND HYBRID NEURAL NETWORK MODELS AMONG CHILDREN BELOW 15 YEARS IN HOMA BAY AND TURKANA COUNTY, KENYA for the period ending : 25/August/2023.**

License No: **NACOSTI/P/22/19567**

**392833**

Applicant Identification Number

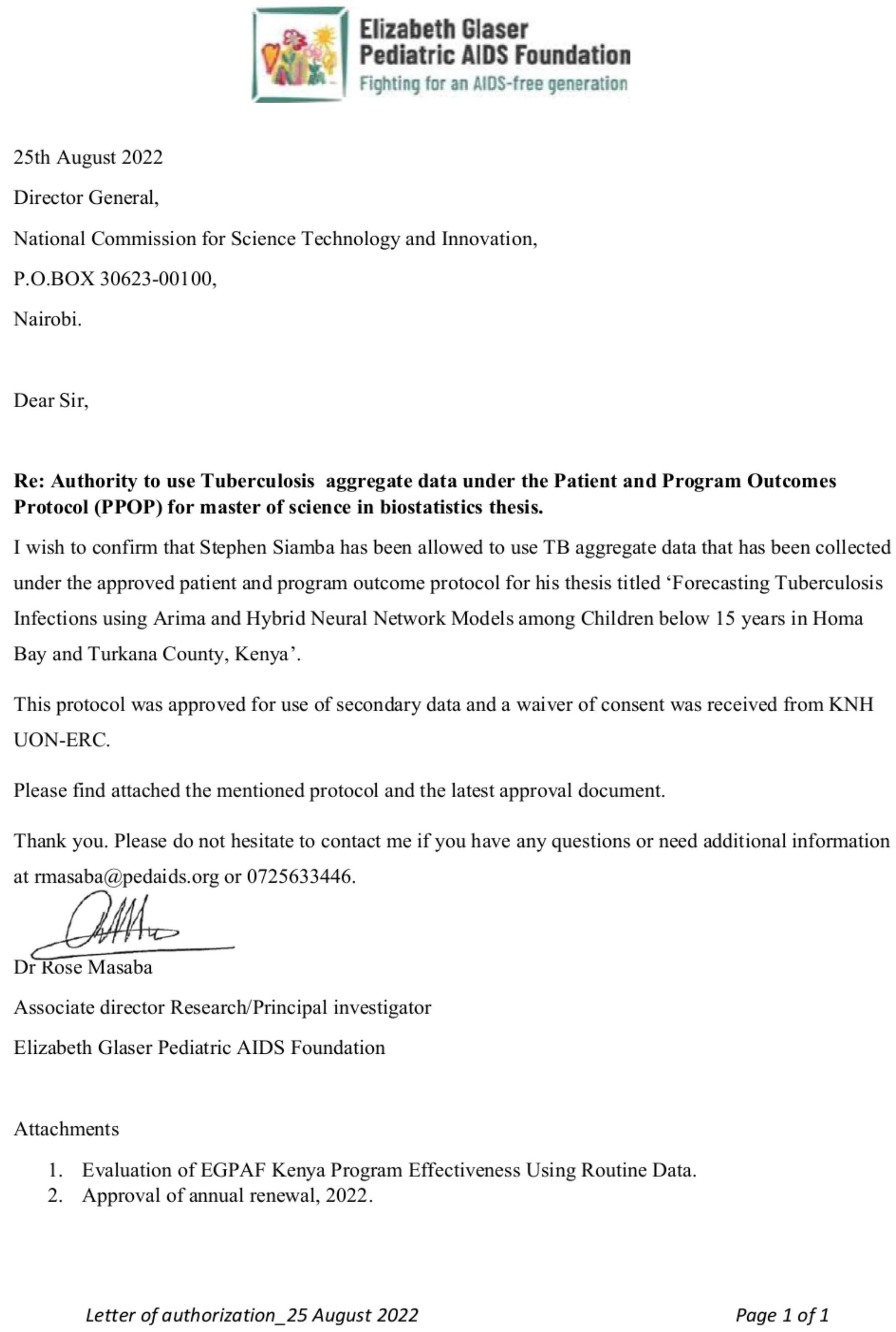
Director General
NATIONAL COMMISSION FOR
SCIENCE,TECHNOLOGY &
INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document,
Scan the QR Code using QR scanner application.

National Commision for Science, Technology and Innovation -

# APPENDIX III: Approval from Elizabeth Glaser Pediatric AIDS Foundation



Elizabeth Glaser
Pediatric AIDS Foundation
Fighting for an AIDS-free generation

25th August 2022

Director General,
National Commission for Science Technology and Innovation,
P.O.BOX 30623-00100,
Nairobi.

Dear Sir,

**Re: Authority to use Tuberculosis aggregate data under the Patient and Program Outcomes Protocol (PPOP) for master of science in biostatistics thesis.**

I wish to confirm that Stephen Siamba has been allowed to use TB aggregate data that has been collected under the approved patient and program outcome protocol for his thesis titled ‘Forecasting Tuberculosis Infections using Arima and Hybrid Neural Network Models among Children below 15 years in Homa Bay and Turkana County, Kenya’.

This protocol was approved for use of secondary data and a waiver of consent was received from KNH UON-ERC.

Please find attached the mentioned protocol and the latest approval document.

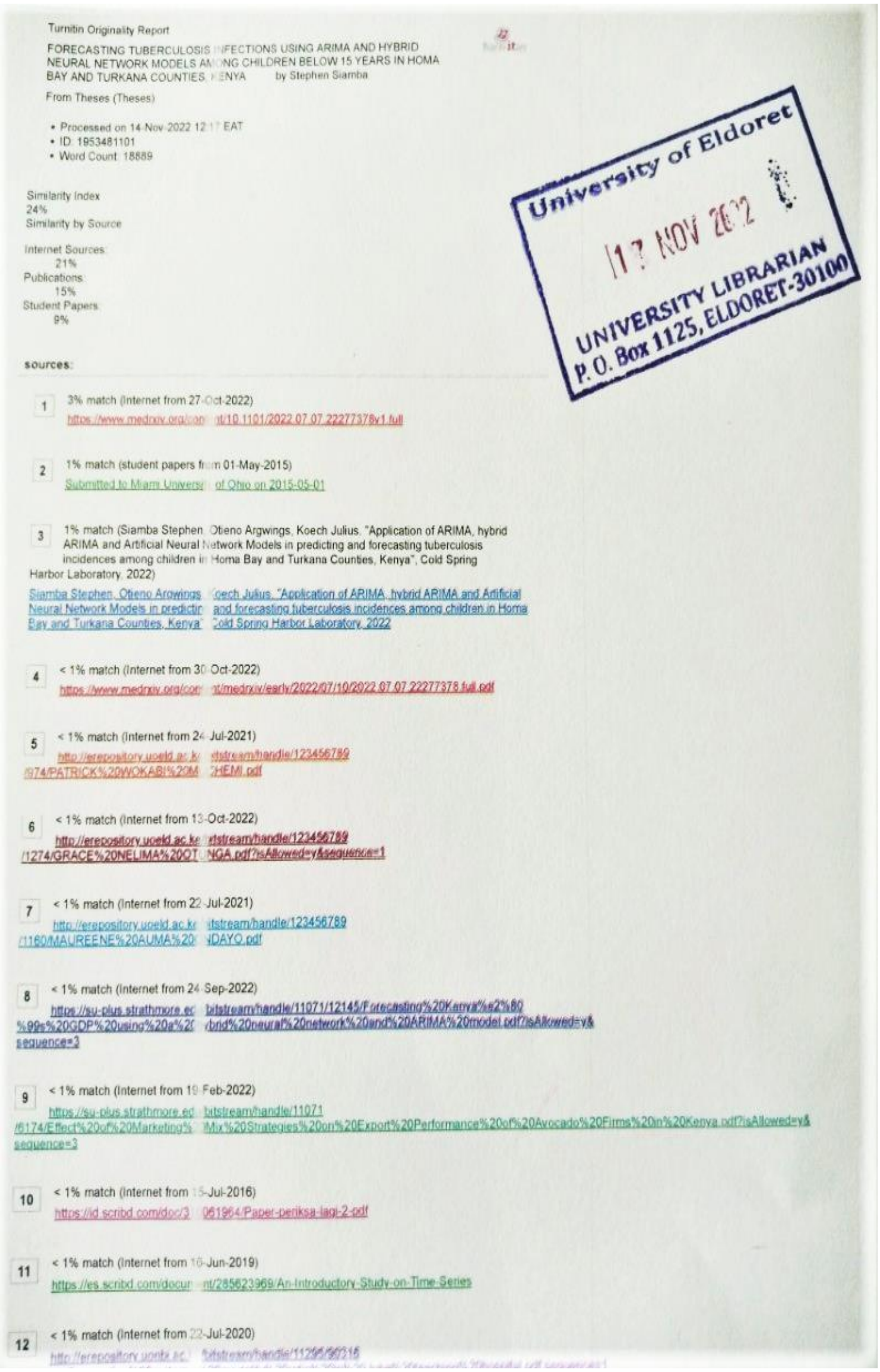
Thank you. Please do not hesitate to contact me if you have any questions or need additional information at rmasaba@pedaids.org or 0725633446.

Dr Rose Masaba
Associate director Research/Principal investigator
Elizabeth Glaser Pediatric AIDS Foundation

Attachments

1. Evaluation of EGPAF Kenya Program Effectiveness Using Routine Data.
2. Approval of annual renewal, 2022.

## APPENDIX IV: Similarity Report



Turnitin Originality Report

FORECASTING TUBERCULOSIS INFECTIONS USING ARIMA AND HYBRID NEURAL NETWORK MODELS AMONG CHILDREN BELOW 15 YEARS IN HOMA BAY AND TURKANA COUNTIES, KENYA by Stephen Siamba

From Theses (Theses)

- Processed on 14-Nov-2022 12:1 EAT
- ID: 1953481101
- Word Count: 18889

University of Eldoret
17 NOV 2022
UNIVERSITY LIBRARIAN
P. O. Box 1125, ELDORET-30100

Similarity Index
24%
Similarity by Source

Internet Sources:
21%
Publications:
15%
Student Papers:
9%

**sources:**

1 3% match (Internet from 27-Oct-2022)
https://www.medrxiv.org/con nt/10.1101/2022.07.07.22277378v1.full

2 1% match (student papers from 01-May-2015)
Submitted to Miami Univers of Ohio on 2015-05-01

3 1% match (Siamba Stephen, Otieno Argwings, Koech Julius. "Application of ARIMA, hybrid ARIMA and Artificial Neural Network Models in predicting and forecasting tuberculosis incidences among children in Homa Bay and Turkana Counties, Kenya", Cold Spring Harbor Laboratory, 2022)
Siamba Stephen, Otieno Argwings, Koech Julius. "Application of ARIMA, hybrid ARIMA and Artificial Neural Network Models in predictin and forecasting tuberculosis incidences among children in Homa Bay and Turkana Counties, Kenya", Cold Spring Harbor Laboratory, 2022

4 < 1% match (Internet from 30-Oct-2022)
https://www.medrxiv.org/con nt/medrxiv/early/2022/07/10/2022.07.07.22277378.full.pdf

5 < 1% match (Internet from 24-Jul-2021)
http://erepository.uoeld.ac.k itstream/handle/123456789/974/PATRICK%20WOKABI%20M HEMI.pdf

6 < 1% match (Internet from 13-Oct-2022)
http://erepository.uoeld.ac.ke itstream/handle/123456789/1274/GRACE%20NELIMA%20OT NGA.pdf?isAllowed=y&sequence=1

7 < 1% match (Internet from 22-Jul-2021)
http://erepository.uoeld.ac.ke itstream/handle/123456789/1160/MAUREENE%20AUMA%20 NDAYO.pdf

8 < 1% match (Internet from 24-Sep-2022)
https://su-plus.strathmore.ed bitstream/handle/11071/12145/Forecasting%20Kenya%e2%80%99s%20GDP%20using%20a%20 brid%20neural%20network%20and%20ARIMA%20model.pdf?isAllowed=y&sequence=3

9 < 1% match (Internet from 19-Feb-2022)
https://su-plus.strathmore.ed bitstream/handle/11071/6174/Effect%20of%20Marketing% Mix%20Strategies%20on%20Export%20Performance%20of%20Avocado%20Firms%20in%20Kenya.pdf?isAllowed=y&sequence=3

10 < 1% match (Internet from 15-Jul-2016)
https://id.scribd.com/doc/3 061964/Paper-periksa-lagi-2-pdf

11 < 1% match (Internet from 16-Jun-2019)
https://es.scribd.com/docum nt/285623969/An-Introductory-Study-on-Time-Series

12 < 1% match (Internet from 22-Jul-2020)
http://erepository.uonbi.ac. bitstream/handle/11295/90316
[illegible]