BINDURA UNIVERSITY OF SCIENCE EDUCATION

FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF STATISTICS AND MATHEMATICS

TIME SERIES FORECASTING OF PLASMODIUM FALCIPARUM MALARIA EPIDEMIC: A COMPARATIVE ANALYSIS OF ARIMA AND INTEGRATED ARTIFICIAL NEURAL NETWORKS - A CASE STUDY OF THE MINISTRY OF HEALTH AND CHILD-CARE, MT DARWIN DISTRICT.

BY

NYASHA DZAPASI (B213204B)

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR BSc.HONOURS IN STATISTICS AND FINANCIAL MATHEMATICS

SUPERVISOR: Mr. BASIRA

JUNE 2025

Authorship Declaration Statement

Title of the Thesis: Time Series Forecasting of Plasmodium Falciparum Malaria Epidemic: A Comparative Analysis of ARIMA and Integrated Artificial Neural Network - A Case Study of the Ministry of Health and Child-Care, Mt Darwin District.

Author: Nyasha Dzapasi

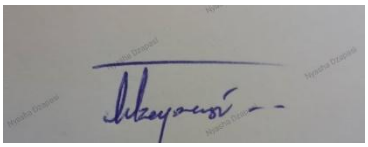
Program: Honors Bachelor of Science Degree in Statistics and Financial Mathematics

I, the undersigned author of the above-mentioned thesis, hereby declare that:

1. This thesis is my original work and has been prepared by me in accordance with the institution's requirements.
2. All sources, data, and references used in this thesis have been acknowledged and cited appropriately.
3. This thesis has not been submitted elsewhere for any degree or diploma.
4. I have obtained all necessary permissions for the inclusion of third-party content where applicable.

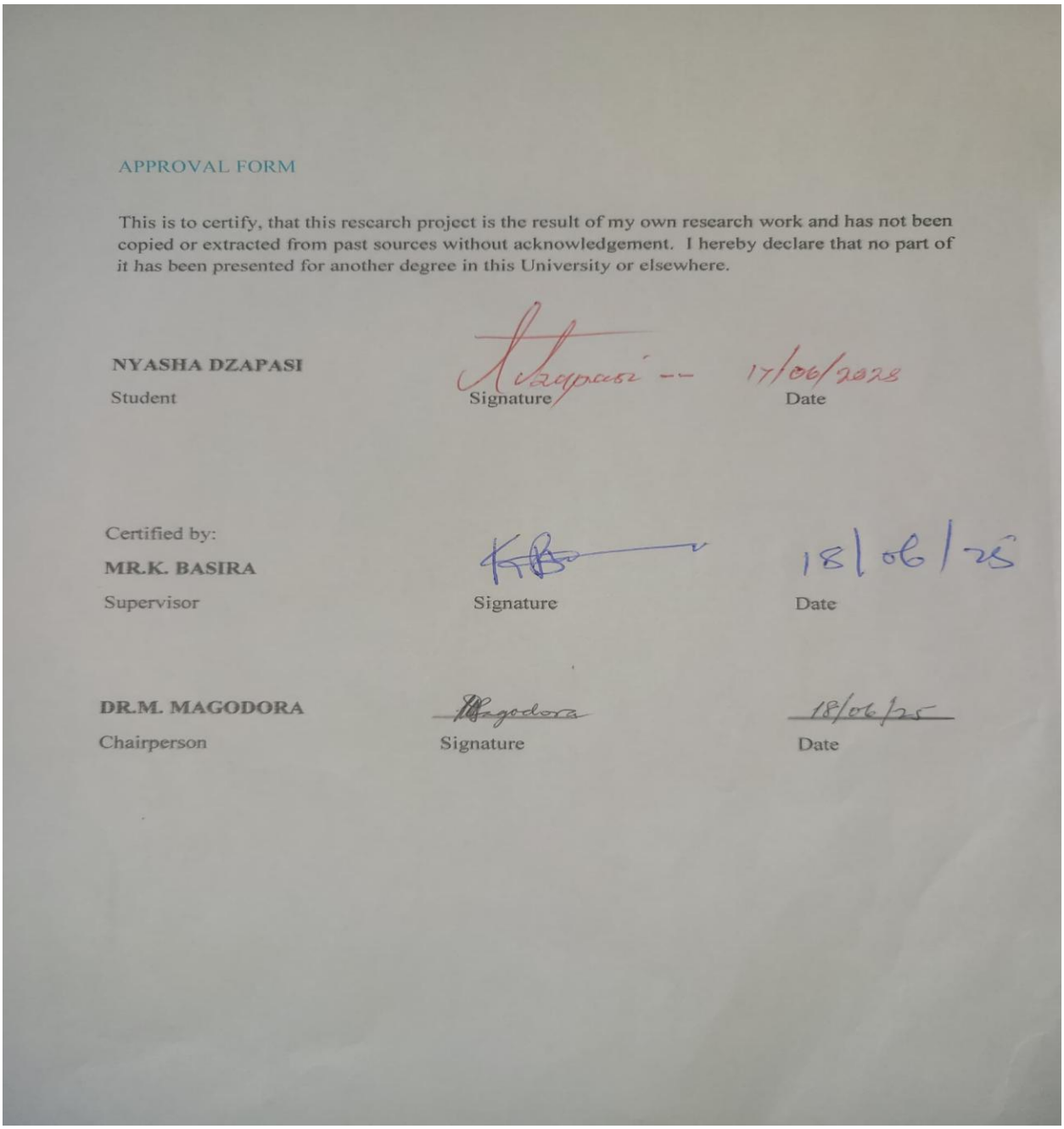
I affirm that this declaration is made with full integrity and in compliance with the institution's policies and academic practice.

Author's Signature:



Date: 14/06/2025

APPROVAL FORM



APPROVAL FORM

This is to certify, that this research project is the result of my own research work and has not been copied or extracted from past sources without acknowledgement. I hereby declare that no part of it has been presented for another degree in this University or elsewhere.

NYASHA DZAPASI — Student — Signature — 17/06/2025 — Date

Certified by:

MR.K. BASIRA — Supervisor — Signature — 18/06/25 — Date

DR.M. MAGODORA — Chairperson — Signature — 18/06/25 — Date

DEDICATION

This research project is respectfully dedicated to the memory of my beloved aunt, Ms. Sarah Mawonga, who tragically lost her life to malaria in 2010, and to the 338 lives lost during the devastating malaria epidemic that struck Zimbabwe between February and March of 2017. Their untimely deaths serve as a solemn reminder of the continued burden malaria imposes on families and communities across the nation. It is my hope that this study contributes meaningfully to the ongoing fight against malaria and supports and protect future generations from the pain and loss we have endured.

ACKNOWLEDGEMENTS

I extend my heartfelt gratitude to the professionals and experts in the field who generously shared their knowledge and expertise throughout the development of this dissertation. I am deeply grateful to my supervisor, Mr. K. Basira, for his invaluable guidance, unwavering support, and insightful feedback, which have been instrumental in refining my research. I also wish to express my sincere appreciation to Madam P. Hlupo, her coordination and continuous encouragement have made a significant impact on this journey. The Department of Statistics and Mathematics at Bindura University of Science Education has provided essential research facilities and the Mt Darwin District Hospital's administrative support, for which I am truly thankful. Furthermore, I am grateful to God for granting me strength, wisdom, and the opportunity to pursue my education at this level.

ABSTRACT

Malaria caused mainly by Plasmodium falciparum is still a significant public health problem in rural areas of Zimbabwe. This study develops and contrasts time series forecasting models to predict monthly malaria incidence and mortality in Mt Darwin District, using historical data from January 2013 to December 2023. Two model approaches were employed, a standard Autoregressive Integrated Moving Average (ARIMA) model and an integrated computational intelligence strategy incorporating Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Feedforward Neural Networks (FFNN). ARIMA model selection was consistent with the Box-Jenkins method, while the hybrid neural network was trained with a 12-month sliding input window. Model performance was assessed with a held-out 2024 test set with Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). Optimal ARIMA specifications (2,1,3) for incidence and (1,1,1) for mortality achieved moderate accuracy ($R^2 = 0.83$ for incidence), but unacceptability for mortality ($R^2 = -0.16$). The CNN+LSTM+FFNN hybrid model performed the best among all models with an MAE of 5.43, RMSE of 96.85, and $R^2 = 0.94$ for mortality, and an MAE of 2, RMSE of 2.16, and $R^2 = 0.91$ for incidence. 2025-2030 projections of malaria case declines from 1,824 in 2025 to 1,703 in 2030 and of deaths from 46 to 33 over the same period, with seasonal highs in February to April. These findings illustrate the strength of hybrid neural networks in modeling nonlinear, intricate patterns of disease in under researched environments. The study recommends that Mt Darwin District Health officials and the Ministry of Health and Child Care coordinate antimalarial procurement with NatPharm, augment bed-net and diagnostic kit distribution in high months, intensify targeted indoor spraying by community health workers, and improve reporting through DHIS2/Impilo systems.

Table of Contents

Chapter 1

1.0 INTRODUCTION

Malaria continues to pose significant health and economic challenges in many parts of the world, with Plasmodium falciparum accounting for the majority of severe cases in sub-Saharan Africa (World Health Organization, 2023). In Zimbabwe's Mt Darwin District, public health officials are tasked with not only responding to active outbreaks but also anticipating future transmission patterns to strengthen the efficiency of resource deployment and targeted interventions. As noted in recent studies, developing accurate models for forecasting malaria incidence and mortality is increasingly vital for shaping timely and effective public health strategies (Chikoko et al., 2021).

This research seeks to build a forecasting model tailored to predict both malaria incidence and mortality associated with Plasmodium falciparum in Mt Darwin District. By applying statistical and computational intelligence methods to historical health data, the study aims to generate evidence based future forecasts that can support decision making by the District Medical Officer (DMO), District Health Authorities (DHA), and the Ministry of Health and Child Care (MoHCC). These projections are intended to improve planning and prioritization of malaria control strategies, particularly in resource constrained rural settings. (Alhassan et al., 2017).

1.1 BACKGROUND OF THE STUDY

Malaria continues to be a major global health problem, putting about 3.3 billion people in 97 countries, including Zimbabwe, at risk. Each year, it causes around 200 million infections and about 600,000 deaths (World Health Organization, 2015). In Zimbabwe, Plasmodium falciparum is the most common cause of malaria, making it a serious public health issue. Studies show that malaria cases are affected by factors such as climate, economic conditions, and access to healthcare (Chikoko et al., 2021).

The District of Mt Darwin, which is found in Zimbabwe’s Mashonaland Central Province, is mostly rural and has different levels of healthcare access with 11 health clinic and 1 hospital. The district frequently experiences malaria outbreaks, especially during the rainy season when mosquito populations rise (Chikoko et al., 2021). Analyzing malaria trends in this area is important for planning effective disease control strategies. The Ministry of Health and Child Care (MOHCC) collects data on malaria cases and deaths using the DHIS2 and Impilo health information systems, making it possible to conduct a detailed analysis of malaria trends.

Although there have been studies on malaria patterns in Zimbabwe, very few have focused on Mt Darwin District. Research highlights the need of Mt Darwin oriented strategies due to the district’s

unique malaria trends. However, there has been little use of time series forecasting methods to predict malaria cases and deaths in the area. Using historical data to forecast malaria trends can help improve response strategies and allow for timely interventions to reduce infections and fatalities.

This study aims to address this gap by using historical malaria data from the MoHCC to build predictive models. By applying artificial intelligence-based time series forecasting, the research will provide useful insights to help improve malaria control efforts in Mt Darwin District. The findings will assist in better resource allocation and the development of targeted malaria prevention programs.

1.2 PROBLEM STATEMENT

Malaria continues to pose a significant public health challenge in Zimbabwe, particularly in the rural district of Mt Darwin. Despite ongoing efforts by the Ministry of Health and Child Care, the district frequently experiences seasonal outbreaks of plasmodium falciparum malaria often resulting in avoidable loss of life. This is because the district lacks predictive tools and predictive statistical insights into the epidemic. Without predictive insight, and planning for future outbreaks, medical supply distribution intervention strategies become fruitless rather than positive. This research therefore proposes the development of a forecasting model using both traditional time series methods such as the Autoregressive Integrated Moving Average (ARIMA) and computational intelligence driven models, specifically integrated artificial neural networks (IANNs) which will see a combination of Feedforward Neural Networks (FFNN) with Long Short Term Memory (LSTM) and Convolutional Neural Networks (CNN) architectures. The study aims to provide statistically driven insights into malaria epidemiology in Mt Darwin, ultimately enhancing the district's epidemic response systems and contributing to improved public health outcomes.

1.3 RESEARCH OBJECTIVES & QUESTIONS

1.3.0 OBJECTIVE

1. To build and apply an ARIMA models using the Box-Jenkins approach.
2. To build and apply IANN (integrated artificial neural networks) using computational intelligence techniques.
3. To compare the forecasting performance of ARIMA and IANN models using appropriate statistical performance metrics.

4. To forecast Plasmodium falciparum malaria incidence and mortality trends in Mt Darwin District for the period 2025 to 2030 using the best performing model.
5. To provide data-driven insights that support malaria surveillance and guide intervention planning for the District Health Administrator (DHA) and the Ministry of Health and Child Care (MoHCC)

1.3.1 QUESTIONS

1. Is there a statistical trend on plasmodium falciparum malaria cases and mortality from the historical data that can be mathematically computed?
2. (a) Can a time series model be developed to estimate future plasmodium falciparum cases incidence?
   (b) Can a time series model be developed to estimate future plasmodium falciparum malaria mortality?
3. How do seasonal variations affect cases and mortality incidence of plasmodium falciparum malaria in the district?
4. Is there a significant difference in forecasting accuracy of traditional time series models compared to computational intelligence driven integrated artificial neural network hybrid models in forecasting of the plasmodium falciparum epidemic?

1.4 SCOPE OF THE STUDY

This research aims to build a forecasting model tailored to predict trends in plasmodium falciparum malaria incidence and related mortality in Mt Darwin District. It utilizes historical records extracted from the Ministry of Health and Child Care's DHIS2 and Impilo health systems, covering previous reporting periods and extending projections through to 2030. The study compares classical ARIMA models with integrated artificial neural network (IANN) approaches to determine which model best captures temporal disease patterns. Although the findings will support evidence based planning for the District Health Authorities (DHA), District Medical Officers (DMO), and the MoHCC, they are designed specifically for Mt Darwin and may not generalize to other settings.

1.5 SIGNIFICANCE OF THE STUDY

The research study seeks to offer new insight into the patterns and trends of plasmodium falciparum malaria incidence and mortality in Mt Darwin District, a region that has received limited attention in previous forecasting research. By applying both traditional statistical models and hybrid computational intelligence approaches, the project aims to support the District Health Administrator and District Medical Officer with tools that can strengthen early detection and guide health planning. Reliable projections can improve the timeliness and efficiency of malaria response strategies in this high burden setting.

In Mt Darwin, where seasonal outbreaks often strain limited health resources, enhancing predictive capacity is essential. This research addresses a practical need for forward looking planning tools in epidemic management. At the national level, the research study supports Zimbabwe's broader health strategy by promoting the integration of local data and forecasting in policy and resource allocation. Its approach to combining computational models with epidemiological data may also offer lessons for other rural malaria endemic regions and contribute to ongoing global efforts toward malaria control and elimination.

Aligned with education 5.0 and the Heritage based curriculum, the study reflects innovation driven problem solving, fostering the integration of cutting edge technologies such as computational intelligence into real world socio-economic challenges faced by Zimbabwe. It also contributes to the academic fields of public health informatics, application of computational intelligence and statistics, offering a valuable case study for future interdisciplinary research in low resource settings.

1.6 ASSUMPTIONS OF THE STUDY

- The study assumes that the secondary data obtained from the DHIS2 and Impilo systems is complete, reliable, and free from reporting errors.
- It is assumed that the time series data used in the analysis does not exhibit autocorrelation that would violate model assumptions.
- No major structural changes such as new variant outbreaks and health system disruptions occurred during the study period.
- The malaria incidence and mortality data are considered homogeneous across Mt Darwin District.

1.7 LIMITATIONS OF THE STUDY

- This research is limited to time series forecasting methods and does not incorporate alternative statistical or machine learning techniques.
- The analysis is restricted to Mt Darwin District therefore; the results may not be generalizable to other regions of Zimbabwe.
- The forecasting model is built only on historical time series data, excluding explanatory variables such as rainfall, temperature, and population mobility which are known to influence malaria dynamics.
- The research study does not account for possible data incompleteness in the DHIS2 and Impilo datasets which could impact model accuracy.

1.8 DEFINITION OF KEY TERMS

1. Malaria: Malaria is a potentially fatal illness resulting from infection by Plasmodium parasites, which are transmitted to humans through bites from infected Anopheles mosquitoes (WHO, 2023).

2. Plasmodium falciparum is the most lethal of the five malaria parasite species known to infect humans. It is highly prevalent in sub-Saharan Africa and accounts for the largest proportion of malaria-related fatalities worldwide (Centers for Disease Control and Prevention, 2023).

3. Time Series: A sequential collection of data points measured at consistent time intervals. (Brockwell and Davis, 2016).

4. Forecasting: Estimating future values based on historical data through the application of statistical models (Hyndman and Athanasopoulos, 2018).

5. ANN: Artificial Neural Networks are computational models structured in layered networks of interconnected processing units. These systems are loosely inspired by the biological brain and are capable of recognizing intricate patterns within data by transmitting signals through multiple weighted connections (Russell and Norvig, 2016).

6. Feedforward Neural Network: A type of ANN where information flows in a single direction from input to output layers, without cycles or loops. (Mapuwei et al., 2023).

7. Convolutional Neural Network: A specialized ANN architecture designed to process grid-like data, especially effective in extracting spatial features through convolutional layers. (LeCun et al., 2015).

8. Long Short-Term Memory: A type of recurrent neural network (RNN) that captures long-term dependencies in sequential data through memory cell structures, overcoming the vanishing gradient problem common in traditional RNNs (Hochreiter and Schmidhuber, 1997).

9. Integrated Artificial Neural Network**:** A hybrid model that combines multiple ANN architectures into a unified structure to enhance prediction accuracy and learn both spatial and temporal features from time series data (Zhang et al., 2021).

10. The Autoregressive Integrated Moving Average (ARIMA) model is a classical approach to time series forecasting that combines autoregressive terms, differencing to ensure stationarity, and moving averages to account for past error patterns (Box & Jenkins, 1976).

11. Mt Darwin District: A rural district in Zimbabwe's Mashonaland Central Province.

1.9 CHAPTER SUMMARY

As malaria continues to pose significant health challenges in Mashonaland Central Province particularly in rural districts like Mt Darwin, there is a need for localized forecasting models that support early detection and better intervention. This chapter introduced the background and validation for the study, outlined the research objectives, discussed the assumptions and limitations guiding the investigation. These elements provide the foundation for the upcoming literature review which will examine the theoretical and empirical basis for applying time series and computational intelligence hybrid neural network models in forecasting malaria incidence and mortality.

Chapter 2

2.0 INTRODUCTION

This chapter reviews theoretical and empirical studies on time series forecasting, focusing on its application in artificial intelligence driven models and traditional forecasting methods for malaria incidence and mortality. It also highlights key findings in the field and provides an overview of time series concepts.

2.1 THEORETICAL LITERATURE

2.2.0 Plasmodium Falciparum Malaria

Malaria is known to result from plasmodium parasites which are spread by a bite of a female anopheles mosquito. This plasmodium has four known species which are plasmodium falciparum, plasmodium vivax, plasmodium ovule, and plasmodium malaria, with plasmodium falciparum responsible for most malaria deaths, especially in Africa. Malaria parasites enter the bloodstream of the host person to destroy essential red blood cells, the destruction leads to fever, flu-like symptoms with vomiting and diarrhea, and if left untreated, the condition may progress to coma and ultimately prove fatal. (Alhassan et al., 2017).

2.2.1 Malaria Epidemiology in Zimbabwe

Zimbabwe has achieved noticeable reductions in malaria cases and mortality rates through various targeted health interventions however, rural districts continue to face persistent challenges due to limited access to healthcare services (Mutambara et al., 2019). Plasmodium falciparum, the dominant species responsible for malaria in the country continues to pose a significant public health burden. The disease's transmission patterns are shaped by a combination of environmental conditions, economic constraints, and the poor healthcare system in the country (Chikoko et al., 2021). Time series forecasting has proven effective in malaria control by enabling timely interventions. Chikoko et al., (2021) demonstrated how forecasting models successfully predicted malaria cases, improving planning and control efforts. Forecasting has also been linked to faster response times during peak transmission seasons, reducing morbidity and mortality (Mavundla et al., 2020).

Despite advancements in data collection and analysis existing forecasting methods often struggle to accurately predict malaria incidence due to the complex interactions between environmental, climatic, and socio-economic factors (Briar et al., 2020). The lack of region specific models further

limits intervention effectiveness, as generic approaches fail to address local epidemiological dynamics (Muller et al., 2019).

2.2.2 TIME SERIES

Box, Jenkins, Reinsel, and Ljung (2016) describe a time series as a series of observations collected sequentially over time. Similarly, Brockwell and Davis (2016) define a time series as a collection of data points, X_T , where each point is recorded at a specific time t. A discrete time series occurs when observations are made at distinct points in time, while a continuous time series is generated when data points are recorded continuously over a time interval, such as when $T_0 = [0,1]$.

2.2.3 TIME SERIES ANALYSES

Mapuwei et al., (2022) explain that time series analysis helps to uncover the underlying processes, understand how data changes over time and assess the impacts of planned or unplanned activities. In addition, Weigend and Gershenfeld (1994) highlight that the three primary goals of time series analysis are forecasting, modeling, and characterizing the time dependent behavior of data.

2.2.4 TIME SERIES FORECASTING

Lim and Zohren (2021) define time series forecasting as the process of predicting future values of a target variable Y_{it} for a specific entity i at time t. These entities represent logical groups of temporal data, such as measurements of vital signs from various patients in medicine, which can all be observed simultaneously. They further explain that in the simplest form of forecasting one-step ahead models the prediction takes the following form:

$$Y_{i,t+1} = f(Y_{i,t-k}, X_{i,t-k}, S_i) \text{ ----------(2.1)}$$

Where, $Y_{i,t+1}$ is the model forecast

$Y_{i,t-k} = ()$, $X_{i,t-k;t} = ()$ are observations of the target and exogenous inputs respectively over- back window k, S_i.

2.2.5 COMPONENTS OF TIME SERIES

Time series data typically exhibit four structural elements which are trend, seasonality, cyclical behavior, and random fluctuations. According to Shumway and Stoffer (2017) recognizing these components is essential for selecting suitable forecasting models. The trend reflects gradual, long-term movements in the data, often driven by macro-level influences such as policy reforms or demographic shifts (Davis and P.J., 2002). Mapuwei et al., (2022) noted that seasonality is the periodic variations that repeat at consistent intervals, like monthly or quarterly changes often shaped by climatic or institutional schedules and cyclic behavior, while also recurring, unfold over longer durations and are typically influenced by broader social or economic dynamics, making them harder to anticipate with precision (R.H. and Stoffer, 2017). Lastly, Woodward et al., (2017) described irregular components as noise or non-recurring anomalies that obscure underlying trends and are not explained by systematic patterns.

2.2.6 ASSUMPTIONS OF TIME SERIES

Several statistical assumptions support time series modeling, and ensuring these conditions are met enhances the reliability of the forecasts. One key assumption is stationarity, where the time series should exhibit no systematic change in its mean or variance, and all seasonal or periodic influences should be removed (Chatfield, 2003). Non-stationary data can often be transformed using differencing or logarithmic adjustments. Common tests for stationarity include the Augmented Dickey-Fuller (ADF) test and root tests (Mapuwei et al., 2022). Tsay (2010) distinguished between strict stationarity, which implies that the entire distribution remains unchanged over time, and weak or second-order stationarity as where only the mean, variance, and auto covariance remain constant. Another critical assumption is normality, which assumes that the data follow a normal distribution and violating this assumption may lead to inaccurate parameter estimates. Assessment testing tools include histograms, box plots, Q-Q plots, and probability distribution visualizations (Das and Imon, 2017). Independence of residuals is also important as it means that autocorrelation should be minimal and this is typically assessed using the Durbin-Watson test, residual plots, and ACF/PACF plots (Mapuwei et al., 2022). Finally, according to Mapuwei et al., 2022 homoscedasticity requires that the residuals maintain constant variance, which can be evaluated using scatter plots that show a consistent spread around a central line with no visible trend.

2.2.7 Models in Time Series Analysis

Throughout data science evolution, there has been a shift from traditional time series models, such as ARIMA and exponential smoothing as noted by Box, G. E. P., and Jenkins, G. M. (1976), to

modern deep learning techniques like ANNs RNNs, LSTMs, and IANN. This research will utilize both methodologies.

2.2.7.1 Traditional Time Series Models

(a) The Moving Average (MA)

Tsay, (2010) defined the moving average (MA) as the average of a specified number of time series values surrounding each point t in the series. An example of a moving average series with order q is denoted as {MA (q)}

$$Y_t = a_t + \theta a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q} \text{ ----------(2.2)}$$

(b) Autoregressive (AR) Model

A sequence is considered autoregressive if its current value is influenced by past values, along with a random shock (DaHye et al., 2021). Thus,

$$Y_t = \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + a_t \text{ ----------(2.3)}$$

Where

Y_t– Current Value, Y_{t-p} is the value at lag p

a_t – White noise error

$\emptyset_{1,}\emptyset_{2,}, \ldots, \emptyset_{p,}$ – Parameter of the model which is estimated from the data.

(c) Autoregressive Moving Averages (ARMA) Model

Tsay (2010) described the ARIMA model as a combination of autoregressive (AR) and moving average(MA) models, compacted to minimize the number of parameters and ensure simplicity in its parameterization. Similarly, Box, Jenkins, Reinsel, and Ljung 2015 referred to ARIMA as a blend of both AR and MA models. They further argued that when the equation of the first-order AR model approaches the starting point it will lead to an infinite moving average. To effectively use the ARMA model, the values p and q values must be determined, the value of p corresponds

to significant terms in the autocorrelation function (ACF), and q represents the number of significant terms in the partial autocorrelation function (PACF). If a time series obeys an ARMA (p, q) model, it is considered to exist.

$$Y_t = \sigma + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_q \in_{t-q} \text{ ----------(2.4)}$$

, $\in_{t-q}$ is considered to be the white noise process.

P.J Brockwell and R. A. Davis (2001) argued that $\{X_t\}$ is said to be ARMA(p, q) process if $\{X_t\}$ is stationary and if for every t,

$$X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \text{ ----------(2.5)}$$

where $\{Z_t\} \sim$ WN (0, σ^2) and the polynomials ($1 - \varphi_1 Z \quad - \cdots - \varphi_p Z^P$) and ($1 + \theta_{1Z} + ... + \theta_Q Zq$) have no common factors.

(d) SARIMA MODEL

A SARIMA, or Seasonal Autoregressive Integrated Moving Average, is described by Ramasubramanian (2015) as a model that can be applied to both seasonal and non-seasonal data. It adjusts for seasonal variations in the data to achieve stationarity. The model is defined as follows:

$$(1 - \emptyset \quad \beta)(1 - \emptyset_1 \beta^s)(1 - \beta)(1 - \beta^s) y_t = \ (1 + \emptyset_1 \beta)(1 + \emptyset_1 \beta^s)\varepsilon_t \text{ -----(2.6)}$$

, s is seasonal lag period

β is the backshift operator

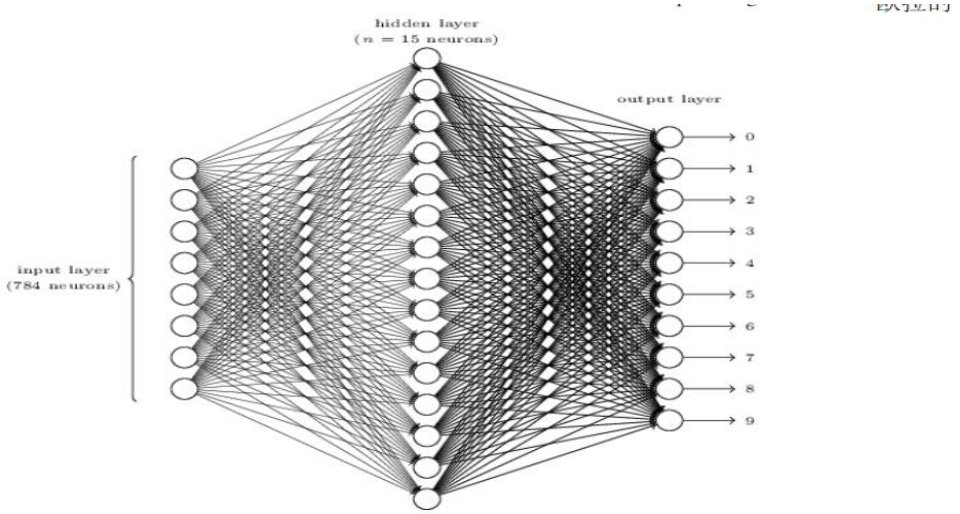
ε_t are noise

2.2.7.2 Artificial Neural Networks

(a) Overview of ANN

Artificial neural networks (ANNs) are a wide-ranging category of machine learning models designed based on how biological neural networks, such as the human brain process information and a make decision. These networks are made up of layers of interconnected nodes, or neurons, which apply various activation functions to transform input data for tasks like predictions and classifications. Bhimala, Patra, Mopuri, and Mutheneni (2021)**.** In a similar vein, Mapuwei et al., (2022) describe an artificial neural network as an information processing system created to generalize mathematical models based on human neural biology as shown by figure 2.2.1 below.

Figure 2.2.1 Artificial Neural Network Visual Architecture



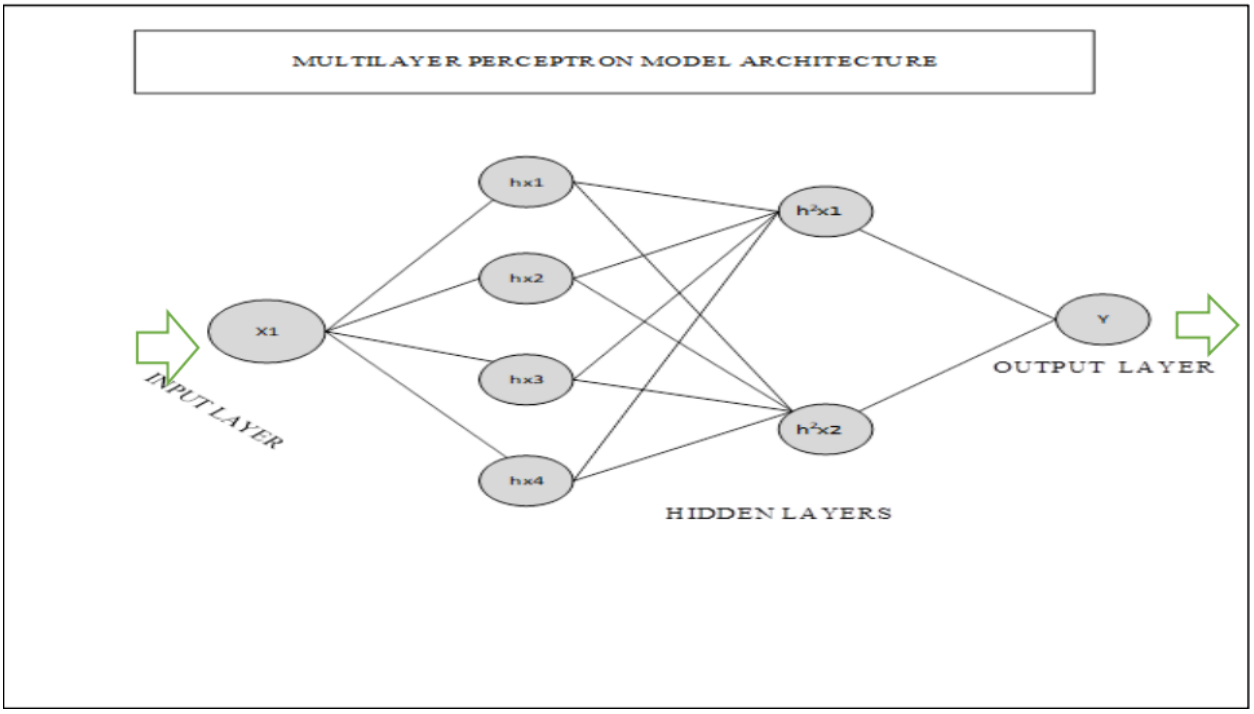
(b) Integrated Artificial Neural Networks (IANN)

Zhao et al., (2020) describe integrated artificial neural networks (IANNs) as advanced hybrid frameworks that bring together different deep learning architectures to enhance forecasting precision. Verma et al., (2021) and Farooq and Bazaz (2021) supported their theoretical potential through applications in epidemic modeling. However, these models were implemented in large, urban populations with rich datasets a contrast to the data scarcity and reporting delays common in Mt Darwin District. This study builds on those frameworks while tailoring them to a rural, malaria endemic setting where seasonality and limited data availability challenge predictive accuracy.

(c) Multilayer Perceptron (MLPs)

The figure shown below is a multilayer perception an artificial neural network as described by Nielsen et al. (2016) are neural network model which has multiple hidden layers (the middle layers), the output layer (the rightmost) which houses the output neurons and final the input layer (the leftmost) which houses the input neurons. A good example of the MLP models is the FFNN feedforward neural networks as shown by the diagram below.

Figure 2.2.2 Multilayer Perceptron Model Architecture



(a) Feed-Forward Neural Networks

FFNN feedforward neural networks are neural networks where the information is always fed forward and never fed back and their other characteristic is that the output from one layer is used

again in the next layer and is different from recurrent neural networks which have feedback loops. (Nielsen et al. 2016)

According to Mapuwei et al., (2022), the structure is defined by the number of hidden and output layers, and the feedforward neural network structure can be generalized with the assistance of the following equation.

$$I - (H_1, H_2, H_3, \dots, H_N) - O, \text{----------} (2.7)$$

Where

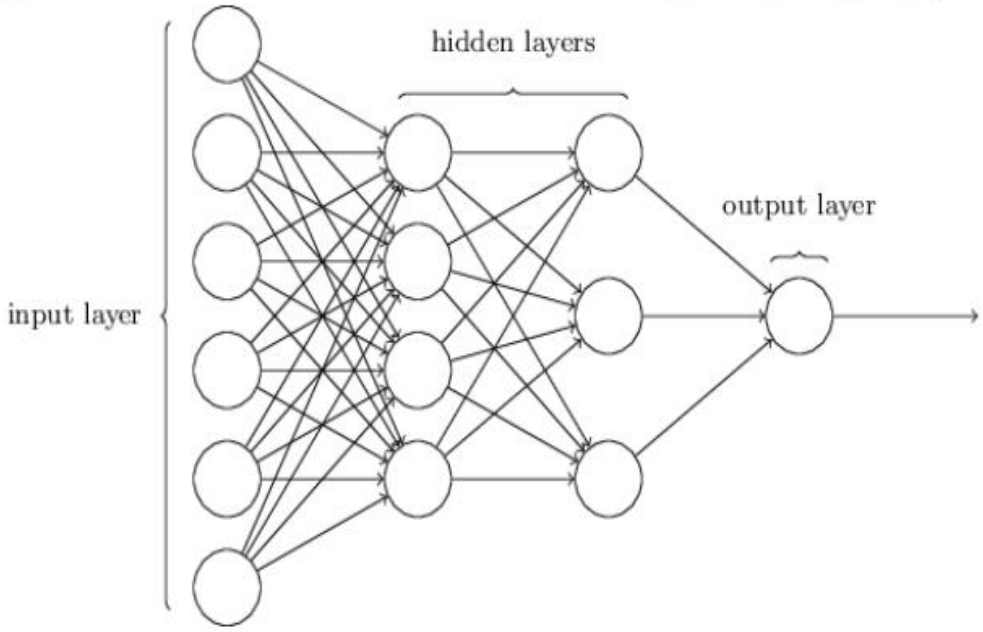
I = input nodes

$\boldsymbol{H_n}$= total of neurons in hidden layer, determined by the formula $\times = f\ (\sum_{j=1}^{n} Wjk\ Yj + \theta)$

and $o =$ the number of neurons in the output layer, determined by $Y = f\ (\sum_{k=1}^{n} Wk + Yk + \theta)$

The figure below shows the architecture of an FFNN (6-(4,3)-1) model.

Figure 2.2.3 FFNN (6-(4,3)-1) Architecture Model



According to Nielsen et al., (2016) the output is based on the perception rule which can be written as

$$\text{Output} = \begin{cases} 0 \; if \; w.x + b \; \leq 0 \\ 1 \; if \; w.x + b \; > 0 \end{cases} \text{----------(2.8)}$$

Where $w * x \equiv \sum_j WjXj$, w and x are vectors whose components are weights and inputs, respectively.

and b represents perception's bias , $b \equiv -threshold$

2.2.7.3 FFNN Model Training and Selection

Training a neural network is the process of teaching it to make accurate predictions and decisions based on its internal parameters which are weights and biases based on examples of input-output data. According to Mapuwei et al., (2022), the whole process is primarily based on determining weights and the number of neurons in the network.

(a) Forward Propagation:

This a training process, where data is fed starting from the input layer moving through the hidden layers to the output layer and in each layer, data is transformed by weights and activation functions, which helps the network learn complex patterns.

(b) Back Propagation:

The method is used to update the weights and biases of the network in order to minimize the loss function by calculating the gradient or rate of change of the loss concerning each weight by applying the chain rule of calculus. This process is done by propagating the error backward through the network (from the output to the input) and updating the weights to reduce the error.

(c) Model Selection

Husseien et al., (2017) assessed the predictive accuracy of different neural network models for malaria incidence using data from Sudan. Their evaluation relied on mean square error (MSE) and root mean square error (RMSE) as performance metrics. The model yielding the lowest error values was considered the most effective for forecasting purposes.

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(Yt - Yt)^2 \text{ ----------(2.9)}$$

$$RMSE = \sqrt{\frac{1}{N}}\sum_{t=1}^{N}(Yt - Yt)^2 \text{ ----------(2.10)}$$

2.2 EMPIRICAL LITERATURE

ARIMA MODELS 2.3.1

A study by Mapuwei et al., (2022), utilized the Box-Jenkins methodology in building an ARIMA model to forecast tobacco production in Zimbabwe. The ARIMA (1,1,0) with no seasonality was identified as the best model. The data was nonstationary as the ADF test failed to reject the null hypothesis as the p-value obtained was $0.6106 > 0.5$ and also evidenced by the absence of constant variation although a decreasing trend in tobacco production was noticed.

In their study, Alhassan et al., (2017) applied the Box-Jenkins approach to develop an ARIMA model for predicting malaria incidence in the Kasena Nankana Municipality. The objective was to identify a reliable model for short-term forecasting, with the ARIMA(1,0,1) configuration emerging as the most suitable after conducting standard adequacy tests. This model was then employed to project monthly malaria cases over a two-year horizon. Findings revealed a steadily increasing trend, shaped by a quadratic growth pattern, prompting the authors to recommend proactive interventions by the Ministry of Health. These included awareness campaigns to address the persistent nature of the disease and resource planning within health facilities in anticipation of future changes in case volumes.

Kumar et al., (2014), noted that ARIMA models were the simplest and yet the most reliable time series analysis tool for malaria forecasts after he had employed them to forecast malaria cases from 2006 to 2013 in the rural areas of Najafgarh, India. ARIMA $(0,1,1)\ (0,1,0)^{12}$, was the best fit with a seasonal component that was evidenced by ACF autocorrelation function which showed a significant peak at a lag of 12 months (autocorrelation = 0.675, Box-Ljung statistics (P=0.000)). Since the seasonal pattern was detected an ordinal R- squared was used as goodness of fit statistics, and it indicated a value of 0.725 meant that the model could explain 72.5% variability in the time series data. While Kumar highlighted ARIMA's simplicity and reliability in capturing short term patterns of malaria incidence, its effectiveness is limited when dealing with nonlinear and nonstationary data, which is often characteristic of malaria outbreaks. In contrast, Bhimala et al., (2021) demonstrated that artificial neural networks outperform traditional statistical models in capturing complex nonlinear relationships in malaria cases. However, the implementation of ANNs requires large datasets and more computational resources, which may not always be available in low resource settings like rural Zimbabwe. This contrast highlights the potential value of hybrid models that balance interpretability with predictive power.

NEURAL NETWORK MODELS 2.3.1

Yamak et al., (2020) conducted a comparative analysis of three different machine learning models in making a time series forecast of bitcoin prices. The models were ARIMA, GLU (gated recurrent units) and LSTM (long short-term memory), the ARIMA gave best results at MAPE =2.76% and RMSE = 302.53 which was outperformed by GLU model however the LSTM was chosen as the best model with 3.97% and 381.34 MAPE and RMSE respectively.

Mapuwei et al., (2020) conducted a comparative study of an ARN (FFNN) feedforward neuron network and (SARIMA) seasonal autoregressive integrated moving average in an effort to model city council ambulance demand. Performance calculation suggested an FFNN with an MAE = 94.0 RMSE = 137.19 and test value $p = 0.493(>0.05)$ was the best model for short-term annual forecasts rather than SARIMA with performance value of 105.71,125.28 and $p= 0.005(<0.05)$, respectively.

Research work on integrated artificial neural networks by Zhang et al., (2021) demonstrated the value of combining neural architectures. In his research, he reviewed hybrid deep learning frameworks, the CNN + LSTM with attention mechanisms outperformed single architecture models in time series application. Similarly, Wang and Li (2020) utilized a CNN-LSTM fusion for traffic flow prediction and their research showed that with convolutional layers to extract local temporal features LSTM layers then model for improved accuracy. Both the research studies although independent provide a strong precedent for integrated artificial neural networks compared to single architecture neural networks.

Epidemic research work by Verma, Mandal, and Gupta (2021) demonstrated that a CNN-LSTM hybrid model outperformed classical approaches in forecasting COVID-19 cases in India by effectively capturing both spatial and temporal dependencies. Similarly, Bhimala et al., (2021) showed that incorporating weather variables into LSTM-based models significantly improved the prediction of COVID-19 case trends.

RESEARCH GAP 2.4

The research gap exists in the lack of localized studies on the application of time series forecasting for Plasmodium falciparum malaria incidence and mortality trends in the Mt Darwin District. Additionally, there is a need for comparative studies between modern computational intelligence driven integrated artificial neural network models and autoregressive integrated moving average (ARIMA) models. Previous studies by Alhassan et al., (2017) and Kumar et al., (2014) were conducted primarily in India, focusing on the municipality of Kasena Nankara and the rural community of Najafgarh, respectively, and only utilized the autoregressive integrated moving average (ARIMA) model.

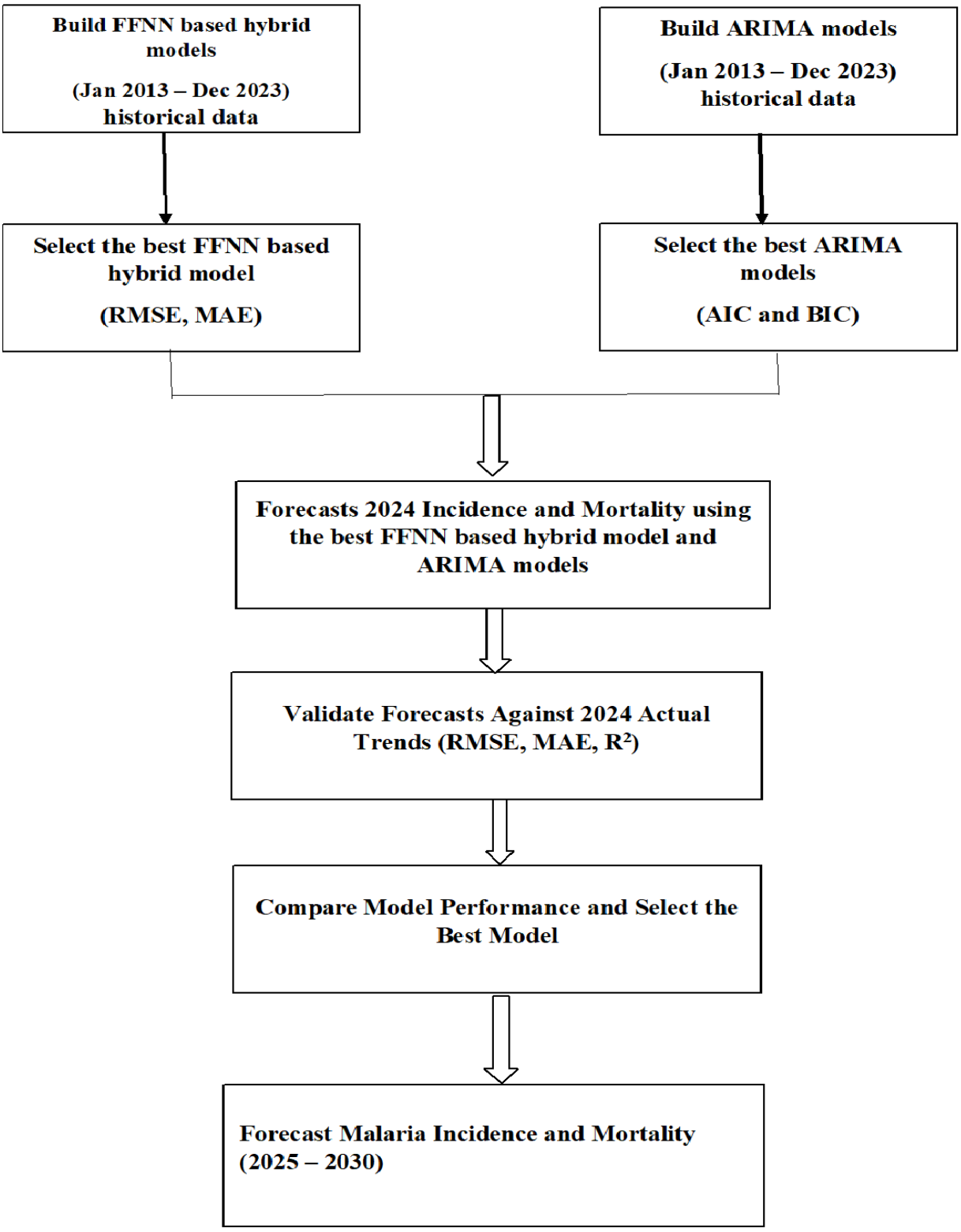
PROPOSED CONCEPTUAL METHOD 2.5

This research suggests building a time series prediction models for forecasting Plasmodium falciparum malaria incidence and mortality trends. The methodology starts with the building of a classical ARIMA model based on the Box-Jenkins method. In tandem, an Integrated Artificial Neural Network (IANN) model, mainly centered on a Feedforward Neural Network (FFNN), will be built and compared. Additionally, advanced hybrid architectures combining FFNN with Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) will be implemented to enhance predictive performance.

Each model will be trained using historical data from January 2013 to December 2023 and evaluated on 2024 data. Performance will be assessed using standard metrics such as RMSE, MAE, and R^2. The best-performing model will then be selected and used to forecast monthly and yearly malaria incidence and mortality from 2025 to 2030.

The proposed conceptual flowchart, presented in the figure below, outlines the methodological framework for model development, evaluation, and forecasting.

Figure 2.2.4 Proposed Conceptual Flowchart Structure


Build FFNN based hybrid models
(Jan 2013 – Dec 2023) historical data
Build ARIMA models
(Jan 2013 – Dec 2023) historical data
Select the best FFNN based hybrid model
(RMSE, MAE)
Select the best ARIMA models
(AIC and BIC)
Forecasts 2024 Incidence and Mortality using the best FFNN based hybrid model and ARIMA models
Validate Forecasts Against 2024 Actual Trends (RMSE, MAE, R^2)
Compare Model Performance and Select the Best Model
Forecast Malaria Incidence and Mortality (2025 – 2030)

CHAPTER SUMMARY 2.6

This chapter reviewed both the theoretical and empirical literature on the application of time series forecasting to plasmodium falciparum malaria incidence and mortality. Traditional time series models such as ARIMA and SARIMA were discussed for their strength in handling linear time series patterns, while the growing role of ANN models driven by computational intelligence was highlighted for their capability in capturing nonlinear dependencies. A research gap was identified in the limited application of ANN computational intelligence driven hybrid models within the rural and malaria endemic district of Mt Darwin. Based on the literature, this study adopts a computational intelligence hybrid approach with the integration of CNN, LSTM, and FFNN architectures due to their demonstrated strength in capturing both short and long term temporal patterns in complex epidemiological datasets. A conceptual framework was also presented to guide subsequent data analysis and interpretation. The next chapter outlines the research methodology and data collection techniques used in this study.

Chapter 3

3.0 INTRODUCTION

This chapter presented the research methodology used in the study, including the research design, data sources, sampling procedures, research tools, and analytical techniques. A quantitative approach was adopted, using time series forecasting methods to analyze historical malaria incidence and mortality data in Mt Darwin District. Specifically, ARIMA models and integrated artificial neural network (IANN) models were applied to generate forecasts. The chapter also addressed ethical considerations regarding the use of secondary health data according to the Public Health Act of Zimbabwe.

3.1 RESEARCH DESIGN

In this study, the researcher used a quantitative research design, a methodology that predicts future outcomes by analyzing patterns and trends in historical data. Quantitative research focuses on collecting numerical data and applying statistical and computational techniques to analyze relationships, test hypotheses, and make future predictions. This research included two variables: the number of malaria incidences at time t and the mortality rate (number of deaths) at time t.

3.2 SECONDARY DATA SOURCES

This research study primarily relied on secondary data which was drawn from the two official electronic health records (EHR) platforms which are Impilo health information systems and dhis2 system. Developed with input from local stakeholders and licensed under the Ministry of Health and Child Care (MoHCC), Impilo is a healthcare management system designed to optimize the management of health data across Zimbabwe by enabling healthcare providers to access, manage, and share patient information more efficiently. (impilo health systems, 2024). According to DHIS2 (2023), the district health information software 2 (dhis2) is an electronic health records management system that is used by district health authorities to collect, store, and analyze health related data to improve decision-making.

3.3 TARGETED POPULATION AND SAMPLING PROCEDURES

The research study primarily focused on the Mt Darwin District which is found in one of the biggest province Mashonaland Central Province. The researcher identified the district as a malaria endemic hotpot area making it more suitable to conduct the study for time series forecasting of plasmodium falciparum malaria incidences and mortality rate. Through the application of the

purposive sampling technique, the study took into account historical data on the number of malaria incidences and mortality from 2013 to 2023 to give a total of 120 monthly data observations points as the sample size for each time series variable.

3.4 RESEARCH INSTRUMENTS

The researcher mainly used computational tools as research instruments. Microsoft Office packages including Excel and Power BI were utilized. Excel was used as the main data birth-base after downloading it from the dhis2 servers and Power BI for visualization of results tables and plots. Then the R 4.4.1 statistical software and python 3.10 in google colab development environment, and visual code development environment were used for advanced data analytic such as data reprocessing, model building, model testing and performance measurements for the ARIMA and IANN (FFNN) based models respectively.

3.5 DATA ANALYSIS PROCEDURES

The main objective of this project was to build an adequate time series model for forecasting future Plasmodium Falciparum malaria incidence and mortality rates for Mt Darwin district. Data for the research was secondary data retrieved from the dhis2 server stationed at Mt-Darwin District Hospital for the years 2013 through to 2024. The 2013 to 2023 data was used for model building whilst 2024 data was used for model performance checking and testing and forecasts are to be made for the year 2025 to 2030. The statistical techniques used in model building and forecasting were the Box-Jenkins methodology for building ARIMA models and integrated artificial neural networks based on a feed-forward neural network an ANN process of model analysis.

3.6 THE BOX-JENKINS METHODOLOGY FOR BUILDING ARIMA MODEL

Box-Jenkins method, named after its creators George Box and Gwilym Jenkins, who introduced it for the first time in the 1970s, is typically used for the prediction of economic, financial, health, and other time-series data. Box-Jenkins offers a structured method to time series data modeling and forecasting mainly by utilizing ARIMA models (Box and Jenkins, 1976). Mapuwei et al., (2022) provide three iterative steps in the method: model identification, parameter estimation, and diagnostic checking while Alhassan et al., (2017) added forecasting as the fourth iterative step. In addition, the method can be applied in datasets with at least 30 observations and was readily adopted by the researcher as the sample database that was utilized in this research study contained 120 observations for each time series variable.

3.6.0 MODEL IDENTIFICATION (SELECTING AN INITIAL MODEL)

The researcher first determined whether the series is stationary or not by considering the graph ACF. According to Alhassan et al., (2017), if the ACF graph values either cut off fairly quickly or die down extremely quickly then it is considered stationary otherwise if the ACF dies down slowly it is considered non-stationary. The researcher learned that the series was not stationary and could be converted to stationarity by differencing the series and once stationary series status was obtained, the form of the model to be used was identified.

The autocorrelation function (ACF) can be calculated using the formula below

$$k = Y_k Y_0 \;\; Covariance\ at\ lag\ k\ variance \text{ ----------(3.1)}$$

PACF is calculated by the formula below

$$K_k = Corr(Y_t, Y_{t-k} | + Y_{t-1}, Y_{t-2}, \ldots\ Y_{t-k+1}) \text{ ----------(3.2)}$$

3.6.1 MODEL ESTIMATION AND EVALUATION

Alhassan et al., (2017) suggest that once the model has been identified, the next stage in the Box-Jenkins methodology chronologically sequence is to estimate parameters, the main method of estimating parameters is the maximum likelihood estimation and with the help of R-console statistical software the researcher utilized the method.

3.6.1.0 MLE ESTIMATION OF ARIMA MODEL OVERVIEW

An ARIMA model is denoted as ARIMA (p, d, q) where:

P is the order of autoregressive (AR) part

d is the degree of differencing required to make the series stationary

q is the order of moving average (MA) part.

3.6.2 MODEL CHECKING (goodness of fit)

In this iterative step, the researcher checked for model adequacy by considering the normality (normal distribution) of the residuals from the ARIMA model. Alhassan et al. (2017) argued that overall model adequacy is done using the Ljung-box statistic given below:

$$Q_m = n(n + 2) \sum_{k=1}^{m} \frac{rk2(e)}{n-k} \sim \chi^2 \, m - r \text{ ----------(3.3)}$$

,Where : e is the residual autocorrelation at lag

n is the number of residual

m is the number of times lags is included in the test.

If the p-value associated with the Q statistic is small, then the model is considered not inadequate, and if else the researcher continued with the analysis.

3.6.3 FORECASTING

According to the Box-Jenkins methodology, forecasting involves determining the expected values at a specific point in time (Alhassan et al., 2017). After confirming that the model fit the data well, the researcher proceeded with multi-step ahead forecasting of future values. While the accuracy of the forecast is generally expected to decrease as the forecast horizon extends, the forecast was based on the model's coefficients and past observed values.

3.7 THE ARTIFICIAL NEURAL NETWORKS METHODOLOGY

DATA PREPROCESSING 3.7.0

Data preprocessing was the first and most critical step in designing an Artificial Neural Network (ANN). This process included data cleaning, coding, normalization (standardization), and splitting the data into training, validation, and test sets according to the preprocessing setup by Mapuwei et al., (2020). The researcher performed data cleaning by addressing missing values, and replacing them with measures of central tendency from the respective rows or columns.

$$\text{Arithmetic mean} = \frac{\sum_{i=1}^{n} X_i}{n} \text{ ----------(3.4)}$$

, X_i = each data point and n is the number of data points

Median = if there is an odd number of data points, it is the middle value in a sorted dataset. If there is an even number of data points, the median is the average of the two middle values.

Mode is the value that occurs most frequently in the dataset.

Data normalization was done to ensure that all input features are on a similar scale preventing some features from dominating others and making the training process more stable and efficient.

Min-max Normalization

$$X_{\text{normalized}} = \frac{X - \min(x)}{\max(x) - \min(X)} \text{ ----------(3.5)}$$

Z-score Standardization

$$X_{\text{standardization}} = \frac{x - \mu}{\sigma} \text{ ----------(3.6)}$$

, μ *is the mean* , σ *is the standard deviation*

3.7.1 MODEL TRAINING AND TESTING SET

At this stage, it is crucial to divide the processed data into a model building (training) set and a testing set, with a larger percentage of the data allocated to the model building set and a smaller percentage to the testing set (Mapuwei et al., 2020). The researcher first used the model-building set to develop the Feed-Forward Neural Network (FFNN) model, while the testing set was used to evaluate the forecasting accuracy.

3.7.2 FEED-FORWARD NEURAL NETWORK ARCHITECTURE

As noted by Mai et al., (2021), selecting the optimal number of hidden layers in a neural network lacks a standardized guideline. In most cases, researchers rely on empirical testing and iterative adjustments to identify a suitable architecture. This process often involves applying a generalized structural formula as a starting point.

$$I - (H_1, H_2, H_3, \dots, H_N) - O, \text{ ----------(3.7)}$$

, Where

I = input nodes

$\boldsymbol{H_n}$ = neurons in the hidden layer, determined by the formula

$$\times = f\ (\sum_{j=1}^{n} Wjk\ Yj + \theta) \text{ ----------(3.8)}$$

and *O*= the number of neurons in the output layer, determined by

$$Y\ = f\ (\sum_{k=1}^{n} Wk + Yk + \theta) \text{ ----------(3.9)}$$

3.7.3 TRAINING A NEURAL NETWORK

The researcher employed the backward pass method, as outlined in two separate studies by Mapuwei et al., (2022) and Mai et al., (2021). This method involves determining the weights and the number of neurons in each layer of the network. In line with their studies, the researcher began by initializing the weights of the neurons randomly and setting the biases to zero. Then, the researcher computed the gradients of the loss function with respect to the weights and biases using backpropagation. This process was repeated until the model converged and stopped improving.

3.8 Integrated-Artificial Neural Network Hybrid Models

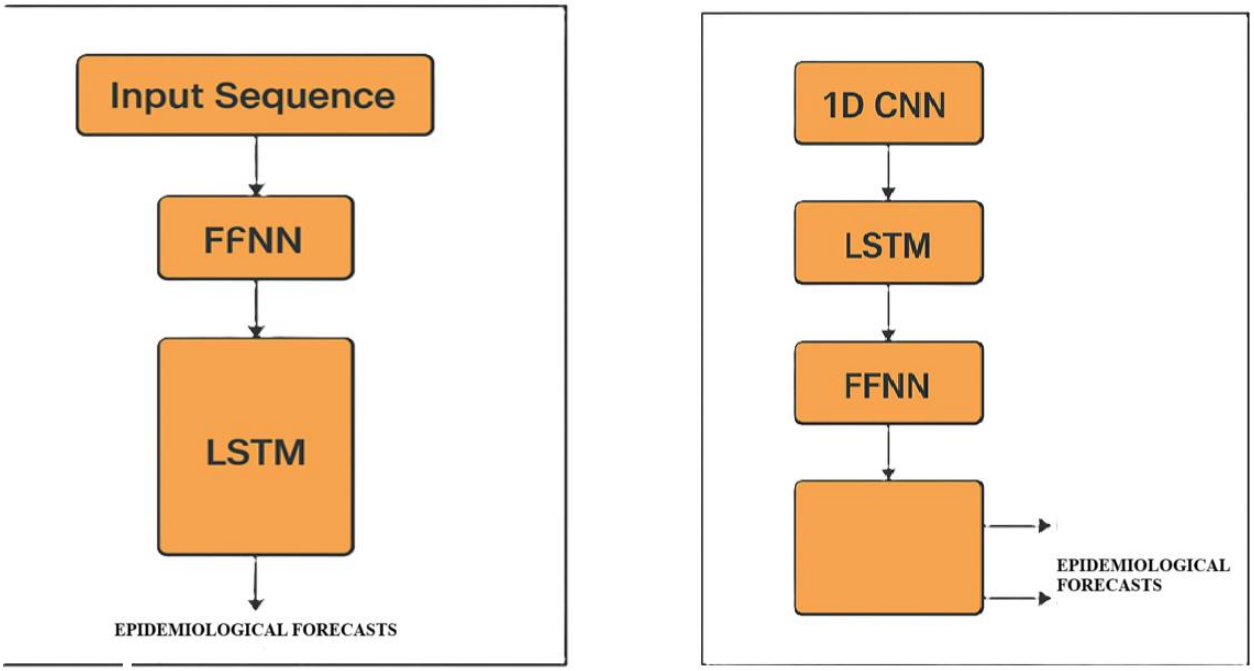
3.8.1 **Epidemiological Data Preprocessing**

The epidemiological data was cleaned out of any duplicates, outliers and disordered values, and missing values. Each time series was then scaled independently to N ~ [0,1] by the Min–Max normalization to align their ranges and stabilize the model building. Finally, the researcher generated supervised learning samples using a 12 month sliding window which is a 12 by 2 matrix of normalized incidence and mortality that will predict the next month's values by feeding the raw last month's values into a parallel FFNN branch in the FFNN+LSTM hybrid model, and in both convolutional and recurrent layers in the CNN+LSTM+FFNN hybrid.

3.8.2 **IANN Architecture**

The research employed two integrated ANN hybrids to capture both short term nonlinearities and long term dependencies in PF incidence and PF mortality. The FFNN + LSTM parallel architecture simultaneously feeds the last month's values into a small feedforward layers and the full 12-month sequence into an LSTM branch. The outputs are then brought together and passed through additional dense layers to produce dual forecasts. On the other hand, the CNN + LSTM + FFNN hybrids first uses a 1dimensional convolutional layer to the 12-month sequence to extract local temporal patterns and feeds the convolved features into an LSTM layer which models sequential dependencies and then route both the LSTM output and the final month's raw values through FFNN layers before the final multi-output layer. Both designs as shown below on figure 4.4.1 aimed at blending convolutional feature extraction, recurrent memory, and dense prediction in a unified model for monthly incidence and mortality forecasting.

Figure 3.8.1 Trained IANN Architecture



3.8.3 NEURAL NETWORKS SELECTION

The researcher evaluated the performance of the neural network models using regression metrics including MSE and RMSE. These metrics provided a basis for comparing different model configurations. The model yielding the lowest RMSE on the test set was selected, as it demonstrated the highest predictive accuracy for malaria trends in Mt Darwin District. The general formulas for MSE and RMSE are shown below, respectively:

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(Yt - Yt)^2 \text{ ----------(3.10)}$$

$$RMSE = \sqrt{\frac{1}{N}}\sum_{t=1}^{N}(Yt - Yt)^2 \text{ ----------(3.11)}$$

3.9 MODEL COMPARISON

The two models, the traditional ARIMA and ARN (FFNN) model were compared using the RMSE (root mean squared error) and MAE (mean absolute error) as model performance measures. The researcher also utilized the R^2 (root-squared) to measure the goodness of fit of the two models.

$$MAPE = \sum\left|Y_t - Y_{tYt}\right| n * 100 \text{ ----------(3.12)}$$

WhereYt=*the actual value*, $Yt\hat{}$*the forecasted value and n the number of observations*

3.10 ETHICAL CONSIDERATIONS

- The researcher abided by the medical data ethics code set by the MoHCC and Medical and Dental Practitioners Council of Zimbabwe (MDPCZ), personal and private information such as names and ages of patients were avoided.
- Data was retrieved from the allowed dhis2 domain abiding to the Public Health Act, Statutory Instrument 154 of 2020.
- Also abiding to the Public Health Act, harmful representations were avoided and some sensitive values were normalized for reputation purposes where necessary.
- There was no conflict of interest to be reported by the researcher.

3.11 CHAPTER SUMMARY

This chapter detailed the methods used to build and evaluate forecasting models for the plasmodium falciparum malaria epidemic in Mt Darwin District. The ARIMA model was employed as it is usually successful in capturing linear trends and seasonality, while FFNN based integrated ANN hybrid models were also designed since there are able detect complex nonlinear patterns, a usually character of epidemics databases. The methodological choice reflects the study's position that a comparison of classical statistical models with neural network hybrids model offers a better forecasting framework, especially in health systems where both structured seasonality and irregular fluctuations are present. The next chapter presents the analysis results and compares model performance using RMSE, MAE, and R^2.

CHAPTER FOUR

DATA PRESENTATION, ANALYSIS AND DISCUSSION

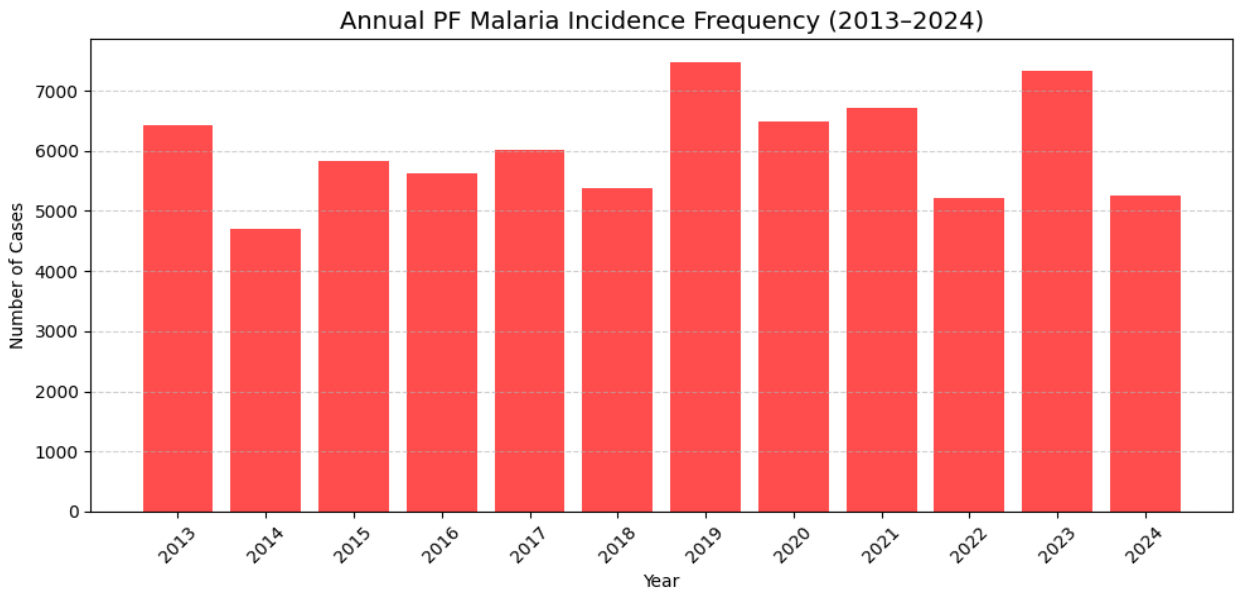
4.0 Introduction

This chapter focuses on data presentation, analysis, interpretation, and discussion of results. This was done to answer research objectives and questions. A time series forecasting of the plasmodium falciparum malaria epidemic was done utilizing integrated artificial neural networks and meaningful results and discussions were obtained.

4.1 Preliminary Analysis

4.1.1. Frequency on monthly incidence and mortality

Figure 4.1.1. Monthly PF Malaria Incidence Frequency (2013 -2024)



The two monthly frequency tables for malaria incidence figure 4.1.1 above and malaria mortality figure 4.1.2 below suggest a clear seasonal trend. The two variables show peaks between January and May and a decline from June to December. This pattern suggests that malaria outbreaks in Mt Darwin are strongly influenced by seasonal climatic factors such as rainfall and temperature which affect the mosquito breeding cycle.

Figure 4.1.2. Monthly PF Malaria Mortality Frequency (2013 -2024)

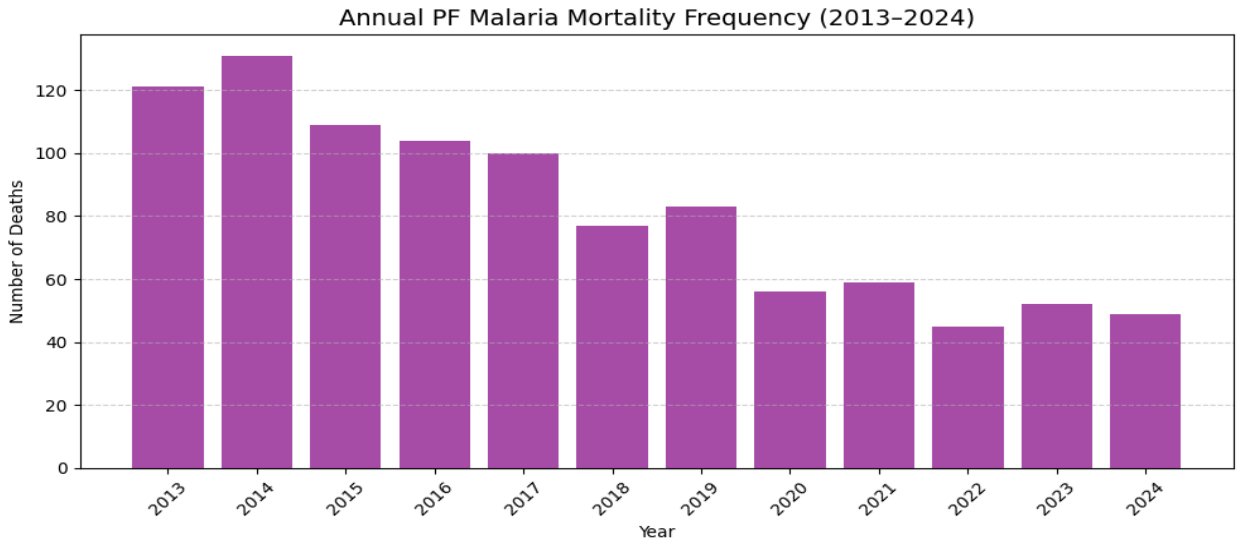


Table 4.1.1 Descriptive Statistics

Variable	Incidence	Mortality
Minimum	35	0
Maximum	69	12
Range	34	12
Sum	7246	707
1st Quartile	45.75	3
Median	50.50	5
3rd Quartile	54	6
Mean	50.32	4.91
Sample Variance	45.07	5.12
Standard Deviation	6.71	2.26
Kurtosis	0.23	-0.10
Skewness	0.23	0.45
Count	144	144

The descriptive statistics for PF malaria between 2013 and 2024 as shown in the figure above show valuable insights into the epidemiological pattern in District. The incidence data show a consistent trend, with a mean of 50 cases per month, with a deviation of about 6.71 cases, and positive skewness 0.23, indicating that while most months recorded case numbers around the mean, there is a peak season of January to May. The mortality data, reveal a lower mean of 4.91 deaths per month and a tighter spread of 2.26 deviation suggesting that deadly cases were generally fewer

and more stable over time. The mortality distribution is slightly skewed right = 0.45 but with negative kurtosis = -0.10, pointing to a flat distribution. The two variables have ranges of 34 for incidence and 12 for mortality, and the relatively low kurtosis and skewness in both cases indicate a likelihood of a normal distribution.

4.2 Pre-tests /Diagnostic tests

Figure 4.2.1 Time series plot of PF Malaria Incidence from 2013 to 2024

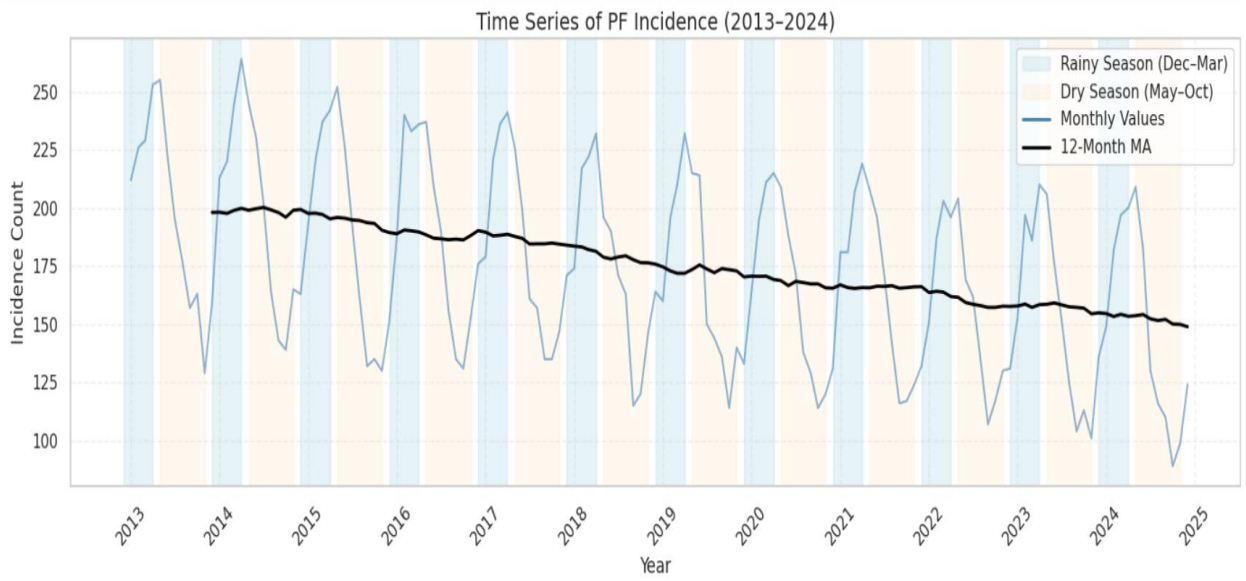
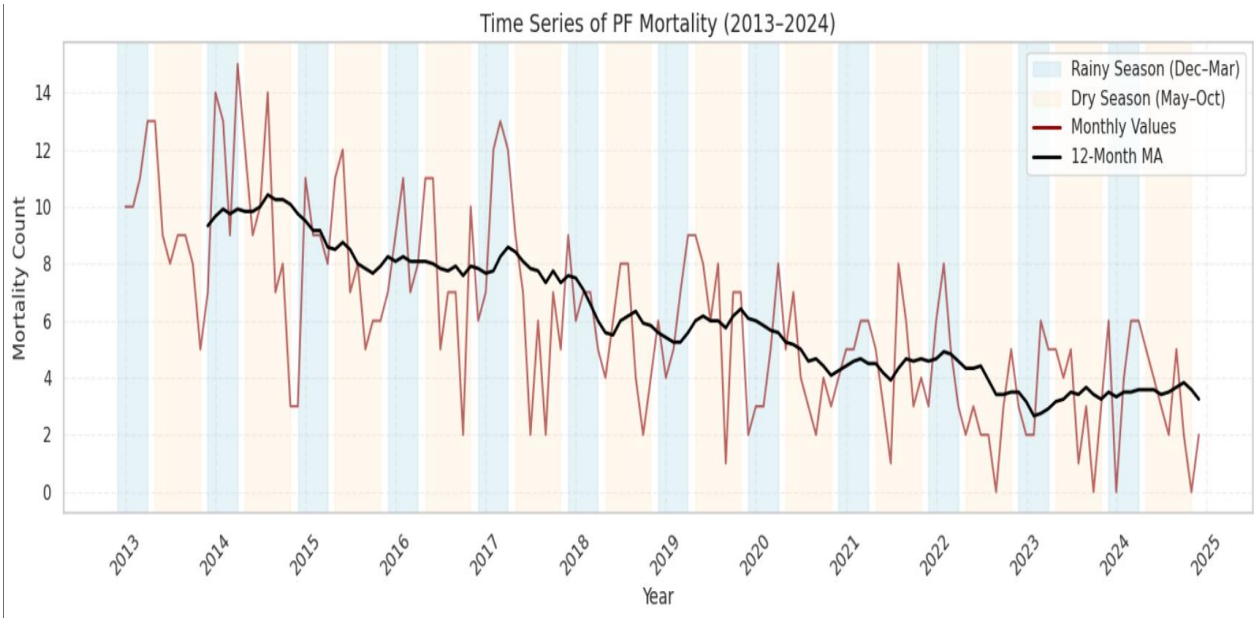


Figure 4.2.2 Time Series plot of PF Malaria Mortality from 2013 to 2024



The two-time series plots for PF Incidence and PF Mortality from 2013 to 2024 as shown above by figure 4.2.1 and figure 4.2.2 respectively show visible non-stationarity, as both series display clear seasonal patterns and long-term trends, particularly with repeated peaks during the rainy seasons which is usually from December to March in Mashonaland Central. The 12-month moving averages indicate persistent upward and downward shifts over time, suggesting that the data's mean and variance change over the years**,** which violates the assumptions of stationarity. These visual showings combined with seasonal fluctuations highlighted the need for further statistical confirmation using tests like the Augmented Dickey-Fuller and autocorrelation (ACF) and partial autocorrelation function (PACF) test before proceeding with model selection.

4.2.3 ADF Test for Trend Stationarity

H_0: The time series trends are non-stationary

H_1: The time series trends are stationary

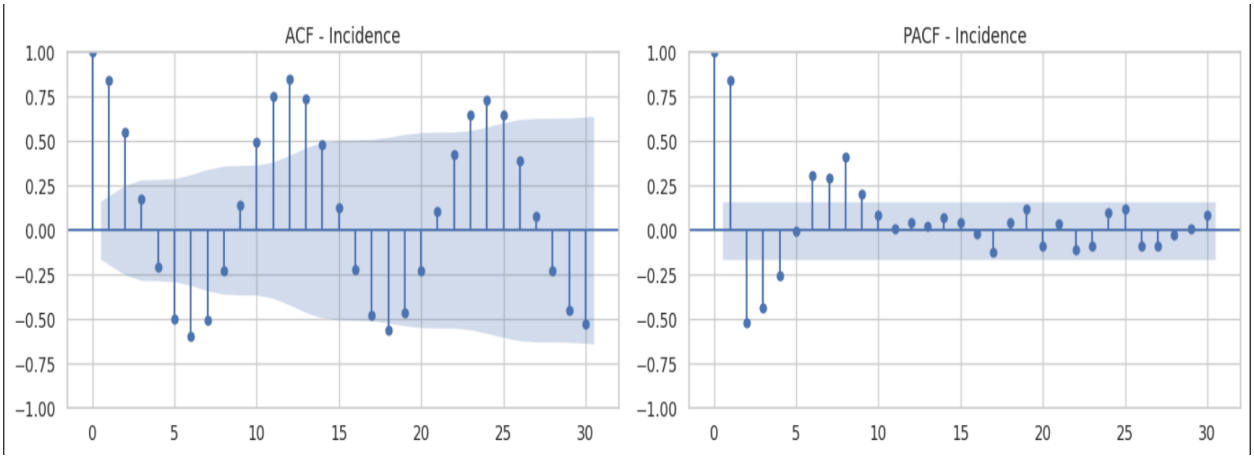
Figure 4.2.3 Augmented Dickey-Fuller Test Results Table

index	Incidence	Mortality
Test Statistic	-0.11195808898040253	-0.886196331653781
p-value	0.94823435898882	0.7924841693576671
# Lags Used	14	11
# Observations Used	129	132
Critical Value (1%)	-3.482087964046026	-3.4808880719210005
Critical Value (5%)	-2.8842185101614626	-2.8836966192225284
Critical Value (10%)	-2.57886438134727 5	-2.5785857598714417
Stationary	No	No

The research study concluded that we have failed to reject the null hypothesis that the time series data trends are indeed non-stationary evinced at 0.05 significance level and the data requires differencing. ACF and PACF plots were also used to confirm the non-stationary as suggested by the Augmented Dickey-Fuller hypothesis test.

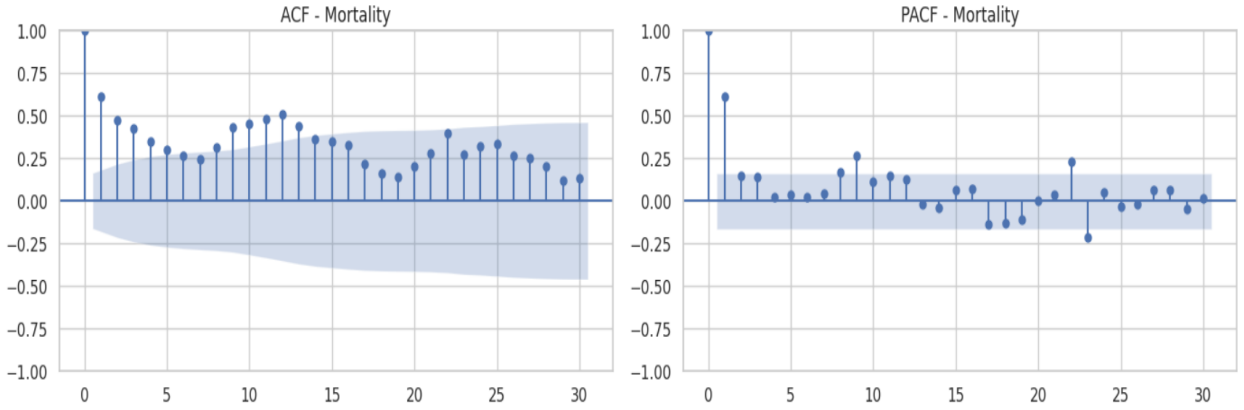
4.2.4 Autocorrelation function (ACF) and Partial autocorrelation function (PACF)

Figure 4.2.4: Incidence's ACF and PACF Plot for Raw Data



The ACF plot for PF incidence shown by figure 4.2.4 above shows a slow, gradual decay, while the PACF exhibits a few significant spikes followed by a cutoff. This pattern is characteristic of a non-stationary time series.

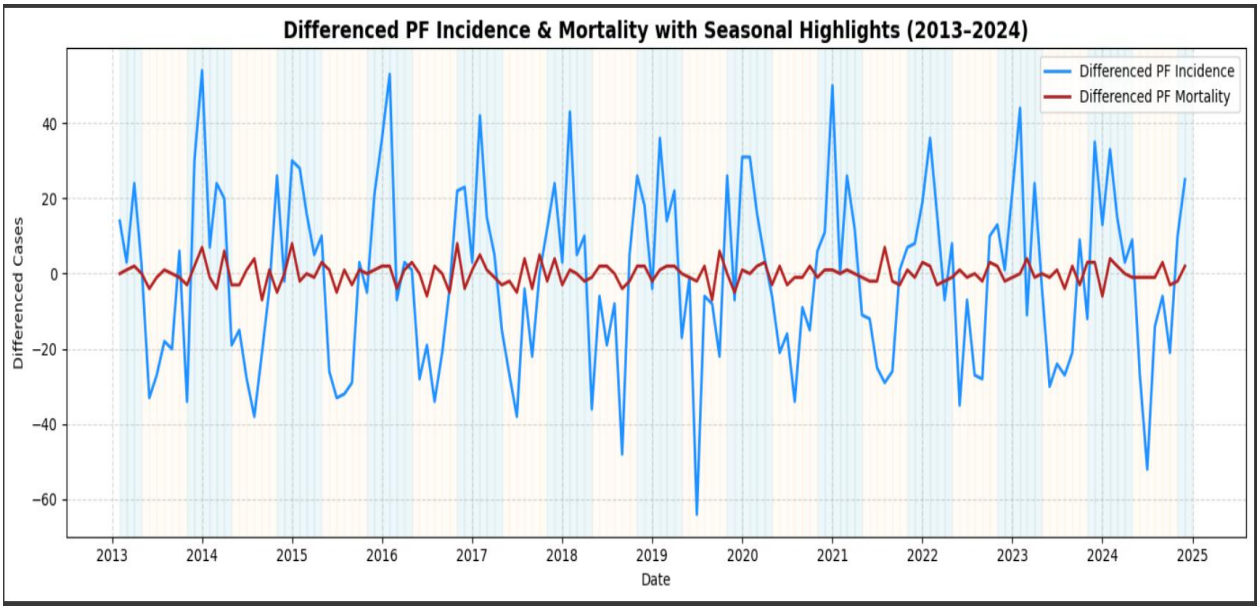
Figure 4.2.5: Mortality's ACF and PACF Plot for Raw Data



Similarly, the ACF of PF mortality shows a persistent autocorrelation with slow decay and several significant lags, while the PACF reveals a cutoff pattern, again showing non-stationarity. The ADF test result also supports this, with a p-value greater than 0.05, confirming that both series are non-stationary. Differencing will be necessary to achieve stationarity for time series modeling

3.8.3 Time Series Differencing

Figure 4.2.6 Differenced Time Series Data



The plot of the differenced time series figures 4.2.6 above shows that the strong trends and seasonality that were visible before differencing have been removed, especially for mortality which now fluctuates closely around zero. Incidence still shows some noticeable spikes but with reduced magnitude and more constant variance over time. These patterns suggest that the data may now be closer to stationarity and formal confirmation was done through the ADF test shown on figure 4.2.7 and checking for remaining autocorrelations using the ACF and PACF on figure 4.3.1 and 4.3.2.

4.2.2: Augmented Dickey-Fuller Test on Differenced Data

H_0: The differenced data time series are non-stationary

H_1: The differenced data time series trends are stationary

Figure 4.2.7 Augmented Dickey-Fuller Test on Differenced Data Results Table

index	Diffi_Incidence	Diffi_Mortality
Test Statistic	-6.715482560201263	-6.255596806267191
p-value	3.5910669031383137e-9	4.3468509908057975e-8
# Lags Used	14	14
# Observations Used	128	128
Critical Value (1%)	-3.4825006939887997	-3.4825006939887997
Critical Value (5%)	-2.884397984161377	-2.884397984161377
Critical Value (10%)	-2.578960197753906	-2.578960197753906
Stationary	Yes	Yes

We conclude that the ADF test results on the differenced data suggest that both series are now stationary**,** evinced at 0.05 (5%) significance level as test statistics are less than the critical values. This successful transformation into stationarity means we can proceed to ARIMA model building following the Box-Jenkins sequence. It also sets a strong foundation for developing **i**ntegrated artificial neural network models since stationarity helps improve the forecasting performance of the hybrid neural networks models by stabilizing the variance, mean and autocorrelation structures of the models.

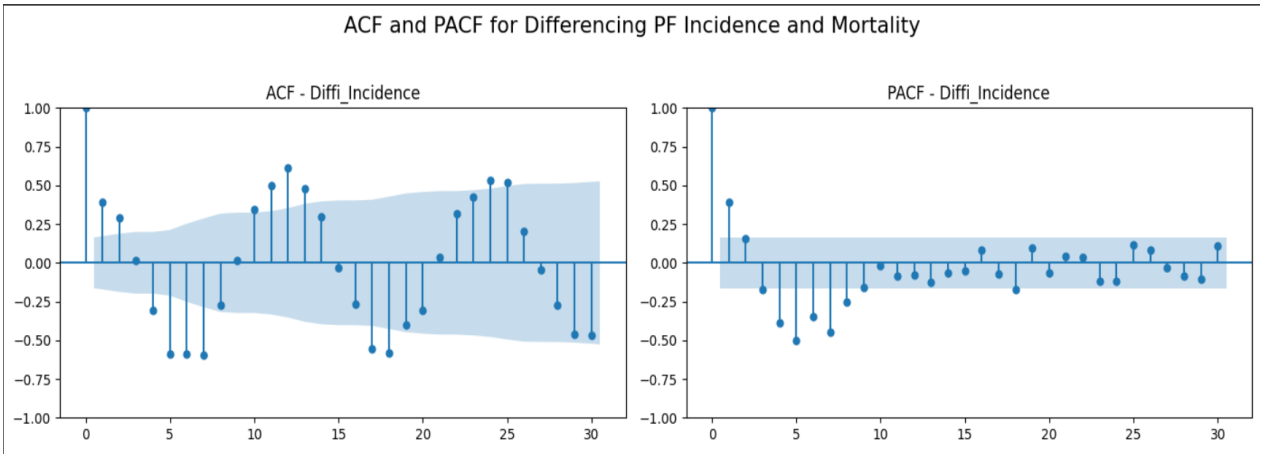
4.3 The Box-Jenkins Methodology

The Box-Jenkins Methodology sequence was followed in developing the ARIMA models adequate for the epidemiological data. The sequence includes model identification, parameter estimation and model checking applied separately to PF incidence and PF mortality after differencing.

4.3.1 Model Identification

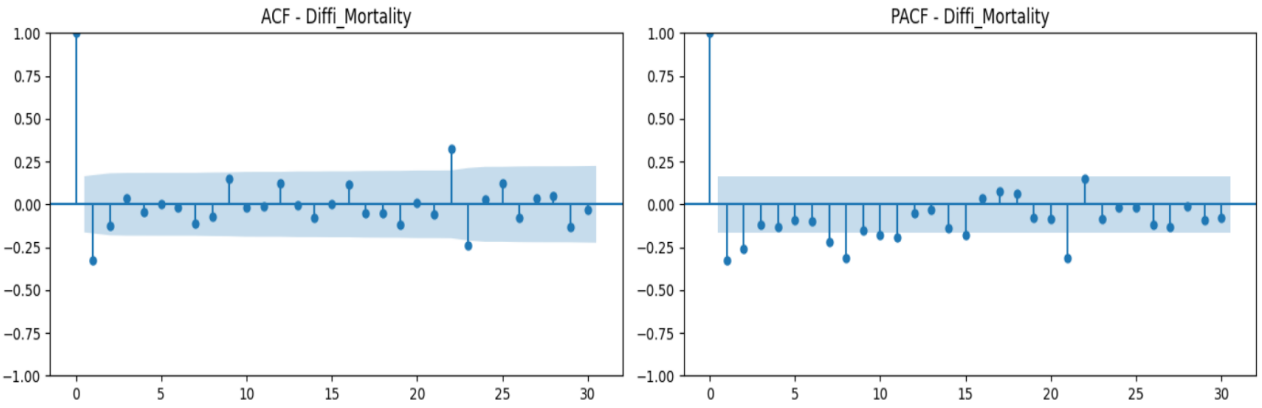
To determine the appropriate orders of the autoregressive (AR) and moving average (MA) components, we examined the ACF and PACF correlograms of the differenced series for incidence and mortality shown by figures 4.3.1 and 4.3.2 respectively.

Figure 4.3.1 ACF and PACF for Differ Incidence



The PACF showed significant spikes at lags 1 and 2 but cut off thereafter, while the ACF decayed gradually over several lags suggesting an AR = 2 component and an MA component of order 3.

Figure 4.3.2 ACF and PACF for Differ Mortality



The ACF exhibited a noticeable spike at lag 1 before clearing out, and the PACF likewise cut off at lag 1 suggesting an AR =1 and MA = 1 model structure.

This model structures insights were confirmed by an auto-arima search via the pmdarima package with the result presented in table 4.2.1. The pmdarima package which selected ARIMA (2,1,3) for incidence and ARIMA (1,1,1) for mortality as the models with the minimum AIC.

Table 4.3.1 Incidence and Mortality Auto-ARIMA AIC Results.

```
Best ARIMA for Incidence: (2, 1, 3) with AIC = 1009.76
Best ARIMA for Mortality: (1, 1, 1) with AIC = 614.94
ARIMA Model AIC Results - Incidence
   ARIMA_order          AIC
23   (2, 1, 3)  1009.763548
30   (3, 1, 2)  1039.485373
22   (2, 1, 2)  1068.195325
31   (3, 1, 3)  1074.138262
26   (3, 0, 2)  1077.043944
19   (2, 0, 3)  1081.295256
18   (2, 0, 2)  1087.840594
27   (3, 0, 3)  1091.831634
29   (3, 1, 1)  1114.372536
25   (3, 0, 1)  1124.274462
ARIMA Model AIC Results - Mortality
   ARIMA_order         AIC
13   (1, 1, 1)  614.940646
6    (0, 1, 2)  615.751046
31   (3, 1, 3)  616.220203
21   (2, 1, 1)  616.893430
14   (1, 1, 2)  616.898812
7    (0, 1, 3)  617.019933
22   (2, 1, 2)  618.383121
29   (3, 1, 1)  618.809376
15   (1, 1, 3)  618.889019
30   (3, 1, 2)  620.881020
```

4.3.2 Parameter Estimation

Having fixed the model orders, we estimated the AR and MA coefficients, along with the constant drift term where they were applicable using maximum likelihood. The resulting parameter estimates are summarized in Table 4.3.2 for incidence and Table 4.3.3 for mortality.

Table 4.3.2 Parameter Estimation for PF Incidence Model

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.7314	0.001	1976.564	0.000	1.730	1.733
ar.L2	-0.9999	0.000	-5405.447	0.000	-1.000	-1.000
ma.L1	-2.5620	0.076	-33.616	0.000	-2.711	-2.413
ma.L2	2.4141	0.143	16.903	0.000	2.134	2.694
ma.L3	-0.8093	0.076	-10.631	0.000	-0.959	-0.660
sigma2	107.1319	13.039	8.216	0.000	81.576	132.688

Table 4.3.3 Parameter Estimation for PF Mortality Model

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.3152	0.095	3.316	0.001	0.129	0.501
ma.L1	-0.9183	0.049	-18.620	0.000	-1.015	-0.822
sigma2	5.9149	0.704	8.405	0.000	4.536	7.294

All retained coefficients were highly significant ($p < 0.05$), confirming their contribution to capturing the autocorrelation structure of the differenced series.

4.3.3 Model Diagnostic Checking

To validate the adequacy of each fitted ARIMA model, we performed a series of residual diagnostics:

4.4.1 Residual Time Series Plot

Figure 4.4.1 Residual Time Series Plot

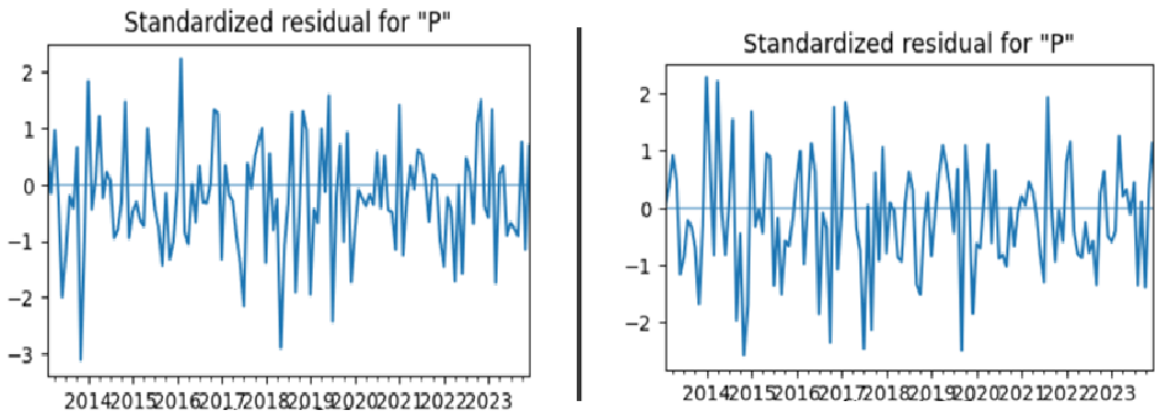
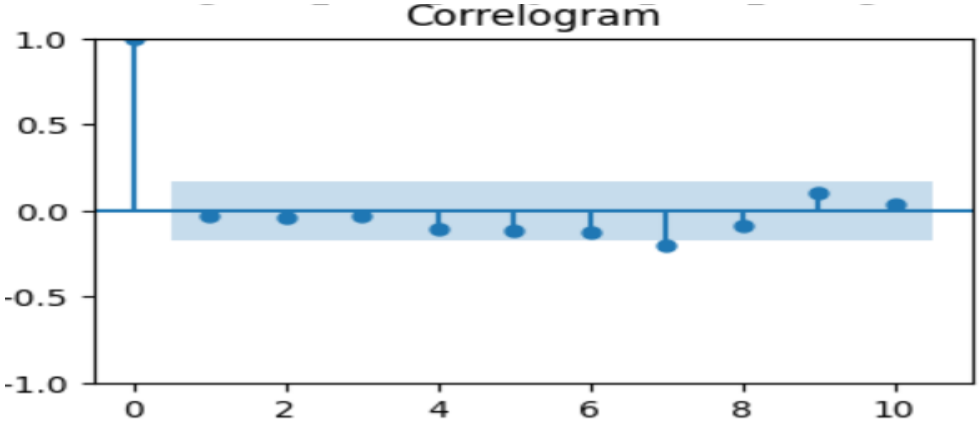


Figure 4.4.1 shows that plotted residuals over time have no remaining trend or obvious seasonality remains. The residual series appeared to fluctuate randomly around zero despite the spikes that within the range of -3 to 2 for both series.

4.3.4 Test of Independence

We examined the residual ACF correlogram as shown on figure 4.4.2 below. No autocorrelation spikes exceeded the 95% confidence bounds expect the one at first lag, indicating the residuals are effectively white noise.

Figure 4.4.2 ACF Correlogram Plot



4.3.4 Test for Normality

Histograms of residuals on figure 4.4.2 and a Q–Q plot on figure 4.4.3 both suggested approximate normality.

Figure 4.4.2 Incidence on the right and mortality on left

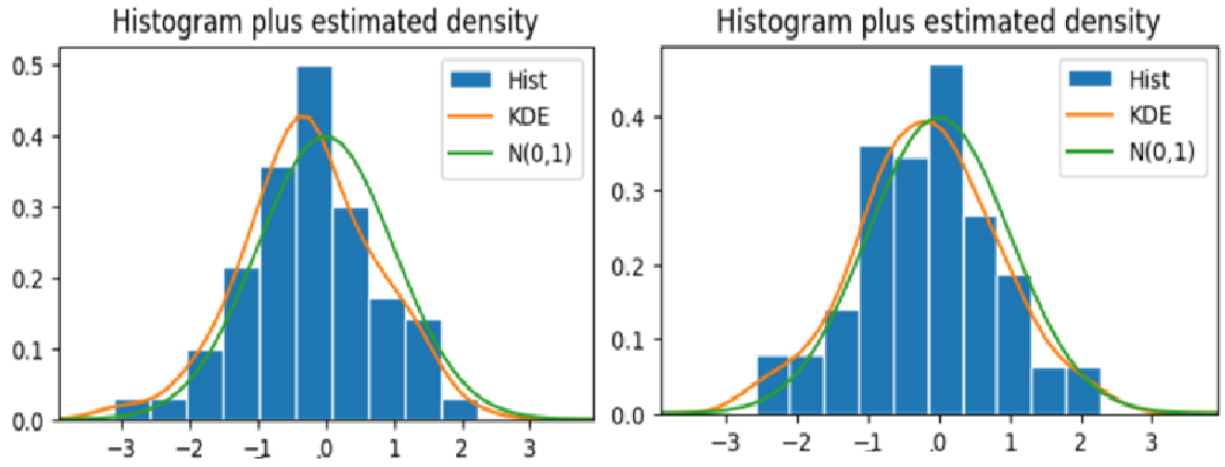
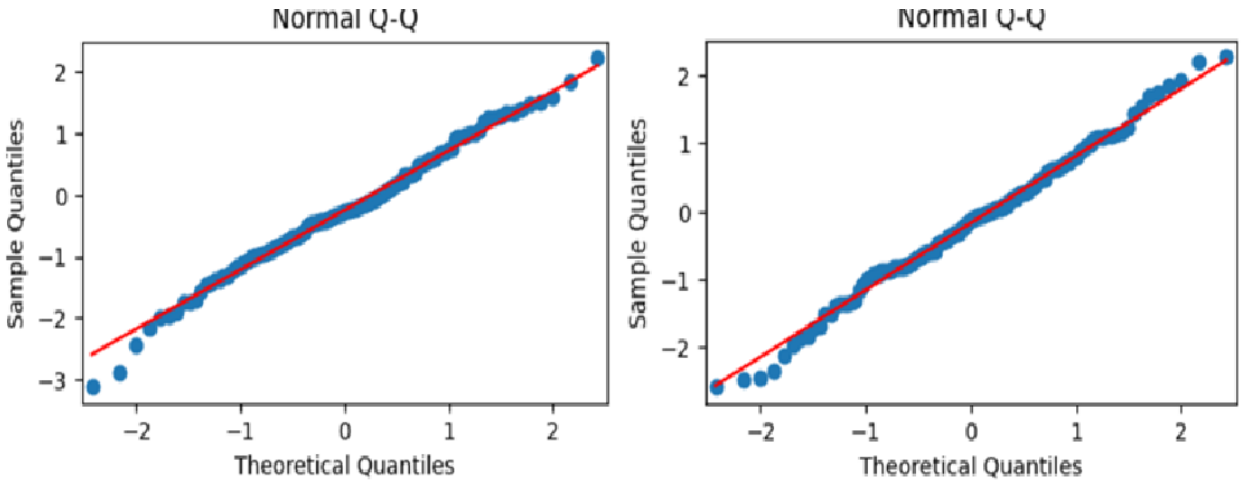


Figure 4.4.3 Q-Q Residual plot for incidence and mortality



The histograms displayed a bell shaped distribution and the Q-Q plots on figure 4.4.3 above showed that most of the residuals lie closer to the line. The suggested residuals are approximately normally distributed. Formal test for normality such as the Kolmogorov Smirnov test, Anderson

Darling, and Ryan Joiner tests returned p-values above 0.05, failing to reject normality as shown on table 4.4.4 below.

Table 4.4.4 Normality Test Results

Test	**Statistic**	**P-value**
Kolmogorov Smirnov	At 0.05	0.225
Anderson Darling	At 0.05	2.295
Ryan Joiner	At 0.05	0.875

These diagnostics confirmed that the ARIMA (2,1,3) and ARIMA (1,1,1) models adequately capture the dynamics of PF incidence and mortality respectively, with residuals that are stationary, uncorrelated, and normally distributed thereby meeting the key Box Jenkins conditions for effective forecasting models.

4.3.5 ARIMA Models Validations

Firstly, the researcher partitioned the data into a training set and a testing set, the training set covered from January 2013 through to December 2023 and the rest was a test set comprising January to December 2023. A one step ahead forecast for each month of 2024 was done using the finalized models, ARIMA (2.1.3) for PF incidence and ARIMA (1.1.1) for PF mortality and compared them against actual observed values (Table4.3.1).

Table 4.3.1 Testing Set 2024 District's PF Malaria Epidemiological Data (Actual Values Vs ARIMA Forecasted)

2024	PF Incidence Testing Set		PF Mortality Testing Set	
Month	**Actual**	**ARIMA (2.1.3)**	**Actual**	**ARIMA(1.1.1)**
January	264	264	0	8
February	393	387	4	7
March	797	780	6	7
April	612	600	6	7
May	216	224	5	6
June	141	150	4	5
July	921	895	3	2
August	605	615	2	1
September	345	340	5	0
October	429	418	2	0
November	697	685	0	0
December	539	530	2	0
Results	***MAE***	***8.67522***	***MAE***	***1.78303***
	RMSE	***125.430663***	***RMSE***	***4.65113***
	R^2	***0.826141***	***R^2***	***-0.156753***

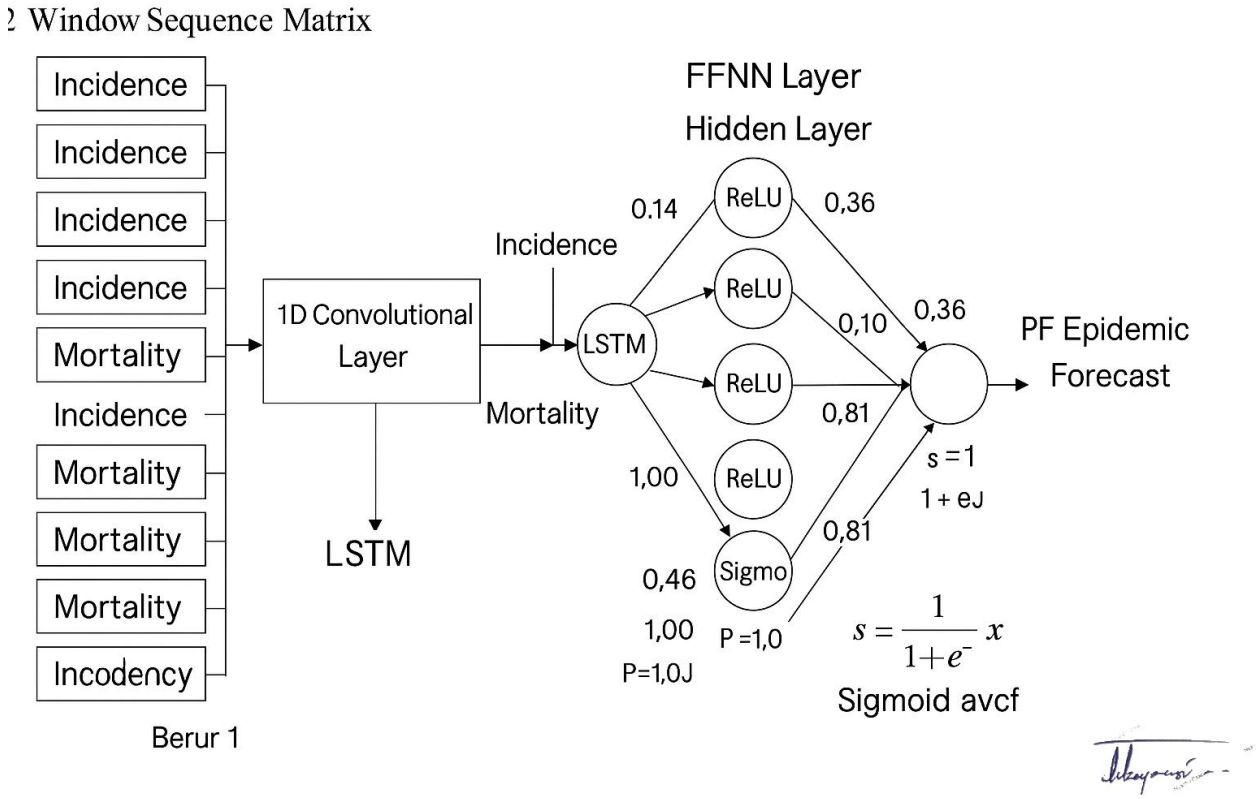
The ARIMA (2.1.3) model achieved an MAE of 8.68 and RMSE of 125.43 cases on the 2024 test set with a strong R^2 of 0. 826. This indicates that the model can explain over 80% of the variations in PF incidence and suggests it can provide reliable forecasts for malaria case counts in the district. However, unlike the ARIMA (2.1.3) the ARIMA (1.1.1) model for mortality underperformed with a negative R^2 of -0.157 which implied that the model totally failed to model death count and it couldn't provide reliable forecasts although it was the best model according to AIC values.

4.4 Integrated-Artificial Neural Network Hybrid Models

The integrated ANN hybrid models demonstrated superior capacity in modeling both short-term fluctuations and long term temporal dependencies in plasmodium falciparum incidence and mortality. By simultaneously feeding both series of the historical data through distinct interconnected layers, the FFNN + LSTM hybrid effectively captured sharp and recent changes using feedforward layers, while the LSTM path modeled sequential dependencies and seasonal trends over the 12-month window. This parallel structure proved particularly effective in balancing immediate outbreak spikes with underlying temporal trends and patterns.

The addition of a 1D Convolutional layer in the CNN + LSTM + FFNN hybrid further improved model performance by extracting local temporal recurring three month dips before passing them to the LSTM layer for memory holding. This layered relationship allowed the model to detect and connect localized and cumulative epidemiological signals, which were then refined through feedforward layers into actionable forecasts. The benefit of this architecture was its ability to fuse pattern recognition, memory, and prediction into an interconnected learning process. The trained model architecture is presented in Figure 4.4.1

4.4.1 Trained IANN Architecture



4.4.2 Optimal Trained Computational Dense Layers Results

Layer (type)	Output Shape	Param #	Connected to
input_layer (InputLayer)	(None, 12, 2)	0	-
conv1d (Conv1D)	(None, 8, 64)	704	input_layer[0][0]
max_pooling1d (MaxPooling1D)	(None, 4, 64)	0	conv1d[0][0]
input_layer_1 (InputLayer)	(None, 2)	0	-
lstm (LSTM)	(None, 64)	33,024	max_pooling1d[0]…
dense_1 (Dense)	(None, 32)	96	input_layer_1[0]…
dense (Dense)	(None, 32)	2,080	lstm[0][0]
dense_2 (Dense)	(None, 16)	528	dense_1[0][0]
concatenate (Concatenate)	(None, 48)	0	dense[0][0], dense_2[0][0]
dense_3 (Dense)	(None, 32)	1,568	concatenate[0][0]
dense_4 (Dense)	(None, 16)	528	dense_3[0][0]
dense_5 (Dense)	(None, 2)	34	dense_4[0][0]

Total params: 38,562 (150.63 KB)
Trainable params: 38,562 (150.63 KB)

The CNN+LSTM+FFNN hybrid could both capture short-term variability and long-term seasonality. Such high performance is the best proof of the intelligent design of the model, which replicates the information processing capability of the human brain. The 64-filter convolutional layer (Conv1D) allowed the model to detect small, repetitive patterns in the 12-window sequence matrix. These features were input to a pooling layer and to the LSTM layer, used to learn from past trends and seasonal outbursts. At the same time, another input layer handled current information such as instant conditions. Outputs from the two branches were combined and input to fully connected layers to generate the output prediction. The entire model had 38,562 trainable parameters, which was complicated enough to be learned from the data but not too large to overfit.

3.8.4 **Model Training and Testing**

The integrated artificial neural network (IANN) models were trained using normalized PF incidence and PF mortality data from 2013 to 2023, with 2024 set out for testing. Supervised learning samples were generated using a sliding 12-month input window to predict the next month's incidence and mortality values. After training, forecasts for the 12 months of 2024 were generated and compared against the actual observed values. The forecasting results were summarized in a table showing the models' ability to track real epidemiological trends across the test period.

Table 4.3.2 Neural Models Validation

MONTH	Actual Incidence	LSTM+FFNN Incidence	CNN+LSTM+FNN Incidence	Actual Mortality	FFNN+LSTM Mortality	CNN+LSTM+FNN Mortality
January	264	260	262	0	1	0
February	393	388	391	4	4	3
March	797	790	794	6	6	5
April	612	605	608	6	5	5
May	216	221	218	5	4	4
June	141	138	140	4	3	3
July	921	915	918	3	3	2
August	605	610	606	2	2	2
September	345	340	343	5	5	4
October	429	425	427	2	1	1
November	697	692	695	0	1	0
December	539	533	537	2	1	1
MAE		***4.25***	***2***		***6.71***	***5.43***
RMSE		***4.40***	***2.16***		***116.93***	***96.85***
R^2		***0.8138***	***0.9132***		***0.9132***	***0.9396***

The performance of each model was evaluated using three statistical metrics which are the mean absolute error, root mean squared error, and the coefficient of determination. On the 2024 test set, the FFNN+LSTM hybrid model produced an MAE of 6.71, an RMSE of 116.93, and an R^2 value of 0.91, suggesting strong alignment between predicted and actual values. The CNN+LSTM+FFNN model achieved slightly better results, with an MAE of 5.43, an RMSE of 96.85, and an R^2 of 0.94. These results indicate that incorporating both convolutional and recurrent layers contributed to improved forecasting performance when compared to the simpler FFNN+LSTM model.

3.9 Models Comparison

The effectiveness of both traditional time series and computational intelligence approaches, the ARIMA, FFNN+LSTM hybrid, and CNN+LSTM+FFNN hybrid models was evaluated by key performance metrics such as the mean absolute error (MAE), root squared error (RMSE), and R^2 score. While ARIMA showed good forecasting capability in PF incidence although it was very weak in modeling PF mortality, the FFNN+LSTM hybrid showed improvements with the ability to model both, and the CNN+LSTM+FFNN model consistently outperformed both, as it achieved the lowest errors and highest variance explanation. These results showed the significant advantage of deep learning-based hybrids in modeling complex, nonlinear plasmodium falciparum epidemiological trends.

These results are similar to the findings of Wang and LI (2020) and consistent with those of work by Zhang et al., (2021) who both independently noted that integrating convolutional to extract local patterns with LSTM memory and FFNN predictions layers yields better models compared to ARIMA and in some case single architecture ANN.

3.10 Best Model Selection

Based on the model validation results, the CNN+LSTM+FFNN hybrid was selected as the best performing model, having delivered the best accuracy and strength across all evaluation metrics. Therefore, this integrated artificial neural network hybrid model was employed to generate monthly forecasts of PF malaria incidence and mortality from January 2025 to December 2030 due to its ability to forecast future trends of plasmodium falciparum malaria with such a high degree of confidence rather than traditional time series models.

4.7 2025 to 2030 Forecasting

TABLE 4.7.1: 1D CNN+LSTM+FFNN Forecasted Monthly Plasmodium Malaria Incidence and Mortality 2025 – 2030

TABLE 4.7.1: CNN+LSTM-FFNN Forecasted Monthly Plasmodium *falciparum*

		Expected Incidence			Expected Mortality				
Monc	2025	2026	2027	2027	2025	2026	2027	2028	2030
January	149	150	165	146	3,67	3.64	3,32	3,32	2,79
February	176	165	157	152	4,75	4,19	4,48	3,30	2,93
March	188	176	165	162	5,36	5,36	4,56	3,99	3,05
April	188	171	165	159	5,56	5,56	4,98	3,02	3,08
May	180	171	165	145	4,98	4,32	4,48	3,77	3,02
June	149	149	146	140	3,46	3,46	2,95	2,73	2,88
July	149	146	144	140	3,46	3,95	2,95	2,73	2,73
August	131	132	136	133	2,95	2,95	2,09	2,60	2,73
September	119	131	128	133	2,64	2,95	2,50	2,50	2,51
October	116	121	120	133	2,52	2,52	2,50	2,48	2,48
November	121	122	127	133	2,52	2,52	2,49	2,48	2,52
December	132	133	135	133	3,30	3,06	2,50	2,52	2,59
Average	150,8	149,7	143,7	143,7	**3,93**	**3,93**	3,96	2,96	2,96

4.8Discussion of Findings

The forecasts by the CNN+LSTM+FFNN hybrid model show that there is a seasonal cycle on the plasmodium falciparum incidence with peaks every year between February and April which mirrors the rainy-season spikes trends documented by Kumar et al., (2014). For this research, it means the hybrid model had successfully learned the district annual transmission trend. The area chart on figure 4.8.3 further supports that, even incidence levels decline over time from the forecasted 1824 cases in 2025 to 1703 by 2030 with seasonal peaks remaining marked each year.

Figure 4.8.1 CNN+LSTM+FFNN Time Series Plot of Forecasted Monthly Plasmodium Malaria Incidence

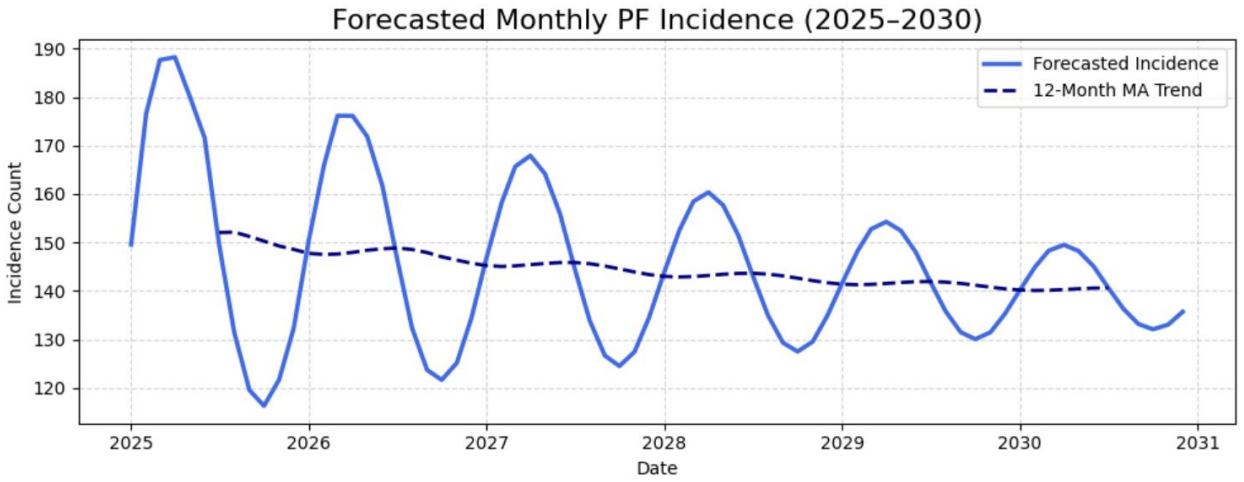
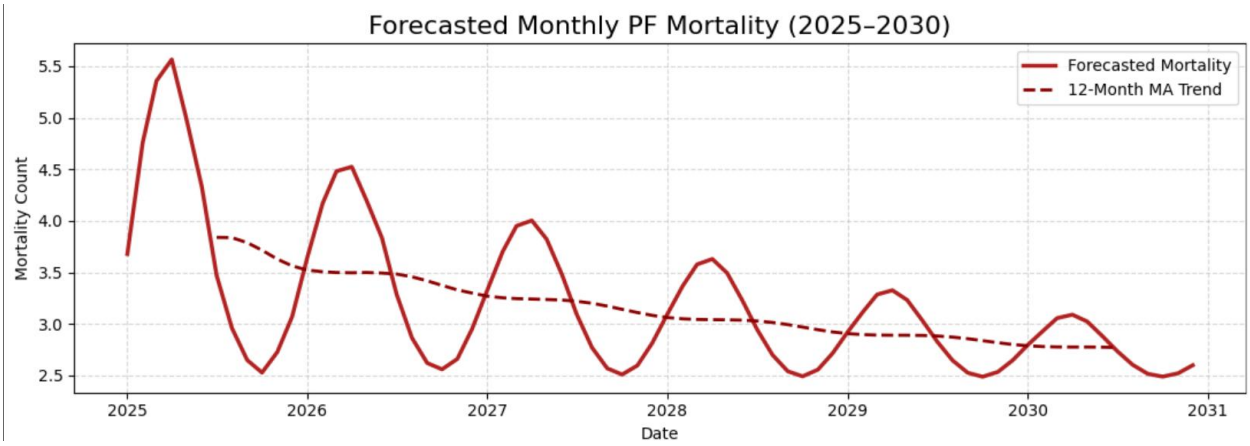
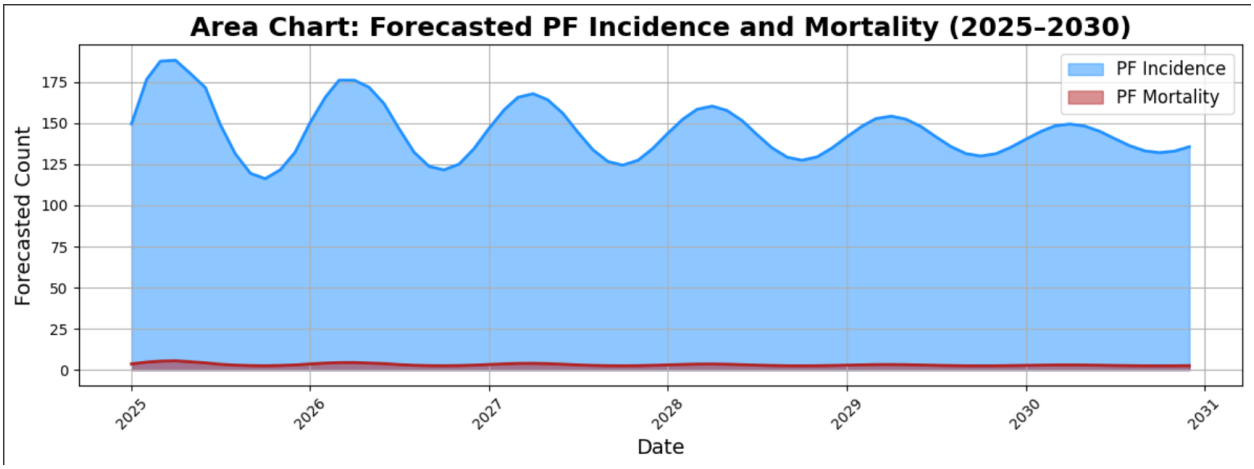


Figure 4.8.2 CNN+LSTM+FFNN Time Series Plot of Forecasted Monthly Plasmodium Malaria Mortality



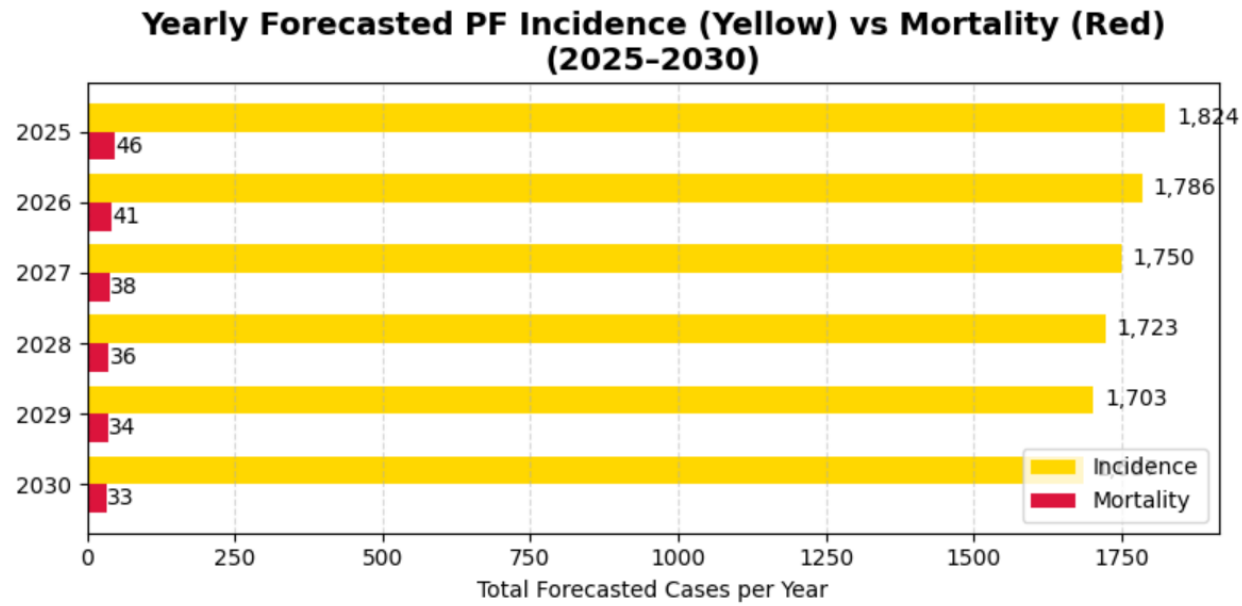
On the other hand, mortality forecasts show a flatter trend on monthly deaths with a minimum of 2 and a maximum of 6 and a gradual downward drop from the annual total of 46 deaths expected in 2025 to 33 in 2030. This decline aligns with the research work of Alhassan et al., (2017) on improving case management in northern Ghana which further suggested that, even if malaria case numbers are seasonally volatile there is need for enhanced treatment and prevention efforts to reduce the risk.

Figure 4.8.3 Area Chart of Forecasted PF Incidence and Mortality (2025-230)



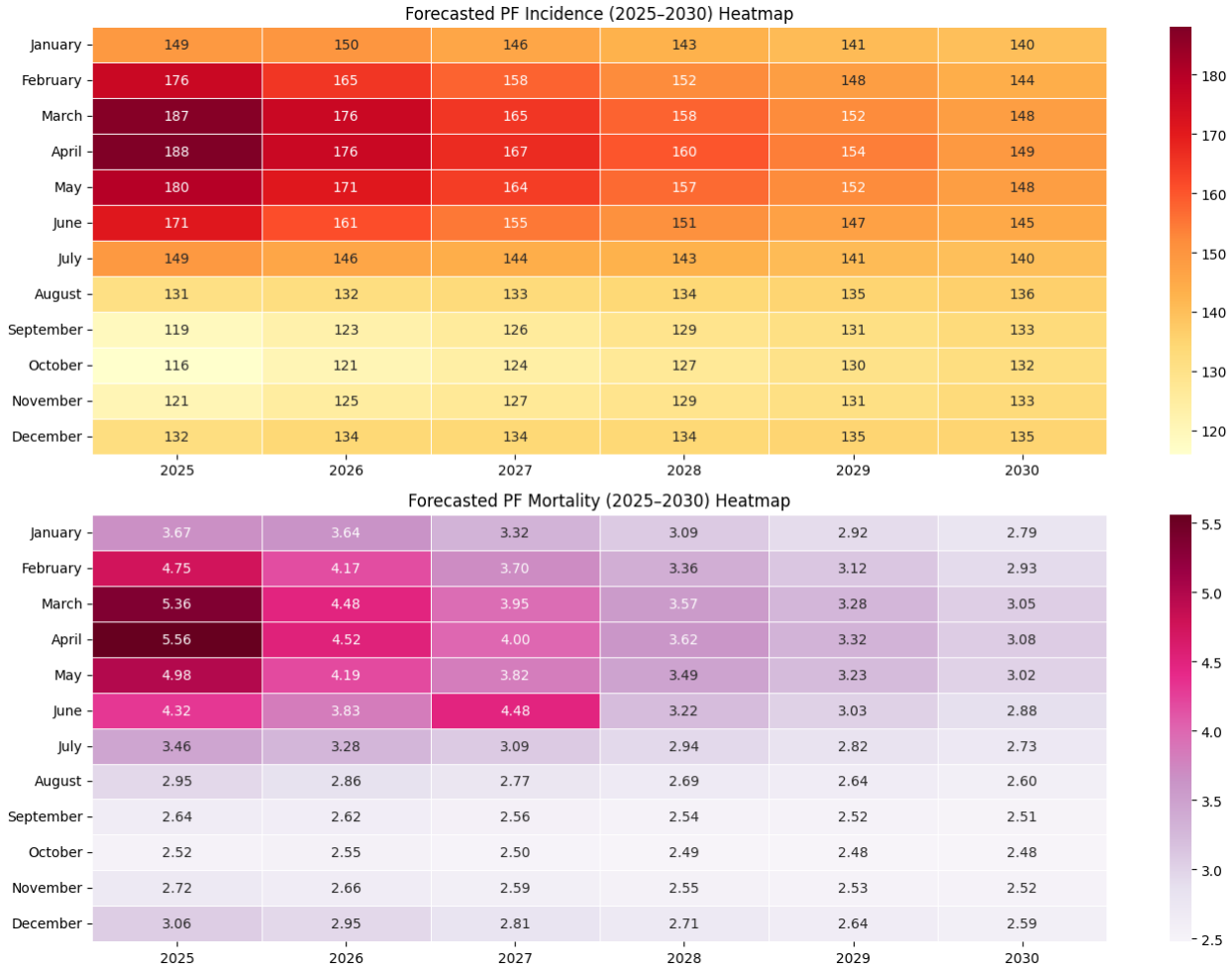
4.8.1 Statistical and Predictive Insights to MoHCC, DHA and DHO

Figure 4.8.4 Horizontal bar chart of Forecasted PF Incidence and Mortality (2025-230)



The horizontal bar chart shown by figure 4.8.4 illustrates that, even by 2030 the Mt Darwin district is expected to face over 1700 malaria cases annually underlining the need for sustainable control measures to be done by MoHCC and DHA while mortality is expected to fall below 40 deaths per year. The MoHCC and DHA should consider action for timely distribution of mosquito nets, having a reliable adequate medicine supply chain of antimalarial drugs from NatPharm, and malaria health campaigns each pre-rainy season.

Figure 4.8.5 Forecasted Plasmodium Falciparum Heat-map Charts (2025-2030)



The time series heatmaps of the forecasted Plasmodium falciparum incidence and mortality shown by the figure above offer a visual summary of seasonal malaria patterns from 2025 to 2030. These heatmaps highlight months of consistently high transmission, especially from February to April of 2025,2026 and 2027. This will allow health authorities to forestall pressure points in the healthcare system. For decision like DHA, MoHCC, and NatPharm, this visualization shows complex forecasts into actionable windows for intervention whether it's scaling up diagnostic efforts, allocating frontline health workers, and distribution of medicines and preventive supplies.

4.9Chapter Summary

This chapter outlined data presentation, analysis, and discussion of results of the forecast on plasmodium falciparum malaria in the district of Mt Darwin. The best performing model which was found to be the CNN+LSTM+FFNN hybrid model was utilized. Visualizations which were done in this chapter provides health predictive insights to the DHA, MoHCC and the National pharmaceuticals company (Natpharm). The following chapter will be on the research recommendations and conclusion.

CHAPTER FIVE

FINDINGS, CONCLUSIONS AND RECOMMENDATIONS

5.0 Introduction

This chapter provides a comprehensive summary of the research study focused on time series forecasting of Plasmodium falciparum malaria epidemics in Mt. Darwin District. The study utilized an integrated artificial neural network hybrid modeling approach to come up with accurate epidemic predictions. In addition to highlighting the key findings, the chapter presents practical recommendations for the district health authorities, the Ministry of Health and Child Care (MoHCC), and other interested stakeholders. It also outlines directions for future research, all aimed at enhancing malaria surveillance, preparedness, and response strategies.

5.1 Summary of Study

This research study focused on forecasting Plasmodium falciparum malaria epidemics in Mt. Darwin District using an integrated artificial neural network (IANN) hybrid modeling approach. The study compared the effectiveness of traditional time series forecasting techniques with modern, data driven computational intelligence models in predicting both malaria incidence and mortality. While traditional models such as ARIMA were initially applied specifically ARIMA (2,1,3) for incidence and ARIMA (1,1,1) for mortality they demonstrated clear limitations. Although ARIMA models were able to reflect general trends and seasonal peaks commonly observed during the rainy season, typically from December to April, they struggled to capture sudden fluctuations in case numbers and consistently underperformed in forecasting months with low mortality. These shortcomings were evident in their relatively poor performance evaluation results compared to those of the neural network based hybrid models.

In contrast, the introduction of the IANN hybrid framework offered a stronger and adaptive approach. By integrating convolutional layers to detect local outbreak signals, recurrent layers to account for time dependent behavior, and fully connected layers for decision mapping, the hybrid model significantly improved forecast accuracy to nearly 94%. This was especially important for anticipating sharp increases in cases during peak transmission periods and for identifying gradual declines in mortality over the years. The enhanced performance of the IANN hybrid not only addressed the weaknesses of traditional models but also provided more reliable forecasts that can support proactive health planning and early intervention of PF malaria in Mt Darwin district.

The time series forecast of Plasmodium falciparum malaria incidence and mortality extended from January 2025 through December 2030, using the CNN+LSTM+FFNN hybrid model identified as the most effective among all models tested in this study. This advanced model outperformed traditional techniques by successfully capturing both short-term fluctuations and long-term patterns in PF malaria epidemic trends and patterns. Meaningful visualizations in chapter 4 played a crucial role in transforming the model's numerical outputs into clear and actionable insights. The heatmap plot, a color graded, month by month grid spanning the forecast horizon enabled a visual risk assessment of the forecasted situation, highlighting high risk periods in a way that allowed the district healthy to easily identify when and where interventions would be most needed.

The forecast exposed that while both PF malaria incidence and mortality are projected to decline gradually over the next 6 coming years a positive indication of progress in malaria control, the regular seasonal surges, particularly during the rainy months, are expected to continue. These findings underscore the need for sustained vigilance and timely interventions despite the overall downward trend. Ultimately, the shift from traditional ARIMA models to a refined data driven computational intelligence hybrid approach proved crucial in capturing the complicated dynamics of malaria epidemic in Zimbabwe's rural settings.

5.2 Conclusions

This research study concludes that integrated artificial neural network modeling offers a superior approach to time series forecasting of PF malaria epidemic trends and pattern in Mt Darwin District. It also acts a baseline for future research in rural communities especially where health systems face resource limitations and need to act on early warnings. This approach also aligns with the national and global goals on malaria elimination, reinforcing the need to integrate advanced computational intelligence methods and data-driven discussions into routine malaria epidemic surveillance and planning.

5.4 Recommendations

In light of the study's findings, it is recommended that the district health authorities, the Ministry of Health and Child Care, and other relevant stakeholders should pay special attention to the December to April rainy season, when malaria transmission peaks, ensuring adequate preparation through in advance medical supplies in strategic collaboration with NatPharm, enhanced vector control, and intensified public health campaigns. Moreover, investment in local data infrastructure including training of health personnel in digital health informatics systems will be vital to sustaining forecasting and future research studies. By embracing these recommendations, the

district and national health systems can significantly strengthen their preparedness and response capacity, contributing meaningfully to WHO and Zimbabwe's broader malaria elimination goals.

5.5 Areas for Further Research

Future research studies should expand this modeling framework to other districts and provinces and include additional environmental and socio economic variables to improve the power of the model. Moreover, comparative research involving other machine learning architectures such as random forest could further enhance overall modeling and forecasting capabilities.

5.6 REFERENCES

Alhassan, S.M., Antwi, S. and Adams, I., 2017. *Time series analysis of malaria incidence in northern Ghana*. Ghana Journal of Science, 57(1), pp.25–38.

Box, G.E.P. and Jenkins, G.M., 1976. *Time Series Analysis: Forecasting and Control*. Revised Edition. San Francisco: Holden-Day.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M., 2016. *Time Series Analysis: Forecasting and Control*. 5th ed. Hoboken: Wiley.

Briar, R., Musa, H., Mathe, T. and Chaka, L., 2020. *Malaria forecasting in resource-constrained settings: A review*. Journal of Health Informatics in Africa, 7(2), pp.47–54.

Brockwell, P.J. and Davis, R.A., 2016. *Introduction to Time Series and Forecasting*. 3rd ed. New York: Springer.

Centers for Disease Control and Prevention, 2023. *Malaria: Biology*. [online] Available at: https://www.cdc.gov/malaria/about/biology/index.html [Accessed 27 May 2025].

Chikoko, D., Mazvimavi, D. and Chikodzi, D., 2021. *Spatial patterns of malaria incidence and health service accessibility in Zimbabwe: A GIS-based study*. African Health Sciences, 21(1), pp.57–69.

DHIS2, 2023. *District Health Information System 2 – Ministry of Health Zimbabwe*. [online] Available at: https://dhis2.org.zw [Accessed 27 May 2025].

Hochreiter, S. and Schmidhuber, J., 1997. *Long Short-Term Memory*. Neural Computation, 9(8), pp.1735–1780.

Hyndman, R.J. and Athanasopoulos, G., 2018. *Forecasting: Principles and Practice*. 2nd ed. Melbourne: OTexts. Available at: https://otexts.com/fpp2/.

Impilo Health Systems, 2024. *Digital Health Information Platform – MoHCC Zimbabwe*. [online] Available at: https://impilo.health.gov.zw [Accessed 27 May 2025].

Kumar, A., Sharma, M. and Katyal, R., 2014. *Forecasting malaria incidence in rural India using ARIMA model*. Journal of Vector Borne Diseases, 51(3), pp.216–222.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. *Deep learning*. Nature, 521(7553), pp.436–444.

Mai, T., Liu, Y. and Nguyen, T., 2021. *Hybrid deep learning models for disease prediction: A case study on dengue*. Computers in Biology and Medicine, 138, p.104897.

Mapuwei, T., Oliver, B. and Mwambi, H., 2020. *Univariate Time Series Analysis of Short-Term Forecasting Horizons Using Artificial Neural Networks: The Case of Public Ambulance Emergency Preparedness*. Journal of Applied Mathematics, 6(2), pp.11.

Mapuwei, T., Tsododo, M. and Gondo, C, 2022. Demand Modelling of Alcoholic Beverages in Manicaland Province Using Time Series Analysis. International Journal of Innovative Science, Engineering and Technology, Vol. 3 Issue 10.

Mavundla, T.R., Gora, P. and Moyo, S., 2020. *Early warning systems for malaria in Southern Africa: A review of approaches*. South African Medical Journal, 110(12), pp.1187–1192.

Muller, D., Smith, K. and Mavengano, F., 2019. *Forecasting malaria epidemics using spatiotemporal modeling techniques: Evidence from Eastern Zimbabwe*. Tropical Medicine and International Health, 24(11), pp.1316–1325.

Mutambara, J., Mugwagwa, N. and Zhou, M., 2019. *Barriers to malaria elimination in rural Zimbabwe: Lessons from the field*. Malaria Journal, 18(1), p.123.

Russell, S.J. and Norvig, P., 2016. *Artificial Intelligence: A Modern Approach*. 3rd ed. Pearson Education.

Weigend, A.S. and Gershenfeld, N.A., 1994. *Time Series Prediction: Forecasting the Future and Understanding the Past*. Redwood City: Addison-Wesley.

World Health Organization, 2015. *World Malaria Report 2015*. Geneva: WHO Press.

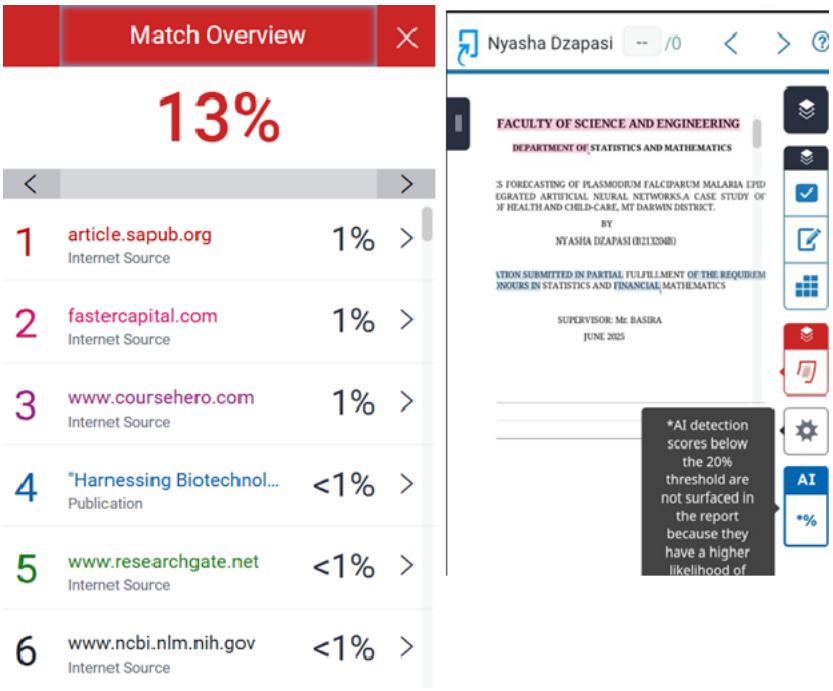
World Health Organization, 2023. *World Malaria Report 2023*. Geneva: WHO. Available at: https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2023 [Accessed 27 May 2025].

Wang, L. and Li, X., 2020. *Hybrid deep learning architecture for predicting epidemic trends*. International Journal of Forecasting, 36(2), pp.867–879.

Zhang, Z., Liu, Q., Li, R. and Zhao, Y., 2021. *Deep learning-based hybrid models for infectious disease forecasting: An application to malaria*. Computers in Biology and Medicine, 134, p.104504.

APPENDICES

TURNIT IN REPORTS



ARIMA Python Code

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import warnings
import statsmodels.api as sm
from statsmodels.tsa.arima.model import ARIMA
from itertools import product
warnings.filterwarnings("ignore")
df = pd.read_csv("/content/drive/MyDrive/malaria_data.csv", parse_dates=['Date'])
df.set_index('Date', inplace=True)
train_df = df.loc['2013-01-01':'2023-12-31']
incidence = train_df['Pf_Incidence']
mortality = train_df['Pf_Mortality']
# ================================
def arima_grid_search(series, label='Series', p_range=range(0, 4), d_range=range(0, 2),
q_range=range(0, 4)):
    results = []
    for order in product(p_range, d_range, q_range):
        try:
            model = ARIMA(series, order=order).fit()
            results.append((order, model.aic))
        except:
            continue

   results_df = pd.DataFrame(results, columns=["ARIMA_order", "AIC"]).sort_values(by='AIC')
    best_model = results_df.iloc[0]

    print(f"\n🔍 Best ARIMA for {label}: {best_model['ARIMA_order']} with AIC =
{best_model['AIC']:.2f}")
    return results_df, best_model

# ================================

inc_results, inc_best = arima_grid_search(incidence, label="Incidence")
mort_results, mort_best = arima_grid_search(mortality, label="Mortality")
# ================================

print("\n📊 ARIMA Model AIC Results - Incidence")
print(inc_results.head(10))

print("\n📊 ARIMA Model AIC Results - Mortality")
print(mort_results.head(10))

# ================================

final_inc_model = ARIMA(incidence, order=inc_best['ARIMA_order']).fit()
final_mort_model = ARIMA(mortality, order=mort_best['ARIMA_order']).fit()

# Optional: Plot residuals
final_inc_model.plot_diagnostics(figsize=(10, 6))
plt.suptitle("Diagnostics - Best Incidence ARIMA")
plt.show()

final_mort_model.plot_diagnostics(figsize=(10, 6))
plt.suptitle("Diagnostics - Best Mortality ARIMA")
plt.show()
```

```
# ARIMA Model Evaluation and Forecasting for Malaria Incidence and
Mortality

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.metrics import mean_absolute_error, mean_squared_error,
r2_score
import warnings
from statsmodels.tsa.arima.model import ARIMA

warnings.filterwarnings("ignore")

df = pd.read_csv("/content/drive/MyDrive/malaria_data.csv",
parse_dates=['Date'])
df.set_index('Date', inplace=True)
df = df.sort_index()

# Split into training (2013-2023) and test (2024)
train = df.loc['2013-01-01':'2023-12-31']
test  = df.loc['2024-01-01':'2024-12-31']

model_inc = ARIMA(train['Pf_Incidence'], order=(2,1,3)).fit()
model_mort= ARIMA(train['Pf_Mortality'], order=(1,1,1)).fit()

dates_2024 = test.index
fore_inc_2024 = model_inc.forecast(steps=12)
fore_inc_2024.index = dates_2024
fore_mort_2024 = model_mort.forecast(steps=12)
fore_mort_2024.index = dates_2024

df_2024 = test.copy()
df_2024['Inc_Forecast']  = fore_inc_2024
df_2024['Mort_Forecast'] = fore_mort_2024
metrics = {
    'Metric': ['MAE','RMSE','R2'],
    'Incidence': [
        mean_absolute_error(df_2024['Pf_Incidence'],
df_2024['Inc_Forecast']),
        mean_squared_error(df_2024['Pf_Incidence'],
df_2024['Inc_Forecast']), #,squared=False),
        r2_score(df_2024['Pf_Incidence'], df_2024['Inc_Forecast'])
    ],
    'Mortality': [
        mean_absolute_error(df_2024['Pf_Mortality'],
df_2024['Mort_Forecast']),
```

```
        mean_squared_error(df_2024['Pf_Mortality'],
df_2024['Mort_Forecast']),),
        r2_score(df_2024['Pf_Mortality'], df_2024['Mort_Forecast'])
    ]
}
metrics_df = pd.DataFrame(metrics)
print("\nARIMA 2024 Performance Metrics:")
print(metrics_df)

# Plot Actual vs Forecast (Monthly)
plt.figure(figsize=(12,6))
plt.plot(df_2024.index, df_2024['Pf_Incidence'], label='Actual Incidence')
plt.plot(df_2024.index, df_2024['Inc_Forecast'], label='Forecast
Incidence')
plt.plot(df_2024.index, df_2024['Pf_Mortality'], label='Actual Mortality')
plt.plot(df_2024.index, df_2024['Mort_Forecast'], label='Forecast
Mortality')
plt.title('ARIMA Model: Actual vs Forecast (2024)')
plt.xlabel('Month')
plt.ylabel('Count')
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()
fore_mort_2024.index = dates_2024
```

IANN Python

```
# CNN + LSTM + FFNN Hybrid Model for Malaria Incidence and Mortality
Forecasting

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import mean_squared_error, mean_absolute_error,
r2_score
from tensorflow.keras.models import Model
from tensorflow.keras.layers import Input, Dense, LSTM, Conv1D,
MaxPooling1D, Flatten, Concatenate
from tensorflow.keras.callbacks import EarlyStopping
import datetime
df = pd.read_csv("/content/drive/MyDrive/malaria_data.csv",
parse_dates=['Date'])
df.set_index('Date', inplace=True)
df = df.sort_index()
```

```
df = df.loc['2013-01':'2024-12']

# Normalize features
scaler = MinMaxScaler()
df_scaled = pd.DataFrame(scaler.fit_transform(df), columns=df.columns,
index=df.index)
def create_dual_inputs(data, seq_length):
    X_seq, X_ffnn, y = [], [], []
    for i in range(len(data) - seq_length):
        X_seq.append(data[i:i+seq_length])
        X_ffnn.append(data[i+seq_length - 1])
        y.append(data[i+seq_length])
    return np.array(X_seq), np.array(X_ffnn), np.array(y)

sequence_length = 12
X_seq, X_ffnn, y = create_dual_inputs(df_scaled.values, sequence_length)

train_end_loc = df_scaled.index.get_loc('2023-12')

if isinstance(train_end_loc, slice):
    train_end = train_end_loc.stop -1
else:
    train_end = train_end_loc

train_end_index =
df_scaled.index.get_indexer([df_scaled.index[train_end]])[0]
X_seq_train, X_seq_test = X_seq[:train_end_index - sequence_length + 1],
X_seq[train_end_index - sequence_length + 1:]
X_ffnn_train, X_ffnn_test = X_ffnn[:train_end_index - sequence_length +
1], X_ffnn[train_end_index - sequence_length + 1:]
y_train, y_test = y[:train_end_index - sequence_length + 1],
y[train_end_index - sequence_length + 1:]

# ------------------ Model Architecture ------------------ #

# CNN + LSTM branch
input_seq = Input(shape=(sequence_length, 2))
x = Conv1D(filters=64, kernel_size=5, activation='relu')(input_seq)
x = MaxPooling1D(pool_size=2)(x)
x = LSTM(64, return_sequences=False)(x)
x = Dense(32, activation='relu')(x)

# FFNN branch
input_ffnn = Input(shape=(2,))
y_ff = Dense(32, activation='relu')(input_ffnn)
y_ff = Dense(16, activation='relu')(y_ff)
```

```
# Merge
merged = Concatenate()([x, y_ff])
z = Dense(32, activation='relu')(merged)
z = Dense(16, activation='relu')(z)
output = Dense(2)(z)

model = Model(inputs=[input_seq, input_ffnn], outputs=output)
model.compile(optimizer='adam', loss='mse')
model.summary()

# ------------------ Model Training ------------------ #
early_stop = EarlyStopping(monitor='val_loss', patience=10,
restore_best_weights=True)
model.fit([X_seq_train, X_ffnn_train], y_train, epochs=500, batch_size=16,
validation_split=0.2, callbacks=[early_stop], verbose=1)

# ------------------ Model Testing ------------------ #
y_pred = model.predict([X_seq_test, X_ffnn_test])
y_pred_inverse =
scaler.inverse_transform(np.hstack((np.zeros((len(y_pred), df.shape[1] -
2)), y_pred)))[:, -2:]
y_test_inverse =
scaler.inverse_transform(np.hstack((np.zeros((len(y_test), df.shape[1] -
2)), y_test)))[:, -2:]
mae = mean_absolute_error(y_test_inverse, y_pred_inverse)
rmse = mean_squared_error(y_test_inverse, y_pred_inverse) #,
r2 = r2_score(y_test_inverse, y_pred_inverse)

print("\nModel Evaluation (2024):")
print(f"MAE: {mae:.2f}")
print(f"RMSE: {rmse:.2f}")
print(f"R^2 Score: {r2:.2f}")
dates_2024 = df.index[-12:]
plt.figure(figsize=(12, 6))
plt.plot(dates_2024, y_test_inverse[:, 0], label='Actual Incidence')
plt.plot(dates_2024, y_pred_inverse[:, 0], label='Predicted Incidence')
plt.plot(dates_2024, y_test_inverse[:, 1], label='Actual Mortality')
plt.plot(dates_2024, y_pred_inverse[:, 1], label='Predicted Mortality')
plt.legend()
plt.title("Actual vs Predicted Malaria Incidence and Mortality (2024)")
plt.xlabel("Month")
plt.ylabel("Count")
plt.grid(True)
plt.tight_layout()
plt.show()
```

```
# ------------------ Forecasting: 2025-2030 ------------------ #
def forecast_future(model, history, steps):
    forecast = []
    current_seq = history.copy()
    for _ in range(steps):
        ffnn_input = current_seq[-1]
        pred = model.predict([current_seq[np.newaxis, :, :],
ffnn_input[np.newaxis, :]])[0]
        forecast.append(pred)
        current_seq = np.vstack([current_seq[1:], pred])
    return np.array(forecast)

last_seq = df_scaled.values[-12:]
future_preds = forecast_future(model, last_seq, 72)
future_padded = np.hstack((np.zeros((future_preds.shape[0], df.shape[1] -
2)), future_preds))
future_unscaled = scaler.inverse_transform(future_padded)[:, -2:]

future_dates = pd.date_range(start='2025-01-01', periods=72, freq='MS')
future_df = pd.DataFrame(future_unscaled,
columns=['Pf_Incidence_Forecast', 'Pf_Mortality_Forecast'],
index=future_dates)
future_df.plot(figsize=(12, 6), title='Forecasted Monthly Malaria
Incidence and Mortality (2025-2030)', grid=True)
plt.ylabel("Count")
plt.tight_layout()
plt.show()

yearly_summary = future_df.resample('Y').sum()
print("\nYearly Forecast Summary (2025-2030):")
print(yearly_summary)

yearly_summary.plot(kind='bar', figsize=(10, 5), title='Yearly Forecasted
Incidence and Mortality', rot=45, grid=True)
plt.ylabel("Total Count")
plt.tight_layout()
plt.show()
```

DATA SOURCE LINK

https://apps.mohcc.gov.zw/impilo-dhis/dhis-web-commons/security/login.action

Nyashadzapasi b213204b