HAWASSA UNIVERSITY

Predicting malaria incidence using case load and metrological data in Sidama Regional State

MSc Thesis

BY

Eyoel Nebiyu

HAWASSA, ETHIOPIA

November, 2024

HAWASSA UNIVERSITY

HAWASSA UNIVERSITY INSTITUTE OF TECHNOLOGY

SCHOOL OF POST GRADUATE STUDIES

FACULTY OF COMPUTER SCIENCE

THESIS ON

Predicting malaria incidence using case load and metrological data in Sidama Regional State

BY

Eyoel Nebiyu

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

Advisor: Dr. Degif Teka **(PhD)**

HAWASSA, ETHIOPIA

October, 2024

DECLARATION

I declare that this thesis is my original work and has not been submitted as a partial requirement for a degree in any university. All the resources used for this thesis work are cited and acknowledged.

Name: Eyoel Nebiyu

Signature: _____

Date: _____

HAWASSA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

ADVISORS' APPROVAL SHEET

This is to certify that the thesis entitled "***Predicting malaria incidence using case load and metrological data in Sidama Regional State***" submitted in the partial fulfillment of the requirements for the degree of Master's in Computer Science, the graduate program of the School of Computer Science, Hawassa University Institute of Technology has been carried out
By Eyoel Nebiyu under our supervision. Therefore, we recommend that the student has fulfilled the requirements and hence hereby can submit the thesis to the faculty for defense.

Dr. Degif Teka	_____	_____
Name of Major Advisor	Signature	Date

HAWASSA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

EXAMINERS' APPROVAL SHEET

We, the undersigned, members of the Board of Examiners of the final open defense by Eyoel Nebiyu have read and evaluated his/her thesis entitled "***Predicting malaria incidence using case load and metrological data in Sidama Regional State***" , and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirements for the degree.

Name of Major Advisor	Signature	Date
Andargachew Mekonnen		02/12/2024
Name of Internal Examiner-I	Signature	Date
Name of Internal Examiner-II	Signature	Date
Siraj Sebhatu(Ph.D)		29/11/2024
Name of External examiner	Signature	Date
SGS Approval	Signature	Date

Final approval and acceptance of the thesis is contingent upon the submission of the final copy of the thesis to the School of Graduate Studies (SGS) through the Department/School Graduate Committee (DGC/SGC) of the candidate's department.

Stamp of SGS Date: _____

Table of Contents

LIST OF TABLE

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to begin by expressing my heartfelt gratitude to Almighty God and his mother saint Marry for all the blessings in my life, especially as I embarked on this research journey. I am deeply thankful to my advisor, Dr. Degif Teka, for his invaluable feedback and guidance from the initial stages to the finalization of this thesis. My appreciation also extends to the Hawassa University Institute of Technology, the instructors, and the entire Computer Science Department for their support, which played a crucial role in the success of my research project. I would like to acknowledge the Sidama Region Public Health Institution for providing malaria case data and the Hawassa Metrology Agency for supplying the meteorological data essential for my study. Finally, I would like to sincerely thank my parents, Mr. Nebiyu Melaku and Dr. Achamyelesh Gebretsadik, for their kindness, constructive criticism, invaluable advice, and unwavering support throughout my research. Additionally, I am grateful to my friends for their expertise and assistance during this process.

ACRONYMS

ANN.....Artificial Neural Network

AI.....Artificial Intelligence

CSV.....Comma-Separated Values

CV.....Cross Validation

KNN.....K-Nearest Neighbor

ML.....Machine Learning

MAE.....Mean Absolute Error

MSE.....Mean Squared Error

NB.....Naïve Bayes

NDVI.....Normalized Difference Vegetation Index

R2.....R- Squared

RF.....Random Forest

ReLU.....Rectified Linear Units

RMSE.....Root Mean Squared Error

SVM.....Support Vector Machine

WHO.....World Health Organization

XGBoost.....Extreme gradient boosting

ABSTRACT

Malaria remains a significant public health challenge, particularly in tropical regions like Ethiopia's Sidama Regional State, where climatic factors heavily influence transmission dynamics. This study utilizes an ANN feed forward model to integrate meteorological data (minimum temperature, maximum temperature, and rainfall) with historical malaria case records (2017–2022) to construct a predictive model for malaria incidence. Data were obtained from the Ethiopian National Meteorological Agency and the Sidama Regional Health Bureau. Four districts Boricha, Dale, Hawassa Zuria, and Shebedino were used to validate the model. To determine the most effective machine learning technique for malaria prediction, this study compared the ANN feed forward model with Random Forest and Decision Tree models. Among these, the ANN feed forward model demonstrated superior predictive accuracy, achieving the lowest RMSE values across districts, with Shebedino (0.4787) and Hawassa Zuria (0.7359) performing best. However, challenges remain in capturing short-term fluctuations, particularly in Boricha (RMSE: 2.610).The results emphasize the importance of incorporating meteorological factors into malaria prediction models and highlight the ANN model's potential as a robust early warning system. By enabling public health officials to forecast outbreaks and allocate resources more effectively, predictive models like ANN can significantly enhance malaria prevention efforts. Future research should focus on improving model accuracy by integrating additional variables and exploring advanced machine learning techniques to handle complex transmission scenarios

CHAPTER 1: INTRODUCTION

1.1 Background

Malaria is primarily seen in tropical areas and is a potentially deadly disease. It can be prevented and treated. However, if a diagnosis is not made promptly and efficiently, a case of uncomplicated malaria can develop into a severe form of the sickness, which can occasionally be fatal if treatment is not received.

Malaria can be obtained through the bite of a female Anopheles mosquito, but it is not contagious and cannot spread from person to person. The two parasite species that pose the biggest risk among the five that can cause malaria in humans are Plasmodium falciparum and Plasmodium vivax. Of the more than 400 species of Anopheles mosquitoes, about 40 are thought to be vector species and have the ability to transmit disease[1].

Malaria continues to pose a serious threat to world health, especially in tropical and subtropical areas. Fever, chills, and a flu-like illness are some of the symptoms of this condition, which in extreme situations can cause complications and even death. Insecticide-treated bed nets, indoor residual spraying, and antimalarial drugs have all been used in the fight against malaria. In recent decades, the prevalence of malaria has declined due to effective control measures [2].

According to the latest World Malaria Report, there was a slight increase in global malaria cases, with numbers rising from 245 million in 2020 to 247 million in 2021. In Ethiopia, the battle against malaria continues to be challenging, as evident from the 2.9 million confirmed cases of malaria and 4,782 recorded deaths in 2019. What's particularly concerning is the recent surge in cases within Ethiopia, with a notable 32.5% increase in confirmed malaria cases between 2021 and 2022, jumping from 1.1 million to 1.5 million. The situation has taken a more alarming turn in 2023, with a staggering 150% increase in recorded malaria cases compared to the same periods in 2021, and a 120% increase over 2022. This underscores the urgent need for enhanced malaria control and prevention efforts in the region. [3].

In order to implement more effective control measures, there is an increasing need for methods that allow for forecasting, early warning, and timely case detection in areas of unstable transmission, such as the African highlands, due to the severe health effects of malaria epidemics [4]. Research on malaria epidemics in these regions has linked the disease to high temperatures, excessive precipitation, and a density of vegetation indicated by the normalized difference

vegetation index (NDVI). This is demonstrated by the clear relationship between rainfall and Anopheles mosquito abundance [5], increased transmission and temperature and vegetation density [6], and seasonality of malaria. As a result, estimating the incidence of malaria cases using data from weekly case reports is a crucial step in monitoring and controlling the spread of the illness in the area.

Machine learning models have proven to be immensely valuable in predicting malaria outbreaks and understanding disease dynamics. These models utilize a range of input features, including climate data, environmental factors, and historical disease patterns, to make accurate predictions. By processing large datasets and identifying complex relationships, machine learning algorithms can forecast areas at high risk of malaria transmission, allowing for targeted interventions. Additionally, machine learning enables early detection of outbreaks, facilitating timely response and resource allocation. The ability to adapt and learn from new data further enhances the models' effectiveness in an ever-changing disease landscape. By using the power of machine learning, public health authorities can take proactive measures to mitigate the impact of malaria, ultimately contributing to the global effort to reduce its burden on affected populations [4].

1.2 Motivation

The battle against malaria persists as a formidable challenge, particularly in regions like Ethiopia, where recent surges in malaria cases underscore the pressing need for more effective control and prevention strategies. In response to this urgent demand, emerging technologies such as ANN, specifically the feed forward type, offer a promising avenue for intervention. By harnessing the power of ANN and leveraging meteorological data, including temperature, humidity, and rainfall patterns, these advanced predictive models can significantly enhance our ability to anticipate and mitigate malaria outbreaks.

Through the integration of sophisticated ANN algorithms with comprehensive meteorological datasets, public health authorities gain invaluable insights into the complex dynamics of malaria transmission. By analyzing historical weather patterns alongside epidemiological data, these models can identify key environmental factors that contribute to malaria proliferation, allowing for proactive interventions and targeted control measures. Moreover, the predictive capabilities of ANN enable health officials to forecast future malaria trends with greater accuracy, empowering

decision-makers to allocate resources more efficiently and implement timely interventions to curb the spread of the disease.

Furthermore, the deployment of ANN-based predictive models within existing public health systems holds immense potential for the establishment of early warning systems tailored to specific geographical regions. By continuously monitoring meteorological conditions and integrating real-time disease surveillance data, these systems can provide timely alerts and actionable insights to frontline healthcare workers, enabling them to implement preventive measures and allocate resources preemptively to high-risk areas. Additionally, the integration of ANN-driven predictive analytics into malaria control programs facilitates adaptive response strategies, allowing for agile adjustments in intervention strategies based on evolving epidemiological trends and environmental conditions.

In essence, the utilization of ANN feed forward models in conjunction with meteorological data represents a pivotal advancement in malaria prediction and control efforts. By leveraging the predictive capabilities of ANN algorithms and harnessing the wealth of information contained within meteorological datasets, we can enhance our capacity to preemptively identify and mitigate malaria outbreaks, ultimately advancing the global fight against this deadly disease.

1.3 Statement of the Problem

According to the WHO's most recent report on malaria, there are estimated to be 245 million cases and 627 000 deaths from the disease globally in the year 2020. This amounts to 69 000 more fatalities and approximately 14 million more cases in 2020 compared to 2019. Inadequate malaria prevention, diagnosis, and treatment during the pandemic were responsible for about 27% of these excess deaths (47 000). [5]

There are over 200 million cases of malaria worldwide each year, and there are over 500,000 fatalities. In Sub-Saharan Africa, malaria accounts for about 90% of all fatalities. Malaria affects children under the age of five more than other age groups. Malaria claims a youngster every two minutes or so [5].

Malaria is a major public health and economic issue in Ethiopia. Depending on altitude, rainfall patterns, and climate, the distribution differs from place to place. One of the main causes of sickness and mortality in the nation, it is a serious problem. Malaria affects 75% of Ethiopia's terrain areas below 2000 meters above sea level.The major epidemics occur cyclically in every 5–

8 years in Ethiopia, but focal epidemics are occurring every year. Ethiopia has a population of above 127 million, of which approximately 68% of the population is at risk of the disease. About 2.9 million cases of malaria and 4,782,000 related deaths have been reported annually, and the rate of morbidity and mortality dramatically increases during epidemics [6].

As per the malaria expert working in Sidama regional health bureau malaria cases currently increasing continuously since the last three years. A study done in Boricha district showed that the rate of asymptomatic malaria prevalence was 6.1% [7]

Malaria remains a significant public health concern, particularly in regions like the Rift Valley area, where the disease poses a substantial threat to the local population. The Rift Valley region is characterized by a unique set of environmental and climatic factors that influence the prevalence and distribution of malaria. Despite efforts to combat the disease, the region continues to experience high malaria transmission rates, impacting the health and well-being of its residents.

The problem at hand is the need for effective and timely malaria prediction in the Rift Valley area. Predictive models that utilize climate information can offer valuable insights into the spatiotemporal dynamics of malaria transmission. However, the existing research on malaria prediction in this region is limited[8], [9], and the accuracy of predictions can be further improved [8].

Malaria is a major public health problem in Ethiopia. Forecasting malaria cases is essential for allocating appropriate preventive control measures and eventual elimination strategies. Unfortunately, there are no practical tools to predict malaria epidemics based on climate forecasts. Such tools would be useful in making efficient use of the limited resources for malaria control. It also supports monitoring and evaluation of the disease progress.

This research therefore aims to provide an insight to health workers, policy makers, and decision makers to predict the malaria case load ahead of time .The following research questions can be drawn and addressed in this study.

1. What features are more important for developing efficient prediction model for malaria transmission prediction?
2. What machine learning techniques are most effective in predicting malaria transmission, and how do environmental factors influence the accuracy of these prediction models?

1.4 Objectives of the Study

1.4.1 General Objective

To develop a prediction model using feed forward ANN for malaria incidence in the Sidama region by studying the association between environmental variables and disease dynamics.

1.4.2 Specific Objectives

The specific objectives to attain the general objective and aiming at answering the research question are:

- Review related literature for a better understanding of the problem and research design.
- To Gather data on malaria cases and metrological data
- Preprocess the data by detecting the outliers, missing values and inconsistencies in the data.
- Identify features relevant to improve the accuracy of malaria incidence prediction.
- Design effective machine learning model using feed forward architectures and ANN to forecast the incidence of malaria.
- Assess the model's effectiveness in malaria prediction by evaluating its accuracy through evaluation metrics

1.5 Scope and Limitation of the Study

This study focuses on the analysis of meteorological and clinical malaria case data from 2017 to 2022. It involves the utilization of weekly malaria case reports from the Sidama Regional Health Bureau and concurrent meteorological data from the Ethiopian National Meteorology Agency for the same period. The primary objective is to develop a predictive models to forecast malaria incidence, and based on the findings, provide recommendations and insights to inform future actions.

1.6 Methodology

1.6.1 Literature Review

The research has chosen effective methods for malaria prediction using a feed forward ANN model by reviewing a variety of literatures, including books, journals, research papers, and materials connected to the subject. This procedure has successfully identified the issue and comprehended the state of the art for developing prediction models for malaria.

1.6.2 Data Collection

The data gathering necessary to accomplish the research objectives has been finished. This includes collecting weekly data on malaria cases from the Sidama Regional Health Bureau, which covers four different zones in the region, from 2017 to 2022. At the same time, meteorological information encompassing the same time period and zones that was necessary for the investigation was obtained from the Ethiopian National Meteorology Agency.

1.6.3 Design and Development Tools

In the design and development phase, we leverage specialized tools and frameworks to implement ANN models for prediction, particularly suitable for our continuous data. Unlike traditional regression models, ANNs excel at capturing complex nonlinear relationships present in our dataset, offering superior predictive capabilities. This is crucial for our analysis, as our data involves continuous numerical variables with intricate interdependencies. To accomplish this, we utilize a variety of software libraries and platforms tailored for building, training, and evaluating neural networks. Popular options include Tensor Flow, Keras, PyTorch, and scikit-learn, which provide comprehensive APIs for constructing ANN architectures, handling data preprocessing, and optimizing model performance.

1.6.4 Evaluation Technique

During the evaluation phase, various techniques were employed to assess the efficiency and accuracy of the prediction model. The RMSE, MSE, MAE, and R-squared (R2) were among the frequently utilized metrics. By measuring the average difference between expected and actual values, these metrics provide a quantitative understanding of the model's performance. This comprehensive evaluation approach ensured a thorough assessment of the model's strengths and limitations, enhancing the overall robustness of the predictive framework.

1.7 Significance of the Study

The findings of this research have the capacity to significantly influence stakeholders at different levels to take decisive action and lessen the impact of incomplete evidence. Beyond its immediate effects, it provides a useful chance to identify dominant patterns in the burden of disease, allowing for the efficient use of resources and focused interventions. Furthermore, by providing grassroots workers with easily available and useful data-driven insights, this study represents a ray of hope for them. Through the utilization of research findings, policymakers can take a proactive approach

to addressing new health issues and promote readiness and resilience in their communities. Additionally, this study's longitudinal approach makes it easier to develop long-term, evidence-based policies to tackle endemic diseases. This initiative protects public health and cultivates a culture of data-driven innovation through cooperative efforts and well-informed decision-making, providing the foundation for revolutionary change on a local and global level.

1.8 Organization of the Thesis

This section presents an overview on the contents of all chapters. This research work is organized into five different chapters.

A brief introduction to this study is given in the first chapter. It offers the overall framework for this investigation. As a result, it gives the reader enough background knowledge to comprehend the purpose of the study and the goals the researcher hopes to achieve. The chapter gives a summary of the entire research project.

In Chapter Two, literature is reviewed on the concept of different research work using machine learning and deep learning is presented.

In Chapter Three, a detailed description of the proposed system is discussed. It explains the various components integral to the system, encompassing dataset preparation, preprocessing, training, testing and predictive modeling.

In chapter four, experimental evaluation of the proposed model for malaria prediction is described in detail. The proposed model's implementation and the dataset used are both fully detail.

In Chapter five conclusions, future work, and the contributions of this research work are presented.

CHAPTER 2: LITERATURE REVIEW AND RELATED WORKS

This chapter reports review of existing research on malaria prediction, including studies on spatial modeling, seasonal patterns, predictive mapping, and machine learning techniques. It also examines related works conducted in various regions, discussing and reporting different methodologies, findings, and limitations.

2.1 Introduction

This section provides a technical assessment of malaria prediction and current methodologies for utilizing meteorology data to predict malaria occurrences. It offers an analysis of pertinent works relevant to the objective of predicting malaria outbreaks using ANN feed-forward models. Initially, it presents a summary of machine learning techniques, particularly focusing on ANN feed-forward models, and an overview of malaria risk factors associated with meteorological variables. Additionally, it discusses the significance of predicting malaria outbreaks and the potential of machine learning approaches in this context.

2.2 Malaria

Malaria is a potentially deadly disease caused by parasites and transmitted to people by female Anopheles mosquitoes carrying the infection, according to the World Health Organization. It is preventable and treatable. Two of the five parasite species that cause malaria in humans, Plasmodium falciparum and Plasmodium vivax, are the most dangerous. Certain categories of people are more susceptible to malaria than others; these groups include young children under five, pregnant women, persons living with HIV/AIDS, non-immune migrants, people who move about a lot, and travelers[10].

2.3 Malaria Cases

A malaria case occurs when there is evidence of the disease or illness in humans, which is verified by parasitological testing. Malaria cases can be categorized as autochthonous, indigenous, induced, introduced/imported, or relapsing depending on where the illness originated. Until the parasitological confirmation, a suspected case of malaria is not considered confirmed. A malaria case occurs when a proven malaria infection occurs in an area where malaria is being eliminated, regardless of whether symptoms are present. In a malaria control context, a malaria case occurs when a confirmed malaria infection occurs[10].

2.4 Lifecycle of Malaria

Malaria is typically spread by female mosquitoes. Additionally, malaria can be spread by mosquito bites. A parasite enters our bloodstream by the bite of an infected Anopheles mosquito. Additionally, a mosquito that bites an infected person is infected and spreads the parasite. Through the bite, the parasite—which is found in Anopheles saliva—enters the bloodstream. The parasite enters the bloodstream and travels to the liver. Within 48 to 72 hours, the parasite grows and multiplies in the liver. After maturing, the parasite enters the bloodstream and begins infecting blood cells, most often red blood cells. Within two to three days of the parasite entering the red blood cell, it begins to grow, causing a cell to burst. This infection is then spread to additional blood cells. Infected blood cells rupture at regular intervals, adding to the parasite population in the blood. Infected blood cells have a 48–72 hour bursting cycle. A person experiences chills, sweating, and fever whenever their cells burst. Figure 1 depicts the malaria parasite infection cycle. The parasites convince the infected hepatocyte to separate, which allows it to move to the liver sinusoid, where merosomes—vesicles loaded with parasites—sprout (Figure: Label 1). Within erythrocytes, the new merozoites divide quickly, occasionally concurrently during fever and chill cycles (Figure: Label 2). Few parasites differentiate into male and female gametocytes (Figure: Label 3), which are the forms that remain dormant in the circulatory system for a week, in response to an unidentified stimulus. Gametocytes swiftly transition into initiated male and female gametes after entering the mosquito through the bloodstream (Figure 1: Label 4). The ookinete, a transient and mobile diploid parasite frame, exits the bloodstream (Figure: Label 5), crosses the peritrophic lattice, and enters the mid-gut partition, where it forms an oocyst (Figure: Label 6)Numerous sporozoites are formed inside the oocyst following a meiotic reduction in the chromosome (Figure: Label 7). In order to prepare for exchange with the vertebrate host, the sporozoites move to the salivary organ when the oocyst splits. Malaria infections can progress to anemia, cerebral malaria, or hypoglycemia because the thick blood blocks the capillaries that deliver blood. This occurs when the parasite is resistant to drugs or when appropriate medications are unavailable. One of the main causes of lifelong learning difficulties is cerebral malaria, which can also result in coma and even death.

[11], [12].

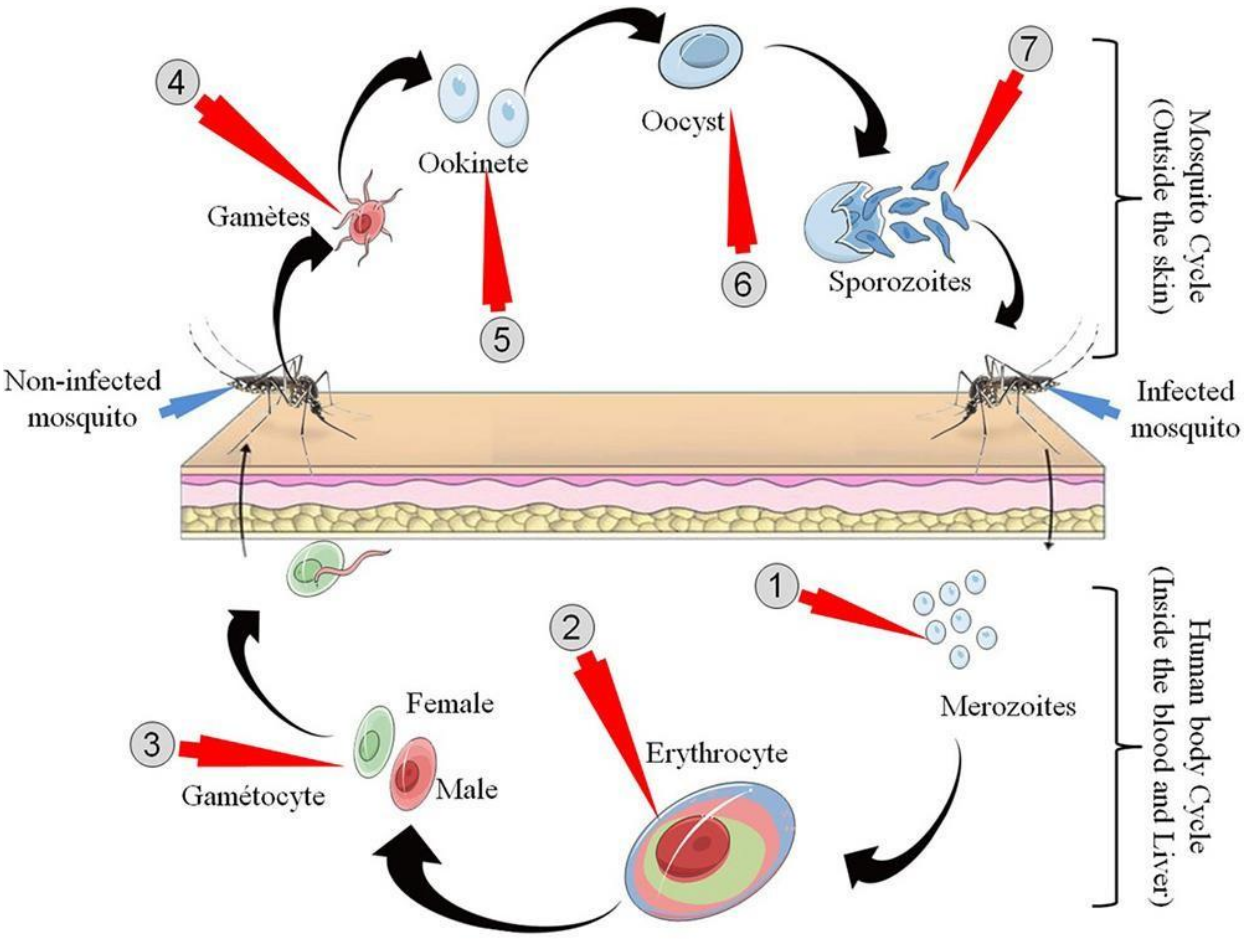

Figure 2.1: Malaria parasite life cycle

2.5 Overview of Malaria Prediction

Malaria remains a significant public health concern, particularly in regions where meteorological conditions play a crucial role in vector breeding and disease transmission. Malaria prediction involves forecasting the occurrence and spread of the disease based on various factors, including meteorological variables such as temperature, humidity, rainfall, and vegetation indices. These environmental factors influence the breeding habitats and life cycle of mosquitoes, the primary vectors of malaria parasites. By analyzing historical malaria data alongside meteorological data, researchers aim to develop predictive models that can anticipate malaria outbreaks and inform preventive measures.

2.6 Overview of Artificial Neural Network (ANN) Feed-Forward Models

In a variety of domains, such as epidemiology and healthcare, ANNs have become highly effective tools for predictive modeling. The architecture and operation of biological neural networks in the human brain served as the paradigm for artificial neural networks ANNs. It is composed of networked nodes, or neurons, arranged in layers. The three main kinds of layers in an ANN are the input layer, hidden layers, and output layer.

Data is entered into the network through the input layer. Generally speaking, the amount of input features supplied to the network equals the number of neurons in the input layer. A particular input value or feature is delivered to every neuron in the input layer.

The layers that lie between the input and output layers are referred to as hidden layers. The network's capacity to represent complex relationships and identify patterns in the incoming data is greatly aided by these layers. The difficulty of the issue being solved determines the design decision regarding the number of neurons in each hidden layer.

Based on the data processed by the earlier levels, the output layer presents the ultimate outcomes or predictions. The type of problem being solved determines how many neurons are in the output layer. Regression tasks usually consist of a single neuron in the output layer, with the objective being to predict a continuous Value. Weights are used to represent the connections between neurons and indicate how strongly two neurons are connected. Every neuron in a layer is coupled to every other layer's neuron in a completely connected feed-forward network. During the training phase, these weights are iteratively modified to maximize the network's performance; initially, they are assigned random values. The network learns from a labeled dataset during training, modifying the weights to reduce the variation between the true labels and the anticipated outputs. Usually, an algorithm known as back propagation is used for this optimization process. It determines the gradient of the network's error and modifies the weights correspondingly.

The network can be utilized in the testing phase with fresh, untested data once it has been trained. The network processes the input data, and the output layer predicts classifications using the relationships and patterns that were learned during training. An artificial neural network's design, which includes the number of layers, the number of neurons in each layer, and the kinds of connections between layers, affects the network's behavior and performance. The behavior of the network is also greatly influenced by the selection of activation functions, which control each neuron's output.

Among the different architectures of ANNs, the feed-forward neural network is widely used for regression and classification tasks. In the context of malaria prediction, feed-forward ANNs can learn complex relationships between meteorological variables and malaria occurrences, enabling accurate forecasting of disease outbreaks. Figure 2.2 shows fully connected multilayer feed-forward neural network.

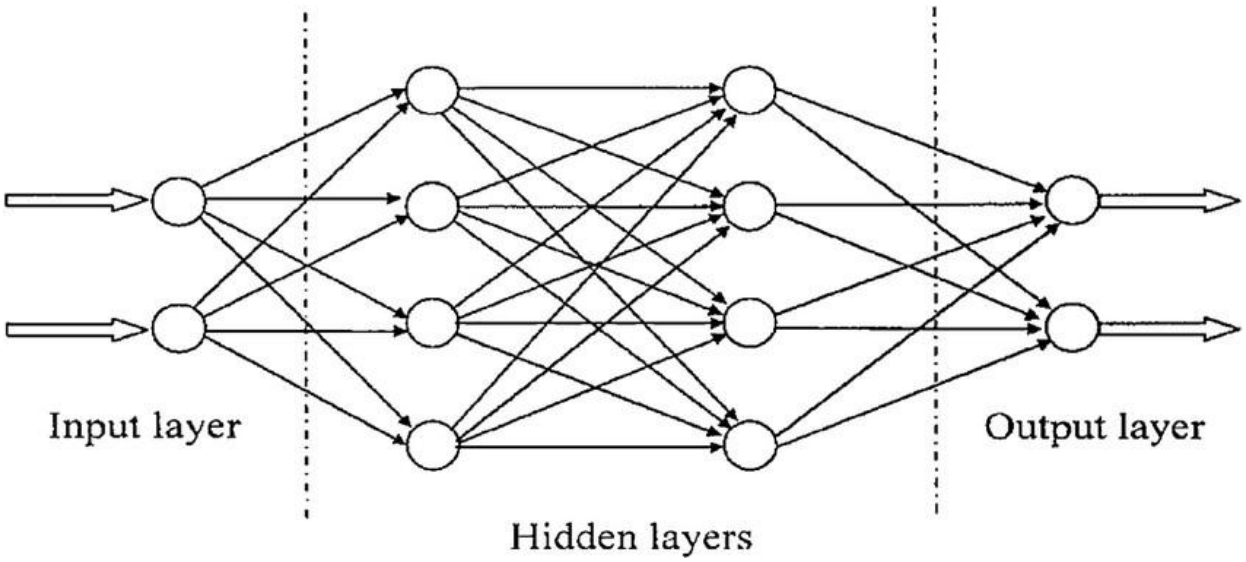

Figure 2.2: Fully connected Multilayer Feed-Forward Neural Network Architecture

2.6.1 Supervised Learning with Feed-Forward ANNs

Supervised learning with feed-forward ANNs involves training the network on labeled input-output pairs, where meteorological variables serve as inputs and malaria occurrence (e.g., presence or absence) serves as the output. During the training process, the network learns to map input meteorological data to the corresponding malaria outcomes, optimizing its internal parameters through back propagation and gradient descent algorithms. By iteratively adjusting the network weights and biases, the ANN can accurately predict malaria occurrences based on meteorological data[13].

2.6.2 Preprocessing Meteorological Data for ANN

Modeling Before training an ANN for malaria prediction, preprocessing of meteorological data is essential to ensure the quality and relevance of input features. This may involve data cleaning, normalization, and feature selection techniques to handle missing values, scale variables, and identify the most informative predictors of malaria outbreaks. Additionally, temporal aggregation

of meteorological data may be performed to capture seasonal patterns and long-term trends that influence malaria transmission dynamics.

2.6.3 Evaluation and Validation of ANN

Models After training, feed-forward ANNs for malaria prediction require thorough evaluation and validation using appropriate performance metrics and validation techniques tailored for prediction tasks. While classification models focus on categorizing data into discrete classes (e.g., presence or absence of malaria), prediction models aim to estimate continuous numerical values (e.g., malaria incidence rates or disease severity).

For prediction models, common evaluation metrics include:

MSE: it measures the average squared difference between the actual and predicted values. Lower MSE values indicate better predictive performance, with a value of 0 indicating perfect predictions.

RMSE: is the square root of the MSE, providing a measure of the average magnitude of prediction errors in the same units as the target variable. Lower RMSE values suggest more accurate predictions.

MAE: it measures the average absolute difference between the actual and predicted values, providing a more interpretable metric compared to MSE. Like MSE and RMSE, lower MAE values indicate better predictive performance.

R2 Score: it represents the proportion of the variance in the target variable that is predictable from the input variables. It ranges from 0 to 1, where higher values indicate better model fit to the data.

In addition to these metrics, cross-validation techniques such as k-fold cross-validation or time-series cross-validation can assess the generalization performance of the prediction model on unseen data. K-fold cross-validation involves splitting the dataset into k subsets (folds), training the model on k-1 folds, and evaluating its performance on the remaining fold. This process is repeated k times, with each fold serving as the validation set once. Time-series cross-validation is suitable for temporal data, where the dataset is split into consecutive time periods for training and validation.

Moreover, sensitivity analysis can assess the model's sensitivity to variations in input variables, helping identify influential factors and potential sources of uncertainty. Uncertainty quantification techniques, such as bootstrapping or Monte Carlo simulation, can provide insights into the variability of predictions and estimate confidence intervals for predicted outcomes.

By employing these evaluation and validation techniques, researchers can assess the predictive performance, robustness, and reliability of feed-forward ANN models for malaria prediction, enabling informed decision-making and intervention planning in malaria-endemic regions.
By synthesizing the knowledge from these key areas, researchers can develop robust ANN-based predictive models for malaria outbreaks, leveraging meteorological data to enhance the accuracy and timeliness of malaria risk assessments and interventions.

2.7 Approaches to prediction Algorithms

Depending on the type of data and the issue at hand, there are numerous approaches and applications for different prediction-making techniques. Several popular techniques include:

2.7.1 Random Forest

Random Forest is a popular ensemble learning method for applications involving both regression and classification. During the training phase, the approach constructs numerous decision trees and aggregates their outputs to arrive at a final prediction. Whereas the average of all the trees' predictions is used for regression, the ultimate decision for classification is determined by the majority vote of all the trees. This method aids in getting over the drawbacks of a single decision tree, like high variance and over fitting[14].

The two main strategies used by Random Forest bagging, or bootstrap aggregation, and feature randomness are what give it its strength. By using replacement sampling to create several subsets of the training data, bagging entails creating a decision tree for each subset. By taking into account only a random subset of characteristics at each split in the tree, feature randomness increases diversity even more. This reduces correlation between individual trees, leading to a more robust model that is less likely to over fit [15].

Random Forest's versatility in a range of applications lies from its capacity to handle big datasets with high dimensionality, which is one of its key advantages. It can also reveal feature relevance by highlighting the factors that have the greatest influence on the forecast. Nevertheless, even with its benefits, Random Forest can be computationally demanding and slower to produce results, particularly when working with a big number of trees. In addition, compared to a single decision tree, the model's outcomes can be more challenging to understand even when they are extremely accurate.

Here is an example of previous literature conducted by researchers using Random Forest.

Forecasting Severe Weather with Random Forests

Hill, Herman, et al. [16] Addresses the application of Random Forest (RF) methodology for generating calibrated probabilistic forecasts of severe weather across the contiguous United States (CONUS). The study utilized nine years of historical forecast data from April 2003 to April 2012, collected from NOAA's Second Generation Global Ensemble Forecast System (GEFS/R). The data was split into training and testing sets, where 80% was used for training the models and 20% for testing. The RF models were tested over unseen data from April 2012 to December 2016. The study demonstrates that the RF models exhibit impressive probabilistic forecast skill, significantly outperforming equivalent Storm Prediction Center (SPC) outlooks for Days 2 and 3, as well as for significant severe events on Day 1. The overall Brier Skill Scores (BSS) indicate that the RF model achieved a score of 0.105 for severe wind, 0.079 for hail, and 0.029 for tornadoes. The skill decreases with increasing lead time, with BSS values of 0.108 and 0.089 for Day 2 and Day 3 RF outlooks, respectively. Notably, a weighted blend of RF and SPC outlooks significantly outperformed the SPC forecasts across all phenomena and lead times.

2.7.2 Support Vector Machines (SVM)

Support Vector Machines (SVM) represent a prominent algorithm in machine learning, primarily utilized for classification tasks, SVM can also be applied to regression problems through a variant known as Support Vector Regression. Both linear and non-linear classification problems respond well to this method. The main purpose of an SVM is to determine the best hyper plane inside the feature space that efficiently divides distinct classes of data points. The margin, or the distance between the hyper plane and the nearest data points from each class, is maximized while choosing this hyper plane, which is an important decision. SVMs aim to improve generalization by maximizing this margin, which increases the model's ability to correctly classify new, unseen cases. Support Vector Machines (SVM) do not use activation functions like neural networks. Instead, they focus on finding an optimal hyper plane that maximizes the margin between different classes of data points through mathematical optimization. In linear SVMs, the data is linearly separable, allowing the model to utilize a hyper plane directly. For non-linear classification tasks,

SVMs employ the kernel trick to transform the input data into a higher-dimensional space, enabling effective separation without the need for activation functions [17].

SVMs are effective at solving classification problems when the data points are linearly separable. However, they can also be applied to non-linear classification tasks. The original data can be mapped using SVMs into a higher-dimensional feature space where it may become linearly separable, thanks to the kernel trick. Support Vector Machines (SVM) do not use activation functions like neural networks. Instead, they focus on finding an optimal hyper plane that maximizes the margin between different classes of data points through mathematical optimization. In linear SVMs, the data is linearly separable, allowing the model to utilize a hyper plane directly. For non-linear classification tasks, SVMs employ the kernel trick to transform the input data into a higher-dimensional space, enabling effective separation without the need for activation functions. To enable this transformation, a variety of kernel functions including sigmoid, polynomial, linear, and Gaussian Radial Basis Function (RBF) kernels are frequently used in SVM implementations [18].

SVMs have many notable advantages, such as their great generalization skills, robustness against over fitting, and skill at handling high-dimensional datasets. But it's crucial to understand that SVMs can become computationally demanding, especially when dealing with big datasets, which could affect how effective they are in practical applications.

Here is an example of previous literature conducted by researchers using Support Vector Machines (SVM).

Atmospheric Temperature Prediction using Support Vector Machines

Radhika and Shashi [19] focuses on applying machine learning methods, specifically Support Vector Machines (SVM) and Multi-Layer Perceptron (MLP), to forecast daily maximum temperatures. The study used a dataset covering weather data from the University of Cambridge, spanning from 2003-2007 for training, and tested it on data from January to July 2008. The dataset included observations taken every half-hour, which were then processed to forecast the highest temperatures for each day based on the values from the previous n days. According to the results, the SVM model performed better than the MLP model for all window spans (orders), consistently obtaining lower MSE values. The SVM's MSE ranged from 7.07 to 7.56, whereas MLP's MSE ranged from 8.07 to 10.26. Overall, the study found that SVM outperformed MLP in terms of accuracy and stability when it came to managing the non-linear aspects of the atmospheric data.

2.7.3 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a straightforward yet effective algorithm used for classification and regression tasks in machine learning. It functions according to the idea that comparable data points are usually found near to one another in the feature space. By examining the 'k' closest neighbors using a selected distance metric such as the Manhattan distance or the Euclidean distance the algorithm forecasts the class or value of a new data point. In regression tasks, KNN averages neighbor values to forecast the outcome, whereas in classification tasks, it selects the class that occurs most frequently among the closest neighbors. The simplicity and convenience of implementation of KNN are one of its main advantages. KNN essentially maintains the complete training dataset for upcoming predictions, negating the need for an explicit training process. For those who are new to machine learning, this makes it very appealing. Additionally, KNN is versatile, capable of handling both classification and regression problems. It works best when the classes are well-separated because this enables it to make precise predictions in a variety of situations [20].

KNN does have certain limits, though. When working with huge datasets, the approach may become computationally expensive and slow because it needs to calculate the distance to every training point for every prediction. Because KNN keeps the complete training dataset, this may result in excessive memory use. Furthermore, KNN's sensitivity to noise and extraneous characteristics might skew distance estimates and impair prediction accuracy. The algorithm also suffers from the "curse of dimensionality," where the efficiency of distance measures diminishes as the number of features grows. Choosing the appropriate value of 'k' is critical for KNN's performance. While a larger value helps smooth predictions but may result in under fitting, where the model fails to catch meaningful patterns, a small value of 'k' might cause over fitting, making the model unduly sensitive to noise in the data. In order to achieve a balance between bias and variance, practitioners frequently use techniques like as cross-validation to identify the best value of 'k' [21].

KNN has a wide range of applications, including recommendation systems, medical diagnosis, image recognition, and customer segmentation. It is a useful tool for a variety of predictive modeling jobs because of its capacity to provide predictions based on local trends in the data. All things considered, K-Nearest Neighbors is a flexible and easily understandable technique that can

successfully handle a variety of machine learning problems, particularly in small to medium-sized datasets.

Here is an example of previous literature conducted by researchers using K-Nearest Neighbors (KNN).

Used Car Price Prediction using K-Nearest Neighbor Based Model

Samruddhi & Kumar [22] explores the use of the K-Nearest Neighbor (KNN) algorithm for predicting the price of used cars. 14 variables, including car attributes like mileage, engine type, transmission, and pricing, are included in the dataset, which was obtained via Kaggle. The authors trained the KNN model with three distinct train-test split ratios: 75%-25%, 80%-20%, and 85%-15%, after preprocessing the dataset (e.g., converting category variables to numerical values).In order to prevent over fitting and assess the model's performance, the study also used 5-fold and 10-fold cross-validation, or K-Fold Cross-Validation. The results showed that the KNN model performed best with a K value of 4, achieving an accuracy of 85%, an RMSE of 4.01, and an MAE of 2.01. Ten-fold cross-validation was used to confirm these findings, and the model's accuracy and RMSE were 82% and 4.73, respectively. The study comes to the conclusion that the KNN-based model outperforms other models, such linear regression, which only attained 71% accuracy, because it is tuned for this dataset. The study suggests applying cutting-edge machine learning approaches in further work to enhance the model's prediction accuracy and further optimize it.

2.7.4 Decision Trees

Decision Trees are a widely used machine learning algorithm that serves both classification and regression tasks. They function by repeatedly dividing a dataset into smaller groups according to the input feature values, resulting in a decision tree-like model. Decision trees are made up of nodes, branches, and leaves. An internal node in a decision tree represents a characteristic or attribute, a branch a decision rule, and a leaf node the result or prediction. This intuitive approach is simple to grasp and comprehend since it closely resembles how humans make decisions. The functioning of decision trees involves several key steps. The whole dataset is covered by the root node, where the algorithm starts initially. It assesses possible splits at each node according to many qualities and chooses the one that maximizes information gain or reduces impurity, which is frequently quantified in classification problems by Gini impurity or entropy. The goal of regression jobs is usually to reduce the MSE. The data is split into subsets based on the decision rule when a

split is created, and this process is repeated recursively until a stopping requirement is satisfied, like reaching a certain depth or having a minimum amount of samples in a node [23].

The interpretability of decision trees is one of its main benefits. Decision trees are a popular option for applications in industries such as business and healthcare because of their clear visual depiction, which enables users to grasp the decision-making process. Furthermore, decision trees simplify preprocessing because they do not require feature scaling because they are invariant to the scale of the data. They don't require extra transformations because they are adaptable enough to handle both numerical and categorical data. Additionally, by tracking several branches for absent data points during training and generating predictions based on the information at hand, decision trees are capable of managing missing values with effectiveness. However, decision trees also have their disadvantages. They are prone to over fitting, which can result in poor generalization on unobserved data, particularly when they grow too deeply. Because decision trees can pick up noise in the training set, they become less accurate predictors and lead to over fitting problems. Additionally, they have the potential to be unstable; even a slight alteration to the training dataset might cause noticeable changes to the tree's structure. Decision trees may also show bias in favor of features with higher levels, which could lead to less-than-ideal splits and predictions.
In conclusion, decision trees are a powerful and intuitive machine learning technique suitable for both classification and regression tasks. They are a popular choice in many applications because of their capacity to represent complex relationships in an understandable and straightforward manner. To improve prediction performance, practitioners frequently use strategies like pruning or integrating decision trees into ensemble approaches like Random Forests, but they must be aware of the possibility of over fitting and instability[24].

In conclusion, decision trees are a powerful and intuitive machine learning technique suitable for both classification and regression tasks. They are a popular choice in many applications because of their capacity to represent complex relationships in an understandable and straightforward manner. To improve prediction performance, practitioners frequently use strategies like pruning or integrating decision trees into ensemble approaches like Random Forests, but they must be aware of the possibility of over fitting and instability.

Here is an example of previous literature conducted by researchers using Decision trees.

Using Decision Tree Algorithm to Predict Student Performance

Apolinar-Gotardo [25] focuses on the use of data mining techniques—specifically the J48 Decision Tree Algorithm—to predict the academic performance of 2nd-year BSIT students in a Data Structures and Algorithms course. The 108 examples in the dataset were taken from student grades in the academic year 2015–2016.Lab exercises/projects (LEP), quizzes (Q), midterms (M), and finals (F) were the main study variables. The accuracy of the model was assessed using metrics such as the ROC Curve, and the study employed the 10-fold cross-validation approach for model evaluation. In a 10-fold cross-validation, the data is divided into 10 equal parts (folds), with each part used once as a test set while the remaining nine parts are used for training. The results showed that the model achieved the following accuracies: Pass: 85.31%, Conditional: 79.41% and Failed: 91.67%.Additionally, the decision tree model identified Finals as the most significant factor in predicting whether a student would pass, fail, or receive a conditional result, requiring a grade higher than 72.30% to pass the course. The study suggests that using a data-driven approach can help educational institutions track student performance and implement timely interventions.

2.7.5 Deep Learning Techniques

ANNs, which are modeled after the architecture and operation of the human brain, are used in deep learning. As additional instances and data are added, as well as the number of layers in the network increases, these networks perform better. Aside from their scalability, one of the main benefits of deep learning models is their capacity to automatically extract features from unprocessed data without requiring human feature engineering [26]. A variety of deep learning designs, such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), deep belief networks (DBNs), deep neural networks (DNNs), and deep Boltzmann machines (DBMs), are used for a variety of artificial intelligence (AI) tasks.

Deep Neural Networks (DNNs) are neural networks that consist of more than two layers, enabling them to model complex non-linear relationships. DNNs are popular tools for handling complex and varied datasets since they can be utilized for both regression and classification tasks. They learn hierarchical representations of the incoming data by processing information through several layers of sophisticated mathematical models.

Deep Belief Networks (DBNs) are probabilistic, unsupervised learning models composed of multiple layers of Restricted Boltzmann Machines (RBMs). Each RBM layer in a DBN only interacts with the layers next to it, creating a hierarchical model that, as more levels are stacked, may extract progressively abstract features from the input data. This ability to automatically learn multi-level representations makes DBNs effective for a range of tasks, including dimensionality reduction and feature extraction [27].

Deep Boltzmann Machines (DBMs), like DBNs, are unsupervised generative models, but they differ in that they feature undirected connections between layers. DBMs are able to record intricate interconnections between hidden units across layers thanks to these links. The structure of DBMs is a bipartite graph, in which the odd and even layers interact with the nearby layers but are independent of one another. For jobs requiring the modeling of intricate probability distributions, DBMs are quite helpful [28].

Recurrent Neural Networks (RNNs) are designed to handle sequential data, such as text, audio, and time series. RNNs, feature feedback loops that allow them to retain knowledge from earlier stages and apply it to subsequent inputs. For tasks like speech recognition, natural language processing, and time-series forecasting, RNNs' ability to capture temporal connections in data is crucial. RNNs are essential for applications involving sequential data interpretation, but they train more slowly than ordinary neural networks due to their increased computational complexity[29].

ANNs, particularly deep learning architectures like Deep Neural Networks (DNNs), Deep Belief Networks (DBNs), Deep Boltzmann Machines (DBMs), Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs), are highly effective for both prediction and classification tasks. These models can automatically learn complex patterns and hierarchical representations from raw data, making them ideal for a wide range of applications. DNNs and CNNs are often used for image and pattern recognition, as well as for classification tasks, where they model non-linear relationships and extract features from data through multiple layers. DBNs and DBMs are valuable for unsupervised tasks such as feature extraction and dimensionality reduction, helping improve model accuracy in complex scenarios. RNNs, with their ability to capture temporal dependencies, are particularly suited for prediction in time series data, such as forecasting trends or events, as well as in natural language processing and speech recognition. Together, these models

are crucial for both predictive modeling and classification in various fields, including healthcare, finance, and AI-driven systems[30].

2.8. Related work

In this section related works are reviewed targeting at what has been done, the models used, research data, the experimentation process their results and showing the limitations under each case.

The research carried out in Rwanda [31] included machine learning approaches, such as Random Forest, Naïve Bayes, Decision trees, support vector machines, K-nearest neighbor, and logistic regression, to forecast malaria outbreaks by taking into account environmental factors and past malaria data. The study used historical malaria case data from 2016 to 2019 as well as metrological data, totaling around 2080 observations for the 10 districts included in the study. The data set was split into two parts including the training set and test set, the training set was composed of 80% of the whole data while the test was made by remained part of 20% of the whole data set. Out of all the models that have been suggested, the random forest performs better than the others, with 90.75% accuracy, 90.69% precision, 90.88% recall, and a 90.75% F-score. The study also classifies in to low, medium, high and outbreak. However, the study was limited by the size of the data collection, and the results are more classification than prediction.

The study done on Nigeria [32] used 5 supervised machine learning techniques (Naive Bayes, Support Vector, Linear Regression, Logistic Regression, and K-Nearest Neighbor) to predict malaria outbreaks by considering the past malaria data which is from 2010-2020. The dataset was split into 70% for the training data and 30% for testing. Python was the programming language used for the research, which was conducted using the Scikit-learn Library that was imported into the Anaconda IDE. The average accuracy of Naïve Bayes is 79.1%, which is higher than all other models employed in the study. The study also classifies in to low, mild, high and outbreak. Even yet, the accuracy was somewhat lower when compared to other models that other researchers have used. Furthermore, it solely considers historical malaria data for the forecast and ignores other factors.

The study conducted in a six African countries [33] that uses a machine learning algorithm called Extreme Gradient Boosting (XGBoost), which is a machine learning method that is used to solve regression and classification problems. Mostly in the form of trees, it offers outcomes in a prediction model. Additionally, they made use of A vast number of data points are divided into a limited number of clusters using the K-means clustering technique. By calculating how similar they are in terms of distance, it puts the objects in one group based on shared qualities. This study employed K-means clustering to identify outliers and purify the dataset. The dataset contains a total of 28 records for each of the six countries, resulting in a combined total of 168 records covering the period from 1990 to 2017. The dataset split into a ratio of 70:30 for training and testing, where 70% (18 records) comprise the training set and 30% (8 records) constitute the test set. Each record consists of 6 attributes: precipitation, surface radiation, temperature, atmospheric pressure, relative humidity, and the target variable indicating changes in malaria incidence this study used the above methods to predict malaria incidence in the six African countries. The malaria incidence prediction model classifies malaria incidence into high and low target classes based on climate variability. They also compared and discussed different prediction models like XGboost which ranges an accuracy from 0.93 to 0.98, SVM which ranges from 0.74 to 0.81, naïve Bayes which ranges from 0.71 to 0.76, logistic regression which ranges from 0.78 to 0.82 .These accuracy values show the performance of different machine learning models in predicting outcomes for various countries. XGBoost and Logistic Regression (LR) tend to perform better. Yet, because of the absence of weekly malaria case data, the researchers had to use an annual dataset, which might not capture the seasonal variations in the predicted cases.

The research study carried out in Africa [34] employed six distinct machine learning techniques, including Support Vector Machine, K-Nearest Neighbor, Random Forest, Decision Tree, Logistic Regression, and Naïve Bayes, to forecast malaria occurrences by considering both weather-related and non-climatic factors. The dataset comprises key features including sanitation and drinking water access percentages, average temperature and rainfall, and yearly reported malaria cases. The data set was divided into two parts: the training set and the test set. 80% of the whole data set made up the training set, and the remaining 20% made up the test set. The study systematically assessed and compared the performance of all six techniques and determined that Random Forest exhibited

the highest level of accuracy with 97.72% and Area under Curve (AUC) 98% in predicting malaria cases.

In a recent research conducted in Burundi [8], an attempt was made to predict malaria using roo Neural Network models. The initial models leveraged climate-change-related factors like temperature, rainfall, and relative humidity but did not yield the desired predictive accuracy. As a result, the study transitioned to a different approach by adopting the Long Short-Term Memory (LSTM) model, a type of deep learning model. The dataset collected from IGEBU spans from 2010 to 2022 and was structured on a monthly scale. It comprises data from five provinces and includes parameters such as relative humidity, rainfall, and temperature, along with their corresponding maximum and minimum values Therefore, the overall dataset contains 780 records. 80% of the dataset was used for training in this experiment, while the 20% was used for testing. Province-level malaria case forecasts for the period from September 2020 to September 2022. The Univariate LSTM predictions show a curve trend that is similar to the observed cases, with the exception that most of the time the observed cases slightly exceed the expected values. In Bujumbura, for example, the Multivariate LSTM forecasts with RMSE 16777.17 whereas the Univariate LSTM predicts 4868.69 RMSE. With the implementation of LSTM, more favorable results were achieved, however the model has a deficiency of handling multiple variables at once because of the high RMSE in the multivariate LSTM model.

In a malaria prediction study conducted in China [9], researchers used a deep neural network model called the long short-term memory sequence-to-sequence (LSTMSeq2Seq) to predict the recurrence of malaria cases between 2004 and 2016. In all 31 Chinese provinces, they gathered monthly data on malaria cases between January 2004 and December 2016.Ten meteorological variables—pressure, average temperature, maximum temperature, wind speed, minimum temperature, wind direction, precipitation, average relative humidity, sunshine duration, and minimum relative humidity—were preserved with no missing values in all meteorological data features. The data set includes four classes of Plasmodium species: P. falciparum, P. vivax, P. malariae, and other Plasmodium species. In this study, they partitioned the dataset into training, validation, and test sets. 70% of the data was allocated for training, 15% for validation to prevent over fitting, and the remaining 15% for testing model evaluation. The study employed various

models to forecast the resurgence of malaria cases in China, considering the impact of climatic factors. These models encompassed the Long Short-Term Memory Sequence-to-Sequence (LSTMSeq2Seq) deep neural network, Extreme Gradient Boosting (XGBoost), Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM). In descending order of performance throughout the study, the models rank as follows: LSTMSeq2Seq, LSTM, GRU, and XGBoost. LSTMSeq2Seq consistently yielded the lowest RMSE values, measuring 0.0252, 0.0107, 0.0586, and 0.0077 for P. falciparum, P. vivax, P. malariae, and other plasmodia, respectively. The forecasts were predicated on how climate variables would affect things. But it's important to note that making precise predictions proved difficult in some of the study's provinces. In addition the study used a bit older data.

The research carried out in northwest Ethiopia [35] offers a comprehensive exploration of malaria, with a specific focus on spatial modeling, seasonal patterns, and predictive mapping. The study utilizes a dataset comprising 916,204 malaria cases reported between 2014 and 2017 and adopts a cross-sectional study design along with the Kulldorf method, a tool for identifying spatial clusters or anomalies in data. Through spatial modeling, the study pinpoints areas with heightened disease prevalence, providing vital insights for targeted intervention strategies. It also delves into the characterization of seasonal variations, aiding the efficient allocation of resources during peak transmission periods. Moreover, the research generates predictive maps, serving as a valuable resource for public health authorities to proactively combat malaria. These findings carry significant implications for public health, underscoring the role of data-driven decision-making in disease control and offering a model for comprehending and mitigating infectious diseases. Nevertheless, it's worth noting that the data utilized in this study spans from 2014 to 2017, representing relatively older information. Given the dynamic nature of disease transmission, maintaining accuracy and relevance necessitates regular updates.

Research conducted by using past morbidity data [36], aimed to assess the effectiveness of several approaches for predicting malaria prevalence in areas with unstable transmission. This study used five different forecasting techniques to cover 20 regions in central and northwest Ethiopia. The differences between actual incidence and projections covering prediction periods of up to 12 months served as the foundation for the accuracy evaluation. Interestingly, the best accurate

projections were obtained using a seasonal adjustment strategy that used the mean deviation of the last three observations from expected seasonal values. With this strategy, longer observation periods of three years produced better results, especially in the wet months of June through August. Forecasting from December to February, which are typically dry months, showed less accuracy. This study emphasizes the drawbacks of basing incidence forecasting only on past morbidity trends and emphasizes the need to improve epidemic early warning systems by including outside variables like weather.

A research conducted in south Ethiopia [37] investigated if it was possible to forecast malaria in various regions of the country using meteorological data. The illness malaria is transmitted by mosquitoes and is associated with variations in the weather. Over a nine-year period, researchers gathered information from 42 locations, including the number of cases of malaria and the local climate. They discovered that, in some regions, historical instances of malaria may be used to forecast new ones. The study demonstrated the difficulty of developing a single model that is suitable for all situations. However, this study is dated because it was carried out between 1998 and 2007. More recent statistics and study are desperately needed, given the dynamically shifting nature of malaria transmission.

2.9 Research Gaps

There have been studies on the prediction of malaria, however not many of them make use of ANN for predictive modeling. The ability of ANN to capture complex relationships between malaria cases and meteorological variables remains largely unexplored in the field of malaria prediction.

Relevant and Up-to-Date Data Are Essential for ANN Models to maximize the accuracy of malaria prediction models, it is recommended to incorporate very current and up-to-date data. Research is still lacking on the incorporation of meteorological and real-time malaria case data into ANN models, which can enhance the accuracy and efficiency of malaria predictions.

Previous research may not have fully captured the seasonal fluctuations in malaria transmission because they relied on annual data sets. This limitation can be overcome by collecting and analyzing more exact and temporally defined data, such as weekly or monthly data, which will allow for the accurate capturing of the nuances and fluctuations in malaria incidence over time.

The proposed work aims to fill these research gaps and develop malaria prediction systems by utilizing ANN models and integrating meteorological and current malaria case data. Through the use of ANN models and the integration of current malaria case and meteorological data, the proposed study seeks to close these research gaps and develop malaria prediction approaches. The accuracy and efficiency of ANN-based malaria prediction will be better understood by this study, which also has the potential to guide the development of timely and focused interventions to stop the spread of malaria.

CHAPTER 3: MATERIALS AND METHODS

3.1 Introduction

This section describes the methodology used in the study to use an ANN and meteorological data to predict malaria incidences. It also includes a detailed overview of the proposed model. It also describes the data sources in detail, including where they came from, how they were collected, and what preprocessing measures were performed to get the data ready for model building.

3.2 proposed system architecture

The proposed system model for predicting malaria incidences consists of four main components: data collection, preprocessing, feature extraction, and model training and prediction.

The first component is data collection, meteorological data relevant to malaria incidences were collected from Hawassa meteorological station, and these data include variables such as minimum temperature, maximum temperature and rainfall, which are known to influence the spread of malaria. And Malaria weekly data within this period 2017–2022 were collected from Sidama health region bureau

In the second step, the acquired meteorological and past malaria data undergo preprocessing to prepare them for further analysis. This involves tasks such as data cleaning, normalization, and feature scaling. Missing values are handled, outliers are detected and removed, and the data are standardized to ensure consistency and comparability across different variables.

In the third step, Feature Extraction were performed to identify relevant patterns and relationships in the meteorological data that may be indicative of malaria incidences. This step involves extracting meaningful features from the raw data, such as temperature trends, rainfall patterns. Feature selection techniques may also be applied to identify the most informative variables for predicting malaria.

Finally Model Training and Prediction involves training an ANN feed forward model with a single hidden layer using the preprocessed meteorological data. The ANN model learns to map the input features to the target variable, which is the occurrence or prevalence of malaria incidences. During training, the model adjusts its parameters to minimize prediction errors and optimize performance. Once trained, the model can be used to predict future malaria incidences based on new meteorological data inputs.

In summary, the proposed architecture utilizes meteorological data and an ANN feed forward model to predict malaria incidences. By leveraging the relationships between weather variables and malaria transmission, the model aims to provide accurate and timely predictions to support malaria control and prevention efforts.

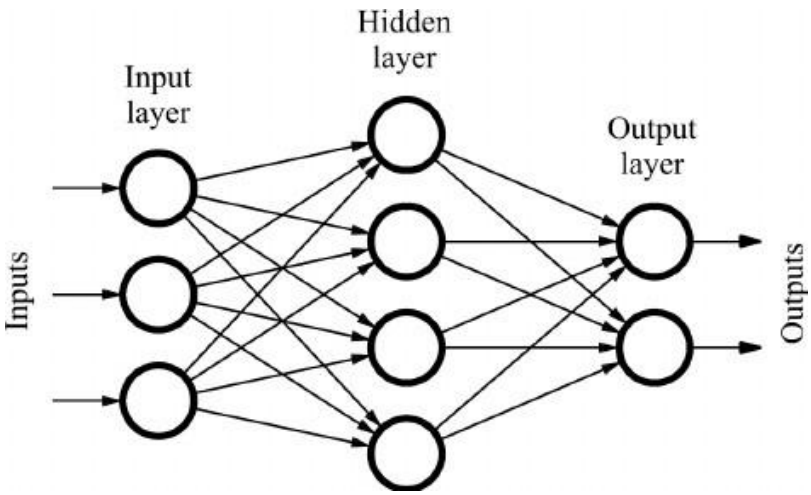

Figure 3.1: Summary of the ANN feed forward model

3.3 Dataset preparation

Meteorological data relevant to malaria incidences were collected from the Hawassa meteorological station spanning 2017-2022, covering four different Woredas in the region (Boricha, Dale, Hawassa Zuria, and Shebedino). These data include variables such as minimum temperature, maximum temperature, and rainfall, known to influence malaria spread. Weekly malaria data for the same period (2017–2022) were collected for the four Woredas from the Sidama Health Region Bureau .The dataset for each Woredas were 312 records and the total dataset size equals to 1248 records. In this experiment, 80% of the dataset was allocated for the training phase , where the model learns patterns and relationships from the data while the remaining 20% was designated for testing, allowing for the assessment of the model's performance on unseen data. The meteorological and malaria data were integrated into a single dataset, aligning observations by timestamps and specific geographic locations. This integration ensures that each data point corresponds to the same timeframe and geographic region, facilitating effective analysis of their relationship.

Table 1: Hawassa zuria rainfall

Zone	Wereda	Longtude	Latitude	Altitude	Element (in mm)	Year	Month	w1(rainfall)	w2(rainfall)	w3(rainfall)	w4(rainfall)
Sidama	Hawassa Zuri	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Rainfall	2017	Jan	0.0	0.0	0.0	0.0
Sidama	Hawassa Zuri	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Rainfall	2017	Feb	0.9	0.6	8.8	2.3
Sidama	Hawassa Zuri	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Rainfall	2017	Mar	0.0	0.0	3.5	6.7
Sidama	Hawassa Zuri	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Rainfall	2017	Apr	0.8	0.9	0.0	8.4
Sidama	Hawassa Zuri	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Rainfall	2017	May	4.4	14.8	6.2	5.8
Sidama	Hawassa Zuri	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Rainfall	2017	Jun	2.1	3.1	0.6	2.0
Sidama	Hawassa Zuri	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Rainfall	2017	Jul	3.8	6.7	5.5	5.9
Sidama	Hawassa Zuri	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Rainfall	2017	Aug	7.4	5.1	0.6	2.4
Sidama	Hawassa Zuri	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Rainfall	2017	Sep	4.1	2.2	5.9	11.1
Sidama	Hawassa Zuri	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Rainfall	2017	Oct	1.0	5.8	4.2	5.3
Sidama	Hawassa Zuri	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Rainfall	2017	Nov	0.5	0.0	0.0	0.3
Sidama	Hawassa Zuri	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Rainfall	2017	Dec	0.0	0.0	0.0	0.0

Table 2: Hawassa Zuria max temperature

Station Name	Class	Wereda	Longtude	Latitude	Altitude	Element (Daily)	Year	Month	w1(max.temp)	w2(max.temp)	w3(max.temp)	w4(max.temp)
Hawassa	Synoptic	Hawassa Zuriya	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Max.Temp.	2017	Jan	27.6	28.2	29.6	30.1
Hawassa	Synoptic	Hawassa Zuriya	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Max.Temp.	2017	Feb	30.1	30.4	29.6	29.6
Hawassa	Synoptic	Hawassa Zuriya	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Max.Temp.	2017	Mar	31.5	33.1	32.8	29.8
Hawassa	Synoptic	Hawassa Zuriya	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Max.Temp.	2017	Apr	30.5	32.2	32.6	31.2
Hawassa	Synoptic	Hawassa Zuriya	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Max.Temp.	2017	May	29.2	29.7	27.5	27.6
Hawassa	Synoptic	Hawassa Zuriya	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Max.Temp.	2017	Jun	27.3	29.2	27.3	27.3
Hawassa	Synoptic	Hawassa Zuriya	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Max.Temp.	2017	Jul	26.4	26.2	25.8	26.0
Hawassa	Synoptic	Hawassa Zuriya	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Max.Temp.	2017	Aug	27.0	25.6	27.4	26.5
Hawassa	Synoptic	Hawassa Zuriya	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Max.Temp.	2017	Sep	25.9	25.9	27.2	26.3
Hawassa	Synoptic	Hawassa Zuriya	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Max.Temp.	2017	Oct	27.3	28.3	28.1	28.1
Hawassa	Synoptic	Hawassa Zuriya	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Max.Temp.	2017	Nov	28.0	29.0	28.3	28.7
Hawassa	Synoptic	Hawassa Zuriya	038° 28' 59.2" E	07° 03' 53.8" N	1694	Daily Max.Temp.	2017	Dec	28.6	27.6	28.3	28.6

Table 3: Hawassa Zuria min temperature

Station Name	Class	Wereda	Longtude	Latitude	Altitude	Year	Month	w1(min.temp)	w2(min.temp)	w3(min.temp)	w4(min.temp)
Hawassa	Synoptic	Hawassa Zu	038° 28' 59.2" E	07° 03' 53.8" N	1694	2017	Jan	6.8	6.6	11.4	12.8
Hawassa	Synoptic	Hawassa Zu	038° 28' 59.2" E	07° 03' 53.8" N	1694	2017	Feb	14.9	13.7	13.9	15.5
Hawassa	Synoptic	Hawassa Zu	038° 28' 59.2" E	07° 03' 53.8" N	1694	2017	Mar	13.3	13.8	15.3	15.3
Hawassa	Synoptic	Hawassa Zu	038° 28' 59.2" E	07° 03' 53.8" N	1694	2017	Apr	14.0	14.3	15.6	15.2
Hawassa	Synoptic	Hawassa Zu	038° 28' 59.2" E	07° 03' 53.8" N	1694	2017	May	15.3	15.4	16.4	16.1
Hawassa	Synoptic	Hawassa Zu	038° 28' 59.2" E	07° 03' 53.8" N	1694	2017	Jun	16.1	16.5	14.4	15.1
Hawassa	Synoptic	Hawassa Zu	038° 28' 59.2" E	07° 03' 53.8" N	1694	2017	Jul	16.4	15.1	15.9	16.5
Hawassa	Synoptic	Hawassa Zu	038° 28' 59.2" E	07° 03' 53.8" N	1694	2017	Aug	15.3	15.2	14.9	16.4
Hawassa	Synoptic	Hawassa Zu	038° 28' 59.2" E	07° 03' 53.8" N	1694	2017	Sep	15.9	15.7	16.0	14.8
Hawassa	Synoptic	Hawassa Zu	038° 28' 59.2" E	07° 03' 53.8" N	1694	2017	Oct	14.4	15.1	16.7	14.5
Hawassa	Synoptic	Hawassa Zu	038° 28' 59.2" E	07° 03' 53.8" N	1694	2017	Nov	15.3	11.8	8.8	11.0
Hawassa	Synoptic	Hawassa Zu	038° 28' 59.2" E	07° 03' 53.8" N	1694	2017	Dec	9.1	8.3	10.3	9.8

Table 4: Hawassa Zuria malaria cases

Region	Zone	Woreda	Year	Month	w1(malaria_case)	w2(malaria_case)	w3(malaria_case)	w4(malaria_case)
SNNPR	Sidama	Hawasa Zuriya	2017	January	5	6	8	5
SNNPR	Sidama	Hawasa Zuriya	2017	February	10	10	4	10
SNNPR	Sidama	Hawasa Zuriya	2017	March	4	3	3	9
SNNPR	Sidama	Hawasa Zuriya	2017	April	3	3	5	4
SNNPR	Sidama	Hawasa Zuriya	2017	May	1	1	3	0
SNNPR	Sidama	Hawasa Zuriya	2017	June	4	1	3	3
SNNPR	Sidama	Hawasa Zuriya	2017	July	1	2	3	5
SNNPR	Sidama	Hawasa Zuriya	2017	August	0	2	4	4
SNNPR	Sidama	Hawasa Zuriya	2017	September	4	4	8	2
SNNPR	Sidama	Hawasa Zuriya	2017	October	12	7	5	6
SNNPR	Sidama	Hawasa Zuriya	2017	November	0	4	2	7
SNNPR	Sidama	Hawasa Zuriya	2017	December	3	5	5	5

Table 5: Sample of integrated dataset

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
Station Name	Wereda	Altitude	Year	Month	w1(min.temp)	w2(min.temp)	w3(min.temp)	w4(min.temp)	w1(max.temp)	w2(max.temp)	w3(max.temp)	w4(max.temp)	w1(rainfall)	w2(rainfall)	w3(rainfall)	w4(rainfall)	w1(malaria_case)	w2(malaria_case)	w3(malaria_case)	w4(malaria_ca
Hawassa	Hawassa Zu	1694	2017	Jan	6.8	6.6	11.4	12.8	27.6	28.2	29.6	30.1	0.0	0.0	0.0	0.0	5	6	8	5
Hawassa	Hawassa Zu	1694	2017	Feb	14.9	13.7	13.9	15.5	30.1	30.4	29.6	29.6	0.9	0.6	8.8	2.3	10	10	4	10
Hawassa	Hawassa Zu	1694	2017	Mar	13.3	13.8	15.3	15.3	31.5	33.1	32.8	29.8	0.0	0.0	3.5	6.7	4	3	3	9
Hawassa	Hawassa Zu	1694	2017	Apr	14.0	14.3	15.6	15.2	30.5	32.2	32.6	31.2	0.8	0.9	0.0	8.4	3	3	5	4
Hawassa	Hawassa Zu	1694	2017	May	15.3	15.4	16.4	16.1	29.2	29.7	27.5	27.6	4.4	14.8	6.2	5.8	1	1	3	0
Hawassa	Hawassa Zu	1694	2017	Jun	16.1	16.5	14.4	15.1	27.3	29.2	27.3	27.3	2.1	3.1	0.6	2.0	4	1	3	3
Hawassa	Hawassa Zu	1694	2017	Jul	16.4	15.1	15.9	16.5	26.4	26.2	25.8	26.0	3.8	6.7	5.5	5.9	1	2	3	5
Hawassa	Hawassa Zu	1694	2017	Aug	15.3	15.2	14.9	16.4	27.0	25.6	27.4	26.5	7.4	5.1	0.6	2.4	0	2	4	4
Hawassa	Hawassa Zu	1694	2017	Sep	15.9	15.7	16.0	14.8	25.9	25.9	27.2	26.3	4.1	2.2	5.9	11.1	4	4	8	2
Yirgalem	Dale	1786	2022	Jan	11.2	7.1	8.0	6.5	28.6	28.6	29.1	30.0	0.0	0.0	5.3	0.0	32	33	25	39
Yirgalem	Dale	1786	2022	Feb	5.4	5.9	5.7	6.4	28.9	30.3	31.1	31.4	0.0	0.0	0.0	5.4	29	34	28	31
Yirgalem	Dale	1786	2022	Mar	6.0	5.2	5.8	6.7	31.0	31.1	30.2	29.1	0.0	5.1	9.3	8.6	34	30	20	205
Yirgalem	Dale	1786	2022	Apr	11.2	12.4	9.8	9.6	29.2	27.4	26.7	26.6	10.4	6.2	11.3	6.1	96	47	22	21
Yirgalem	Dale	1786	2022	May	10.7	12.4	10.5	11.4	26.6	26.5	26.1	26.5	2.2	5.2	1.4	8.1	28	26	43	37
Yirgalem	Dale	1786	2022	Jun	9.6	8.4	10.2	11.1	25.2	25.4	23.9	24.8	9.1	2.9	10.5	2.3	26	42	59	53
Yirgalem	Dale	1786	2022	Jul	12.2	13.3	13.1	12.5	25.2	24.4	24.9	23.6	0.0	3.5	3.5	3.2	55	47	46	90
Yirgalem	Dale	1786	2022	Aug	12.5	12.5	12.4	12.4	24.7	25.0	24.7	25.2	11.8	13.4	11.7	9.1	30	28	34	53
Yirgalem	Dale	1786	2022	Sep	12.7	13.3	11.8	13.3	24.9	25.0	25.4	25.3	23.5	20.6	20.6	19.2	54	85	45	53
Yirgalem	Dale	1786	2022	Oct	13.3	11.5	11.1	11.2	25.4	25.3	25.9	27.2	8.6	6.0	4.6	5.0	53	52	44	119
Yirgalem	Dale	1786	2022	Nov	12.1	11.0	10.8	10.6	25.7	27.4	26.0	26.6	1.5	0.0	0.0	1.5	54	67	56	294
Yirgalem	Dale	1786	2022	Dec	8.2	8.1	8.3	6.8	26.4	27.0	27.4	28.0	0.0	0.0	1.2	0.0	162	148	115	121

Natures of the data

Meteorological Data:

- Continuous Variables: Meteorological data typically consists of continuous variables such as temperature (minimum and maximum), rainfall, humidity, wind speed, and atmospheric pressure. These variables are measured at regular intervals (e.g., hourly, daily) and can exhibit a wide range of numerical values.
- Temporal Variation**:** Meteorological variables exhibit temporal variation, meaning they change over time due to factors such as diurnal cycles, seasonal changes, and long-term climate trends. Understanding the temporal patterns of meteorological variables is essential for identifying their impact on malaria transmission dynamics.
- Spatial Variation**:** Meteorological conditions vary spatially based on geographic location, altitude, proximity to water bodies, and other geographical factors. Capturing the spatial variability of meteorological data is crucial for assessing its influence on malaria risk across different regions.

Malaria Incidence Data:

- Discrete Variables: Malaria incidence data typically consists of discrete variables representing the number of malaria cases reported over a specific time period (e.g., weekly, monthly). These variables are often aggregated at the regional or district level and recorded as counts or incidence rates.
- Temporal Trends: Malaria incidence data exhibit temporal trends, including seasonal variations, epidemic outbreaks, and long-term trends influenced by factors such as climate, population dynamics, and public health interventions.
- Spatial Patterns: Malaria incidence rates vary spatially, with certain regions experiencing higher transmission intensity compared to others. Understanding the spatial distribution of malaria cases is essential for targeting control measures and allocating resources effectively.

Integrated Dataset:

- Time Series Data: Integrating meteorological and malaria incidence data results in a time series dataset, where each observation is time stamped and corresponds to a specific time period (e.g.,

weekly, monthly) and geographic location (e.g., district, region). This dataset captures the dynamic relationship between meteorological factors and malaria transmission over time.

- Multivariate Data: The integrated dataset contains multiple variables, including meteorological parameters (e.g., temperature, rainfall) and malaria incidence indicators (e.g., number of cases). Analyzing the interactions between these variables allows for the identification of environmental factors associated with changes in malaria risk.

3.4 Preprocessing

Preprocessing refers to the data preparation steps performed on raw or unprocessed data to make it suitable for analysis or modeling.it have 3 different components like data cleaning, feature engineering and data splitting.

1. Data Cleaning

   In preparing the meteorology data for malaria prediction, missing values, outliers, and inconsistencies were carefully addressed to ensure the reliability and accuracy of the analysis. Any missing values were identified and replaced using mean imputation, which is a simple technique for handling missing data in datasets. It involves replacing missing values with the mean (average) value of the non-missing values in the same column. Several normalization techniques are available in machine learning to standardize numerical quantities to a common scale without changing or jeopardizing data. To normalize the malaria case data in this study, the researchers decided to use the min-max normalization technique.

2. Feature Engineering

   To prepare the data for malaria prediction using meteorology data, we first integrated the past malaria data, which had been converted from daily to weekly intervals, into our dataset. In order to accommodate for the variations in the Gregorian calendar—some months having 31 days, while others have 28 or 30 days—the data utilized in this study were transformed into 48 weeks. The length of each week was set at seven days to maintain uniformity. Some weeks had more than seven days because the extra days at the end of the month were merged with the remaining days when a month had more than twenty-eight days. This strategy was used to preserve consistency throughout the dataset and guarantee precise model assessment and evaluation.

   In this study, Pearson correlation analysis was employed to assess the relationship between meteorological factors such as temperature and rainfall and the incidence of malaria cases. This

analysis included ideas from a variety of recent publications and highlighted the importance of specific features that are regularly used in related research projects. Thus, the model for prediction was chosen to incorporate the lowest and highest temperatures, rainfall, and past weekly incidences of malaria. This choice was guided by their established, high correlation with the incidence of malaria, which increased the model's anticipated accuracy. To visualize the relationships between the meteorological factors and malaria incidence, a correlation heat map was generated, as shown in Figure 3.2.

We identified key features that reflected important environmental factors affecting the spread of malaria. These data include variables such as minimum temperature, maximum temperature, and rainfall, known to influence malaria spread. Weekly malaria data for the period (2017–2022) were collected for the four Woredas from the Sidama Health Region Bureau. The dataset consists of data from 4 woredas (districts), with each woreda contributing 288 records. This number comes from collecting data over 6 years, with each year having 48 records per woreda. The 48 records per year are calculated by gathering data every month, and each month provides 4 weekly records (12 months × 4 weeks = 48 records). Since the data was collected for 6 years, for each woreda, the total dataset equals 288 records (48 records per year × 6 years). Across the 4 woredas, the total dataset sums up to 1152 records (288 records × 4 woredas).

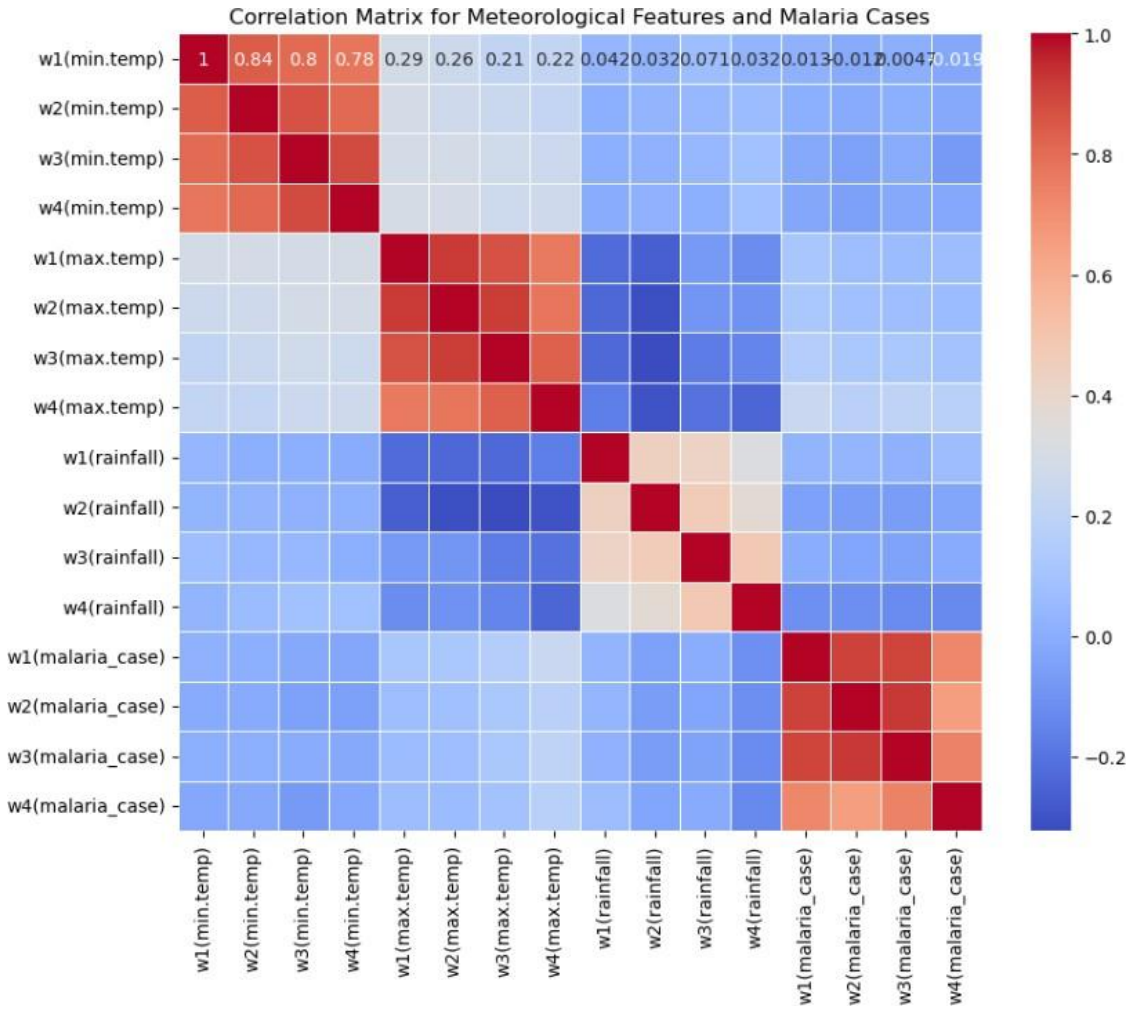

Figure 3.2 Correlation Matrix for Metrological Features and Malaria Cases

3. Data Splitting:

We divided the preprocessed dataset into training and testing sets using an 80-20 split, allocating 80% of the data for training the model and reserving the remaining 20% for testing. This split ensures that the model is trained on a sufficient amount of data to generalize well to unseen samples, while also providing a separate dataset for evaluating the model's performance independently.

3.5 malaria prediction

The process of malaria prediction involves training an ANN feed forward model to analyze patterns and relationships within the input data, enabling it to forecast malaria incidences. During the feature learning phase, the ANN consists of multiple layers stacked on top of each other

according to a defined architectural layout. The model undergoes training, validation, and testing phases to learn and evaluate its performance. In the subsequent sections, we present the proposed architecture and operations for each phase, detailing how features are learned and malaria incidences are predicted using the ANN feed forward model.

The feed forward ANN model was chosen for predicting malaria incidence due to its ability to capture complex, non-linear relationships between meteorological factors and disease transmission. Malaria outbreaks are heavily influenced by variables such as temperature, humidity, and rainfall, which interact in a highly non-linear manner. Traditional machine learning models, such as decision trees and logistic regression, often struggle to model such complexity. Feed forward ANNs, on the other hand, have demonstrated superior performance in capturing these intricate relationships. As noted by Xia, J., Zhang, C., & Yang, L. (2019) [38], feed forward neural networks can effectively handle complex, non-linear data structures, outperforming simpler models in epidemiological tasks involving environmental factors. Additionally, literature supports the effectiveness of ANNs in disease prediction, particularly in cases where environmental factors play a crucial role. Kar, M., & Bansod, S. (2016)[39] showed that ANNs were successfully applied to predict dengue outbreaks using environmental data, achieving higher accuracy compared to other machine learning models. Similarly, Yang, W., Cowling, B. J., & Lau, E. H (2015)[40] demonstrated that ANN models were effective in forecasting influenza spread based on weather patterns. These examples further validate the choice of ANN feed forward models for malaria prediction, as both malaria and dengue are vector-borne diseases influenced by climatic conditions.

3.5.1 Training using ANN feed forward model

The malaria prediction model is trained using an ANN feed forward architecture with specific configurations tailored to the nature of the dataset and prediction task. The dataset comprises meteorological variables such as minimum temperature, maximum temperature, and rainfall recorded at weekly intervals alongside corresponding malaria case counts for 2017-2022 period. Each feature set includes historical data from past weeks, denoted by prefixes 'w1' to 'w4', indicating the temporal aspect of the dataset. The target variable comprises both weekly meteorological data and malaria case counts ('w1' to 'w4') to predict future malaria incidences. To prepare the data for training, features are rounded to maintain consistency, and numpy arrays are

constructed for both features and targets. The dataset is splited using TimeSeriesSplit with five splits, considering the temporal nature of the data and to ensure robust model evaluation.

Adam (Adaptive Moment Estimation) is the optimization technique utilized in this model. Because of its minimal memory needs and high computational efficiency, Adam has gained popularity as a neural network training tool. Adam is a gradient descent modification that ensures faster and more stable convergence by using both momentum and adaptive learning rates for each parameter. By building up gradients, the momentum component speeds up learning, and adaptive learning rates let you make changes dependent on how big the gradient is. These qualities make Adam particularly good at managing noisy and sparse datasets, which is excellent for time-series data like malaria occurrences. Because Adam can adjust learning rates dynamically during training, it offers a balanced optimization strategy that allows the model to converge fast without overshooting, improving overall performance in predictive tasks.

The model in this study is trained for 50 epochs, meaning that during the training phase, the network processes the whole dataset 50 times. Since it controls the length of time the model is exposed to the training set, an epoch is a crucial idea. If there are insufficient epochs in the data, the model may not be sufficiently trained to identify patterns, leading to under fitting. On the other hand, if there are too many epochs, the model may become over fit, which occurs when it memorizes the training data rather than applying it to new data. Experimentation led to the selection of fifty epochs to provide adequate learning while accounting for over fitting. In an effort to reduce over fitting even more, early stopping is used to track the validity loss. And halt the training if the model's performance on unseen data stops improving, ensuring that the model generalizes well beyond the training set.

One important hyperparameters that determines how much the model's weights are adjusted in relation to the loss gradient at each update is the learning rate. This model uses a learning rate of 0.001, which is a typical deep learning task default. The model can proceed steadily and consistently in the direction of minimizing the loss function thanks to this value. A higher learning rate might lead to faster convergence, but it raises the danger of skipping over the ideal solution, resulting to poor performance. Conversely, a lower learning rate would cause learning to proceed too slowly and would cause the model to become trapped in local minima. The chosen learning rate ensures that weight updates are neither too large nor too small, maintaining the right balance between speed and accuracy during the learning process.

The model is trained using MSE as the loss function, which is well-suited for regression tasks where the goal is to predict a continuous output. In order to guarantee that greater errors are penalized more harshly than smaller ones, MSE computes the average of the squared differences between the predicted values and the actual values. This trait is especially helpful for forecasting malaria occurrences, as it minimizes significant differences in expected case counts, which is essential for precise forecasting. The model seeks to increase forecast accuracy and reliability by lowering the total error across all predictions through the use of MSE. Moreover, MSE guarantees that the model's primary goal is to minimize outlier errors, which is crucial for practical applications in public health where accurate predictions can drive effective intervention strategies. The batch size defines the number of training samples processed before the model's internal parameters (weights) are updated. 64 is the batch size utilized in this model; this is a widely used size that provides a fair trade-off between memory restrictions and learning efficiency. By processing data in batches rather than feeding the entire dataset at once, computational efficiency is increased, improving the model's ability to generalize without overwhelming memory. The model's rate of convergence is highly dependent on the batch size; larger batches produce more stable gradients but may cause slower convergence, while smaller batches may add noise into the gradient estimates while preventing over fitting. In this instance, a batch size of 64 ensures that the model efficiently updates its weights while maintaining a balance between stability and generalization.

The selected configuration of the ANN model, including the Adam optimizer, 50 epochs, learning rate of 0.001, MSE loss function, and batch size of 64, was carefully chosen to strike a balance between complexity, computational efficiency, and predictive accuracy. By making these decisions, the model is able to identify intricate patterns in both the historical weather data and the malaria case records, guaranteeing accurate and dependable forecasts. With the use of this methodology, the model is able to predict future cases of malaria with a high degree of accuracy, which helps to build early warning systems that are essential for proactive disease management and resource distribution in impacted areas.

3.6 Tools

The suggested thesis work is designed and implemented for this study using a variety of development tools. Modeling tools as well as additional research-related resources. A brief description of these development tools is provided in the sections that follow.

A. Design Tools

Design tools are the means by which design concepts are developed, displayed, and understood. The suggested approach is designed using lucid charts. Among other things, it is a sophisticated and lightweight visual design tool for making flowcharts and network diagrams that appear professional.

B. Hardware and Software tools

Hardware tools:

Table 6: hardware tools are going to be used in the research implementation

No table of figures entries found.	Tools	Used for
1.	GPU	To increase the computation and to fasten the training
2.	Hard Disk	Used as storage for large datasets.
3.	RAM	To accelerate the training process cooperatively with GPU

Software tools:

A description of the various software and coding tools that will be utilized to implement the research through coding is provided below.

Anaconda: in an application used to install the up-to-date version of python with its different modules and IDEs, for implementing the proposed solution an anaconda application version 1.9.7 with 64-bit support will be used.

Jupyter Notebook: is the most widely used and convenient Python IDE for researchers studying AI and deep learning. In this thesis, Jupiter Note-book 6.0.0 is used.

Python: - This thesis is implemented and demonstrated using Python programming, which necessitates the use of numerous drivers for configuring and packaging devices for computer installation.

Tensor flow: is an open-source deep learning framework by Google, offering flexibility and scalability for building and deploying machine learning models. Keras, integrated into Tensor Flow, is a high-level API that simplifies neural network prototyping with its user-friendly interface, making experimentation and customization straightforward.

Keras: is a Python-based high-level neural networks API that may be used with Microsoft Cognitive Toolkit (CNTK), Tensor Flow, or Theano. It runs smoothly on CPUs and GPUs, supports both convolutional and recurrent networks, and makes prototyping simple and quick.

3.7 Evaluation of model

Model evaluation is a crucial component of machine learning because it helps determine which model best fits the data that will be used for future predictions. Rather than evaluating model performance on training to prevent over fitting, model evaluation uses test sets. Various performance metrics are used to evaluate machine learning models, depending on the type of machine learning algorithms used. Additionally, it is preferable to use multiple evaluation metrics for a single model because one model may perform well when using one evaluation metric and poorly when using another. In this study, common evaluation metrics were utilized for prediction models, including MSE, RMSE, MAE, and R2 Score. MSE measures the average squared difference between actual and predicted values, with lower values indicating better performance. RMSE, the square root of MSE, represents the average magnitude of prediction errors in the same units as the target variable. MAE measures the average absolute difference between actual and predicted values, with lower values indicating better performance. The R2 score indicates the proportion of variance in the target variable that is predictable from the input variables, with higher values suggesting better model fit to the data.

CHAPTER 4: RESULTS AND DISCUSSION

4.1. Introduction

This chapter discusses experimental evaluation of the proposed model for the malaria prediction. We present the results of our experimental assessment of the feed forward ANN model that has been proposed for malaria prediction. This evaluation's main goal is to determine how effectively our suggested architecture performs in terms of accurately projecting the occurrence of malaria. The dataset used and the implementation results obtained from different evaluation techniques are described.

4.2. Data Collection and Preparation

Meteorological data relevant to malaria incidences were collected from the Hawassa meteorological station spanning 2017-2022, covering four different Woredas in the region (Boricha, Dale, Hawassa Zuria, and Shebedino). These data include variables such as minimum temperature, maximum temperature, and rainfall, known to influence malaria spread. Weekly malaria data for the same period (2017–2022) were collected for the four Woredas from the Sidama Health Region Bureau. The dataset consists of data from 4 woredas (districts), with each woreda contributing 288 records. This number comes from collecting data over 6 years, with each year having 48 records per woreda. The 48 records per year are calculated by gathering data every month, and each month provides 4 weekly records (12 months × 4 weeks = 48 records). Since the data was collected for 6 years, for each woreda, the total dataset equals 288 records (48 records per year × 6 years). Across the 4 woredas, the total dataset sums up to 1152 records (288 records × 4 woredas).

In order to accommodate for the variations in the Gregorian calendar—some months having 31 days, while others have 28 or 30 days—the data utilized in this study were transformed into 48 weeks. The length of each week was set at seven days to maintain uniformity. Some weeks had more than seven days because the extra days at the end of the month were merged with the remaining days when a month had more than twenty-eight days. This strategy was used to preserve consistency throughout the dataset and guarantee precise model assessment and evaluation. The data is partitioned into 80/20 for training and testing dataset.80 percent of the data is assigned for training and 20 percent of the data is allocated for testing.

Table 7: sample of the dataset (to show data collection and preparation)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
Station Name	Wereda	Altitude	Year	Month	w1(min.temp)	w2(min.temp)	w3(min.temp)	w4(min.temp)	w1(max.temp)	w2(max.temp)	w3(max.temp)	w4(max.temp)	w1(rainfall)	w2(rainfall)	w3(rainfall)	w4(rainfall)	w1(malaria_case)	w2(malaria_case)	w3(malaria_case)	w4(malaria_ca
Hawassa	Hawassa Zu	1694	2017	Jan	6.8	6.6	11.4	12.8	27.6	28.2	29.6	30.1	0.0	0.0	0.0	0.0	5	6	8	5
Hawassa	Hawassa Zu	1694	2017	Feb	14.9	13.7	13.9	15.5	30.1	30.4	29.6	29.6	0.9	0.6	8.8	2.3	10	10	4	10
Hawassa	Hawassa Zu	1694	2017	Mar	13.3	13.8	15.3	15.3	31.5	33.1	32.8	29.8	0.0	0.0	3.5	6.7	4	3	3	9
Hawassa	Hawassa Zu	1694	2017	Apr	14.0	14.3	15.6	15.2	30.5	32.2	32.6	31.2	0.8	0.9	0.0	8.4	3	3	5	4
Hawassa	Hawassa Zu	1694	2017	May	15.3	15.4	16.4	16.1	29.2	29.7	27.5	27.6	4.4	14.8	6.2	5.8	1	1	3	0
Hawassa	Hawassa Zu	1694	2017	Jun	16.1	16.5	14.4	15.1	27.3	29.2	27.3	27.3	2.1	3.1	0.6	2.0	4	1	3	3
Hawassa	Hawassa Zu	1694	2017	Jul	16.4	15.1	15.9	16.5	26.4	26.2	25.8	26.0	3.8	6.7	5.5	5.9	1	2	3	5
Hawassa	Hawassa Zu	1694	2017	Aug	15.3	15.2	14.9	16.4	27.0	25.6	27.4	26.5	7.4	5.1	0.6	2.4	0	2	4	4
Hawassa	Hawassa Zu	1694	2017	Sep	15.9	15.7	16.0	14.8	25.9	25.9	27.2	26.3	4.1	2.2	5.9	11.1	4	4	8	2
Yirgalem	Dale	1786	2022	Jan	11.2	7.1	8.0	6.5	28.6	28.6	29.1	30.0	0.0	0.0	5.3	0.0	32	33	25	39
Yirgalem	Dale	1786	2022	Feb	5.4	5.9	5.7	6.4	28.9	30.3	31.1	31.4	0.0	0.0	0.0	5.4	29	34	28	31
Yirgalem	Dale	1786	2022	Mar	6.0	5.2	5.8	6.7	31.0	31.1	30.2	29.1	0.0	5.1	9.3	8.6	34	30	20	205
Yirgalem	Dale	1786	2022	Apr	11.2	12.4	9.8	9.6	29.2	27.4	26.7	26.6	10.4	6.2	11.3	6.1	96	47	22	21
Yirgalem	Dale	1786	2022	May	10.7	12.4	10.5	11.4	26.6	26.5	26.1	26.5	2.2	5.2	1.4	8.1	28	26	43	37
Yirgalem	Dale	1786	2022	Jun	9.6	8.4	10.2	11.1	25.2	25.4	23.9	24.8	9.1	2.9	10.5	2.3	26	42	59	53
Yirgalem	Dale	1786	2022	Jul	12.2	13.3	13.1	12.5	25.2	24.4	24.9	23.6	0.0	3.5	3.5	3.2	55	47	46	90
Yirgalem	Dale	1786	2022	Aug	12.5	12.5	12.4	12.4	24.7	25.0	24.7	25.2	11.8	13.4	11.7	9.1	30	28	34	53
Yirgalem	Dale	1786	2022	Sep	12.7	13.3	11.8	13.3	24.9	25.0	25.4	25.3	23.5	20.6	20.6	19.2	54	85	45	53
Yirgalem	Dale	1786	2022	Oct	13.3	11.5	11.1	11.2	25.4	25.3	25.9	27.2	8.6	6.0	4.6	5.0	53	52	44	119
Yirgalem	Dale	1786	2022	Nov	12.1	11.0	10.8	10.6	25.7	27.4	26.0	26.6	1.5	0.0	0.0	1.5	54	67	56	294
Yirgalem	Dale	1786	2022	Dec	8.2	8.1	8.3	6.8	26.4	27.0	27.4	28.0	0.0	0.0	1.2	0.0	162	148	115	121

4.3 Development Environment

The experimental setup for our study utilized Keras with Tensor Flow as the backend, operating on an AMD Ryzen 5 5500u 2.1GHz CPU with 16 GB of RAM, running on a 64-bit Microsoft Windows 10 system. We conducted our experiments for malaria prediction using Jupyter Notebook, an interactive web application known for its versatility in coding and data visualization. Leveraging the compatibility of Jupyter Notebook with both local computing resources and cloud-based services allowed us to enhance our computational capabilities. In our research toolkit, we relied on essential tools such as Keras and Tensor Flow. Tensor Flow, renowned for its efficiency, provided a robust framework for deep learning tasks, while Keras, known for its user-friendliness, facilitated the implementation of optimized algorithms and popular machine learning techniques, particularly for prediction tasks. Within Tensor Flow, we employed the 'tf.keras. Sequential' module to construct our feed forward ANN model, specifying the architecture with the desired number of layers and neurons. This module facilitated the sequential stacking of layers, allowing us to define the input layer, hidden layer(s), and output layer with ease. Furthermore, we utilized key Python libraries including NumPy, Pandas, and Matplotlib, which played crucial roles in data manipulation, analysis, and visualization throughout our study.

4.4 IMPLEMENTATION

The proposed model has a batch size of 64 and is trained for 50 epochs, with a starting learning rate of 0.0001. In order to ensure that weight changes happen gradually, it is helpful for the optimization process to converge smoothly towards the ideal solution while using a lower learning rate. By reducing the possibility of exceeding the ideal solution, this method produces training behavior that is more dependable and consistent. Accurate forecasting of malaria occurrences depends on the model's ability to catch complicated patterns in the data, which is made possible by the loss function's steady lowering over time.

The dataset is partitioned into an 80/20 split, where 80% of the data is allocated for training the model and 20% is reserved for testing. This guarantees that the model gets tested on an unknown test set to gauge its generalization ability and provides an adequate amount of data from which to learn. Important factors that are known to affect the spread of malaria are included in the data, including rainfall, minimum and maximum temperatures, and historical weekly malaria case loads. These characteristics are essential for simulating the historical and environmental elements that contribute to the spread of malaria. These characteristics enable the model to account for the direct and indirect effects of historical caseload patterns and climatic change on upcoming malaria outbreaks.

Furthermore, by incorporating meteorological variables with historical caseload data, the model is able to take into consideration trends in the cycles of human disease transmission in addition to accounting for the seasonality and environmental drivers of malaria. This all-encompassing strategy increases the forecast accuracy of the model and helps to establish a strong early warning system, both of which can help public health authorities' better target interventions. The model can offer significant insights into malaria patterns because to a well-organized training procedure and thoughtful feature selection, which eventually helps to slow the disease's spread.

4.5 performance evaluation and Testing

In this study we have used different evaluation metrics to evaluate the performance of the malaria prediction proposed model. To measure the performance of the model we have used MSE, Root RMSE, MAE, and R2 Score.

MSE: it measures the average squared difference between the actual and predicted values. Lower MSE values indicate better predictive performance, with a value of 0 indicating perfect predictions.

RMSE: it is the square root of the MSE, providing a measure of the average magnitude of prediction errors in the same units as the target variable. Lower RMSE values suggest more accurate predictions.

MAE: it measures the average absolute difference between the actual and predicted values, providing a more interpretable metric compared to MSE. Like MSE and RMSE, lower MAE values indicate better predictive performance.

R2 Score: it represents the proportion of the variance in the target variable that is predictable from the input variables. It ranges from 0 to 1, where higher values indicate better model fit to the data.

Here is the formula for the above evaluation metrics[41]

1. Mean Squared Error

$$MSE = (1/n) * \Sigma (y\_i - \hat{y}\_i)^2$$

2. Root Mean Squared Error

$$RMSE = \sqrt{[(1/n) * \Sigma (y\_i - \hat{y}\_i)^2]}$$

3. Mean Absolute Error

$$MAE = (1/n) * \Sigma |y\_i - \hat{y}\_i|$$

4. R-squared

$$R^2 = 1 - [\Sigma (y\_i - \hat{y}\_i)^2 / \Sigma (y\_i - \bar{y})^2]$$

Where:

n is the number of observations.

y_i is the actual value.

$\hat{y}\_i$ is the predicted value.

$\bar{y}$ is the mean of the actual values.

4.6 Hyperparameters tuning

Hyperparameters are configuration settings used to structure and control the learning process. The number of hidden layers in a neural network, the batch size, the learning rate, the number of epochs, and the regularization parameters are a few examples. In contrast to model parameters, which are learned during training, hyperparameters are established before the learning process starts. Because it directly impacts the machine learning model's effectiveness and performance, fine-tuning these hyperparameters is essential[42]. The purpose of the code's hyperparameters tuning procedure is to find the optimal combination of hyperparameters in order to maximize the performance of the malaria prediction model. In this study, hyperparameters such as the learning

rate, number of epochs, batch size, and the number of units in the dense layers were manually tuned through a trial-and-error process. The number of units in the first and second dense layers was varied between 32 and 128 for the first layer and 16 to 64 for the second layer, while the learning rate was tested with values of 0.001 and 0.0001. The model was trained with a batch size of 64 across 50 epochs. Following several iterations, the model's performance was used to determine the best configuration, with a special emphasis on minimizing the validation loss, or more precisely, the MSE. To make sure each model generalizes well to new data, its performance was verified on a different validation set after it was trained using this configuration. These personally selected hyperparameters were then used to build the final model, guaranteeing that it was optimized for the particular dataset. The model was eventually well-suited for predicting malaria cases based on the provided dataset because to this manual approach to hyper parameter tuning, which allowed for a more intuitive grasp of how various configurations effect the model's prediction accuracy and resilience.

The learning path of the model can be observed by the plot of the training and validation loss across 50 epochs in Figure 4.1. Gradually declining training and validation losses show that the model is becoming more and more effective. There is little variation between the losses as they converge, indicating that the generalization to new data is effective. The loss values peak after about 40 epochs, indicating that more training does not result in noticeable improvements. This shows that the selected hyper parameters (learning rate, architecture, etc.) are appropriate for the task at hand and that the model is neither over fitting nor under fitting.

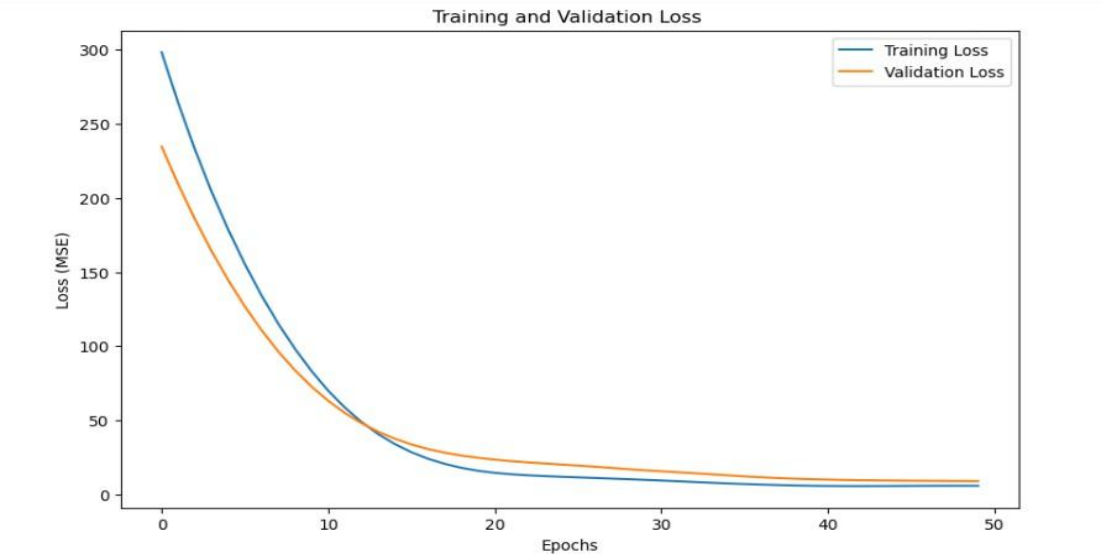

Figure 4.1 Training and Validation Loss curve

Table 8 : Summary Table for the Tuned Hyperparameters

Hyper parameter	Tested Values	Selected Value
Learning Rate	0.001, 0.0001	0.0001
Batch Size	32, 64	64
Number of Epochs	50, 100	50
Units (First Layer)	32, 64, 128	64
Units (Second Layer)	16, 32, 64	32
Loss Function	MSE	MSE

4.7 Evaluation of the model

The performance of the ANN feed forward model was assessed using various evaluation metrics, including RMSE, MSE, MAE, and the coefficient of determination (R^2). These metrics were calculated for the training, cross-validation, and testing phases for each region in the study: Boricha, Dale, Hawassa Zuriya, and Shebedino. The results are presented in the table below.

Table 9: Evaluation of ANN Feed forward model

	ANN Feed forward model											
provinces	RMSE			MSE			MAE			R^2		
	Training	Cross validation	Testing	Training	Cross validation	Testing	Training	Cross validation	Testing	Training	Cross validation	Testing
Boricha	2.45	2.65	2.610	6.25	7.02	6.8125	1.01	1.18	1.1458	0.9915	0.9910	0.9921
Dale	1.263	1.3	1.2416	1.44	1.69	1.5416	0.72	0.77	0.75	0.9923	0.9917	0.9928
Hawassa Zuriya	0.70	0.76										
Shebedino	0.45	0.504	0.4787	0.2025	0.25	0.22916	0.21	0.24	0.22916	0.9973	0.9950	0.99749

The evaluation of the ANN feed forward model across different provinces reveals varying levels of predictive accuracy. The training and testing values for RMSE are continuously low throughout the regions, suggesting that the model does a good

Job of minimizing the error between the values that are predicted and the actual values. For example, the RMSE values in Boricha are 2.45, 2.65, and 2.610 for testing, cross-validation, and training, respectively. This shows that the model's fit during the training and testing stages is balanced. Dale and Hawassa Zuriya both demonstrate similar consistency, with RMSE values in all stages below 1.5; Shebedino has the lowest testing RMSE, at 0.4787.
The MSE and MAE values, which both show low values throughout all provinces, especially in Hawassa Zuriya and Shebedino, further strengthen the model's outstanding performance. Shebedino, for instance, has an MSE and MAE of 0.22916 during the testing phase, indicating the model's accuracy in predicting outcomes with a low variance.

The R^2 values for all regions are above 0.99, indicating a strong goodness-of-fit, meaning the model explains more than 99% of the variance in the data across all phases. The model's resilience is further validated in Shebedino, where the highest R^2 value is recorded throughout the testing phase, reaching 0.99749.

The table's cross-validation results offer crucial information about how well the ANN feed forward model generalizes across other provinces. The cross-validation values of RMSE, MSE, and MAE show that the model retains its predictive accuracy when applied to unobserved data; they are marginally higher than the training values but still in the vicinity of the testing metrics. For instance, in Boricha, the testing RMSE of 2.610 is marginally lower than the cross-validation RMSE of 2.65. The model's capacity to explain the majority of the variation even during the validation phase is demonstrated by the cross-validation R2 values, which, although slightly lower than in training, are nevertheless strong across all regions and have values over 0.99. These findings imply that the model does not over fit and that it adapts well to fresh data.
Which is critical for reliable future predictions.

In summary, the ANN feed forward model demonstrates high predictive accuracy and generalization across different phases and provinces. The consistently low error rates and high R^2 values across training, cross-validation, and testing suggest that the model effectively captures the underlying patterns of the data and can be considered reliable for further analysis and decision-making.

4.8 Results of the model

The prediction results using the feed forward ANN model were evaluated by comparing the actual malaria case data with the predicted values. The model's performance is demonstrated in these various contexts by means of comparisons for the Shebedino, Dale, Hawassa Zuria, and Boricha regions. The graphs showing how well the ANN model predicts the temporal patterns of malaria incidence highlight the distinctive difficulties and trends that each region has to offer. These assessments offer a thorough rundown of the model's predictive power, which is crucial for comprehending how it might be used in actual situations.

The model's initial forecasts in Boricha match up well with the real instances of malaria, accurately reflecting the overall pattern from week 0 to week 20. However, the model struggles to anticipate more turbulent peaks and troughs during the mid-period weeks of 20 to 30. In spite of these problems, it becomes accurate again, closely mirroring the significant increase in cases approaching week 48. While there is still space for improvement in terms of capturing rapid, short-term oscillations, the model's high performance in identifying large trends suggests that it could be useful with further tuning for peak prediction accuracy. Figure 6.

The graph shows that the ANN model performed accurately in the Dale region over the course of 48 weeks. With very few departures from actual instances, the model faithfully replicates the early fluctuation trend from week 0 to week 20. Interestingly, the noteworthy peak at week 20 is accurately represented, despite a small overestimation. After then, the model closely mimics the real data and does a good job of adjusting to the oscillations seen through week 40. Despite small variations in peak magnitudes, the model is remarkably good at adapting to abrupt changes in trend, as evidenced by the astonishing accuracy with which the steep spike in malaria cases toward the end of the period is anticipated. Figure 7.

After an initial phase of underestimating, the feed forward ANN model shows a significant association with the real malaria case trends in Hawassa Zuria. The model finds it difficult to match the sharp early reduction in malaria cases from weeks 0 to 10. But starting in week 10, the model's predictions astonishingly match the real data, even with small changes being precisely captured. The model accurately predicts the significant increase in malaria cases around week 40, but it somewhat overestimates the incidence of cases in the final weeks. This performance indicates that early calibration improvements are necessary, but it also highlights the model's ability to align trends over the medium to long run. Figure 8.

The ANN model demonstrates a high degree of accuracy in predicting malaria cases in Shebedino, as illustrated in Figure 9. The model closely tracks the observed malaria trends throughout the 48-week period, especially during the early weeks when there is a significant rise in cases. Around weeks 5 to 15, the predicted values align almost perfectly with the actual malaria cases, capturing both the peaks and the subsequent declines. Following the peak malaria season, the model continues to provide accurate predictions, even as the case numbers drop steadily from week 20 onward. This close correspondence between predicted and actual values highlights the model's capacity to capture the intricate patterns of malaria incidence, particularly in regions where environmental conditions drive transmission dynamics. Minor deviations occur during sharp peaks, but overall, the prediction curve remains tightly aligned with the actual data, reinforcing the robustness of the ANN feed forward model for malaria forecasting in Shebedino.

As a whole, all regions exhibit a strong predictive performance from the feed forward ANN model, with varied degrees of success in predicting peak malaria incidences and quick shifts. Although the model shows great performance in areas such as Shebedino and Dale, in Hawassa Zuria and Boricha, there is need for improvement in initial response and peak fluctuations. These findings support the model's effectiveness in predicting malaria trends and offer a strong foundation for additional improvement and modification to boost accuracy in subsequent uses.

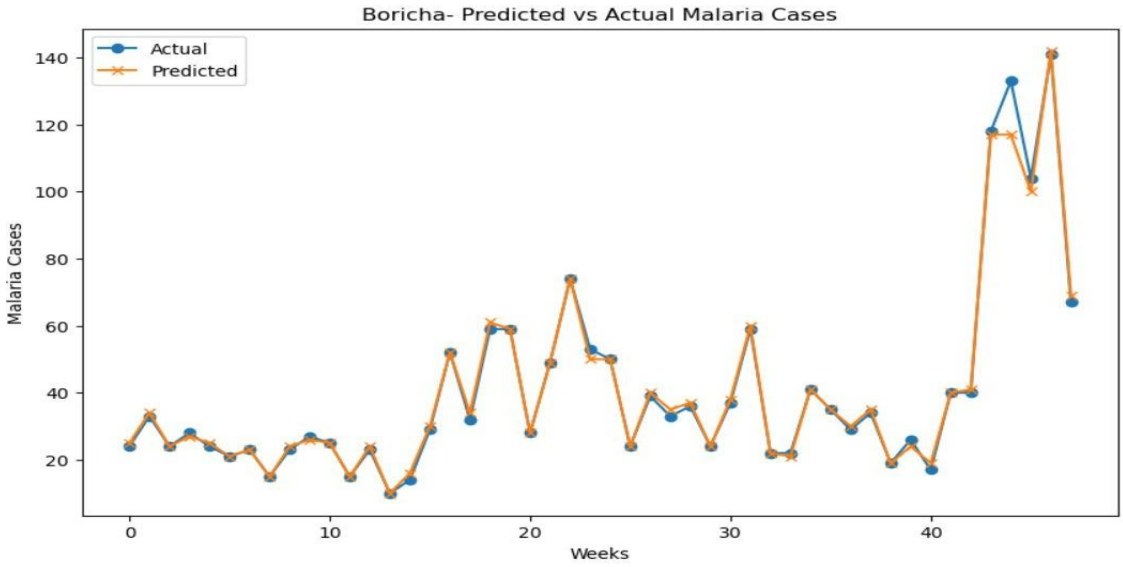

Figure 4.2: Boricha District Malaria Prediction

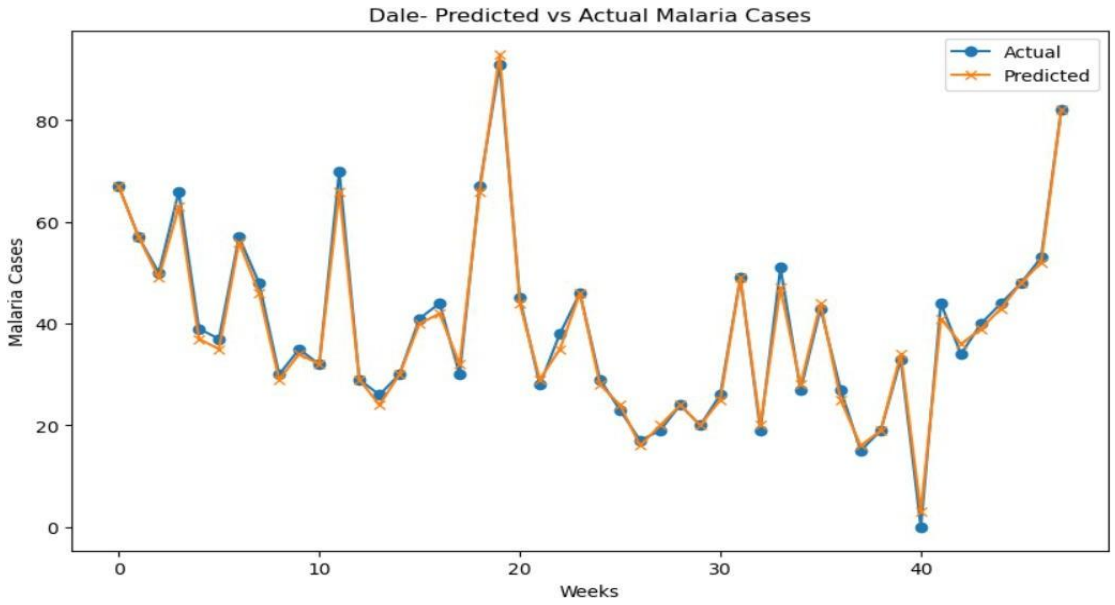

Figure 4.3: Dale District Malaria Prediction

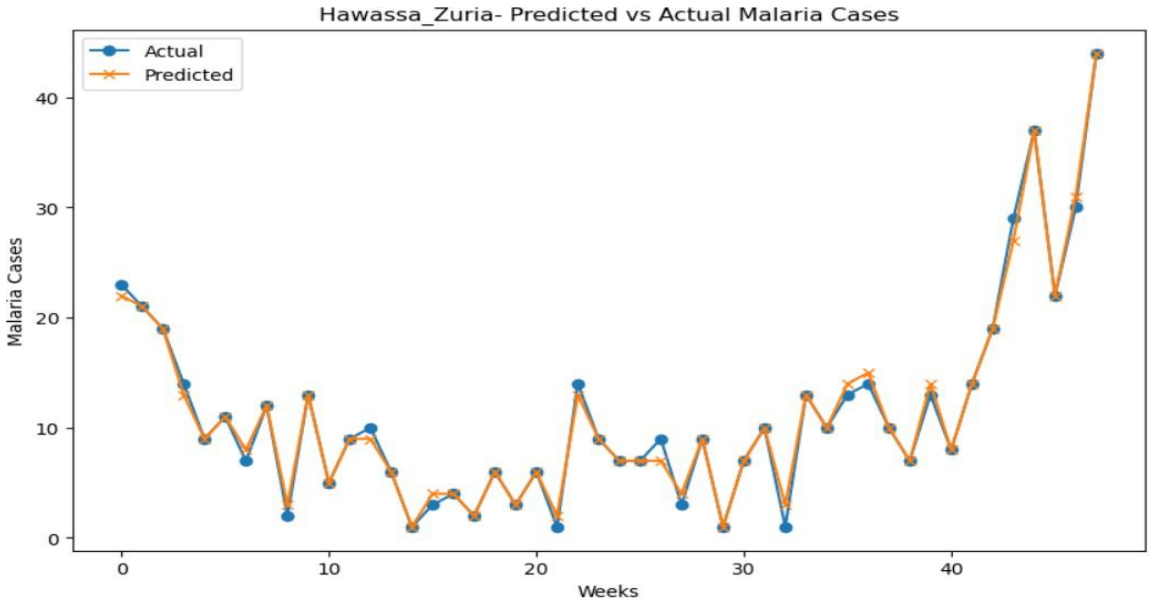

Figure 4.4: Hawassa Zuriya District Malaria Prediction

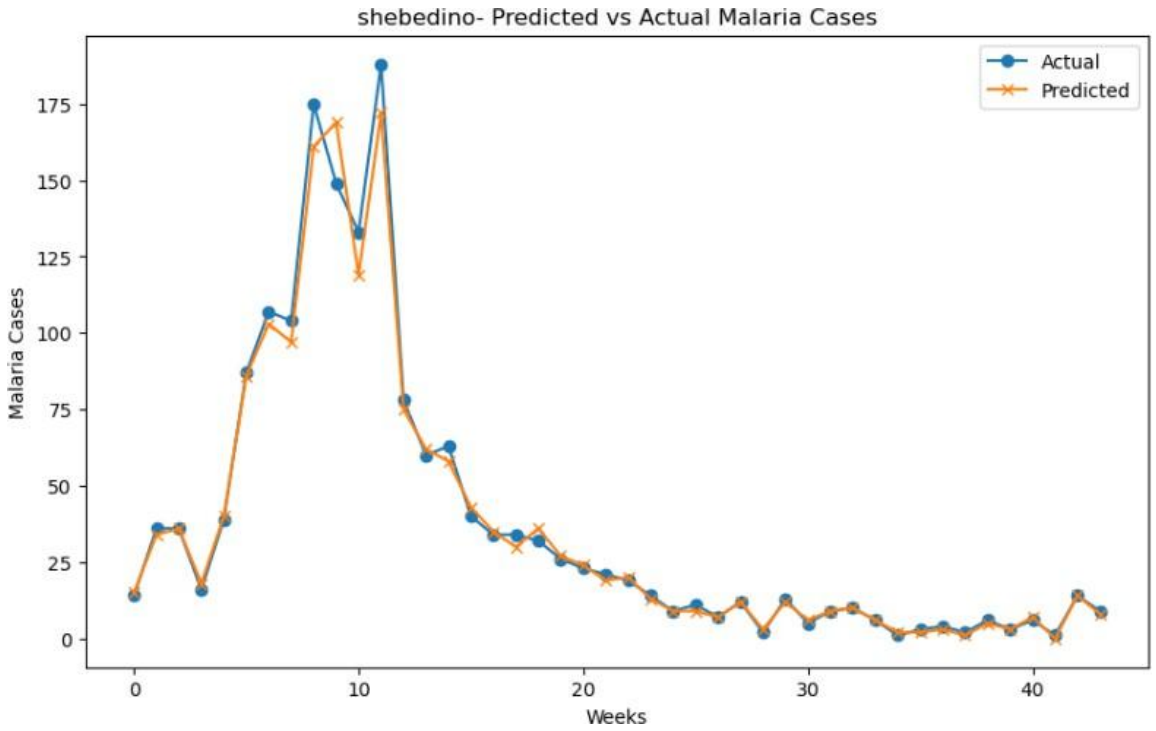

Figure 4.5: Shebdino District Malaria Prediction

4.9 Comparison of the Proposed Model with other models

In this study, the ANN Feed-forward model was used to predict malaria incidence across four districts: Boricha, Dale, Hawassa Zuriya, and Shebedino. Throughout the testing phase, the model's performance was measured using the RMSE.

Performance of the Proposed Model

The RMSE results for the proposed ANN Feed-forward model in the testing phase are as follows:

Table10: Performance of the Proposed Model

Districts	RMSE
Boricha	2.610
Dale	1.2416
Hawassa Zuriya	0.7359
Shebedino	0.4787

These results show that the model performed effectively; Shebedino had the lowest RMSE and Boricha had the highest. All things considered, the model's performance points to dependable forecasting skills, especially for areas with lower rates of malaria transmission.

Hyperparameters for the other models

In this work, the Random Forest model, Decision Tree and the ANN feed forward model were contrasted. The identical dataset, split ratio, and feature extraction methods used for the ANN model were also used by the researcher to tune the RF and DT model using the grid search method.

Grid Search: This methodical technique finds the best hyperparameters for a machine learning model. Using this approach, models are constructed and assessed for each possible combination of the hyper parameter values that have been given. Based on a predetermined evaluation metric, the model that performs the best is chosen [42]. Grid search is a flexible approach that works with different machine learning models. This method was implemented for this study using the GridSearchCV() function from the sklearn library.

As seen in Table 11, the grid search procedure required adjusting particular RF model hyperparameters. Every other hyper parameter that wasn't part of the grid search was kept at its default settings. A fair comparison with the ANN feed forward model is made possible by this careful technique, which guarantees that the RF model is optimized for the prediction job.

Table 11: Hyper parameter for Random Forest and Decision Tree

Models	Hyperparameters used in grid search space					
RF	Parameters	n-estimator	min_samples_leaf	min_samples_split	Max_depth	criterion
	Values	50, 100, 150,200	1, 2, 4,5,10,20,50,100	2, 5, 10	4,5,6,7,8,10,14, 20, None	MSE, MAE
DT	Parameters	min_samples_leaf		min_samples_split	Max_depth	criterion
	Values	1, 2, 4,5,10,20,50,100		2, 5, 10	4,5,6,7,8,10,14, 20, None	MSE, MAE

Random Forest (RF)

The researchers developed a predictive model using a Random Forest algorithm. To optimize the model's performance, various hyper parameter configurations were evaluated. The determination of the optimal hyperparameters was achieved through the application of the grid search method during the model training process. Following a series of experiments, the grid search identified the most effective combination of hyperparameters. The selected parameters are detailed in Table 12.

Table 12: Selected hyperparameters used for RF model

Parameters	n-estimator	min_samples_leaf	min_samples_split	Max_depth	criterion
Values	200	2	2,	4	MSE

So the experiment was done using the above-selected parameters, which were optimized using grid search methods.

Evaluation of Random Forest

The performance of the random forest model was assessed using various evaluation metrics, including RMSE, MSE, MAE, and the coefficient of determination (R^2). These metrics were calculated for the training, cross-validation, and testing phases for each region in the study: Boricha, Dale, Hawassa Zuriya, and Shebedino. The results are presented in the table below.

Table 13: Evaluation of Random Forest model

	Random Forest Model											
provinces	RMSE			MSE			MAE			R^2		
	Training	Cross validation	Testing	Training	Cross validation	Testing	Training	Cross validation	Testing	Training	Cross validation	Testing
Boricha	18.75	20.10	21.61	351.56	404.01	466.95	9.50	10.25	11.39	0.6574	0.5841	0.5125
Dale	8.3256	9.50	10.78	72.25	90.25	116.208	6.75	7.58	8.41	0.74689	0.6254	0.4819
Hawassa Zuriya	4.25	4.85	5.17	18.06	23.52	26.72	3.26	3.45	3.81	0.72365	0.6545	0.59247
Shebedino	3.80	4.05	4.29	14.44	16.40	18.42	2.85	3.10	3.38	0.8345	0.7848	0.7514

Results of Random Forest Model

The prediction results using the Random forest model were evaluated by comparing the actual malaria case data with the predicted values. The model's performance is demonstrated in these various contexts by means of comparisons for the Shebedino, Dale, Hawassa Zuria, and Boricha regions. The graphs showing how well the Random forest model predicts the temporal patterns of malaria incidence highlight the distinctive difficulties and trends that each region has to offer. These assessments offer a thorough rundown of the model's predictive power, which is crucial for comprehending how it might be used in actual situations. The result for the four woreda is listed below.

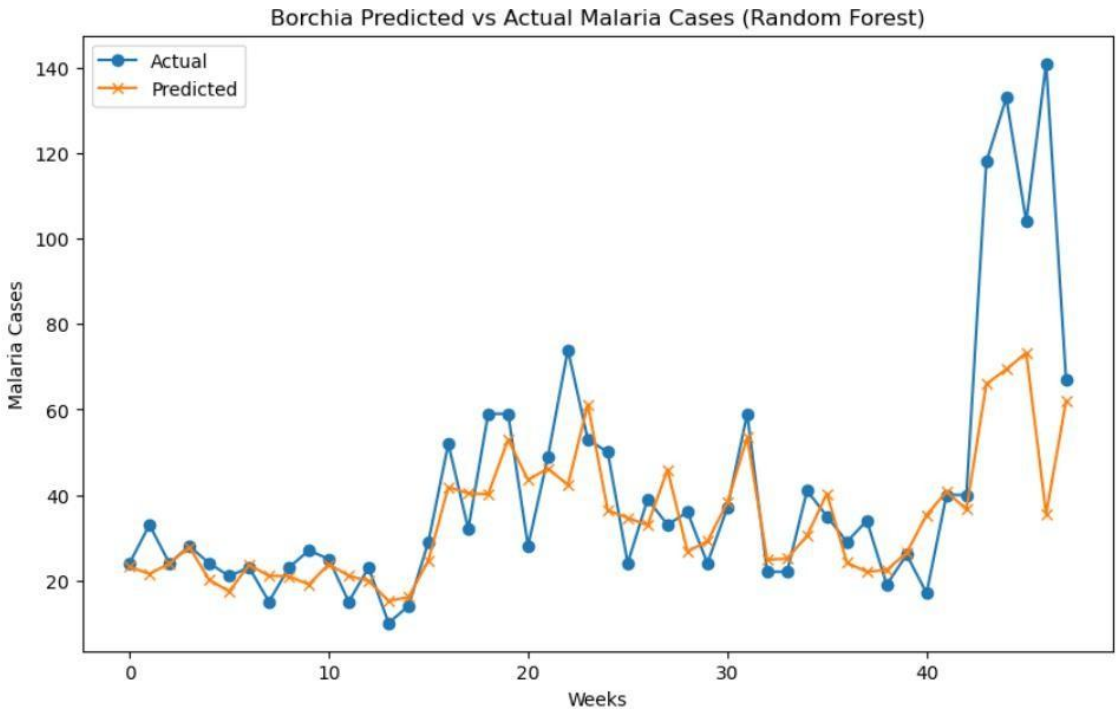

Figure 4.6: Boricha District Malaria Prediction (Random Forest)

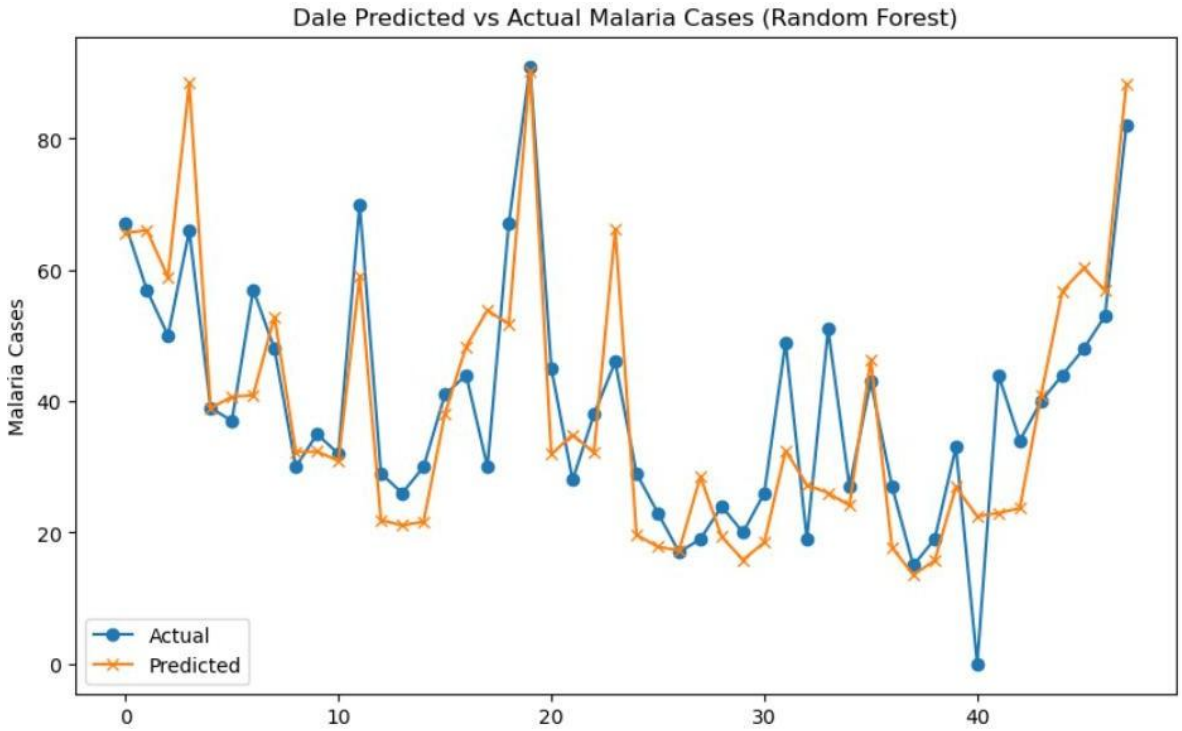

Figure 4.7: Dale District Malaria Prediction (Random Forest)

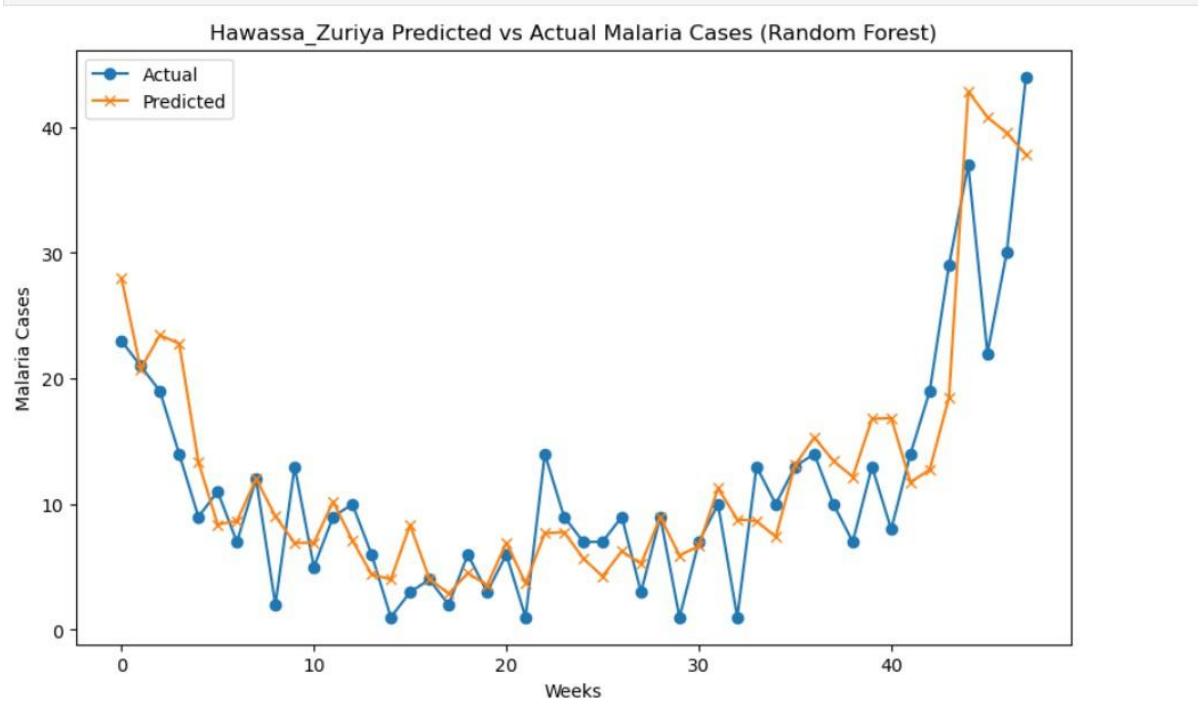

Figure 4.8: Hawassa Zuria District Malaria Prediction (Random Forest)

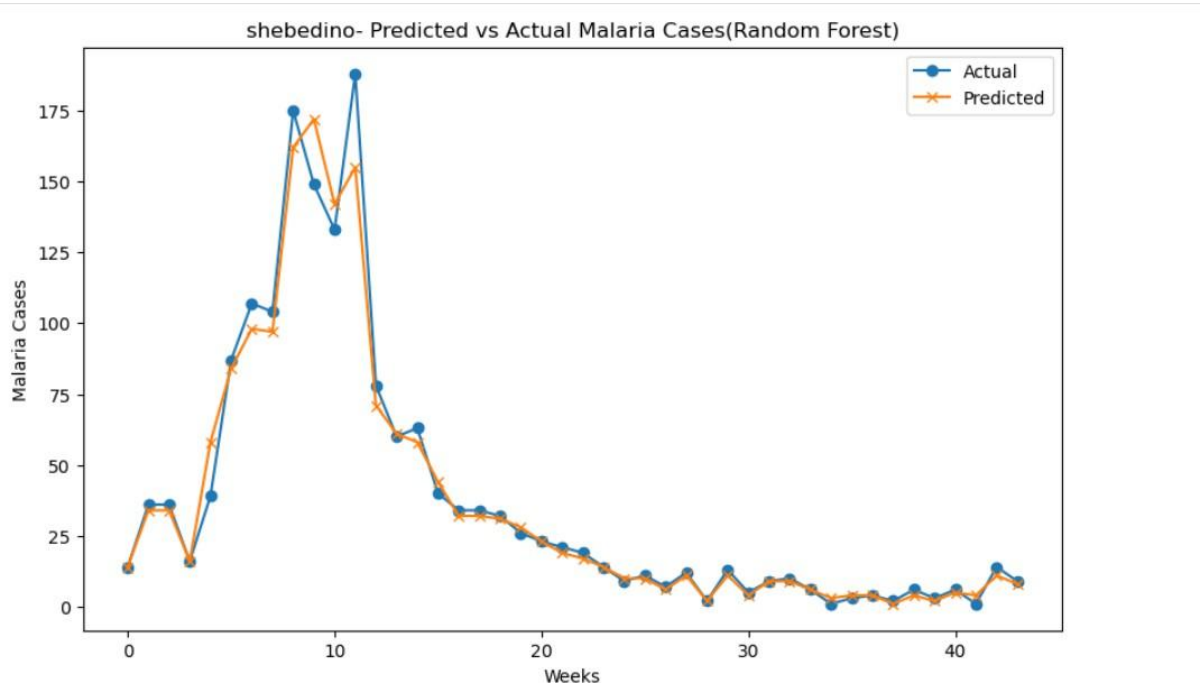

Figure 4.9: Shebedino District Malaria Prediction (Random Forest)

Decision Tree

The researchers developed a predictive model using a Decision Tree algorithm. To optimize the model's performance, various hyper parameter configurations were evaluated. The determination of the optimal hyperparameters was achieved through the application of the grid search method during the model training process. Following a series of experiments, the grid search identified the most effective combination of hyperparameters. The selected parameters are detailed in Table 13.

Table 14: Selected hyperparameters used for DT model

Parameters	min_samples_leaf	min_samples_split	Max_depth	criterion
Values	1	2,	5	MSE

Evaluation of Decision Tree

The performance of the DT model was assessed using various evaluation metrics, including RMSE, MSE, MAE, and the coefficient of determination (R^2). These metrics were calculated for the training, cross-validation, and testing phases for each region in the study: Boricha, Dale, Hawassa Zuriya, and Shebedino. The results are presented in the table below.

Table 15: Evaluation of Decision Tree model

	Decision Tree Model											
provinces	RMSE			MSE			MAE			R^2		
	Training	Cross validation	Testing	Training	Cross validation	Testing	Training	Cross validation	Testing	Training	Cross validation	Testing
Boricha	17.35	20.56	22.7503	300.76	423.12	517.5781	10.48	11.84	12.9618	0.7213	0.5927	0.4664
Dale	11.8923	13.746	14.5977	141.5258	189.00	213.0938	9.6734	10.491	11.1458	0.3421	0.2415	0.1909
Hawassa Zuriya	7.1946	8.115	8.8272	51.7535	65.86	77.919	4.9184	5.34	5.6736	0.3203	0.2439	0.1848
Shebedino	9.8721	10.48	11.19	97.455	109.97	125.29	5.829	6.15	6.47	0.7154	0.6812	0.6479

Results of Decision Tree Model

The prediction results using the Decision Tree model were evaluated by comparing the actual malaria case data with the predicted values. The model's performance is demonstrated in these various contexts by means of comparisons for the Shebedino, Dale, Hawassa Zuria, and Boricha regions. The graphs showing how well the Decision Tree model predicts the temporal patterns of malaria incidence highlight the distinctive difficulties and trends that each region has to offer. These assessments offer a thorough rundown of the model's predictive power, which is crucial for comprehending how it might be used in actual situations. The result for the four woreda is listed below.

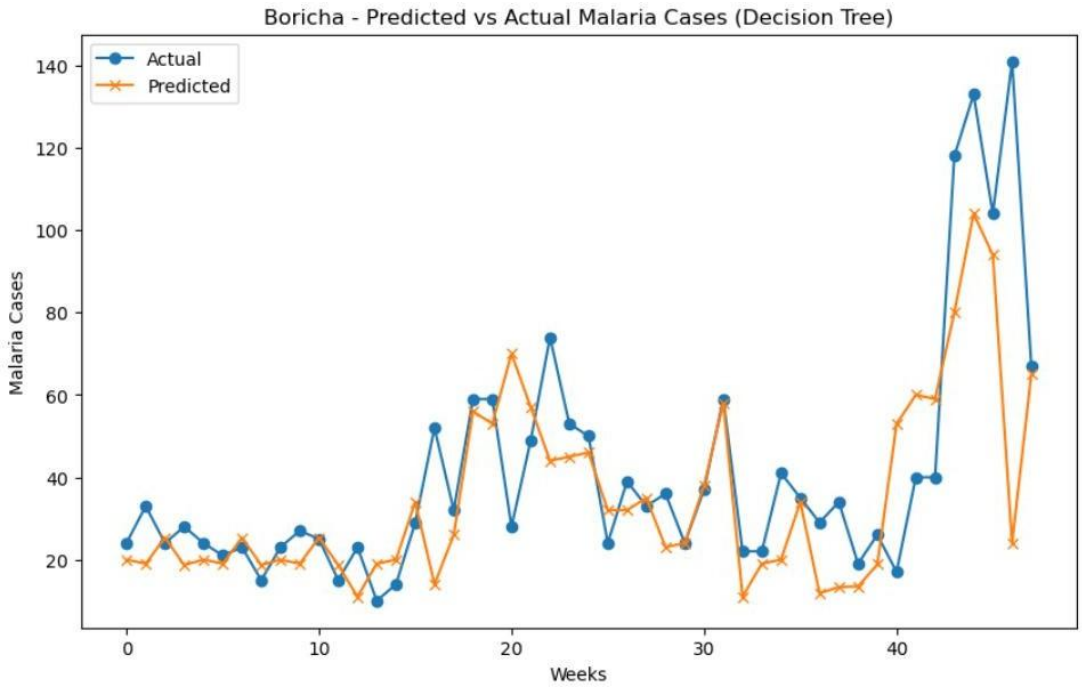

Figure 4.11: Boricha District Malaria Prediction (Decision Tree)

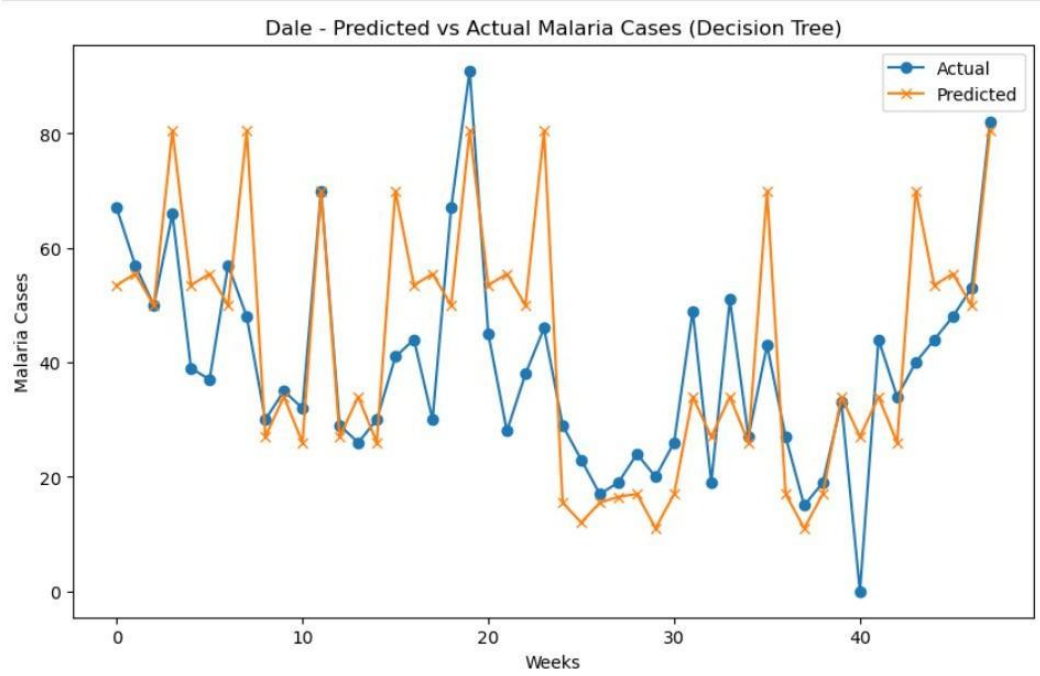

Figure 4.12: Dale District Malaria Prediction (Decision Tree)

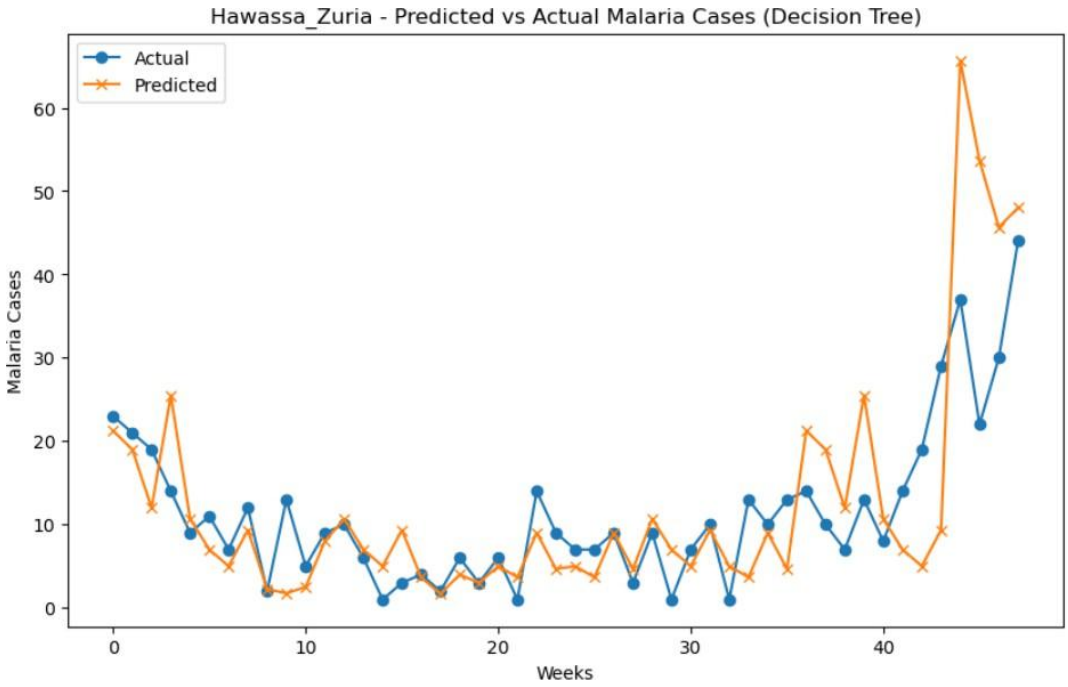

Figure 4.13: Boricha District Malaria Prediction (Decision Tree)

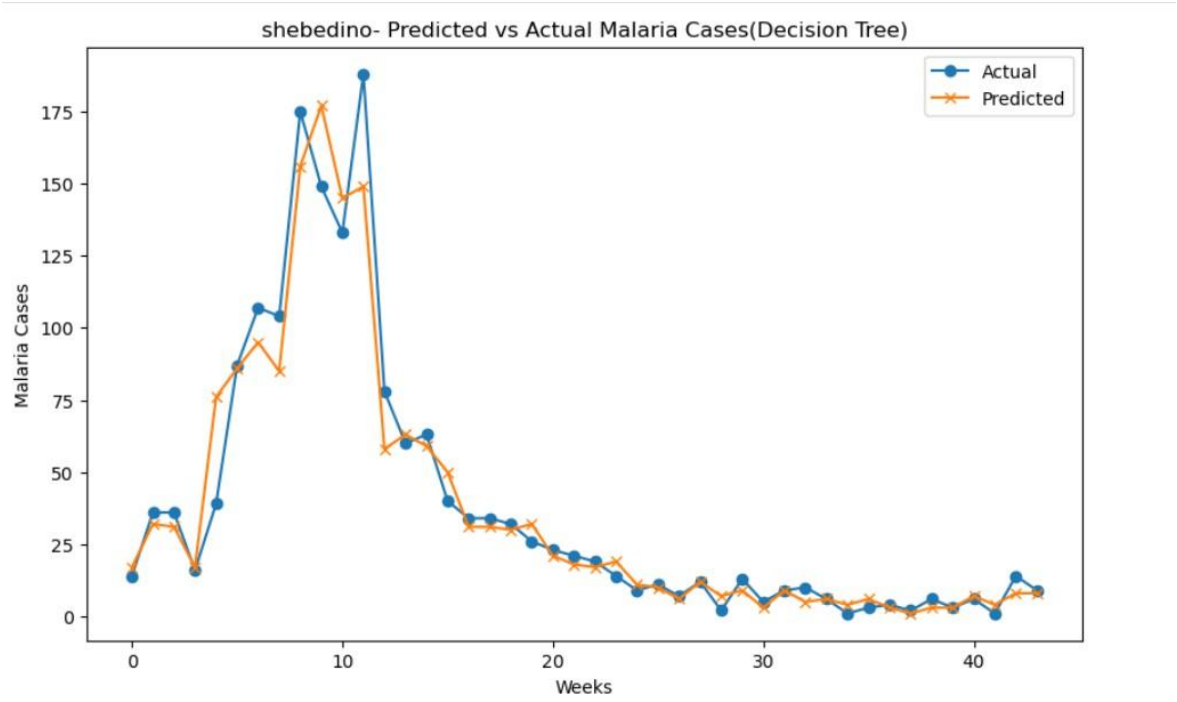

Figure 4.14: Shebedino District Malaria Prediction (Decision Tree)

Result discussions

The study involved the development of three machine learning models, including an Artificial Neural Network (ANN) feed-forward model, a Decision Tree model, and a Random Forest model, using a data split ratio of 80/20 for training and testing. These models were evaluated based on performance metrics such as Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). The performance of the models was then compared to determine their predictive accuracy and practical suitability for forecasting malaria cases. The ANN feed-forward model, the Decision Tree model, and the Random Forest model were tested on the same dataset under consistent conditions to ensure a fair comparison. The results of this analysis are presented in the following table, showcasing how each model performs across the evaluation metrics, which are crucial for selecting the most appropriate model for malaria incidence forecasting.

Table 16: Summary for the Comparison of the models

Districts	**Algorithms**	**RMSE**	MSE	MAE	R^2
Boricha	**ANN(feed-forward)**	2.610	6.812	1.145	0.9921
	RF	21.61	466.95	11.39	0.5125
	DT	22.75	517.57	12.96	0.4664
Dale	**ANN(feed-forward)**	1.2416	1.541	0.75	0.9928
	RF	10.78	116.208	8.41	0.4819
	DT	14.59	213.09	11.14	0.1909
Hawassa Zuriya	**ANN(feed-forward)**	0.7359	0.5416	0.4166	0.9935
	RF	5.17	26.72	3.81	0.5924
	DT	8.827	77.919	5.673	0.1848
Shebedino	**ANN(feed-forward)**	0.4787	0.2291	0.2291	0.9974
	RF	4.29	18.42	3.38	0.7514
	DT	11.19	125.29	6.47	0.6479

The table presents the comparative performance of the Artificial Neural Network (ANN) feed-forward model, Decision Tree (DT), and Random Forest (RF) models across four districts: Boricha, Dale, Hawassa Zuriya, and Shebedino. The evaluation metrics used are Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). Overall, the ANN model consistently outperforms the other two models, achieving the lowest RMSE, MSE, and MAE values, as well as the highest R^2 values across all districts. While the RF model performs moderately well, the DT model generally exhibits higher errors and lower R^2 values, indicating relatively less predictive accuracy. This comparison highlights the superior performance of the ANN model for malaria incidence forecasting in all districts.

CHAPTER 5: CONCLUSION AND RECOMMENDATION

5.1 Conclusion

This study used an ANN feed forward model to forecast malaria incidence in the Sidama Regional State by integrating case load and meteorological data. The results addressed the first research question, confirming the crucial role of meteorological factors such as rainfall, minimum, and maximum temperatures in malaria transmission. These factors, identified as critical through Pearson correlation analysis, strongly influence malaria incidence and are essential for building accurate predictive models. Including these environmental elements significantly enhances the accuracy of malaria incidence forecasts. To answer the second research question what machine learning techniques are most effective in predicting malaria transmission? The study compared Random Forest, Decision Tree, and ANN feed forward models. The ANN feed forward model emerged as the most effective technique due to its ability to capture complex relationships between meteorological data and malaria case trends. It achieved lower RMSE values compared to the other models, particularly in districts like Shebedino and Dale. However, challenges remain in capturing short-term fluctuations in areas like Boricha and Hawassa Zuria, indicating opportunities for further refinement. The ANN model's accuracy highlighted the significant influence of environmental factors, such as rainfall and temperature variations, on malaria transmission patterns. Overall, ANN-based forecasting provides public health officials with a reliable early warning system. This predictive capability is critical for reducing the impact of malaria epidemics, particularly in resource-limited settings where efficient intervention allocation is vital. By addressing both research questions, this study demonstrates the superiority of ANN models for malaria prediction and their potential to contribute to public health strategies in malaria-endemic areas.

5.2 Recommendations

Improve Data Collection: Meteorological and malaria case data should be regularly enhanced in terms of availability and quality in order to increase the accuracy of predictive models. More temporal precision and the integration of real-time data streams should improve model accuracy Even more, especially for short-term forecasts.

Expand Model Training: Although the ANN model performed well, incorporating other advanced machine learning techniques, such as Long Short-Term Memory (LSTM) networks, could improve the prediction of malaria trends, especially in regions with complex transmission patterns.

Regional Customization: The performance of the prediction model differed between the regions, indicating that models tailored to a particular location might produce superior outcomes. Adapting models to regional climate and epidemiological circumstances may improve forecast accuracy and boost the effectiveness of anti-malarial initiatives.
Use in Public Health Policy: These predictive models need to be incorporated into national malaria control plans by public health experts. In high-risk locations, the deployment of insecticide-treated nets, targeted community awareness campaigns, and vector control measures can all be made easier with the help of early warning systems based on these models.

Research and Development to Come: Future studies should investigate how to incorporate data on socioeconomic status and healthcare accessibility into predictive models in order to create a more comprehensive picture of malaria risk. Furthermore, investigating the possibilities of alternative machine learning frameworks may result in models that are more precise and flexible for varying environmental circumstances.

5.3 Contributions

- Development of a Malaria Prediction Model: In order to anticipate the incidence of malaria, this study created a machine learning-based prediction model that makes use of an ANN feed-forward architecture. The model is constructed using meteorological data, which incorporates environmental elements that affect the transmission of malaria and greatly enhances prediction accuracy**.**
- Integration of Case Load and Meteorological Data: This study illustrates how environmental factors including rainfall, min temperature and max temperature can be important indicators of malaria outbreaks by fusing historical case data with meteorological data. This method provides a more thorough understanding of the dynamics of the disease's transmission in connection to climate variables.

- Application in Public Health: By incorporating the suggested model as an early warning system into public health plans, health officials will be able to predict malaria outbreaks and take prompt action. In high-risk areas, this might significantly lessen the burden of disease and improve resource allocation.

- Regional Customization for Better Accuracy: The study shows that different regions have different predictive performance for the model, indicating that region-specific models can improve prediction accuracy even more. This discovery promotes the creation of regionally tailored models for the prediction of malaria in various meteorological and epidemiological contexts.

Reference

[1] WHO, “Malaria,” https://www.who.int/news-room/questions-and-answers/item/malaria.

[2] WHO, “World malaria report 2015,” World Health Organization, 2016.

[3] Q. Hung, P. J. Vries, P. T. Giao, N. V. Nam, T. Q. Binh, M. T. Chong, et al., “Control of malaria: a successful experience from Viet Nam,” Bull. World Health Organ. vol. 80, no. 8, pp. 660–666, 2002.

[4] Z. H. Zhou, Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC, 2012.

[5] WHO, “World malaria report 2021,” https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2021.

[6] F. A. Kendie, W. T. Hailegebriel, E. N. Semegn, and M. W. Ferede, “Prevalence of malaria among adults in Ethiopia: A systematic review and meta-analysis,” J. Trop. Med., vol. 2021, Mar. 2021, doi: 10.1155/2021/8863002.

[7] D. Dabaro, Z. Birhanu, W. Adissu, et al., “Prevalence and predictors of asymptomatic malaria infection in Boricha District, Sidama Region, Ethiopia: implications for elimination strategies,” Malar. J., vol. 22, p. 284, 2023, doi: 10.1186/s12936-023-04722-z.

[8] D. Sakubu, K. J. G. Sinigirira, and D. Niyukuri, “Predicting malaria dynamics in Burundi using deep learning models,” 2023, doi: 10.48550/arXiv.2306.02685.

[9] E. Kamana, J. Zhao, and D. Bai, “Predicting the impact of climate change on the re-emergence of malaria cases in China using LSTMSeq2Seq deep learning model: A modeling and prediction analysis study,” BMJ Open, vol. 12, no. 3, p. e053922, Mar. 2022, doi: 10.1136/bmjopen-2021-053922.

[10] WHO, “WHO malaria terminology,” http://www.who.int/malaria, 2019.

[11] D. A. Fidock, “Malaria parasite biology: The life and times of Plasmodium falciparum,” Cold Spring Harbor Perspect. Med., vol. 8, no. 7, p. a025569, Jul. 2018, doi: 10.1101/cshperspect.a025569.

[12] J. Sattabongkot and J. W. Barnwell, “Malaria parasite biology: The life and times of Plasmodium vivax and Plasmodium falciparum,” Cold Spring Harbor Perspect. Med., vol. 8, no. 7, p. a025567, Jul. 2018, doi: 10.1101/cshperspect.a025567.

[13] “Supervised machine learning,” JavaTpoint, https://www.javatpoint.com/supervised-machine-learning.

[14] T. Hastie, R. Tibshirani, and J. Friedman, the Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, 2009.
[15] A. Liaw and M. Wiener, “Classification and regression by random forest,” R News, vol. 2, no. 3, pp. 18–22, 2002.
[16] A. J. Hill, G. R. Herman, and R. S. Schumacher, “Forecasting severe weather with random forests,” Department of Atmospheric Science, Colorado State University, Fort Collins, CO, 2020.
[17] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
[18] V. Vapnik, Statistical Learning Theory. Wiley-Interscience, 1998.
[19] Y. Radhika and M. Shashi, “Atmospheric temperature prediction using support vector machines,” Int. J. Comput. Theory Eng., vol. 1, no. 1, pp. 1793–8201, Apr. 2009.
[20] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2nd ed. Wiley-Interscience, 2001.
[21] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann, 2011.
[22] K. Samruddhi and R. A. Kumar, “Used car price prediction using K-nearest neighbor-based model,” Int. J. Innov. Res. Appl. Sci. Eng., vol. 4, no. 2, pp. 629–632, Aug. 2020, doi: 10.29027/IJIRASE.v4.i2.2020.629-632.
[23] J. R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.
[24] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Chapman & Hall/CRC, 1984.
[25] M. Apolinar-Gotardo, “Using decision tree algorithm to predict student performance,” Indian J. Sci. Technol., vol. 12, no. 5, 2019, doi: 10.17485/ijst/2019/v12i5/140987.
[26] J. Schmidhuber, “Deep learning in neural networks: An overview,” Neural Netw. vol. 61, pp. 85–117, 2015.
[27] G. E. Hinton, S. Osindero, and Y. W. Teh, “A fast learning algorithm for deep belief nets,” Neural Comput., vol. 18, no. 7, pp. 1527–1554, 2006.
[28] R. Salakhutdinov and G. Hinton, “Deep Boltzmann machines,” Int. Conf. Artif. Intell. Stat., pp. 448–455, 2009.
[29] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[30] L. Deng and D. Yu, “Deep learning: Methods and applications,” Found. Trends Signal Process. vol. 7, no. 3–4, pp. 197–387, 2014.

[31] A. Dukuzumuremyi, Machine learning-based prediction of malaria outbreak using environment data in Rwanda (Doctoral dissertation). University of Rwanda, 2020.

[32] A. Stephen, P. O. Akomolafe, and K. I. Ogundoyin, “A model for predicting malaria outbreak using machine learning technique,” Ann. Comput. Sci. Ser., vol. 19, no. 1, 2021.

[33] O. Nkiruka, R. Prasad, and O. Clement, “Prediction of malaria incidence using climate variability and machine learning,” Inform. Med. Unlocked, vol. 22, p. 100508, Jan. 2021, doi: 10.1016/j.imu.2020.100508.

[34] Y. A. Adamu, “Malaria prediction model using machine learning algorithms,” Turk. J. Comput. Math. Educ., vol. 12, no. 10, pp. 7488–7496, 2021.

[35] A. A. Taddese, A. G. Baraki, and K. A. Gelaye, “Spatial modeling, prediction, and seasonal variation of malaria in northwest Ethiopia,” BMC Res. Notes, vol. 12, no. 1, p. 273, Dec. 2019, doi: 10.1186/s13104-019-4305-1.

[36] T. A. Abeku, S. J. De Vlas, G. Borsboom, et al., “Forecasting malaria incidence from historical morbidity patterns in epidemic-prone areas of Ethiopia: A simple seasonal adjustment method performs best,” 2002, doi: 10.1046/j.1365-3156.2002.00924.x.

[37] E. Loha and B. Lindtjørn, “Model variations in predicting incidence of Plasmodium falciparum malaria using 1998-2007 morbidity and meteorological data from south Ethiopia,” Malar. J., vol. 9, no. 1, p. 166, Dec. 2010, doi: 10.1186/1475-2875-9-166.

[38] J. Xia, C. Zhang, and L. Yang, “Application of artificial neural networks in climate change and epidemiology: A review,” J. Comput. Sci., vol. 33, pp. 50–65, 2019.

[39] M. Kar and S. Bansod, “Artificial neural network-based rainfall-runoff modeling for dengue disease prediction,” Environ. Monit. Assess. vol. 188, no. 4, p. 209, 2016.

[40] W. Yang, B. J. Cowling, and E. H. Lau, “Forecasting influenza epidemics using neural networks with meteorological data,” PLoS ONE, vol. 10, no. 7, p. e0130466, Jul. 2015, doi: 10.1371/journal.pone.0130466.

[41] G. James, D. Witten, T. Hastie, and R. Tibshirani, an Introduction to Statistical Learning: With Applications in R. Springer, 2013.

[42] M. Feurer and F. Hutter, “Hyper parameter optimization,” in Automated Machine Learning, Springer, 2019, pp. 3–33.

APPENDICES

Appendix A: Sample code for importing necessary packages

```
import pandas as pd
from sklearn.model_selection import TimeSeriesSplit
import numpy as np
import tensorflow as tf
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
from sklearn.preprocessing import MinMaxScaler
import matplotlib.pyplot as plt
```

```
import pandas as pd
from sklearn.model_selection import TimeSeriesSplit
import numpy as np
import tensorflow as tf
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
import matplotlib.pyplot as plt
```

Appendix B: Sample code for Defining features and target

```
# Define features and target
features = data[['w1(min.temp)', 'w2(min.temp)', 'w3(min.temp)', 'w4(min.temp)',
                 'w1(max.temp)', 'w2(max.temp)', 'w3(max.temp)', 'w4(max.temp)',
                 'w1(rainfall)', 'w2(rainfall)', 'w3(rainfall)', 'w4(rainfall)',
                 'w1(malaria_case)', 'w2(malaria_case)', 'w3(malaria_case)', 'w4(malaria_case)']]
target = data[['w1(malaria_case)', 'w2(malaria_case)', 'w3(malaria_case)', 'w4(malaria_case)']]
```

Appendix C: Sample code for preprocessing

```
# Normalize the features to the range [5, 95]
scaler = MinMaxScaler(feature_range=(0.05, 0.95))
X = scaler.fit_transform(features)
y = np.array(target)

n_splits = 5
tscv = TimeSeriesSplit(n_splits=n_split
```

Appendix D: Sample code for Defining the neural network architecture

```
# Define the neural network architecture
model = tf.keras.Sequential([
        tf.keras.layers.Dense(64, activation='relu', input_shape=(X_train.shape[1],)),
        tf.keras.layers.Dense(32, activation='relu'),
        tf.keras.layers.Dense(4) # Output layer with 4 nodes for w1 to w4 malaria case
    ])
```

Appendix E: Sample code for Training the model with early stopping

```
# Train the model with early stopping
model.fit(X_train, y_train, epochs=epochs, batch_size=batch_size, validation_data=(X_val, y_val), callbacks=[early_stopping], verbose=0)
```

Appendix F: Sample code for calculating the Evaluation metrics

```
  # Calculate the RMSE and MAE
rmse = np.sqrt(mean_squared_error(y_val, y_pred))
mse = mean_squared_error(y_val, y_pred)
mae = np.mean(np.abs(y_val - y_pred))
r2 = r2_score(y_val, y_pred)
rmse_values.append(rmse)
mae_values.append(mae)
r2_values.append(r2)
mse_values.append(mse)
print(f"Fold {fold} - RMSE: {rmse},MSE: {mse}, MAE: {mae},R2: {r2}")
```

Appendix G: Sample code for Visualizing the Predicted and Actual Values

```
 # Visualize the predicted and actual values
plt.figure(figsize=(10, 6))
plt.plot(y_val.flatten(), label='Actual', marker='o')
plt.plot(y_pred.flatten(), label='Predicted', marker='x')
plt.title(f'shebedino- Predicted vs Actual Malaria Cases')
plt.xlabel('Weeks')
plt.ylabel('Malaria Cases')
plt.legend()
plt.show()
```