ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

A FRAMEWORK FOR PREDICTIVE ANALYSISOF MALARIA DISPERSION

Abreham Kassahun

A THESIS SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE IN PARTIAL FULFILMENT FOR THE DEGREE OF MASTERS OF SCIENCE IN COMPUTER SCIENCE

June, 2018

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

Abreham Kassahun Shiferaw

Advisor: Solomon Atnafu (PhD)

This is to certify that the thesis prepared by *Abreham Kassahun Shiferaw*, titled *A Framework for Predictive Analysis of Malaria Dispersion* and Submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science compiles with the regulations of the University and meets the accepted standards with respect to originality and quality.
Signed by Examining Committee:

Name		Signature	Date
Advisor:	*Solomon Atnafu (PhD)*	_____	_____
Examiner:	*Dida Midekso (PhD)*	_____	_____
Examiner:	*Fekade Getahun (PhD)*	_____	_____

Abstract

As most researches have shown that health related information is inaccessible in developing counties like Ethiopia, and when it is accessible it is not used for decision making. Analysing health information for decision making has an immense contribution in combating societal health problems especially diseases like Malaria which is deadly in developing countries.

Qualitative and quantitative researches have been made so far on malaria dispersion to model prediction of malaria dispersion based on various determinants. Invariably, the models considers climate as a factor of malaria dispersion but other variables such as population density, surface hydrology, living standard, population income, GDP, life span of mosquito are considered unevenly.

In this work, the original datasets of malaria dispersion, climate, NDVI and population from year 2009 to 2015 are transformed into a dataset of 3671 cases so as to train algorithms to produce a model that helps to predict malaria dispersion. Among the trained algorithms, multilayer perceptron using softplus as activation function is selected as a best algorithm with a correlation coefficient of 0.9503 by using the evaluation outcome of 10 fold cross validation.

The prototype implementation of the system that uses the prediction algorithm has a design pattern to facilitate the construction of an extensible system in order to integrate new algorithms into a system in such a way that modelers can use the system to upload their algorithm and build a specific model by using the existing data. Once the models are built, the system has a capability for users to select a model and to provide inputs so that the system will provide the malaria dispersion result along with the evaluation outcome.

Key Words: - *Malaria, Malaria dispersion, Prediction model, Predictive analysis, Multilayer perceptron*

Acknowledgements

It is with immense gratitude that I acknowledge the support and help of my supervisor Dr. Solomon Atnafu. I came up with a vast problem area and shown me how to deal with problems by narrowing to the level required. His understanding of personal and profession matters.

I owe my deepest gratitude to Dr. Solomon Kibret, a great epidemiologist and teacher, for his encouragement and valuable comments.

I would like to acknowledge the Ethiopian Health Centre Institute and the Ethiopian Metrology Agency for helping me to get the necessary data. I wish also to thank Dr. Mesafint and Tesfaye for their support in facilitating administrative issues.

I would like to thank all of our instructors in the MSc. program. I have learnt a lot from you.

I would like also to thank my family and friends for their support and encouragement.

Above all, I would always like to thank the Almighty God for everything.

Table of Contents

List of Figures

List of Tables

List of Acronyms

AID	Automatic Interaction Detector
AVHRR	Advanced Very High Resolution Radiometer
CART	Classification and Regression Trees
CHAID	Chi-square Automatic Interaction Detection
CRISP-DM	Cross-Industry Standard Process Model for Data Mining
CSA	Central Statistics Agency
EIR	Entomological Inoculation Rate
ELISEE	Exploration of Links and Interactions through Segmentation of an Experimental Ensemble
GCM	General Circulation Model
GDP	Gross Domestic Product
GDPpc	Gross Domestic Product per capita
HMIS	Hospital Management Information System
IPCC	Intergovernmental Panel on Climate Change
IRS	Indoor Residual Spraying
ITN	Insecticide-Treated Nets
JSON	Java Script Object Notation
K-NN	K-Nearest Neighbour
LLIN	Long Lasting Impregnated Net
LOF	Local Outlier Factor
LST	Land Surface Temperature
MAP	Malaria Atlas Project
MCMC	Markov Chain Monte Carlo
MIM	Malaria Impact Models
MIS	Malaria Indicator Survey
MODIS	Moderate Resolution Imaging Spectroradiometer
NCEP	National Centre for Environmental Prediction
NDVI	Normalized Difference Vegetation Index
NMA	National Metrology Agency

NMIS	National Malaria Control Programme
RBM	Roll Back Malaria partners
RBM-MERG	Roll Back Malaria-Measurement and Evaluation Reference Group
RCPs	Representative Concentration Pathways
RDT	Rapid Diagnosis Test
SES	Socioeconomic status
SMO	Sequential Minimal Optimization Algorithm
SNNP	Southern Nation Nationalities and People's
SRES	Special Report on Emissions Scenarios
SVM	Support Vector Machine
UC	University of California
UNDP	United Nation Development Program

Chapter 1: Introduction

1.1 Background

The contribution of information technology (IT) is increasingly versatile and impacting many disciplines and areas of business activities. IT enables to organise, execute business processes effectively, to manage data in a controlled manner, to secure data, to provide fast and reliable solutions, to facilitate communication and much more.

Using IT, a large amount of data that is now considered as a valuable resource to support decision, can be generated. Such collection of data as a dataset can be considered as a source for data analysis to support decision making using predictive analytical algorithms. Large datasets are available in various areas such as in health institutions, bank industries, agriculture sectors, tourism industries, etc. These data sets are usually an accumulated data gathered when any kind of activities are performed and can be an asset to know the current trends and to determine the future trend by applying appropriately predictive analytics algorithms. Thus, the datasets are also used for knowledge extraction and the process to discover knowledge from these dataset is called data mining.

Data mining uses techniques to learn patterns from data. These pattern learning techniques are broadly categorized into two classes: supervised (predictive) and unsupervised (descriptive) learning [1]. Supervised learning uses a regression or classification algorithm to find hidden patterns from labelled data; unsupervised learning tries to find hidden patterns from unlabelled data.

Predictive analysis can build predictive models from dataset sources that can be used to identify patterns, which in turn can be used to propose decisions such as identifying risks and opportunities. One area where organisation can detect possible risks using prediction analytics is within health institution using health dataset. Such an approach could help health institutions to understand the trend or patterns of diseases, utilization of medications, utilization of medical resources, *etc.* to enhance the quality of medical treatment.

Since it will be too wide or broad to study the patterns of diseases in general, we focus only on a disease called malaria that severely harms especially the developing countries like Ethiopia. Malaria is one of the most severe public health problems worldwide with 300 to 500 million

cases and about one million deaths reported to date, 90% of which were reported from Sub Saharan African countries[2]. Malaria is the fourth leading cause of death of children under the age of five years in developing countries [2]. According to Aschalew and Tadesse [3], it is noted that Malaria is a severe disease in Ethiopia, where 75% of the land are prone to malarias' and more than 54 million people are vulnerable. Due to the fact that there is significant loss in productive personnel because of illness, school absenteeism; medical costs and other indirect costs including adult labourers are infected by malaria and stop working.

Marlies Craig et al. [4] describe a simple numerical approach to define the distribution of malaria transmission, based upon biological constraints of climate on parasite and vector development. The distribution of stable malaria transmission model, constructed by a fuzzy logic, introduced a hypothetical approach based on the effect of mean rainfall and temperature on the biology of malaria transmission. The fuzzy distribution model, sigmoidal fuzzy membership curve was used, defined in The GIS raster software IDRISI as:

$$y = \cos^2\left(\frac{x-U}{S-U} - \frac{\pi}{2}\right) \quad (1)$$

y is the fuzzy suitability of climate value x.

In the decreasing curve, fuzzy membership is equal to y, in the increasing curve it is (1 - y). For rainfall, U=0, S=80 mm per month; for average temperature U=18, S=22°C for the increasing curve and S=32, U=40°C for the decreasing curve. A value of 1means that conditions in the average year are suitable; a value of 0 means conditions are unsuitable in the average year, hence transmission should be absent or occur in rare epidemic episodes. Fractions from 0 to 1 indicate increasingly suitable climate, hence increased risk of regular transmission.

To examine the pattern of mean climate, as it relates to different epidemiological settings, monthly rainfall and temperature values were extracted from the climate data surfaces for 20 different sites where malaria transmission has traditionally been regarded as perennial (annual, for more than six months), seasonal (annual, for less than six months), epidemic (transmission not recorded every year) and malaria-free (malaria never recorded).

The examples confirm the researchers that the approximate temperature cut-off point between epidemic and no-malaria zones is indeed around 18°C, and that 22°C allows stable transmission and a rainfall requirement for stable transmission of around 80 mm per month for at least five months.

The model comparison with Kenyan and Tanzanian malaria maps is good but the area southeast of Mount Kenya and Nairobi was historically recorded malarious for three to six months, whereas the model predicts low climatic suitability. On closer inspection, this area is found to be flat, low-lying country, which may receive additional run-off water from the adjoining highlands; a high NDVI, which is a measure of the amount of photosynthesis taking place, and hence relates to the moisture availability, saturation deficit, soil properties and humidity that indicate an abundance of water. Nevertheless, empirical data from this region suggest that malaria transmission is low and sporadic. The discrepancies in the Tana and Pangani river valleys, as well as the Limpopo River, are a result of the model using only rainfall to predict the presence of vectors so that, although rainfall may be low, breeding sites are available and humidity is high along banks and floodplains of major rivers. The researchers noted as they are moving from the hypothetical to the quantifiable approach by considering other datasets such as population to provide improved estimates of people at risks.

In the work of Paul Edward Parham et al. [5] described the first phase of research in developing an integrated modeling framework for evaluating and predicting the likely impact of climate change on malaria transmission. Although the framework will ultimately seek to combine climate modeling with mathematical models as well as the socio-ecology and policy dimensions of disease transmission, the researchers focus on the construction of a generalized, realistic, climate-based malaria transmission model that allows capture of the simultaneous effects of rainfall and temperature on infection dynamics. The researchers follow this up by highlighting how analyses of such models can enable examination of critical and general dynamical issues not addressed to date such as the impact on mosquito populations of changes in climatic conditions, analysis of malaria invasion dynamics in disease-free regions and the effects of seasonal variability in climate variables on endemic prevalence, invasion, and extinction. The researchers also highlight the limitations of current mathematical modeling that address these issues, with the objective of stimulating further interdisciplinary research to improve our understanding.

As their research works stated that ideally malaria models should account for a range of more complex epidemiologic factors that affect disease dynamics such as more refined modeling of the subtleties of human immunity; plasmodium strain heterogeneities; multiple infections and co infections; age and genetic factor that account for human susceptibility and infectivity; human

movement patterns; short-range vector dispersal and heterogeneities in Anopheles species; socioeconomic conditions; and the emergence of drug resistance.

The researchers also noted that an important issue in environmentally driven infectious disease systems is identifying the most appropriate spatial and temporal scale at which to model. However, this will be driven by the questions in hand, data availability for parameterization, and the geographic scale over which the model will be calibrated, validated, and applied. Thus, fine-grained approaches will be important to guide interventions at the local level but may not be required for strategic planning on larger scales. Here, their objective is to adopt a coarse-grained dynamic model to explore the general dynamical impact of climatically driven systems on malaria transmission, with the spatial nature of model output the result of parameterization by local values of temperature and rainfall.

$$R_0 = \frac{M(R,T)a(T^2)b_1b_2l_m(T)}{\gamma\mu(T)N} \qquad (2)$$

where M is the total number of mosquitoes, T and R denote temperature and rainfall, respectively; a is the biting rate per day per mosquito, $b_1$is the proportion of bites by susceptible mosquitoes on infectious humans that produce infection, $b_2$is the proportion of bites by infectious mosquitoes on susceptible humans that produce infection,$l_m(T)$ is the proportion of infected mosquitoes that become infectious, γ is the rate at which infectious humans recover and acquire immunity, μ(T) is the daily mortality rate of adult mosquitoes, and N is the total number of humans whereas the derivation of $R_0$with temperature and rainfall seasonality is considerably more complex [6,7]. Estimation of R_0determines whether malaria outbreaks eventually become endemic (guaranteed in deterministic models when $R_0 > 1$) and the prevalence to which the system tends, given here by $N(R_0 - 1) \div (R_0 + \sigma)$ for the human population where $\sigma = a(T)b_1 \div \mu(T)$ is Macdonald's index of stability [8]. The rate of progression to the endemic state is determined by the invasion dynamics, characterized by the real-time growth rate r.

The researchers investigated and illustrated the role that dynamic process-based mathematical models can play in providing strategic insights into the effects of climate change on malaria transmission. They evaluated a relatively simple model that permitted valuable and novel insights into the simultaneous effects of rainfall and temperature on mosquito population dynamics, malaria invasion, persistence and local seasonal extinction, and the impact of seasonality on transmission. They also illustrated how large-scale climate simulations and

infectious disease systems may be modeled and analyzed and how these methods may be applied to predicting changes in the basic reproduction number of malaria across Tanzania.

Their analysis has provided several novel insights regarding quantifying the impact of climate on Anopheles population dynamics. First, although the effect of temperature on vector abundance has a strong physiologic basis and can thus be meaningfully captured by deterministic population models, the effects of rainfall are less predictable and more difficult to quantify. Second, considering the stochastic climate-driven population processes above, they find that the probability of having M mosquitoes at time t tends to a Poisson distribution with mean $\lambda(R,T) \div \mu(T)$ where $\lambda(R,T)$ and $\mu(T)$ are the birth and per-capita mosquito death rates respectively. Given this, they can also show that the probability of ultimate population extinction is $\exp[-\lambda(R,T) \div \mu(T)]$. This expression, together with climate data from WorldClim [9, 10] allows them to predict mosquito fadeout probabilities over a region. Given that the transmission model here is deterministic, mosquito fadeout probabilities may be used as a very approximate indicator of malaria fadeout probabilities, but future work should consider a full stochastic transmission model with more robust parameterization for more direct estimations. The temporal and spatial heterogeneity in mosquito extinction highlights the strategic importance of assessing control measures driven by regional factors, as well as emphasizing the need to incorporate seasonal climatic variability in transmission models. Environmental drivers other than rainfall and temperature will also drive extinction dynamics, and future transmission models should take a more thorough account of variability in a wider range of atmospheric variables, as well as the potentially strong role of factors such as land use, soil type, local topography, and ongoing vector interventions.

Their analyses of the present model have yielded several new insights regarding the potential impact of climate change on the dynamics of malaria transmission. The first major finding is that by influencing vector abundance, changes in rainfall patterns in particular strongly govern malaria endemicity, invasion, and extinction. In contrast, temperature effects, by affecting in multiple parts of the pathogen life cycle, have a more complex relationship with transmission and a stronger influence on the rate of disease spread, but only when sufficient rainfall exists to sustain vector development and survival. They apply these ideas to the generation of risk maps for Tanzania to highlight the sensitivity of spatiotemporal changes in R_0 to predicted shifts in rainfall and temperature under the A2a and B2a scenarios, while taking into account of seasonal

fluctuations in rainfall. They have also shown that mosquito density is more strongly drives the rate of emergence in non-immune populations than the infection dynamics of Plasmodium species, demonstrating how integrated climate and disease-modeling frameworks allow dissection of spatial heterogeneities in climate-induced malaria transmission and hence the nature of local interventions to mitigate or counter such changes.

They found extinction to be more strongly dependent on rainfall than temperature and identified a temperature window of around 32–33°C where endemic transmission and the rate of spread in disease-free regions is optimized. This window was the same for Plasmodium falciparum and P. vivax, but mosquito density played a stronger role in driving the rate of malaria spread than did the Plasmodium species. The results improved our understanding of how temperature shifts affect the global distribution of at-risk regions, as well as how rapidly malaria outbreaks take off within vulnerable populations. Disease emergence, extinction, and transmission all depend strongly on climate. Mathematical models offer powerful tools for understanding geographic shifts in incidence as climate changes. Nonlinear dependences of transmission on climate necessitate consideration of both changing climate trends and variability across time scales of interest.

Different Bayesian geo-statistical modeling approaches were employed to a historical data attempting to improve the model-based prediction of malaria risk. Sequel to this the Malaria AtlasProject (MAP) in 2007 and 2010 generated a geostatistical model-based global malaria risk map from historical survey data [11, 12]. In 2010, the National Malaria Control Programme with the support of Roll Back Malaria partners (RBM) implemented a nationally representative Malaria Indicator Survey (MIS), which assembled malaria burden and control intervention related data. In the work of Adigun et al. on Malaria risk in Nigeria: Bayesian geostatistical modeling of 2010 malaria indicator survey data describes the analysis of the MIS data gathered, produce a contemporary smooth map of malaria risk and evaluate the control interventions effects on parasitaemia risk after controlling of environmental/climatic, demographic and socioeconomic characteristics [13]. Bayesian geo-statistical models fitted via Markov Chain Monte Carlo (MCMC) simulation were employed for parameter estimation and predictions. Gibbs variable selection incorporating spatial dependency was used in identifying the most parsimonious model.

The researchers collected the data using the standard malaria indicator questionnaires developed by the RBM and the demographic health surveillance programme. The dataset consists of malariometric information, demographic characteristics and socio-economic status on a nationally representative sample of around 6,000 households from about 240 clusters of which 83 are in the urban areas. Environmental and climatic predictors were obtained from satellite sources. The acquired factors used in this analysis are Land Surface Temperature (LST), Normalized Difference Vegetation Index (NDVI), altitude, rainfall and distance to permanent water bodies. Weekly and biweekly values of LST and NDVI, respectively, covering the period from October 2009 to October 2010 were extracted from the Moderate Resolution Imaging Spectroradiometer (MODIS) database [14]. Decadal rainfall data for the same period was downloaded from the Africa Data Dissemination Service database [15]. Annual averages at each location (observed or predicted) were derived for the above predictors. Data on permanent water-bodies was obtained from the Health Mapper database of the World Health Organization (WHO).

Data on measures for preventing malaria, including the possession and use of insecticide-treated nets (ITN) / long lasting impregnated net (LLIN) and implementation of indoor residual spraying(IRS) were collected in the National Malaria Control Programme (NMIS). These data were used to generate the following indicators of intervention coverage as recommended by Roll Back Malaria-Measurement and Evaluation Reference Group (RBM-MERG) [16,17]: (i) the proportion with access to ITN in the household, (ii) proportion in every household that slept under an ITN during the previous night to the survey, (iii) proportion of children under 5 who slept under an ITN during the night preceding survey. Socioeconomic data Information on socioeconomic status (SES) was measured by a wealth index, which was present in the NMIS.

Population density grid data for the year 2010 was extracted from Worldpop [18]. Population structure for the same year was obtained from international database of United State census bureau [19] to calculate the number of children less than five years.

The research work stated that a Bayesian geo-statistical logistic regression model was fitted on the observed parasitological prevalence data. Important environmental/climatic risk factors of parasitaemia were identified by applying Bayesian variable selection within geo-statistical model. The best model was employed to predict the disease risk over a grid of 4 km resolution. Validation was carried out to assess model predictive performance. Various measures of control

intervention coverage were derived to estimate the effects of interventions on parasitaemia risk after adjusting for environmental, socioeconomic and demographic factors.

The researchers were identified normalized difference vegetation index and rainfall as important environmental/climatic predictors of malaria risk. The population adjusted risk estimates ranges from 6.46% in Lagos state to 43.33% in Borno. Interventions appear to not have important effect on malaria risk. The odds of parasitaemia appears to be on downward trend with improved socioeconomic status and living in rural areas increases the odds of testing positive to malaria parasites. Older children also have elevated risk of malaria infection.

The modelling approach followed by the researchers is not only focusing on the most important risk factors in order to build a parsimonious model with the best predictive ability. The result indicated that in Nigeria, rainfall and NDVI are the most important drivers of malaria risk while temperature and altitude do not improve our ability to predict the risk. The analysis showed that variation in the bed net coverage indicators across the country is not related to variation in the parasitaemia risk. A limitation of the survey is that it was carried out after the rainy season and, therefore, estimates may not reflect malaria risk during the highest transmission season.

The predictive prevalence map depicts that malaria morbidity is still high in the entire country and variation in malaria intervention coverage indicators is not associated with variation in parasitaemia risk across the country. The produced maps and estimates of parasitaemic children give an important synoptic view of current parasite prevalence in the country. Control activities will find it a useful tool in identifying priority areas for intervention. The coverage of key malaria interventions is still low and needs scaling up, which requires an increase of health expenditure by the federal government and an increase of awareness by the population on the benefit of bed net use.

In the research work of Paaijmans, K.P. et al., standard climate models are used to characterize climate change at relatively coarse spatial and temporal scales on the research called downscaling reveals diverse effects of anthropogenic climate warming on the potential for local environments to support malaria transmission [20]. However, malaria parasites and the mosquito vectors respond to diurnal variations in conditions at very local scales. They bridge this gap by downscaling a series of standard climate models to provide high-resolution temperature data for different four locations that capture the broad environmental distribution of malaria, including

cool upland locations (Kericho and Kitale), a warmer lower altitude site (Kisumu), and a hot savannah-like environment (Garrisa). For each of the four locations in Kenya they developed daily downscaled temperature projections by adapting an established empirical downscaling procedure utilizing General Circulation Model (GCM) and National Center for Environmental Prediction (NCEP) Reanalysis data, as well as meteorological station observations for daily maximum and minimum temperature at each location. The researchers set a minimum of four observations a day for the data to be included, and the data set is quality controlled for random errors.

The researchers represented malaria transmission using a model for Lifetime Transmission Potential (V), which describes the total contribution of a single mosquito over its lifetime to the transmission potential of the whole mosquito population:

$$V = \frac{a^2 bce^{-gn}}{g^2} \qquad (3)$$

where a is the daily biting rate of the vector, bc vector competence, g the daily probability of mortality of the adult vector and n the extrinsic incubation period of the parasites within the vector [21]. The more comprehensive Vectorial Capacity model describes the transmission potential of a mosquito population and so requires estimates of the equilibrium mosquito density per human [8]. While the vector: host ratio is temperature-sensitive, numerous other factors also affect mosquito density, including rainfall [23,24,25]; access to permanent water bodies [22, 26]; and mosquito control operations [27,28].

Finally, the researchers have shown that although outputs from both the GCM and the downscaled models predict diverse but qualitatively similar effects of warming on the potential for adult mosquitoes to transmit malaria, the predicted magnitude of change differs markedly between the different model approaches. Raw GCM model outputs underestimate the effects of climate warming at both hot (3-fold) and cold (8–12 fold) extremes, and overestimate (3-fold) the change under intermediate conditions. Thus, downscaling could add important insights to the standard application of coarse-scale GCMs for biophysical processes driven strongly by local microclimatic conditions.

In order to take preventive measure for malaria dispersion, predictive model need to be developed using various datasets which has a direct/or indirect relationship with malaria dispersion. Such a model can be used to develop a system that integrate the various source data

and assists the prediction and precaution processes. In addition to creating models, there is a need of facilitating a framework for the creation of malaria dispersion models using algorithms. The framework will be valuable to contain a variety of appropriate algorithms and to allow the creation of models by providing proper datasets so as to create various malaria dispersion predictive models.

1.2 Motivation

In developing countries like Ethiopia, people are suffering of diseases. Despite the fact that minimizing the challenge needs an immense multidisciplinary work and commitment, Information Technology will have its own contribution. The ministry of health of the federal republic of Ethiopia reported in 2005 E.C as there is a need of emphasis on the work of health systems to utilize it for monitoring of health interventions [29]. However, my motivation in this regard is, information technology can provide beyond monitoring of health intervention in a way up to the delivery of decision making by cementing predictive capability.

To make it achievable and provide insight, we want to focus only on a deadly disease called malaria that severely harms the lives of people mostly living in developing countries. In this aspect, it drives the research thesis to study the patterns of malaria disease using technologies of machine learning science to produce models and to create a framework that facilitate a fertile ground to the researchers to produce malaria dispersion models using their preferred algorithm.

1.3 Statement of the problem

A quarterly health bulletin prepared by the ministry of health of the Federal Republic of Ethiopia, at the end of 2005 E.C stated that emphasis should be given to the components of the health system to ensure better access services in Ethiopia and utilization of health information and technology to plan, implement and monitor high impact interventions as part of the strategic recommendation for the health sectors in the future [29].

From the preliminary information gathered from the ministry of health and Addis Ababa Health Bureau, most of the hospitals and health centres all over the country have a hospital management information system (HMIS) in place. However, recent research conducted by H. Tessema [30] concluded that the current usage of HMIS in the health facilities of Bahirdar area of the Amhara region is substandard and far from the target of HMIS in the prospect of providing information to

support decision making. The situation elsewhere in the country is not expected to be significantly different.

The work of Dye, Christopher [31] stated that since the year 2000, the Millennium Development Goals have provided a framework for accelerating the decline of infectious diseases, backed by a massive injection of foreign investment to low-income countries despite the successes of the inhabitants of low-income countries still suffer an enormous burden of disease owing to diarrhoea, pneumonia, HIV/AIDS, tuberculosis, malaria and other pathogens. Adding to the predictable burden of endemic disease, the threat of pandemics is ever-present and global. In this aspect and from the target of HMIS, health information systems should have a capability of providing health information in supporting decision making.

Study the patterns of diseases for decision making is broadly infinite whereby we are interested to focus on malaria which severely harms the developing countries at large. Aschalew and Tadesse [3] pointed out that Malaria is a severe disease in Ethiopia, where 75% of the land is prone to malarias' and more than 54 million people are vulnerable. Due to the fact that there is significant loss in productive personnel because of illness, school absenteeism; medical costs and other indirect costs including adult labourers are infected by malaria and stop working.

Marlies Craig et al. [4] describes a simple numerical approach based upon climate and vector development to define the distribution of malaria transmission but noted also that there is a need to move into quantifiable approach by considering additional datasets such as population to provide improved estimates of people at risks in addition to the hypothetical approach of their research work. Edward Parham et al. [5] introduced an integrated modeling framework for evaluating and predicting the likely impact of climate change on malaria transmission by investigated and illustrated the role that dynamic process-based mathematical models can play in providing strategic insights into the effects of climate change on malaria transmission. The researchers evaluated a valuable and novel insight of the simultaneous effects of rainfall and temperature on mosquito population dynamics, malaria invasion, persistence and local seasonal extinction, and the impact of seasonality on transmission.

In the other research works [11, 12] evaluated a bayesian geo-statistical logistic regression model is fitted on observed malaria parasitological prevalence data based on the Land Surface Temperature (LST), Normalized Difference Vegetation Index (NDVI), altitude, rainfall and distance to permanent water bodies. As a result, the researchers identified that normalized

difference vegetation index and rainfall are an important environmental/climatic predictors of malaria risk and also pointed out that the limitation of the research were carried out after the rainy season which may not reflect a malaria risk during the highest transmission season.

Taking into consideration of the researches done so far, it is understood that there is a need of predictive models to take preventive measures for malaria dispersion which will be the focus of this research work. In addition, a framework will be considered that facilitates a fertile ground to work on predictive models for malaria dispersion as models are basically based on algorithms and data in which new algorithms are merging, inevitability of algorithm modification and due to the emerging of data through time.

1.4 Objective of the Research work

General Objective

The general objective of this research work is to design a framework for predictive analysis of malaria dispersion in different environments using data mining techniques.

Specific Objective

To achieve the general objective the following specific objectives are identified.

- Identify the current data structures, business processes, the usage, and forms used at hospitals and health centers involved in the research.
- Identify the causes of malaria disease dispersion to work out on the data related/causing factor for malaria dispersion.
- Review literatures on algorithms for predictive analysis and select the appropriate methodology to use.
- Build a model using the predictive analysis algorithms and the data collected from Ethiopian case.
- Evaluate the model and select the better algorithms that offer a better model to predict.
- Develop a prototype implementation of the framework based on the proposed architecture.

1.5 Methods

In order to realize this research work, the following methodologies will be used to build a framework, to propose reference architecture for predictive model of malaria dispersion and develop a prototype.

Literature Review

Different literatures will be intensively reviewed to explore related works on architecture and algorithms for predictive analysis, process and methods applicable to predictive model. Based on the review, appropriate tools and algorithms will be selected, used and developed.

Identification of Data source

Data sources will be identified.

Prototype Development

Prototype of a framework with the suggested predictive analysis algorithms will be developed based on the data that will be collected and considered as a cause of malaria dispersion. Testing the prototype and evaluation of the algorithms will be conducted using the sample data. The result will be analysed and evaluated from which a conclusion and further works are recommended.

1.6 Scope and Limitation of the Research Work

This study focuses on developing a framework for predictive analysis of malaria dispersion to train predictive analysis algorithms to create a model so as to use it for malaria dispersion prediction using a unified dataset that combines malaria dispersion along with metrology, NDVI and population datasets.

The research work is limited to the size of data where huge amount of data is required to train the algorithm besides the identification of factors causing malaria dispersion is also limited to the source of datasets.

1.7 Significance of the Research Work

From this work, the Ministry of health and health institutions can use the result to generate prognostic malaria dispersion for decision making. Moreover, it can be used as a reference for further research works.

1.8 Organization of the Rest of the Thesis

The remaining part of the thesis is organized as follows. Chapter Two covers literature review. Chapter three covers related work which shows efforts of different researchers in recognition of predictive analysis related studies on malaria dispersion at different locations. The Fourth chapter broadly explains and discusses the design of a framework for predictive analysis of malaria dispersion along with the preparation. Chapter five presents the evaluation and discussion of the models, algorithms and framework using the implemented prototype. Finally, Chapter Six summarizes our findings and presents future works.

Chapter 2: Literature Review

2.1 Predictive Analysis

Predictive analysis is a branch of data mining that focuses on stipulating patterns or trends of data based on the analysis of the current and past data occurrence. Jiawei Ha [29] generalizes a data mining tasks as a classification of descriptive and predictive where descriptive mining tasks characterize the general properties of the data in the database where as predictive mining tasks perform inference on the current data in order to make predictions. Dean Abbott [30] also defines Predictive analytics as the process of discovering interesting and meaningful patterns in data.

In practice, planning at corporate level is challenging as it requires the capability of prediction for future events will happen in such a way that decision makers aided by predictive analytics. The Johns Hopkins University Applied Physics Laboratory [3] has developed a novel and scalable data mining and fuzzy association rule-making approach to deriving disease incidence predictions several weeks in advance of an outbreak. The research [32] describes that such capabilities provide a new set of information that may be used by decision makers in conjunction with other complementary information about the country (e.g., infrastructure, disease history, agriculture, population, etc.) from a variety of sources. The prediction of the future infectious disease incidence provides the decision maker with enhanced ability to determine whether to enable deployment of predictive measures to increase and focus biosurveillance and/or to plan and enable mitigation efforts to reduce morbidity and mortality well in advance of the start of the outbreak [32, 33].

The Cross-Industry Standard Process Model for Data Mining (CRISP-DM) describes the data mining process in six steps which are business understanding, data understanding, data preparation, modelling, evaluation and deployment [34].

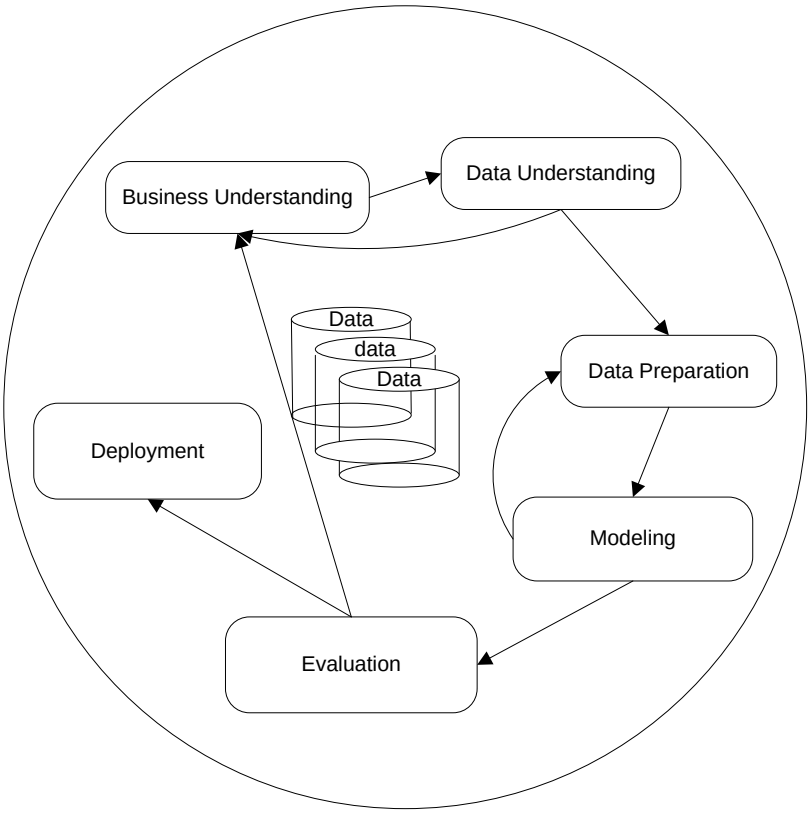

Figure 2.1: The CRISP-DM process model

As it is shown in Figure 2.1, there are iterative activities in the model such as the business objective will be re-defined if there is insufficient data quantity or the model evaluation suggests redefining the business objective.

2.2 Business Understanding

In any project, predictive analytics project has an objective to achieve as CRISP-DM process model shown in the above diagram that business understanding is the first phase in data mining process. Daniel [35] explained also as business understanding or research understanding is the first enunciate project objectives that should be done by the researchers or businesses before formulation of a data mining problem definition and to come-up with preliminary strategy for achieving the objectives. Dean Abbott [30] explains one of the successes of predictive modelling is determined by the collaboration of domain experts, data or database experts and predictive modelling experts. Domain experts are needed to frame a problem properly in a way that will provide value to the organization. Data or database experts are needed to identify what data is

available for predictive modelling and how that data can be accessed and normalized. Gaining business understanding is an interactive procedure in data mining, where the results of various visualization, statistical, and artificial intelligence tools show the user new relationships that provide a deeper understanding of organizational operations so that predictive modellers are able to build the models that achieve the business objectives established in the first phase [36].

2.3 Data understanding and Preparation

Obviously, a data originated from different and multiple sources are highly likely to be a victim of inconsistency, missed and noisy data which leads to unreliable outcome from the predictive models. Problems in the data, such as missing values, outliers, spikes, and high-cardinality, should be identified and quantified so it can be fixed during data preparation otherwise the problem missed during data understanding will come back to haunt the analyst during modelling [30,37]. The first step in predictive modelling is to understand the input data, analyse and prepare in the way that the amount, structure, and format suited to the modelling algorithm. Robert Nisbet, et al [33] describes clearly the basic issues that should be encountered in the pursuit of understanding the data and to make ready for the modelling algorithm.

2.3.1 Data Acquisition

The separate data stores may exist in different departments in the form of spread sheets, miscellaneous databases, printed documents, and handwritten notes. The initial challenge is to identify where the data are and how the information can be obtained. If the data are all in one place (such as in a data warehouse), the best way to access that data must be determined.

Data acquisition technique has various types based on the nature of the business we are dealing with. In social networks, network traffic analysis, Ad-hoc applications and crawling the user graph are mentioned most of the time as data acquisition techniques [38-42]. For the purpose of smart phone volatile memory acquisition for security analysis and forensics investigation, Rokach [43] describes a number of data acquisition techniques such as run mode debugging, stop mode debussing and kernel module. In mobile phone forensics research, various techniques of data acquisition have been studied for GSM phones, the mobile phone's internal flash memory, SIM, android phone and iPhone [44-51]. In general, data could be acquired via APIs', direct access to database or file, streaming, image processing, etc. however all might not be required at

the same time for a specific target of business and hence, careful design of data acquisition is crucial [3].

2.3.2 Data Extraction

Data extraction is the act or process of retrieving relevant information from data sources (like database) in a specific pattern. The process of data extraction involves retrieval of data from dishevelled data sources, and the data extracted are then loaded into the staging area of the database. The extraction logic uses application programming interfaces to query the data [52].

2.3.3 Data Description

Data description is an activity composed of the study of individual attributes in a described and summarized way on which they are measured, how to describe their center as well as the variation using descriptive statistical approaches, and how to make statements about these attributes using inferential statistical methods, such as confidence intervals or hypothesis tests. It also involves the study of relationships between different variables in a number of ways to understand relationships between pairs of variables through data visualizations, tables that summarize the data, and specific calculated metrics [53].

The analysis and evaluation of data description helps to determine on how to prepare the data for modelling. For example, a variable with relatively low mean and a relatively high standard deviation has a relatively low potential for predicting the target. Analysis of the minimum, maximum, and mean might alert to the fact that have some significant outlier values in the data set.

2.3.4 Data Cleansing

Incorrect values are problematic because predictive modelling algorithms assume that every value in each field is completely correct [30]. Cleansing refers fixing problems with values of variables, including incorrect or miscoded values, outliers and missing values. Deletion from the dataset is one and simplest approach to handle missed as well as outliers, however, can result in significant amounts of data loss [54]. Sometimes, the outliers (abnormal values) need to be kept. In fact, some outliers are of primary interest to the modelling of risk, fraud, and other rare events. For models of normal responses, it might be a good idea to remove the extreme outliers. Xiong [55] explained as most existing data cleaning methods focus on removing noise that is the result

of low-level data errors that result from an imperfect data collection process, but data objects that are irrelevant or only weakly relevant can also significantly hinder data analysis. Thus, if the goal is to enhance the data analysis as much as possible, these objects should also be considered as noise, at least with respect to the underlying analysis. Consequently, Xiong[55] explores the three outlier detection techniques which are distance-based, clustering-based, and an approach based on the Local Outlier Factor (LOF) and propose a new method called hyperclique-based data cleaner (HCleaner).

Imputation is also another approach for missing value by replacing with a plausible estimated value and clamp transformation technique can be applied for outliers by setting an upper and lower threshold to correct the offending outliers [54]. Bowman [56] suggests studying the nature of the data before applying smoothing techniques as it is highly dependent on the methods, parameters and settings to be applied for the dataset.

2.3.5 Data Transformation

A data transformation converts a set of data values from the data format of a source data system into the data format of a destination data system [57]. In data transformation, the data are transformed or consolidated into the forms appropriate for mining. In almost all practical cases, it is necessary to perform some transformations on extracted data before it can be used in modeling [58].

2.3.6 Data Abstraction

The purpose of data abstraction is the intelligent interpretation of the raw data so that the derived abstract data are at the level of abstraction corresponding to the given body of knowledge. Abstract data are useful since they can be unified against knowledge that provides useful summaries to a certain situation. Broadly, data abstractions are classified into four groups [59]:

- Qualitative abstraction: a numeric expression is mapped to a qualitative expression. e.g. “a temperature of 41 degrees C” is abstracted to “fever”.
- Generalization abstraction: an instance of an occurrence is mapped to its class. e.g. “halothane is administered” is abstracted to “drug is administered”; the concept “halothane” is an instance of the concept class “drug”.

- Definitional abstraction: an instance in which one data element from one conceptual category is mapped to its counterpart in another conceptual category. An example of definitional abstraction is the mapping of "generalized platyspondyly" to "short trunk". "Generalized platyspondyly" is a radiological concept, the observation of which requires the taking of a radiograph of the spine; platyspondyly means flattening of vertebrae and generalized platyspondyly means the flattening of all vertebrae. "Short trunk" is a clinical concept, the observation of which does not require any special procedure. The knowledge driving such abstractions consists of simple associations between concepts across different categories.
- Temporal data abstraction: The dimension of time adds a new aspect of complexity to the derivation of (temporal) abstractions. It entails temporal reasoning and again Lavrac et al. [59] classified as merge, persistence, trend and periodic abstraction to handle multiple time granularities and temporal relations such as before, overlaps, disjoint, etc. as well as of a specialist nature dealing with persistence semantics of concepts. Thus abstractions referring to the present may need to be modified on the basis of old data that has now become available [60] or abstractions referring to the past are revoked by new data [61].

2.3.7 Data Sampling

Analysts usually use sampling techniques to run models, enable exploration of data, and determine whether more analysis is needed. Data mining techniques that use sampling methods can reveal valuable information and complex relationships in large amounts of data. As Westphal [62] point out, extracting data from a database for the purpose of data mining is based on the sampling techniques routinely used in surveys. Some of the sampling techniques are as follows:

- Simple Random Sampling: Each data record has the same chance of being included in the sample.
- First N Sampling: The first n records are included in the sample.
- Weighted Sampling: In which the inclusion probabilities for each element of the population is not uniform, each element in the population has a different probability of being selected in the sample according to a defined criteria.
- Stratified Sampling : In stratified sampling, one or more categorical variables are specified from the input data table to form strata (or subsets) of the total population by

dividing the area up into a number of strata such that within each of the strata the values of the variable of interest are expected to be relatively similar.

- Proportional Sampling: In probability sampling, every observation in the population from which the sample is drawn has a known probability of being selected into the sample.
- Cluster Sampling: This method builds the sample from different clusters. Each cluster consists of records that are similar in some way. Clustered samples are generated by first sampling cluster unit and then sampling several elements within the cluster unit.
- Multi-Stage Sampling: Multistage sampling is sampling where the elements are chosen in more than one stage. Initially large areas selected then progressively smaller areas within larger area are sampled, this process may continue till a sample of sufficiently small ultimate area units is obtained.

The criteria for selecting appropriate sampling technique is crucial to be applied on the mining algorithm however in the research like this requires to get all available data in related to malaria as the objective is to model the trends of malaria dispersion.

As sampling practice, the data collected will be partitioned into three data tables that will be used for the purpose of training, validation and testing. The training data table is used to train models to estimate the parameters of the model. The validation data table is used to fine tune and select the best model. In other words, based on some criteria, the model with the best criteria value is selected. For example, smallest mean square forecast error is an often-used criterion. The test data is used to test the performance of the selected model. After the best model is selected and tested, it can be used to score the entire database [63].

In data mining, data sampling serves four purposes [33]:

- ✓ It can reduce the number of data cases submitted to the modeling algorithm.
- ✓ It can help to select only those cases in which the response patterns are relatively homogeneous.
- ✓ It can help to balance the occurrence of rare events for analysis by machine learning tools. Machine learning tools like neural nets and decision trees are very sensitive to unbalanced data sets. An unbalanced data set is one in which one category of the target variable is relatively rare compared to the other ones. Balancing the data set involves sampling the rare categories more than average (oversampling) or sampling the common categories less often (under sampling).

2.3.8 Reduction of Dimensionality

Dimensionality reduction is one of the pre-processing steps used in a number of applications to reduce the dimensions of high dimensional data to increase the efficiency of the data analysis [64]. The basic dimensionality of data is to minimize number of parameters needed to account for the observed properties of the data [65]. Dimensionality reduction is important in many domains, since it mitigates the curse of dimensionality and other undesired properties of high-dimensional spaces [66].

The machine learning and data mining techniques may not be effective for high-dimensional data because of the curse of dimensionality and query accuracy and efficiency will degrade rapidly as the dimension increases [67]. The motivation for dimension reduction can be summarized as the identification of a reduced set of features that are predictive outcomes that can be very useful from a knowledge discovery perspective [68]; identification of a reduced set of features as the training and/or classification time increases directly with the number of features [69]; and to minimize the noisy or irrelevant features that have the same influence on training and/or classification as predictive features will impact negatively on accuracy [70].

2.3.9 Data Discretization

Some machine learning techniques can work with only categorical predictor variables, not continuous numeric variables [33]. In such cases the value shall be converted to a continuous numeric variable into a series of categories by assigned sub ranges of the value range to a group of new variables. Data discretization is defined as a process of converting continuous data attribute values into a finite set of intervals with minimal loss of information [70].

The most common approach to handling continuous values is to discretize them into a number of disjoint regions and then use the same set-mining algorithm. Discretization is useful in that it can reduce the number of distinct values, thereby reducing the complexity of the search and the number of mined results [71]. Commonly, the goal of data discretization is maximizing the predictive accuracy for algorithms that cannot handle continuous values. Dougherty et al. [72] showed that discretizing continuous attributes for the naive Bayesian classifier can greatly improve accuracy over a normal approximation.

Stephen D. Bay [71] has noted as discretization should consider the effects on all variables in the analysis and that two regions X and Y should only be in the same interval after discretization if

the instances in those regions have similar multivariate distributions (Fx∼Fy) across all variables and combinations of variables. For example, a variable ranging from 1–100 could be discretized into a range of five categories: 0–20, 21–40, 41–60, 61–80 and 81–100.

2.3.10 Data Derivation

Sometimes independent variables are derived or formulated from a combination of existing variables when it is required to use. In mathematics, the derivation of area of a rectangle can be calculated by multiplying its length and width. Likewise, in data mining, target variables could be derived by following some logical rule.

2.4 Predictive Modelling

The goals of data mining in practice tend to be prediction and description. Prediction involves using some variables to predict unknown or future values of other variables of interest, and description focuses on finding human-interpretable patterns describing the data. In predictive modelling the training data *Dtrain* consists of *pairs* of measurements, each consisting of a vector $\mathbf{x}(i)$ with a corresponding "target" value $y(i)$, $1 = i = n$. Thus the goal of predictive modelling is to estimate (from the training data) a mapping or a function $y = f(\mathbf{x})$ that can predict a value y given an input vector of measured values $\mathbf{x}$ and a set of estimated parameters for the model. [73, 74]

For predictive modelling, the score function is usually relatively straightforward to define, typically a function of the *difference* between the prediction of the model and the true value where the sum is taken over the tuples $(\mathbf{x}(i), y(i))$ in the training data set *Dtrain* and the function d defines a scalar distance such as squared error for real-valued y or an indicator function for categorical y. The actual heart of the data mining algorithm then involves minimizing S which is determined both by the nature of the distance function and by the functional form of $f(\mathbf{x})$ [74].

There are two important distinct kinds of tasks in predictive modelling depending on whether Y is categorical or real-valued. For categorical Y, the task is called *classification* (or *supervised classification* to distinguish it from problems concerned with defining the classes in the first instance, such as cluster analysis), and for real-valued y the task is called *regression* [1].

2.4.1 Decision Trees

A decision tree is considered to be one of the most popular approaches for representing a recursive partition of the instance space [43]. Classification and regression trees are also known

as recursive partitioning or segmentation trees or decision trees are nowadays widely used either as prediction tools or simply as exploratory tools. Their interest lies mainly in their capacity to detect and account for nonlinear effects on the response variable, and especially of even high order interactions between predictors [75].

Decision tree algorithms can handle both nominal and continuous inputs as well as outputs [30]. Morgan et al. [76] proposed the AID (Automatic Interaction Detector) algorithm for growing a binary regression tree in which the outcome variable is quantitative. For categorical dependent variables, a binary method called ELISEE (Exploration of Links and Interactions through Segmentation of an Experimental Ensemble) is proposed by Cellard et al. [77]. And later on Messengeret al. [78] and Morgan et al. [79] extended AID to entertain categorical outcome and Gillo et al. [80, 81] also extended AID for multivariate quantitative outcome variables.

AID algorithm is the first regression tree algorithm published in the literature [82]; later adding the chi-square test to improve variable selection in what became the CHAID algorithm (Chi-square Automatic Interaction Detection), 1980) [83]. Through the evolution of AID leads to algorithm called CHAID, which is introduced by Kass [83] and nowadays the most popular among these earlier statistical supervised tree growing techniques.

Meanwhile, in independent research, two other algorithms were developed. Ross Quinlan [84] developed an algorithm called ID3 in the 1970s, which used Information Gain as the splitting criterion. In 1993, Ross Quinlan [85] improved the algorithm with the development of C4.5, which used Gain Ratio (normalized Information Gain) as the splitting criterion. C4.5 was subsequently improved further in the algorithm Quinlan called C5.0 [85], including improvements in misclassification costs, cross-validation, and boosting (ensembles). It also significantly improved the speed in building trees and built them with fewer splits while maintaining the same accuracy.

Other methods such as CART (classification and regression trees) is proposed by Breiman et al. [86] and had occurred concurrent with the development of ID3, which uses the Gini Index as the primary splitting criterion and it is perhaps the most known and widely used.

Decision trees are nonlinear predictors, where the decision boundary between target variable classes is nonlinear. The extent of the nonlinearities depends on the number of splits in the tree because each split, on its own, is only a piecewise constant separation of the classes. As the tree

becomes more complex, or in other words, as the tree depth increases, more piecewise constant separators are built into the decision boundary and providing the nonlinear separation.

2.4.2 Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success) or 0 (FALSE, failure). The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables [87]. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest. The response variable has only two categories of particular interest and lends itself to an especially nice treatment. This is because, with only two categories, there is essentially only one way to define the odds. If p1 is the probability in the first category and p2 is the probability in the second category, then the odds of getting category one are p1/p2. The odds of getting category two are p2/p1. With a two-category response variable, we will examine models for log (p1/p2). When these models are regression type models, they are called logistic regression models [88].

2.4.3 Artificial Neural Network

Artificial neural networks are mathematical inventions inspired by observations made in the study of biological systems and can be described as mapping an input space to an output space[89]. Artificial neural networks are nonparametric regression models. They can capture any phenomena to any degree of accuracy (depending on the adequacy of the data and the power of the predictors), without prior knowledge of the phenomena. Further, artificial neural networks can be represented, not only as formulae, but also as graphical models. Graphical models can improve the manipulation, and understanding of statistical structures [90]. Burke [90] presents the idea that a neural network is a powerful tool for capturing complex relationships.

Bain [91] and James [92] realized the potential for manmade systems based on neural models. After a number of researches carried out, Werbos [93, 94] developed a learning rule in which the error in the network's output is propagated backwards through the network, and the network's synaptic weights are adjusted by using a gradient descent error-minimization approach and the technique is the back propagation of error algorithm and it is often referred to as back propagation and is currently the most widely used artificial neural network model. In the mid-1980s, Rumelhartet al. [95, 96] and others popularized the back propagation algorithm.

Artificial neural networks are modeled after early observations in biological systems: myriads of neurons, all connected in a manner that somehow distributes the necessary signals to various parts of the body to allow the biological system to function and survive. The neuron can be thought of as a small computing engine that takes in inputs, processes them, and then transmits an output [89]. The artificial neuron with weighted inputs and a corresponding transfer function is:

$$Z = f(\sum_{i=0}^{3} w_i x_i) \qquad (4)$$

The most commonly used transfer function is the sigmoid or logistic function, because it has nice mathematical properties such as monotonicity, continuity, and differentiability, which are very important when training a neural network with gradient descent. Initially, scientists studied the single neuron with a hard-limiter or step-transfer function [97]. Rosenblatt [98] used a hard limiter as the transfer function and termed the hard-limiter neuron a perceptron because it could be taught to solve simple problems. The hard-limiter is an example of a linear equation solver with a simple line forming the decision boundary.

The three different types of transfer functions are step, sigmoid, and linear in unipolar and bipolar formats but the most common transfer function is the logistic sigmoid function, which is given by the equation:

$$output = \frac{1}{1+e^{-(\sum_i w_i x_i)}} \qquad (5)$$

Where *i* is the index on the inputs to the neuron, *xi* is the input to the neuron, *wi* is the weighting factor attached to that input, and *w*0 is the bias to the neuron.

The aggregation and selective use of these decision boundaries is what makes artificial neural networks interesting as it forms these decision boundaries with their associated class regions as

derived from the data. Neural networks can be used to combine these decision regions together to form higher and higher levels of abstraction, which can result in neural networks with some amazing properties. One of the best-known linear classifiers is the perceptron, first introduced by Rosenblatt [98]. The perceptron adjusts its weights using the error vector between a data point and the separating decision line. Changes are made to the weights of the network until none of the training-set data samples produces an error. The multilayer feed forward neural network allows machines to form arbitrarily complex decision regions with multiple hyper planes to solve classification problems as shown below.

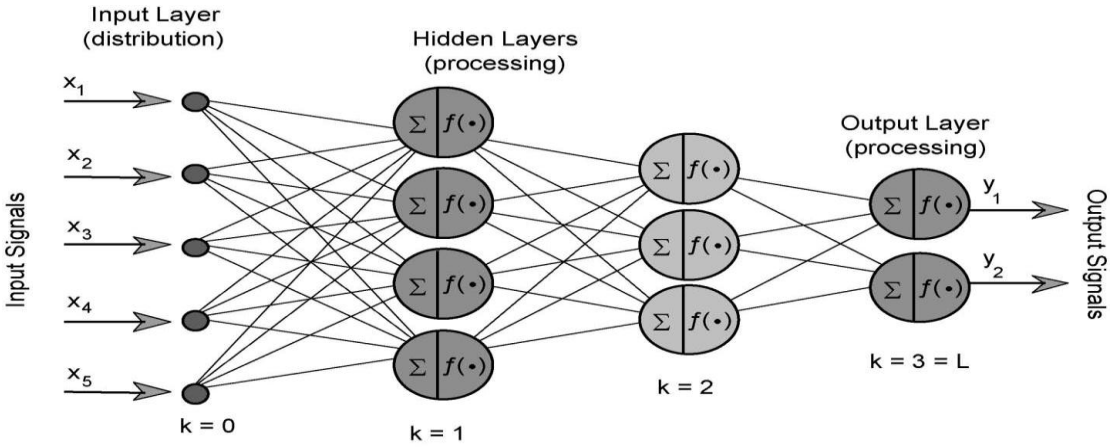

Figure 2.2: Multilayer feed forward neural network. [89]

A modeller can build multiple neural networks with different architectures and determine empirically which is best. A few software implementations build in the ability to build many networks with different architectures. Pruning algorithms helps to remove inputs and neurons are to be identified those that have the least influence on the network predictions. Building neural networks with pruning can take considerable time, however, and therefore it is often left as a final step in the modelling-building process, after a good set of inputs for the model are found and a good baseline architecture is found.

2.4.4 K-Nearest Neighbor

The nearest neighbor algorithm is a non-parametric algorithm that is among the simplest of algorithms available for classification [99]. The nearest neighbor algorithm is a so-called lazy learner, meaning that there is little done in the training stage. In other words, k-NN is merely a

lookup table that is used to predict the target value of new cases unseen during training. The character k refers to how many neighbours surrounding the data point that needs to make the prediction. The mathematics behind the k-NN algorithm resides in how it is computed the distance from a data point to its neighbours.

The more neighbours included in the prediction, the smoother the predictions. For binary classification, it is common to use an odd number of nearest neighbours to avoid ties in voting. However, as the number of nearest neighbours' increases, the classifier becomes less localized, smoother, less susceptible to noise, but also less able to find pockets of homogeneous behaviour in the data, similar to a decision tree with only one split.

The primary distance metric used with the k-NN algorithm is Euclidean distance. Assume there are m records in the training data, and n inputs to the k-NN model. Collect the inputs into a vector called x of length n. The vector of inputs we are intending to predict based on the k-NN model is called y and is also of length n.

The k-NN algorithm usually has few options that need to set besides the value of k. Other common considerations that need to be included are which distance metric to use, how to handle categorical variables, and how many inputs to include in the models.

Euclidean distance is known to be sensitive to magnitudes of the input data; larger magnitudes can dominate the distance measures. In addition, skew in distributions of the inputs can also effect the distance calculations. K-NN is a numerical algorithm requiring all inputs to be numeric. Categorical variables must therefore be transformed into a numeric format, the most common of which is 1/0 dummy variables: one column for each value in the variable.

2.4.5 Naive Bayes

Naive Bayes is a simple technique for constructing classifier models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle. All naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable [100]. The probabilistic model of naive Bayes classifiers is based on Bayes' theorem, and the objective naïve comes from the assumption that the features in a dataset are mutually independent. In practice, the independence assumption is often violated, but naive Bayes

classifiers still tend to perform very well under this unrealistic assumption [101]. Especially for small sample sizes, naive Bayes classifiers can outperform the more powerful alternatives [102].

Using Bayes theorem, the conditional probability can be decomposed as [103]:

$$p(c_k|x) = \frac{p(c_k)p(x|c_k)}{p(x)} \quad (6)$$

In general, using Bayesian probability terminology, the above equation can be written as [104]:

$$posterior = \frac{prior * likelihood}{evidence} \quad (7)$$

Bayes classifiers are used in many different fields such as the diagnosis of diseases and making decisions about treatment processes [105], the classification of RNA sequences in taxonomic studies [106], and spam filtering in e-mail clients [107].

2.4.6 Regression Models

Regression models predict a continuous target variable rather than a categorical target variable. In some ways, this is a more difficult problem to solve than classification. Regression belongs to the supervised learning category of algorithms along with classification. Some modellers call these models continuous valued prediction models; others call them estimation or function estimation models.

The most common algorithm predictive modellers' use for regression problems is linear regression, an algorithm with a rich history in statistics and linear algebra. Another popular algorithm for building regression models is the neural network, and most predictive analytics software has both linear regression and neural networks. Many other algorithms are also used for regression, including regression trees, k-NN, and support vector machines. In fact, most classification algorithms have a regression form.

2.4.7 Linear Regression

The linear regression algorithm makes many assumptions about the data in order to define the relationship between the input variables and the output variable to be linear. If the relationship is not linear, it should create derived variables that transform the inputs so that the relationship to the target becomes linear. The linear fit from a regression model which obviously doesn't fully capture the curvature of the relationship in the data violates the assumption of linearity.

Linear regression methods are the most commonly used in regression, and virtually all other regression methods build upon an understanding of how linear regression works. [108]
The simple linear regression model consists of the mean function and the variance function

$$E(Y|X = x) = \beta_0 + \beta_1 x \quad (8)$$

$$(Var(Y|X = x) = \sigma^2 \quad (9)$$

The parameters in the mean function are the intercept β0, which is the value of E(Y |X = x) when x equals zero, and the slope β1, which is the rate of change in E(Y |X = x) for a unit change in X. By varying the parameters, all possible straight lines can be depicted.
The multiple linear regression generalizes the simple linear regression model by allowing for many terms in a mean function rather than just one intercept and one slope. The general multiple linear regression model with response Y and terms $X_1, \ldots, X_p$ will have the form [109]

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad (10)$$

The symbol X in E(Y |X) means; it is conditioned on all the terms on the right side of the equation. Similarly, when it is conditioned on specific values for the predictors $x_1, \ldots, x_p$ it will collectively called x

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (11)$$

The βs are unknown parameters that need to be estimated. When p = 1, X has only one element where a simple regression problem has realized; When p = 2, the mean function corresponds to a plane in three dimensions; When p >2, the fitted mean function is a hyper plane, the generalization of a p-dimensional plane in a (p + 1)-dimensional space.
Not all response values will have a straight-line. To achieve linearity transformation of variables will be a key tool in extending the usefulness of linear regression models.

2.5 Model Deployment

Deployment of predictive models is the most underappreciated stage of the CRISP-DM process. The deployment phase aims at transferring DM results that meet the success criteria into the business [110]. Model construction is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it [111, 112].

CRISP-DM considers planning for deploying the knowledge at the client site, but it does not regard the development of software installed and accepted in an operational environment as part of this deployment [113].

Finally, a good deployment system plan recognizes that models are not static. They were built from data that represented the history at a particular moment. But that moment will ultimately pass. Models should be monitored and ultimately be replaced by an improved model when needed.

Chapter 3: Related Work

A technical report prepared by Bob Snow et al. [114] explains that the inadequacy of current malaria notification systems in Africa negates their use in defining current disease burdens or the likely impact of drivers of changing infection risk or disease outcome in the next two decades. They noted that a preliminary model of infection risk and disease outcome have been developed that accommodate drivers of malaria transmission (climate and population settlement) and immunological drivers of age and transmission dependent disease risks. They finally put as a conclusion that the future modelling of risk and disease will need to accommodate dynamic projections of changes in the most significant drivers, notably intervention access and coverage, population growth and settlement patterns, changes inland use, and changes in effect modifiers such as HIV, nutritional status and poverty.

In the work of Van Lieshouta, M. et al. [26] used the global model of malaria transmission to estimate the potential impact of climate change on seasonal transmission and populations at risk of the disease (MIASMA v.2.2) [115,116] in their research called climate change and malaria: analysis of the Special Report on Emissions Scenarios (SRES) climate and socio-economic scenarios. The research assessment describes model simulations driven by the latest scenarios from the Intergovernmental Panel on Climate Change (IPCC) [117]. The climate scenarios were derived from the Hadley Centre model HadCM3 runs with four SRES emissions scenarios: A1FI, A2, B1 and B2. The additional population at risk was determined under each of the SRES population scenarios by downscaling national estimates to the 0.5^0x0.5^0scale grid and re-aggregating by region. Additional population at risk due to climate change are projected in East Africa, central Asia, China and areas around the southern limit of the distribution in South America. Decreases in the transmission season are indicated in many areas where reductions in precipitation are projected by the Hadley Centre model, such as the Amazon and in Central America. The outcomes of the malaria model are sensitive to spatial distribution of precipitation projections and population growth in those areas where there is new risk due to climate change. The researchers describe a new method for describing vulnerability to the potential impacts of climate change. The researchers used expert judgment to classify countries according to their current vulnerability and malaria control status. This vulnerability incorporates both socio-economic status, as a measure for adaptive capacity, and climate as malaria at the fringes of its

climate determined distribution is easier to control than malaria in tropical endemic regions. Thus, current malaria control status is used as an indicator of adaptive capacity. For those countries that currently have a limited capacity to control the disease, the model estimates additional populations at risk by 2080s in the range of 90 m (A1FI) to 200m (B2b). The greatest impact under B2 reflects population growth in risk areas in Eurasia and Africa. Climate-induced changes in the potential distribution of malaria are projected in the poor and vulnerable regions of the world. However, climate change is not likely to affect malaria transmission in the poorest countries where the climate is already highly favourable for transmission.

The researchers also noted that as there is a need of further research to get quantitative indicators of vulnerability to malaria and parameterization of rainfall in malaria models. In addition to that the malaria model should primarily be undertaken at regional or national scales to identify more accurately populations at risk.

Caminade, C et al. [118] did comparisons of five statistical and dynamical malaria impact models for health impacts under a future time periods (2030s, 2050s, and 2080s) with climate change using temperature and rainfall. The researchers evaluated based on three malaria outcome metrics at global as well as regional levels, which are climate suitability, additional population at risk and additional person-months at risk across the model outputs. The malaria projections were based on five different global climate models, each run under four Representative Concentration Pathways (RCPs) emissions scenarios developed for the fifth Intergovernmental Panel on Climate Change (IPCC) assessment report and a single population projection.

The LMM_RO [119], MARA [120], and MIASMA [121] models only allow the investigation of climatic suitability for malaria transmission, whereas the fully dynamical VECTRI [122] model considers the impact of climate, surface hydrology, and population densities on malaria distribution. The UMEA [123] model considers the impact of the gross domestic product per capita in combination with climate and population densities to model endemic malaria distribution. The first aim is to compare past distributions of malaria using the malaria models driven by observed climatic and socioeconomic data with “observed” malaria endemicity estimates (which include all potential climatic and socioeconomic effects on malaria distribution). Based on the given different designs and parameterizations of the malaria impact models (MIM) and as the models were originally developed for different regions and scientific objectives, the models might give significant differences from the observed estimates. They

mapped future climate-based distributions of malaria and estimated populations at risk under the new emissions scenarios accounting for population growth and estimate the relative uncertainty and its evolution in time related to the global MIM, the climate model uncertainty inherit in the driving GCMs and the emission scenario uncertainty from the RCPs. An assessment of the different malaria models' sensitivity to climate change was then carried out before providing final conclusions.

The researchers also investigated the modelling uncertainty associated with future projections of populations at risk for malaria owing to climate change. Their findings have shown an overall global net increase in climate suitability and a net increase in the population at risk, but with large uncertainties. The model outputs indicate a net increase in the annual person-months at risk when comparing from RCP2.6 to RCP8.5 from the 2050s to the 2080s. The malaria outcome metrics were highly sensitive to the choice of malaria impact model, especially over the epidemic fringes of the malaria distribution. Our results indicate that future climate might become more suitable for malaria transmission in the tropical highland regions. However, other important socioeconomic factors such as land use change, population growth and urbanization, migration changes, and economic development will have to be accounted for in further details for future risk assessments.

In the work of Lunde, T. M. et al. compared six temperature dependent mortality models for the malaria vector Anopheles gambiaesensustrict [124]. The evaluation is based on a comparison between the models, and observations from semi-field and laboratory settings. The six different parameterization schemes have been developed to describe the mortality rates for adult An. gambiaes.s. These schemes are important for estimating the temperature at which malaria transmission is most efficient. The models can also be used as tools to describe the dynamics of malaria transmission.

Martens 1, the first model, which is called Martens scheme in Ermert et al. [125], and described by Martens et al. [126,127,128] derived from three points, and shows the relationship between daily survival probability (p) and temperature (T). This is a second order polynomial, and is, mathematically, the simplest of the models.

$$p(T) = -0.0016.T^2 + 0.054.T + 0.45 \qquad (12)$$

Martens 2, in 1997 Martens et al. [128] described a new temperature dependent function of daily survival probability. This model has been used in several studies. This model is named Martens 2. Numerically, this is a more complex model than Martens 1, and it increases the daily survival probability at higher temperatures.

$$p(T) = e^{\frac{-1}{-4.4+1.31T-03.T^2}} \quad (13)$$

Bayoh-Ermert, In 2001 Bayoh carried out an experiment where the survival of An. gambiaes.s. under different temperatures (5 to 40 in 5°C steps) and relative humidities (RHs) (40 to 100 in 20% steps) was investigated [129]. This study formed the basis for three new parametrization schemes. In 2011, Ermert et al. [125] formulated an expression for Anopheles survival probability; however, RH was not included in this model. This model is a fifth order polynomial. Overall, this model has higher survival probabilities at all of the set temperatures compared with the models created by Martens.

$$p(T) = -2.123 * 10^{-7} * T^5 + 1.951 * 10^{-5} * T^{-4} - 6.394 * 10^{-4} * T^3 + 8.217 - 3 * T^2 - 1.865 * 10^{-2} * T + 7.238 * 10^{-1} \quad (14)$$

Bayoh-Parham, in 2012, Parham et al. [130] included the effects of relative humidity and parameterized survival probability using the expression shown below. This model shares many of the same characteristics as the Bayoh-Ermert model. The mathematical formulation is similar to the Martens 2 model, but constants are replaced by three terms related to RH (β0, β1, β2).

$$p(T, RH) = e^{(-T^2\beta 2+T\beta 1+\beta 0)^{-1}} \quad (15)$$

Where β0 = 0.00113 · RH2 − 0.158 · RH − 6.61, β1 =−2.32 ·10−4 ·RH2+0.0515 ·RH+1.06, and β2 = 4 ·10-6RH2 − 1.09 · 10−3 · RH − 0.0255. For all models reporting survival probability, they can rewrite p to mortality rates, β according to: β = −ln(p)

Bayoh-Mordecai, Mordecai [131] re-calibrated the Martens 1 model by fitting an exponential survival function to a subset of the data from Bayoh and Lindsay [129]. They used the survival data from the first day of the experiment and one day before the fraction alive was 0.01. Six data points were used for each temperature.

$$p(T) = -0.000828.T^2 + 0.0367.T + 0.522 \quad (16)$$

Bayoh-Lunde, from the same data [129], Lunde et al. [132], derived Lunde, derived an age dependent mortality model that is dependent on temperature, RH, and mosquito size. This model assumes non-exponential mortality as observed in laboratory settings [129], semi-field conditions [133], and in the field [134]. The four other models use the daily survival probability as the measure, and assume that the daily survival probability is independent of mosquito age. The present model calculates a survival curve with respect to mosquito age. Like the Bayoh-Parham model, they have also varied the mosquito mortality rates according to temperature and RH. In their model, the effect of size is minor compared with temperature and relative humidity. Their results have shown how different mortality calculations can influence the predicted dynamics of malaria transmission. With global warming as a reality, the projected changes in malaria transmission will depend on which mortality model is used to make such predictions.

In the work of Beguin, A. et al. [135] describes an empirical model of the past, present and future-potential geographic distribution of malaria which incorporates both the effects of climate change and of socioeconomic development on the research title called the opposing effects of climate change and socio-economic development on the global distribution of malaria. A logistic regression model using temperature, precipitation and gross domestic product per capita (GDPpc) identifies the recent global geographic distribution of malaria with high accuracy (sensitivity 85% and specificity 95%). The researchers have reached in a conclusion that climate factors have a substantial effect on malaria transmission in countries where GDPpc is currently less than US$20,000. Using projections of future climate, GDPpc and population consistent with the IPCC A1B scenario, they estimated the potential future population living in areas where malaria can be transmitted in 2030 and 2050. In 2050, the projected population at risk is approximately 5.2 billion when considering climatic effects only, 1.95 billion when considering the combined effects of GDP and climate, and 1.74 billion when considering GDP effects only. Under the A1B scenario, the researchers showed that climate change has much weaker effects on malaria than GDPpc increase. This outcome is, however, dependent on optimistic estimates of continued socioeconomic development where future socioeconomic development can be influenced to a large degree by unanticipated factors. Even then, climate change has important effects on the projected distribution of malaria, leading to an increase of over 200 million in the projected population at risk.

The objective of the study, an empirical malaria distribution map for West Africa, researched by Immo Kleinschmidt, Judy Omumbo, Olivier Briet, Nick van de Giesen, Nafomon Sogoba, Nathan Kumasenu Mensah, Pieter Windmeijer, Mahaman Moussa and Thomas Teuscher was to produce a malaria distribution map that would constitute a useful tool for development and health planners in West Africa[136]. The resulting model was then used to predict parasite prevalence for the whole of West Africa however a shortcoming in their modeling methodology that they were unable to give an estimation error for the various parts of the map.

Socio-economic development could reduce malaria transmission in a variety of ways. For example, increases in household income of women and poverty reducing measures in general have the potential to reduce exposure to malaria and to improve health seeking behaviour and quality of treatment. However, socio-economic development in a high transmission tropical setting could equally increase malaria transmission because of changes such as forest clearing or the migration of people with little or no immunity into areas of high endemicity. They have been unable to model such factors in our analysis due to the fact that such data for the entire region are currently not available with adequate spatial resolution. It is highly likely that there are other unmeasured, perhaps more local factors that determine variation in parasite prevalence.

A further source of variation that has not been taken into account in this study is variation in prevalence by season and by age. The impact of these factors will differ according to the endemicity level of an area. A regional malaria risk map, such as the one produced in this study, will allow planners to assess the possible health impacts of measures aimed at improving food security through the promotion of large scale irrigation and wetland management projects. Elsewhere in Africa such developments have significantly increased malaria infection and morbidity in epidemic prone areas of unstable malaria. However, the same agricultural production methods are unlikely to affect the malaria risk profile of rural populations living in areas characterized by high parasite prevalence. They also produced estimates of the proportion of each country in the region exposed to various categories of risk to show the impact that malaria is having on individual countries.

Finally, the map will also help guide public health research managers in identifying appropriate study environments for intervention trials as well as assist with the identification of populations potentially benefiting from new interventions. The data they used represents a very large, albeit imperfectly, sampled population of children in West Africa. This study is a first attempt to

produce a malaria risk map of the West African region, based entirely on malariometric data. We anticipate that it will provide useful additional guidance to control programme managers, and that it can be refined once sufficient additional data become available.

A research called modelling and prediction of malaria vector distribution in Bangladesh from remote-sensing data, which is done by RAHMAN, KOGAN, ROYTMAN, M. GOLDBERG and GUO states that moisture and thermal conditions have shown a highest correlation coefficient with malaria cases [137]. Epidemic malaria cases and satellite-based vegetation health indices were investigated and measured by the advanced very high resolution radiometer (AVHRR). Two indices characterizing moisture and thermal conditions were investigated using correlation and regression analysis applied to the number of malaria cases recorded in the entire Bangladesh region and the correlation increased , reaching a maximum value of 0.5 to 0.8 by the middle of the high season. Following these results, regressional equations for the number of malaria cases as a function of VH indices were built and tested independently. They showed that, in the main malaria administrative division (Chittagong) and the entire Bangladesh region, the regressional equations can be used for early prediction of malaria development.

In the research done by Huang, Fang, et al [138], the correlation between malaria and meteorological data has been studied at Motuo County, Tibet. The study was analysed using several statistical methods and spearman correlation analysis was conducted to examine the association between monthly malaria incidence and meteorological variables. Accordingly, correlation coefficient achieved in relative humidity was 0.543 and in rainfall was 0.348. The objective of this work was to analyse the correlation between malaria incidence and meteorological factors in Motou County, in order to seek the particular interventions for malaria control.

A research done by Tesfaye, Solomon, et al [139], claimed as there was no significant association between malaria occurrences and meteorological variables where as in Ayele's PhD study [140], has identified socio-economic, geographic and demographic risk factors of malaria based on the rapid diagnosis test (RDT) survey. In his study, a generalized linear mixed model with spatial covariance structure was used to analyses the data from the Amhara, Oromia and Southern Nation Nationalities and People's (SNNP) regions where the response variable was the presence or absence of malaria using the RDT. As the result, households in the SNNP region were found to be at more risk than Amhara and Oromia regions. Moreover, households which

have toilet facilities clean drinking water, and a greater number of rooms and mosquito nets in the rooms, have less chance of having household members testing positive for RDT.

In the research studied by Eskindir [141], it is observed that higher malaria incidence are found among males, children 5–14 years of age, ITNs non-users, the poor, and people who lived closer to vector breeding places. Rainfall increased and indoor residual spraying with Deltamethrin reduced falciparum incidence. Whereas the researches couldn't observe the mass ITNs distribution as it can reduce falciparum malaria incidences at a village level. Finally, the assessment concluded the potential effects of local meteorological and environmental conditions, indoor residual spraying with insecticides, insecticide-treated nets (ITNs) use at individual and community levels, and individual factors on Plasmodium falciparum malaria incidence in a village in south Ethiopia.

A research studied at Rwanda by Jean [142], shows that malaria is sensitive to both environmental and socio-economic factors, it is equally important to assess the extent to which malaria is the interplay of climate variability and interventions in Rwanda highlands. Correlation coefficients were examined and independent variables are identified that would be the best predictors of malaria using multi linear regression. When using only climate variables, the correlation has shown less than 0.5, which implies that climate variables only explain 27.7 % of variability in malaria but when climate variables were combined with interventions into the model, the correlation has raised to 0.725. This also indicates that malaria has correlation with socioeconomic factors than only the climate variability.

As the research [143] explains, there are habitat characterizations of malaria vector mosquito larvae studies at Gamo Gofa; Ethiopia determines the occurrence and density of larvae. In the study, water depth; habitat size; distance to nearest dwelling; land use within a 10 m and 100 m from the sampling site; and number of domestic animals within 100m are included as environmental variables. There are also physiochemical variables such as water temperature, pH, electric conductivity (EC), total dissolved solids (TDS), dissolved oxygen (DO), turbidity and phosphate are a part also a part of the study. As the result, the study concludes that water depth, temperature and percentage of tall riparian vegetation are important factors to consider when designing a control program for anopheles larvae because of the fact that clearing riparian forest and other tall vegetation is likely to improve growing conditions for anopheles larvae. In

addition, different habitat classes were either exclusively positive or negative for anopheles larvae, irrigation channels in the area not being suitable larval habitats during the time of measurements. The authors suggest that more studies are needed, preferably on a larger set of sampling sites and over a longer period.

Emebet and Suryabhagavan [144] has prepared a malaria-risk map of Mecha district of Ethiopia by establishing the relationship of various climatic and non-climatic factors related to the disease using regression analysis. Accordingly temperature, rainfall, altitude, distance from streams, distance from swamps and ponds, population density, health facilities and land-use/land-cover patterns were used to prepare malaria-risk areas. As the result, most of the study area (99.01%) was found to belong to high and moderate malaria-risk based on the derived four categories of malaria-risk ranging from very high to low. In addition to that, the researcher has shown as there is a correlation between the factors and malaria incidents such as rainfall is the most dominant factor for the prevalence of malaria, whereas altitude has limited effect for the prevalence of the disease as altitude has negative correlation with temperature; NDVI has also higher level of correlation with malaria incidence as the relationship of NDVI to Entomological Inoculation Rate (EIR) is highly correlated; population density has a significant positive correlation with the incidence of malaria and to the contrary, distance from health facilities and altitude has a negative correlation with the incidence of malaria. As a conclusion, the research paper has given a proof of concept as there is a correlation between climatic and non-climatic factors with the incidence of malaria.

In the related works reviewed, researchers have shown the various factors using various techniques that cause malaria dispersion to occur such as sensitivity to climate viabilities, socioeconomic status, population density, surface hydrology, living standard, population income, GDP, life span of mosquito, etc. Moreover, they have also put their results in terms of correlation coefficient, mathematical formula, logical reasoning, etc. and suggest their recommendations. However, we haven't observed the need of a generic framework required for such studies where it will be a place for a number of algorithms will be trained to create a model. And also we haven't observed a highest correlation coefficient between the factors and malaria incidents. Hence, our work focuses on addressing the two main gabs identified in the related research works.

Chapter 4: Data Preparation and Analysis

4.1 Data Preparation

Around thirty three variables are used by the ministry of health in related to malaria data. Most of the variables are used for administrative purposes such as total malaria outpatient cases, total malaria inpatient cases, number of government HPs reported, number of government HCs reported, number of government hospitals reported, number of government NGOHF reported, etc. In the aspect of selecting the variables from the dataset of malaria is location (region, zone, woreda, year, month, week, and total malaria confirmed clinically). Based on the researches [11, 12, 136, 138, 139, 140, 141, 143, 144,] and expert opinions, malaria is correlated with many factors such as land surface temperature (LST), normalized difference vegetation index (NDVI), altitude, rainfall, distance to permanent water bodies, socio-economic information, geographic information, demographic information, water depth, habitat size, distance to nearest dwelling, water temperature, gender, age category, insecticide treated nets (ITNs) non users, people who lived closer to vector breeding places, indoor residual spraying with Deltamethrin, etc. Independent variables, which are considered highly relevant to the dispersion of malaria based on researches done so far and expert opinion, have been identified, collected and organized from the Ministry of Health, Ethiopian Metrology Agency, census data of Central Statistics Agency and raster data of Landsat satellite image of Ethiopia based on data availability.

Malaria dataset from the Ministry of Health contains approximately 127,104 cases that are organized in a weekly cumulative fashion from the year 2009 to 2015, has been collected using 33 fields called region name, zone name, woreda name, year, epidemic week, month, total malaria confirmed clinical, total malaria outpatient cases, total malaria inpatient cases, total malaria suspected fever examined, post malaria RDT or microscopy PF outpatient cases, post malaria RDT or microscopy PV outpatient cases, number of government HPs expected by RHB, number of government HCs expected by RHB, number of government hospitals expected by RHB, number of government NGOHF expected by RHB, number of government other HFs expected by RHB, number of government HPs reported, number of government HCs reported, number of government hospitals reported, number of government NGOHF reported, number of government other HFs reported, all total sites reported, all total sites expected by RHB, total

government sites reported, date sent, date received, date week expected, timeliness, Ethiopian month, Ethiopia month number and Ethiopia year.

The malaria dataset covers 11 regions, 105 zones and 920 woredas. Limited attributes are considered as the intention of gathering this data is to know the number of confirmed malaria cases in a given location. Location in the form of region, zone and woreda name are considered; date in terms of year, month and weeks are considered; and the total confirmed malaria cases are considered for this study.

Metrology dataset is organized into five categorical data sets, which are relative rainfall, minimum temperature, maximum temperature, relative humidity and wind speed. The dataset is approximately 1364 summarized in a monthly average value for each category from the year 2006 to 2015, has been collected from the Ethiopian Metrology Agency based on 19 fields called name, zone, elevation, longitude, latitude, element, year, January, February, March, April, May June, July, August, September, October, November and December.

The name attribute is representing the location name where the metrology data collection station is located and most of these names are the same as the woreda name. Each month is considered as a field value and the category such as rainfall, humidity, etc. is identified using element field and the values are distributed along by the name of each months. Rearrangement has been made in order to make the data integration with malaria dataset. Each category is considered to be as a field value by the Ethiopian Metrology Agency where values of each category are distributed along with those mentioned month fields. A single field called month is assumed to represent fields assigned to each months. The data re-arrangement is done in a form of name (woreda), zone, relative rainfall average, relative humidity average, temperature maximum average, temperature minimum average, month and year.

Data integration is made between the two re-arranged datasets of malaria and metrology datasets using zone, woreda, year and month as a source of integration point. And we have created a dataset that combines malaria and metrology datasets containing a field of region name, zone name, woreda name, elevation, date and total malaria confirmed cases.

The third dataset considered is the raster data of Landsat satellite image of Ethiopia from May 01, 1998 to December 10, 2015. The raster data of Landsat satellite image contains 635 images; consecutive images have 10 days interval. The raster data is converted into vector data in such a way that it is possible to access point or pixel data information to do analysis.

In order to characterize the vegetation conditions, the parameters of the seasonal vegetation patterns observed from MODIS-NDVI time series were used. Value is given from 0 to 255 for each pixel values and accordingly normalized difference vegetation index (NDVI) value is calculated. In addition, co-ordinate values are incorporated on the vector data to consider as one of data integration point with a dataset of malaria and metrology datasets. The NDVI data set is prepared in a monthly fashion by taking the average value of the NDVI data taken three times per month. A dataset of malaria along with metrology and NDVI datasets is created and the common fields of year, month, longitude, and latitude are used as an integration point.

Population dataset is considered from the census data of central statistics agency (CSA) taken during 1994. Accordingly, population estimates for each consecutive years is calculated based on the study result of UNDP Emergency Unit of Ethiopia that reports on the zones and woreda census data projections. The CSA estimates that the growth of population is 3 % per annum [145]. The United Nations Population Division projected that Ethiopia's population in the year 2000 was 62.9 million which is estimated to be 89.8 million by 2015 [146]. Finally, a dataset of malaria, metrology and NDVI along with population dataset is created by combining the common fields of year and woreda as a point of data integration. During the data integration process, the dataset is treated by removing the records which is assumed irrelevant such as data-entry errors, data without associated values during merging datasets, etc.

After the final data integration process, we managed to have 3671 malaria cases with the relevant associated fields such as relative rainfall, relative humidity, temperature maximum, temperature minimum, NDVI and population data of 221 woredas located at 14 zones from 5 regions of Ethiopia (SNNPR, Tigiray, Oromia, Amhara and Ethio-Somalia regions). We have made the following pre-processing of data:

- ***Handling missing values:*** When we examined the original data file from malaria distribution and metrology data, we found several missing values and we have removed the rows containing missing values.
- ***Data cleaning:*** From the four datasets, which are the Malaria dataset, NDVI, Meteorology and population, we have removed the row values that don't have matching the values from one dataset to the other datasets. Population growth rate is calculated to grow by 0.25% monthly by considering the 3 % annum growth rate studied by UNDP so that we can correlate with other datasets which is formulated in monthly fashion.

- ***Data reduction:*** To ensure the robustness of the regression model and to have a more accurate result, only the intersection values are considered. Malaria distribution data was prepared in weekly manner whereas climate data is prepared in monthly manner and hence, all available data were organized in a monthly manner by reducing the dimension.
- ***Data transformation:*** We transformed the Landsat raster data into a vector data to get the NDVI data result of each pixel so that it will be mapped into the other datasets.

4.2 Analysis

Various machine learning algorithms are deployed to train the malaria prediction model. Some of the algorithms used will be discussed along with the results obtained. All algorithms deployed for analysis considered the same environment such as number of instances, test mode, attributes and selected class for classification. The information set that is made available to the analysis are 3671 number of instances; 10 fold cross validation test mode; 11 attributes of dataset (i.e. region name, zone name, woreda name, total malaria confirmed clinically, elevation, rainfall, relative humidity, minimum temperature, maximum temperature, NDVI and population); and selection of total malaria confirmed clinically attributed as a class of prediction.

4.2.1 Support vector machine (SVM)

Support vector machine is among the best supervised machine learning algorithms developed for numeric prediction to produce a model that can usually be expressed in terms of a few support vectors and can be applied to nonlinear problems using kernel functions [147].We deployed the sequential minimal optimization algorithm (SMO) to train a support vector classifier using polynomial or Gaussian kernels. SMOreg implements the sequential minimal optimization algorithm for regression problems [147].

We have tested on kernels including Normalized Polynomial Kernel, Polynomial Kernel, Pre-computedKernelMatrix Kernel, Pearson VII function-based universal Kernel and RBF Kernel. The best performing kernel was Normalized Polynomial Kernel and Polynomial Kernel. The result is as follows:

- NormalizedPolyKernel as an activation function and RegSMOImproved as a regOptimize.
    - Correlation coefficient 0.9078

- Mean absolute error 156.2604
- Root mean squared error 795.1569
- Relative absolute error 27.1906 %
- Root relative squared error 42.7435 %
- Total Number of Instances 3671

- PolyKernel as an activation function and RegSMOImproved as a regOptimize.
  - Correlation coefficient 0.8997
  - Mean absolute error 167.6441
  - Root mean squared error 859.8848
  - Relative absolute error 29.1714 %
  - Root relative squared error 46.223 %
  - Total Number of Instances 3671

PolyKernel is a function used in weka to represent polynomial kernel. Polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other models, supporting kernel function that represents the similarity of vectors in a feature spaces over polynomials of the original variables that permits learning of non-linear models [148].
For degree-d polynomials, the polynomial kernel is defined as

$$K(x, y) = (x^T y + c)^d \quad (17)$$

where x and y are vectors in the input space, i.e. vectors of features computed from training or test samples and $c \geq 0$ is a free parameter trading off the influence of higher-order versus lower-order terms in the polynomial. When c = 0, the kernel is called homogeneous [149].

Normalized polynomial kernel is a polynomial kernel followed by some normalization (such that the normalized kernel is never outside of (-1, 1) for odd exponents and (0, 1) for even exponents.

$$K(x, y) = \frac{\text{PolyKernel(x,y)}}{\text{sqrt(PolyKernel(x,x) PolyKernel(y,y))}} \quad (18)$$

4.2.2 Neural Network

Neural network is composed of two or more layers, although most networks consist of three layers: an input layer, a hidden layer, and an output layer. There may be more than one hidden layer; however, most networks contain only one, which is sufficient for most purposes [150].

The multilayer perceptron is the most known and most frequently used type of neural network. On most occasions, the signals are transmitted within the network in one direction from input to output. There is no loop observed. The output of each neuron does not affect the neuron itself. Such architecture is called feed forward. We used Sigmoid and back propagation algorithms to train the MultilayerPerceptron classifier. The result is as follows:

- Correlation coefficient 0.842
- Mean absolute error 473.5946
- Root mean squared error 1157.3383
- Relative absolute error 82.4093 %
- Root relative squared error 62.2125 %
- Total Number of Instances 3671

We used Softplus and ApproximateSigmoid algorithms to train the MLPRegressor classifier.

MLPRegressor is a multilayer perceptron with one hidden layer using WEKA's Optimization class by minimizing the loss function of ApproximateAbsoluteError or SquaredError plus a quadratic penalty with the BFGS method / Gradient descent for faster training time in case of many parameters. Softplus and ApproximateSigmoid are used as activation function.

The following is the result of using Softplus as an activation function and ApproximateAbsoluteError as a loss function.

- Correlation coefficient 0.9503
- Mean absolute error 133.8818
- Root mean squared error 584.7895
- Relative absolute error 23.2849 %
- Root relative squared error 31.4279 %
- Total Number of Instances 3671

The following is the result of using ApproximateSigmoid as an activation function and SquaredError as a loss function.

- Correlation coefficient 0.8895
- Mean absolute error 200.161
- Root mean squared error 915.2444
- Relative absolute error 34.8296 %
- Root relative squared error 49.1988 %
- Total Number of Instances 3671

4.2.3 Gaussian Process

A Gaussian Process is a collection of random variables, any finite number of which have (consistent) joint Gaussian distributions and fully specified by its mean function m(x) and covariance function k(x, x0). This is a natural generalization of the Gaussian distribution whose mean and covariance is a vector and matrix, respectively. The Gaussian distribution is over vectors, whereas the Gaussian process is over functions.

$$f \sim GP(m, k) \quad (19)$$

Meaning: the function f is distributed as a Gaussian Process with mean function m and covariance function k [145].

We used PolyKernel without hyperparameter tuning to train the Gaussian Processes classifier.

The result is as follows:

- Correlation coefficient 0.862
- Mean absolute error 333.806
- Root mean squared error 962.4009
- Relative absolute error 57.6618 %
- Root relative squared error 51.5213 %
- Total Number of Instances 3671

4.2.4 Linear Regression

Linear Regression performs standard least-squares linear regression and can optionally perform attribute selection, either by greedily using backward elimination or by building a full model from all attributes and dropping terms one by one in decreasing order of their standardized

coefficients until a stopping criteria is reached. The implementation has two further refinements: a mechanism for detecting collinear attributes (which can be turned off) and a ridge parameter that stabilizes degenerate cases and can reduce over fitting by penalizing large coefficients. Technically, Linear Regression implements ridge regression, which is described in standard statistics texts [147].

We used linear regression for prediction that deals with weighted instances. The result is as follows:

- Correlation coefficient 0.7764
- Mean absolute error 435.1748
- Root mean squared error 1173.8986
- Relative absolute error 75.7239 %
- Root relative squared error 63.1027 %
- Total Number of Instances 3671

4.2.5 Linear Least Square Regression

Linear Least Square Regression is a robust linear regression method that minimizes the median (rather than the mean) of the squares of divergences from the regression line. It repeatedly applies standard linear regression to subsamples of the data and outputs the solution that has the smallest median-squared error.

We used linear least square regression with the lowest median squared error for prediction that deals with weighted instances. The result is as follows:

- Correlation coefficient 0.3309
- Mean absolute error 356.7021
- Root mean squared error 1871.0988
- Relative absolute error 62.0691 %
- Root relative squared error 100.5806 %
- Total Number of Instances 3671

As it is shown the results of the models above, it indicates that there is a significant correlation between malaria dispersion with a combination of factors such as location, altitude, average

monthly population, average monthly NDVI, average monthly minimum temperature, average monthly maximum temperature, average monthly relative humidity and average monthly rainfall. Depending on the algorithms which results the models, MultilayerPerceptron algorithm using Softplus as activation function offers 0.9503, which is the best correlation result with a relative absolute error of 23.2849 %.

In this research work, we have considered a number of algorithms to create the predictive models but we have selected to show eight preferred models. As a framework, it is not limited to the algorithms used in this work; rather it is designed to let also other algorithms to be used so that the framework will build a predictive model based on the malaria dispersion data. The framework has also a capability to re-build the already existed predictive models when the data size of the malaria dispersion grows.

Chapter 5: Design of a Framework for Predictive Analysis of Malaria Dispersion

5.1 Overview of a Framework for Predictive Analysis of Malaria Dispersion

The malaria dispersion analysis ecosystem envisaged encompasses a number of stakeholders such as health institutions, environmental institutions, geospatial institution, statistics agency, weather forecasting service providers, researchers, decision makers, donors, etc. Figure 5.1 presents the different components and source of factors that affect the dispersal of malaria.

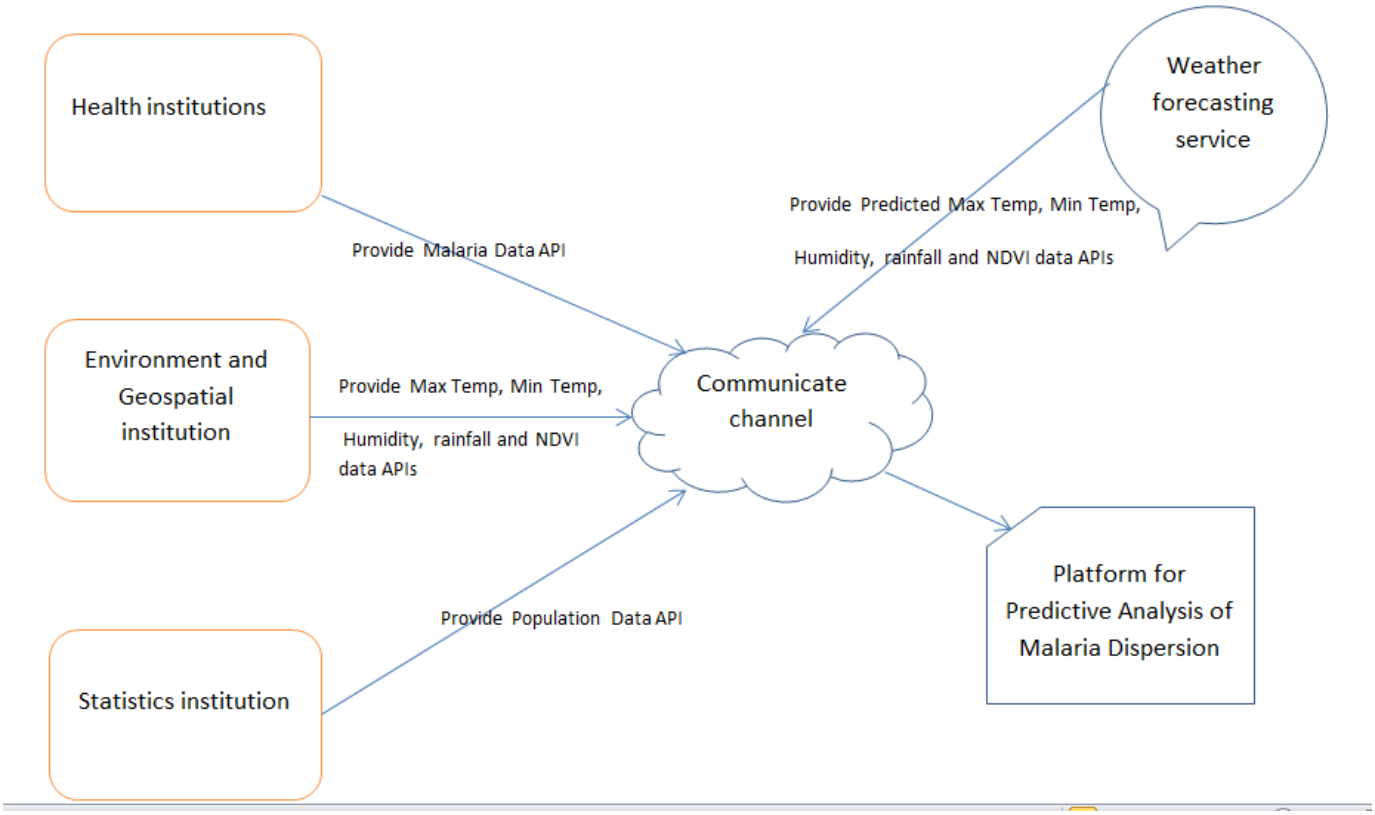

Figure 5.1: Ecosystem of malaria dispersal

As it can be illustrated in the Figure 5.1 in the case of Ethiopia, the Ethiopian health centre institute, the Ethiopian national metrology agency and the Ethiopian statistics agency are entities which are supposed to provide the data that affect the dispersal of malaria, weather including

NDVI and population data respectively. These enables the predictive analysis of malaria dispersion platform for the model training and prediction using weather forecasting services which also supposed to be provided by the Ethiopian metrology agency.

The reference architecture is modular and distributed in nature as it is designed to integrate the Ethiopian public health institute with National Metrology Agency and Central statistical agency so as to support the prediction analysis result to a better level of precision. The architecture contains important components to augment the future capacity of prediction for malaria distribution. Data services are the components that will reside in the organization to provide data by exposing services in such a way that the extraction module, which will be as part of the integral components of predictive analysis, will consume the services to fetch the required data and store in staging database for further processing.

After the extraction of malaria, weather, NDVI and population data, data cleansing activity will be carried out to avoid data with unqualified information such as missing values, values of one source without a relation or association with other sources. Cleansing is a very important stage as it is a component where clean and appropriate data can be established with each dataset or data group.

Three of the datasets shall be interrelated with each other by common variables such as location and time period by merging some of the dataset to fit into one form of dataset. The output of the data transformation component is to have a single dataset what will serve for predictive analytics.

The predictive analytics component is built as a framework to support multiple algorithms to be used whenever needed and to let modellers to test their algorithms using the framework and the data acquired for this purpose. In addition, users can select an algorithm of their choice to predict the malaria distribution by comparing the evaluation result of each model.

The proposed architecture also encompasses a component that has a capability to consume forecasting information of weather and NDVI data so that it will provide data for the inputs necessary for the prediction.

5.2 Framework for Predictive Analysis for Malaria Dispersion

The general framework of predictive analysis for malaria dispersion is presented in Figure 5.2.

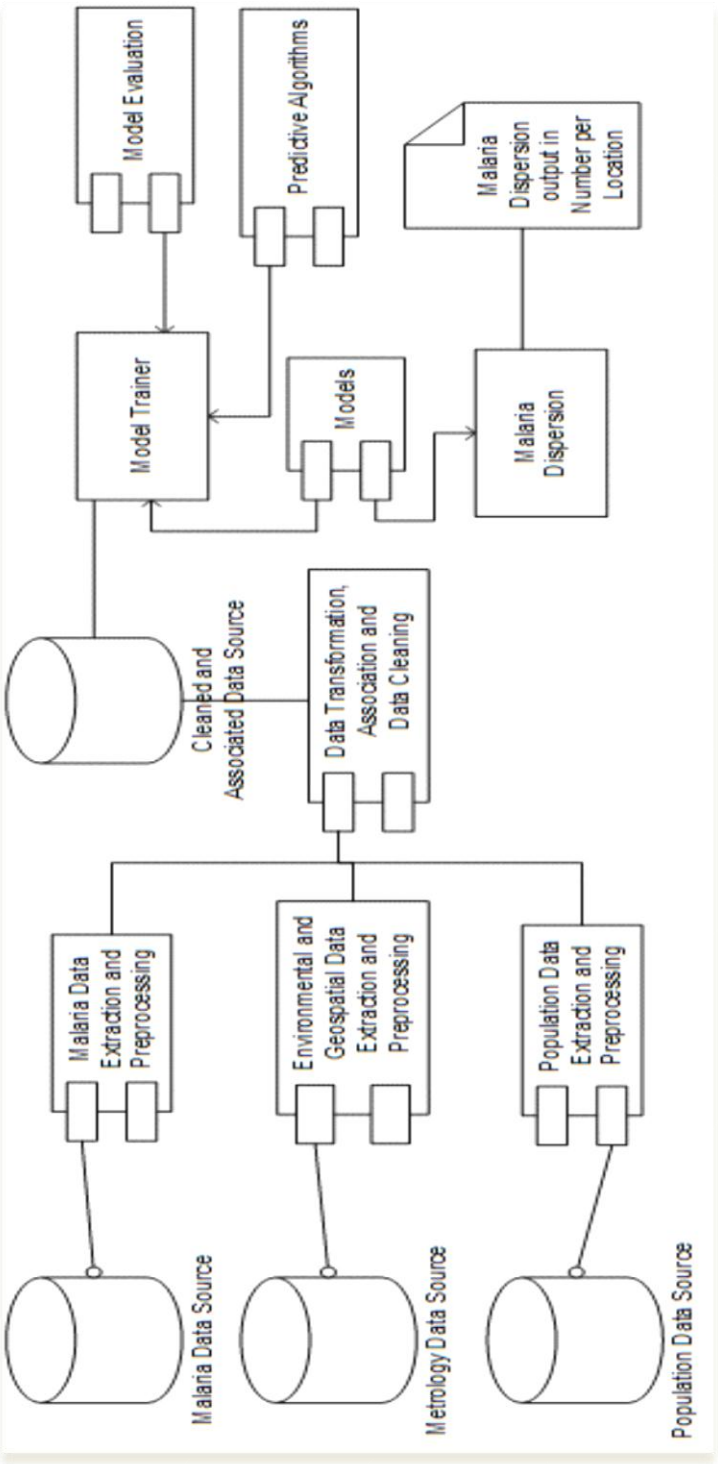

Figure 5.2: A general framework for predictive analysis of malaria dispersal

The framework is based on data intensity which enforces to define a method to extract very important and related information. In this section of the architecture, it is a place where malaria data is extracted from Ethiopian Public Health institute; weather and NDVI data from National Metrology Agency (NMA); and population data from Central Statistical Agency (CSA). In the proposed architecture, we propose the JSON RESTful API as most programming languages have an easy way to convert their standard data structure into Java Script Object Notation (JSON), and to convert JSON into their standard data structures. The methods further comprise determining a boundary of data record by letting the client application to determine the request date by analysing the extracted data.

Malaria Data Pre-Processing, NMA Data Malaria Data Pre-Processing, and Population Data Malaria Data Pre-Processing are services which will be deployed at the respective institutions to expose a service to retrieve the respective data to be consumed by the general framework of predictive analysis for malaria dispersal.

The request and response to fetch the malaria data is as depicted in the Table 5.1. There will be a URL: http://localhost:8080/malariaData/api/getMalariaDatathat accepts a date range request in the form of *{"requestDatefrom":DATE,"requestDateTo":"DATE"}* and to fetch the malaria data in the format as shown in the *Response JSON(Case-1)* in the Table 5.1.

Table 5.1: Request and response for malaria data service

Method	POST
URL	http://localhost:8080/malariaData/api/getMalariaData
Request json	{"requestDatefrom":DATE,"requestDateTo":"DATE"}
Response json (Case-1)	{"malariaDataList":[{"regionName":value,zoneName":value,"woredaName":value,"Year":value,"epidemicWeek":value,"month":value, "totalMalariaConfirmedClinical":value,"totalMalariaOutpatientCases ":value,"totalMalariaInpatientCases":value,"totalMalariaSuspectedFeverExamined":value,"postMalariaRDTOrMicroscopyPFOutpatientCases":value,"postMalariaRDTOrMicroscopyPVOutpatientCases":value,"numberOfGovernmentHPsExpectedByRHB":value,"numberOfGovernmentHCsExpectedByRHB":value,"numberOfGovernmentHospitalsExpectedByRHB":value,"numberOfGovernmentNGOHFExpectedByRHB":value,"numberOfGovernmentOtherHFsExpectedByRHB":value,"numberOfGovernmentHPsReported":value,"numberOfGovernmentHCsReported":value,"numberOfGovernmentHospitalsReported":value,"numberOfGovernmentNGOHFReported":value,"numberOfGovernmentOtherHFsReported":value, "allTotalSitesReported":value,"allTotalSitesExpectedByRHB":value,"totalGovernmentSitesReported":value,"dateSent":value,"dateReceived":value,"dateWeekExpected":value,"timeliness":value,"ethiopianMonth":value, "ethiopiaMonthNumber":value,"ethiopiaYear":value}],"status":"1", "description":""}

Response Json (Case-2)	{"malariaDataList":[],"status":"0", "description":""}

The request and response to fetch the NMA data is as depicted in the Table 5.2 and Table 5.3. There will be a URL: http://localhost:8080/metrologyData/api/getMetrologyData and http://localhost:8080/NDVIData/api/getNDVIDatathat accepts a date range request in the form of *{"requestDatefrom":DATE,"requestDateTo":"DATE"}* and to fetch the NMA data in the format as shown in the *Response JSON(Case-1)* in the Table 5.2 and Table 5.3.

Table 5.2: Request and response for NMA data services for weather data

Method	POST
URL	http://localhost:8080/metrologyData/api/getMetrologyData
Request json	{"requestDatefrom":DATE,"requestDateTo":"DATE"}
Response json (Case-1)	{"metrologyDataList":[{"Name":value,"Zone":value,"Elevation":value,"Longitude":value,"Latitude":value,"Element":value,"Year":value,"January":value,"February":value,"March":value,"April":value,May":value,"June":value,"July":value,"August":value,"September":value,"October":value,"November":value,"December":value}],"status":"1", "description":""}
Response Json (Case-2)	{"metrologyDataList ":[],"status":"0", "description":""}

Table 5.3: Request and response for NMA data services for NDVI data

Method	POST
URL	http://localhost:8080/NDVIData/api/getNDVIData
Request json	{"requestDatefrom":DATE,"requestDateTo":"DATE"}
Response json (Case-1)	{"ndviDataList":[{"regionName":value,zoneName":value,"woredaName":value,"year":value,"month":value,"ndvi":value }],"status":"1", "description":""}
Response Json (Case-2)	{"ndviDataList":[],"status":"0", "description":""}

The request and response to fetch the population data is as depicted in the Table 5.4. There will be a URL: http://localhost:8080/populationData/api/getPopulationDatathat accepts a date range request in the form of *{"requestDatefrom":DATE,"requestDateTo":"DATE"}* and to fetch the population data in the format as shown in the *Response JSON(Case-1)* in the Table 5.4.

Table 5.4: Request and response for population data service

Method	POST
URL	http://localhost:8080/populationData/api/getPopulationData
Request json	{"requestDatefrom":DATE,"requestDateTo":"DATE"}
Response json (Case-1)	{"populationDataList":[{"regionName":value,zoneName":value,"woredaName":value,"year":value,"month":value,"totalPopulation":value }],"status":"1", "description":""}
Response Json (Case-2)	{"populationDataList":[],"status":"0", "description":""}

Because of few constraints, a component couldn't be built to fetch the data using the technology proposed but the data is gathered with the same format manually. Likewise, the NDVI data is highly sensitive to get from the national meteorology agency. Hence, we used MODIS-NDVI time series data of 16 years prepared by Twente University for the purpose of crop micro insurance project. The MODIS –NDVI data is prepared in the form of raster data. More than 600 raster data is converted into vector data to correlate the co-ordinates along with the pixel value of the NDVI data.

The data cleaning component will filter the relevant data by removing the missing values. It focuses on identifying inaccurate, irrelevant and incomplete record set and removing it from the dataset. In principle, data cleaning features requires also to do replacing and modifying the inaccurate, irrelevant and incomplete records by doing extra efforts such as by re-checking with other source of data, expert judgment, doing research, etc. for instance. If the record of data, from malaria dataset of Ethiopian public health institute, is found to be not complete, then it requires to be re-checked by analysing the data sources collected at the health centers of a given woreda. But, from the point of this research work, data cleaning component focuses on getting clean data by removing the missed values and irrelevant fields.

The cleaned data of each respective data source will be consolidated and presented as one data source by the component called association of relevant data in which it consolidates by disregarding mismatched values from each source. The first task is to change each dataset into the same date grouping. Malaria dataset is grouped weekly, weather dataset is grouped monthly, NDVI dataset is prepared in 10 days interval, and population dataset is found only once where the population census is conducted. In addition, the way the dataset structure of weather dataset constructed by the national metrology agency is a bit different because each month of the year is assumed to be fields of weather dataset. As a result, each dataset is consolidated as one and ready to be used for malaria dispersion. The following activities are part of the association of relevant data component.

1. The weather data structure represented at the NMA is not in proper way such as each month is an attribute, weather elements are only represented by only one attribute. Hence, the structure of the weather dataset should be converted in proper way where date is considered as one attribute and each weather elements such as rainfall, humidity, min_temperature and max_temperature should be represented as an attribute (*see Annex A*).
2. Grouping malaria dataset in monthly fashion by aggregating the number of malaria cases collected in weekly bases (*see Annex B*).
3. Grouping NDVI dataset in monthly fashion by taking the average result of the NDVI data collected in 10 days interval.
4. Taking the population of regions, zones and woredas from the report of central statistical agency and calculate the yearly growth by 3% based on UNDP [145] study. Considering the study, we have calculated monthly population for each woredas (*see Annex C*).
5. Merging the weather, malaria, NDVI, population datasets using region, zone, woreda, year and month (*see Annex D*).
6. Finally, the data transformation component will maintain the final dataset of malaria into the Analysis data source.

Malaria dispersion prediction framework is a component where it uses algorithms, which will be provided by users in order to train the algorithm and create a predictive model of malaria dispersion. This framework also allows users to use it for malaria dispersion by providing the input variables after selecting the predictive models. The class and database diagram is depicted in the Figure 5.3 and 5.4 respectively.

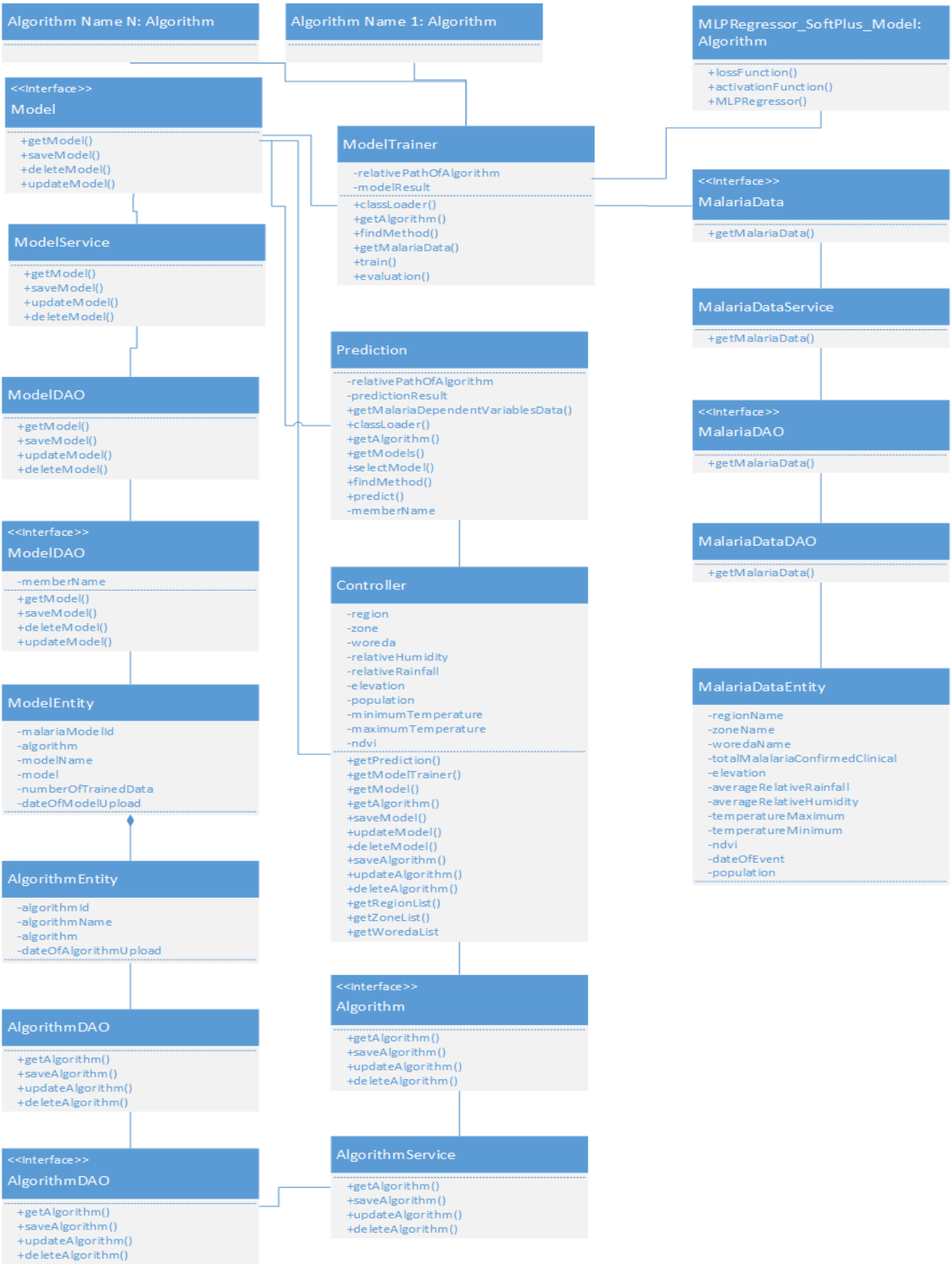

Figure 5.3: Class Diagram for a framework for predictive analysis of malaria dispersion

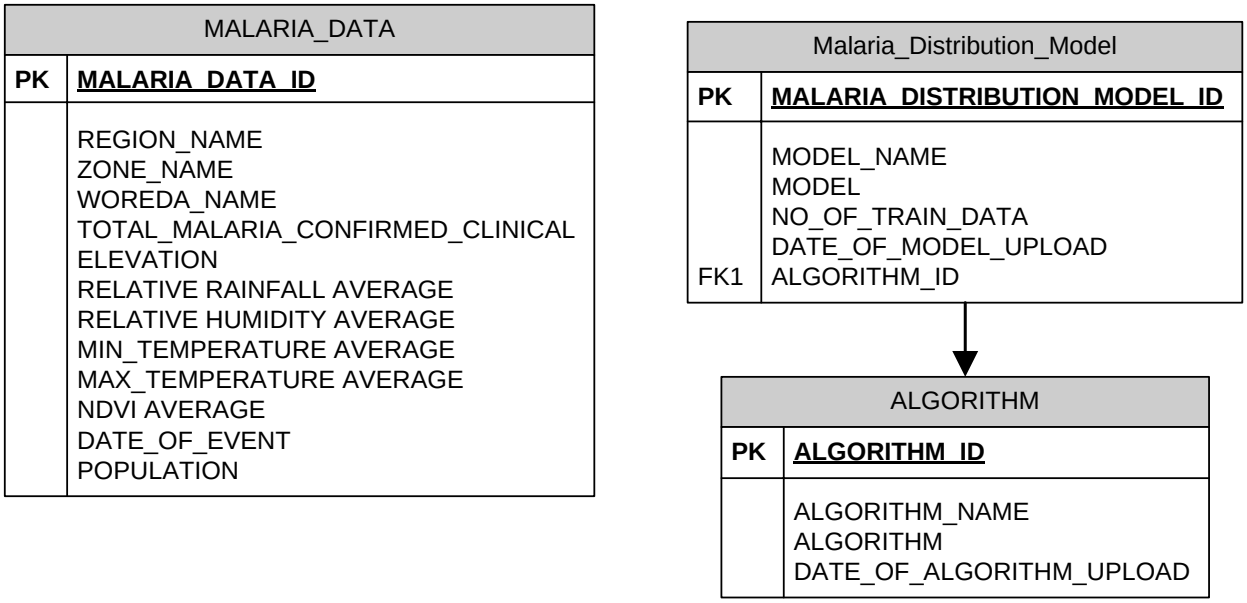

Figure 5.4: Database diagram for a framework for predictive analysis of malaria dispersion

As can be seen in the class diagram, there are classes that can manage the algorithms, malaria data and models to upload, fetch, save, update, etc. The model trainer class uses malaria data service and algorithms to train a model and maintains it with the evaluation result as well as there is prediction class that uses the model class to predict based on the inputs. For the initial implementation, the input is provided by users but when NMA launches the API for weather information, the input will be provided from the API instead.

A standard to be followed to write algorithms and to be executed by the framework built is as follows.

```
public class Model {
/* A method to build a model using the data
Method name is train.
Instances are an argument for the method train.
Vector is a return value which contains a model, evaluation and data used for modelling.
*/
public static Vector train(Instances data) throws Exception {
System.out.println("Training...");
```

/* (Algorithm is supposed to be developed here or it shall be implemented in shall be implemented in another location and shall be instantiated here)*/

```
Evaluation evaluation = new Evaluation(data);
evaluation.crossValidateModel(instantiated class of algorithm , data, 10, new Random(1));
System.out.println(evaluation.toSummaryString());
Vector v = new Vector();
v.add(instantiated class of algorithm);
v.add(new Instances(data));
v.add(evaluation);
/*  SerializationHelper.write(FILENAME, v);// if the model is supposed to be written in a file.
System.out.println("Training finished!");
*/
return v;
}
```

/* A method used to predict to the malaria distribution of a specific woreda.
Method name is "predict".
Instances are an argument for the method "predict" which contains input values of the independent variables.
Vector is an argument for the method "predict" which contains a model, evaluation and data used for modelling. */

```
public static ArrayList<String> predict(Instances data, Vector v) throws Exception {
Classifier cl = (Classifier) v.get(0); // Model
Instances header = (Instances) v.get(1); // Data used for modeling
Evaluation eval = (Evaluation) v.get(2);  //Model Evaluation Result
ArrayList<String> result = new ArrayList<>();
String predictionResult = "Actual -> Predicted";
System.out.println("actual -> predicted");
for (int i = 0; i <data.numInstances(); i++) {
Instance curr = data.instance(i);
```

```
Instance inst = new DenseInstance(header.numAttributes());
inst.setDataset(header);
for (int n = 0; n <header.numAttributes(); n++) {
        Attribute att = data.attribute(header.attribute(n).name());
        if (att != null) {
                if (att.isNominal()) {
                        if ((header.attribute(n).numValues() > 0) && (att.numValues() > 0)) {
                                String label = curr.stringValue(att);
                                        int index = header.attribute(n).indexOfValue(label);
                                                if (index != -1) {
                                                        inst.setValue(n, index);
                                                        }
                                }
                        } else if (att.isNumeric()) {
                                inst.setValue(n, curr.value(att));
                        } else {
                                throw new IllegalStateException("Unhandled attribute type!");
                                }
                        }
                }
doublepred = cl.classifyInstance(inst);
predictionResult = predictionResult  + " -> " + pred;
}
System.out.println("Predicting finished!");
result.add(0, predictionResult); // Prediction result
result.add(1, cl.toString());        // Model
result.add(2, eval.toSummaryString()); // Evaluation Summary of the Model
return result;
}
```

Chapter 6: Implementation

6.1 Prototype Development

A prototype is developed using Java as a programming language. Development Frameworks such as spring and hibernate are considered to ease the development and object relational mapping. Oracle is used as a database. Below is the user interfaces with one scenario for illustration.

Oromia region, Ilu Aba bura zone, Bedele Zuriya with a population of 160593 and elevation 2011 meter is selected as a scenario to be considered. Attributes considered for prediction are:

- Rainfall: 305.5mm
- Relative humidity: 81.2 g/m^2
- Max. Temp: 24.4^0c
- Min, Temp: 12.4^0c
- NDVI: 0.733336

The prediction result as shown from the Figure 6.1 is 590 but the actual confirmed malaria case in the woreda is 596.

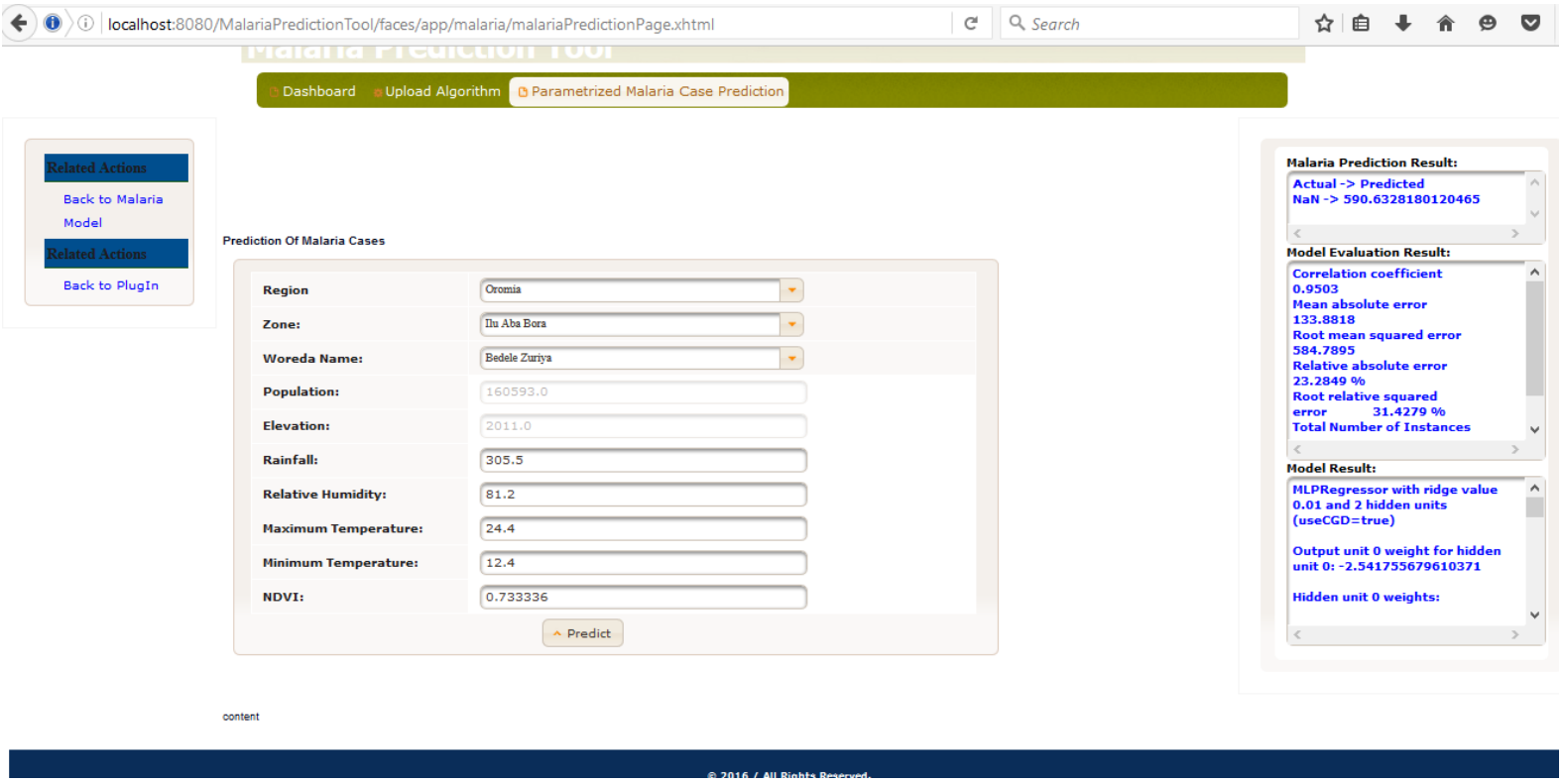

Figure 6.1: Parameterized malaria dispersion prediction interface

The Prediction result from the System is:

- Actual -> Predicted
- None -> 590.6328180120465
- Model Evaluation Result is:
- Correlation coefficient 0.9503
- Mean absolute error 133.8818
- Root mean squared error 584.7895
- Relative absolute error 23.2849 %
- Root relative squared error 31.4279 %
- Total Number of Instances 3671

Besides, the framework can accept a model from other modellers as shown in Figure 6.2.

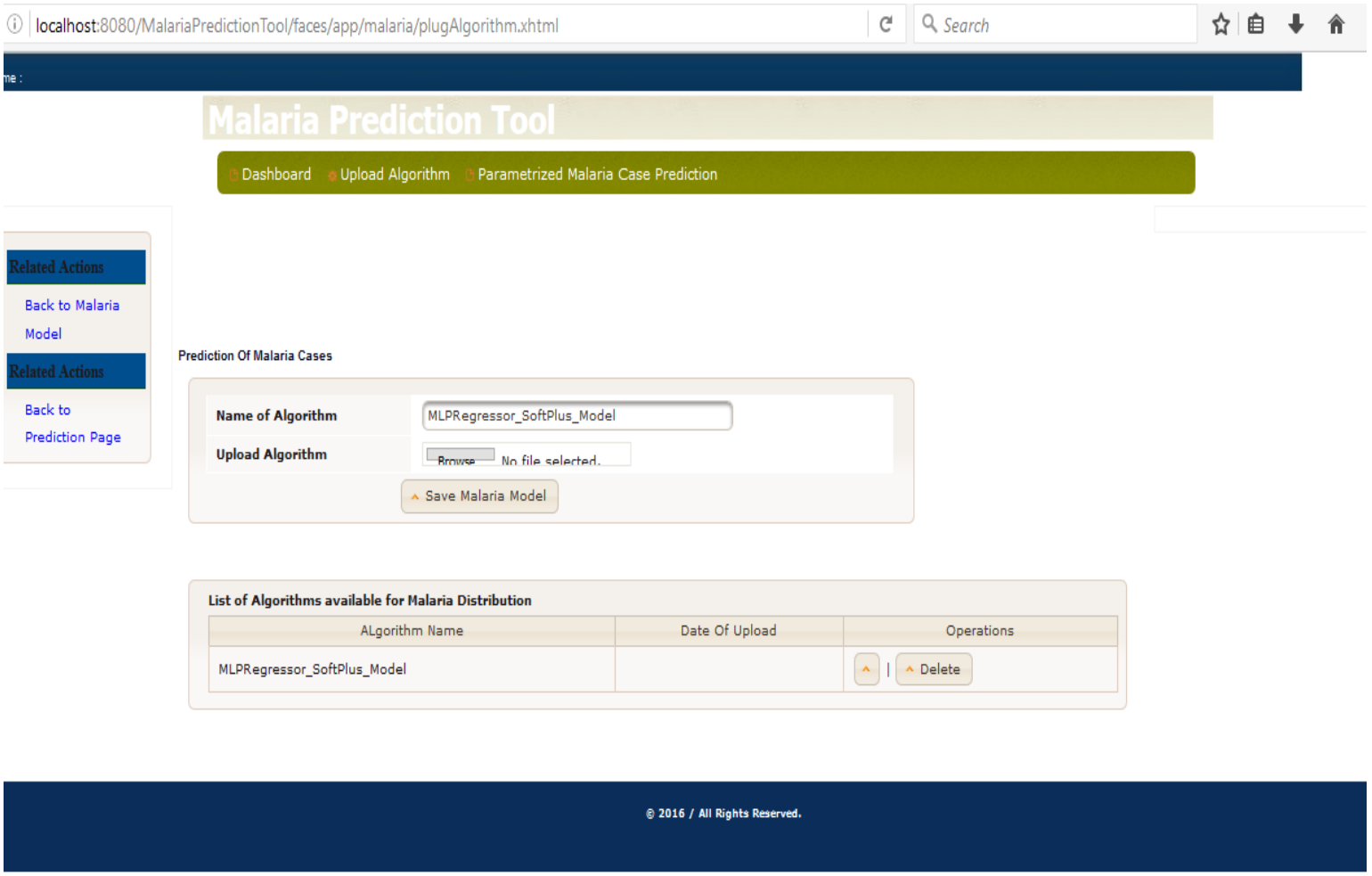

Figure 6.2: Algorithm upload interface

The algorithm shall be written in JAVA language and after compilation, it can be uploaded into a system so that the algorithm will be ready to be used by the model trainer to create a model by using the algorithm and the malaria cases built so far, which are around 3671 cases.

After the algorithm is uploaded, the following user interface on Figure 6.3 appears for algorithm selection so that we will have two options to go. If the algorithm is already trained with the available data, it will go to prediction page otherwise the selection of algorithm will lead to train a model.

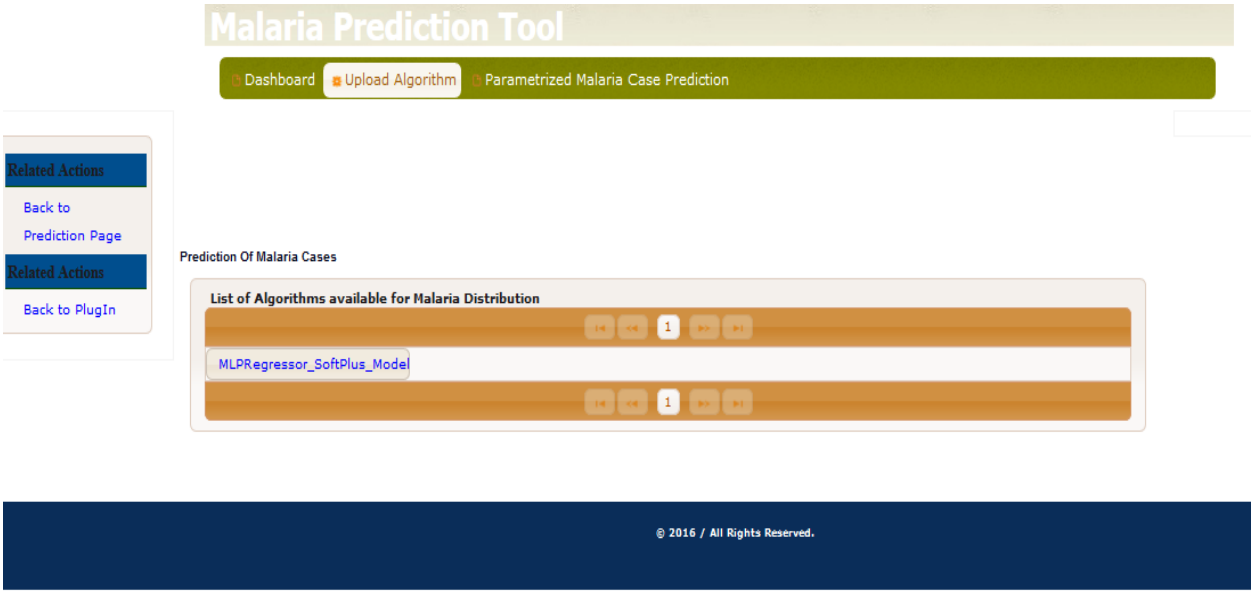

Figure 6.3: Algorithm list view interface

6.2 Evaluation and Discussion

The original datasets of malaria dispersion, climate information, NDVI and population are transformed into a dataset of 3671 clinically confirmed malaria cases labelled for training data to find the most optimal model using different kinds of algorithms that helps to predict the possibility of malaria dispersion existence at woreda level. We have got a correlation coefficient of 0.9503 for MultilayerPerceptron using Softplus as activation function and ApproximateAbsoluteError as a loss function. This model has the best correlation coefficient on predicting the malaria dispersion based on the inputs provided which are mainly location and metrology information.

In the work of Huang Fang et.al [138], the correlation has been done between malaria incident and few elements of meteorology data, which is rainfall and humidity, however the correlation is unsatisfactory compared to the correlation result found in our research work. There is also a research work by Tesfaye, Solomon, et al [139], which is disproved by our research work and other similar research works where by the research paper claims as there is no significant association between malaria incident and meteorological variables.

As indicated in the result of our analysis, there is a significant correlation between malaria incidents with location, altitude, average monthly population, average monthly NDVI, average monthly minimum temperature, average monthly maximum temperature, average monthly relative humidity and average monthly rainfall. We have also observed the correlation becomes less reduced during the analysis when we disregard some of the factors. The correlation becomes 0.75 when we disregard location as a factor which indicates that there are still factors to be considered to identify the causes of malaria incidents as a substitution of location. As it is stated in the PhD work of Ayele [140], additional factors of malaria incidents shall be considered such as socio-economic, geographic and demographic information as an independent variable. In his research work, households which have toilet facilities clean drinking water, and a greater number of rooms and mosquito nets in the rooms have less chance of having household members testing to be positive for rapid diagnosis test. Likewise, a research paper [141] indicated that gender, age category, ITNs non users, people who lived closer to vector breeding places, indoor residual spraying with Deltamethrin, etc. have correlation to the incident of malaria.

The work of Bizimana [142] has shown that malaria is sensitive to environmental and socio-economic factors and also malaria incident intervention programs. When the researchers use only

climate variables, the correlation has shown less than 0.5, which implies that climate variables only explain 27.7 % of variability in malaria but when climate variables were combined with interventions into the model, the correlation has raised to 0.725. This also indicates that malaria has correlation with socioeconomic factors than only the climate variability. And the research output is far lesser than our research work.

As it is stated in the research work of Loha and Brent [141], limited research works are done in relation to the study of malaria incident correlation with associated factors. However, in our research work, it is found a highest correlation coefficient between malaria incident and factors associated with it. In addition, most research papers[143,144] suggest to consider more variables such as water depth; habitat size; distance to nearest dwelling; land use within a 10 m and 100 m from the sampling site; and number of domestic animals within 100m,water temperature, pH, electric conductivity (EC), total dissolved solids (TDS), dissolved oxygen (DO), NDVI, turbidity, phosphate, etc.

10-fold cross validation is applied to evaluate the effectiveness of the proposed models. We have got a correlation coefficient of 0.9503 for MultilayerPerceptron using Softplus as activation function and ApproximateAbsoluteError as a loss function. This model has the best correlation coefficient on predicting the malaria distribution based on the inputs provided which are mainly location and metrology information. Mean absolute error, root mean squared error, relative absolute error and root relative squared error are used to measure the accuracy of the model. Errors $|e_i| = |y_i - x_i|$, where y_i is the prediction and x_i is the true value. Mean absolute and root mean squared error measures accuracy in terms of value whereas relative absolute and root relative squared error measures in terms of percentage. A complete evaluation summary is given in Table 6.1.

Table 6.1: Comparison of Prediction Models for Malaria Dispersion

Models Used	**Correlation coefficient**	**Mean absolute error**	**Root mean squared error**	**Relative absolute error**	**Root Relative squared error**
Support vector machine using NormalizedPolyKernel as an activation function	0.9078	156.2604	795.1569	27.1906 %	42.7435 %
Support vector machine using PolyKernel as an activation function	0.8997	167.6441	859.8848	29.1714 %	46.223 %
MultilayerPerceptron using sigmoid and backpropagation algorithms.	0.842	473.5946	1157.3383	82.4093 %	62.2125 %
MultilayerPerceptron using Softplus as activation function and ApproximateAbsoluteError as a loss function.	0.9503	133.8818	584.7895	23.2849 %	31.4279 %
MultilayerPerceptron using ApproximateSigmoid as activation function and SquaredError as a loss function.	0.8895	200.161	915.2444	34.8296 %	49.1988 %
GaussianProcesses classifier using PolyKernel without hyperparameter tuning	0.862	333.806	962.4009	57.6618 %	51.5213 %
Linear Regression	0.7764	435.1748	1173.8986	75.7239 %	63.1027 %
linear least square regression with the lowest median squared error	0.3309	356.7021	1871.0988	62.0691 %	100.5806 %

The prototype implementation of our system uses a design pattern to facilitate the construction of an extensible system in order to allow uploading algorithms into a system in such a way that modelers can upload their algorithm and built a model by train the algorithm using the existing data and system. Those who need to see the distribution of the malaria will select a model provided and get the result along with the precisions.

Chapter 7: Conclusion and Future works

7.1 Conclusion

A lot of qualitative and quantitative researches have been predicting the likely dispersion of malaria based on various determinants. However, researches done so far are unsatisfactory in terms of the results of the predictive models of malaria dispersion as well as the maturity of predictive analysis work of malaria dispersion in general. The notion of this work is to create a framework of predictive analysis for malaria dispersion to device a generic framework to let users or modellers to create their own models using algorithms and to provide a best malaria dispersion model to let end users to use it as well as to serve as a baseline for modellers.

In this work, malaria incidence, weather, NDVI and population dataset are considered. In order to collect the data, the framework suggests deploying a pre-processing and data extraction components at the premises of the data source to integrate with the framework to extract the required data after pre-processing it. The malaria dispersion data, which is used for this work comprises of the above mentioned datasets after doing pre-processing and data transformation. Our model uses MultilayerPerceptron algorithm and achieves a correlation coefficient of 0.9503 for malaria dispersion, which shows a higher correlation as compared to other works.

In addition, a prototype has been built to show the core feature of the general framework to build a flexible and extendible system to allow modellers to upload their algorithms on the system and to train it by the data source prepared for this purpose and to generate a model and to evaluate it so that users of the system can have alternatives of model for prediction to go for the best choice.

The main contribution of this work is identification of parameters/factors that affects malaria dispersion, create a dataset vital for malaria dispersion from various data sources, design of a framework for predicting malaria dispersion, perform analysis of models that can create higher correlation coefficient between the parameters used and the result of the prediction, development of selected algorithm that shows the higher correlation, development of prototype, train the algorithm to create a model for malaria dispersion prediction and evaluation of the proposed model.

7.2 Future works

This work can be further extended to examine the validity of the models on larger datasets besides recommending the solution by enhancing to a full scale implementation so that it can assist decision makers to provide a better malaria intervention service. Moreover, it can extend the framework to provide visual presentation or map as it takes some ingenuity to produce a representation that is easily understood by managers who are not quantitatively oriented.

References

[1] Weiss, S. M., & Kulikowski, C. A. (1991). *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. Morgan Kaufmann Publishers Inc.

[2] Hoog, A. (2011). *Android forensics: investigation, analysis and mobile security for Google Android*. Elsevier.

[3] Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer.

[4] Craig, M. H., Snow, R. W., & le Sueur, D. (1999). A climate-based distribution model of malaria transmission in sub-Saharan Africa. *Parasitology today*, *15*(3), 105-111.

[5] Parham, P. E., & Michael, E. (2009). Modeling the effects of weather and climate change on malaria transmission. *Environmental health perspectives*, *118*(5), 620-626.

[6] Bacaër, N. (2007). Approximation of the basic reproduction number R 0 for vector-borne diseases with a periodic vector population. *Bulletin of mathematical biology*, *69*(3), 1067-1091.

[7] Bacaër, N., & Ouifki, R. (2007). Growth rate and basic reproduction number for population models with a simple periodic factor. *Mathematical Biosciences*, *210*(2), 647-658.

[8] MacDonald, G. (1957). The epidemiology and control of malaria. London: Oxford Univ. Pr.

[9] Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology*, *25*(15), 1965-1978.

[10] WorldClim. WorldClim – Global Climate Data. 2009.

[11] Hay, S. I., Guerra, C. A., Gething, P. W., Patil, A. P., Tatem, A. J., Noor, A. M., ... & Smith, D. L. (2009). A world malaria map: Plasmodium falciparum endemicity in 2007. *PLoS medicine*, *6*(3), e1000048.

[12] Gething, P. W., Patil, A. P., Smith, D. L., Guerra, C. A., Elyazar, I. R., Johnston, G. L., ... & Hay, S. I. (2011). A new world malaria map: Plasmodium falciparum endemicity in 2010. *Malaria journal*, *10*(1), 378.

[13] Adigun, A. B., Gajere, E. N., Oresanya, O., & Vounatsou, P. (2015). Malaria risk in Nigeria: Bayesian geostatistical modelling of 2010 malaria indicator survey data. *Malaria journal*, *14*(1), 156.

[14] MODIS Products Table, LP DAAC : NASA Land Data Products and Services. [https://lpdaac.usgs.gov/products/modis_products_table]

[15] FEWSNET Data Portals. [http://earlywarning.usgs.gov/fews/downloads/index.php?regionID=af&productID=3&periodID=6]

[16] Evaluation M. Household Survey Indicators for Malaria Control. Unicef: World Health Organization; 2013.

[17] Kilian, A., Koenker, H., Baba, E., Onyefunafoa, E. O., Selby, R. A., Lokko, K., & Lynch, M.

(2013). Universal coverage with insecticide-treated nets–applying the revised indicators for ownership and use to the Nigeria 2010 malaria indicator survey data. *Malaria journal*, *12*(1), 314.

[18] Worldpop Download Dataset. [http://www.worldpop.org.uk/data/summary/?contselect=Africa&countselect=Nigeria&typeselect=Population]

[19] International Programs Region Summary U.S. Census Bureau. [http://www.census.gov/population/international/data/idb/region.php?N=%20Results%20&T=15&A=separate&RT=0&Y=2010&R=115&C=NI]

[20] Paaijmans, K. P., Blanford, J. I., Crane, R. G., Mann, M. E., Ning, L., Schreiber, K. V., & Thomas, M. B. (2014). Downscaling reveals diverse effects of anthropogenic climate warming on the potential for local environments to support malaria transmission. *Climatic change*, *125*(3-4), 479-488.

[21] Smith, D. L., & McKenzie, F. E. (2004). Statics and dynamics of malaria infection in Anopheles mosquitoes. *Malaria journal*, *3*(1), 13.

[22] Koenraadt, C. J. M., Githeko, A. K., & Takken, W. (2004). The effects of rainfall and evapotranspiration on the temporal dynamics of Anopheles gambiae ss and Anopheles arabiensis in a Kenyan village. *Acta tropica*, *90*(2), 141-153.

[23] Mordecai, E. A., Paaijmans, K. P., Johnson, L. R., Balzer, C., Ben-Horin, T., de Moor, E., ... & Lafferty, K. D. (2013). Optimal temperature for malaria transmission is dramatically lower than previously predicted. *Ecology letters*, *16*(1), 22-30.

[24] Parham, P. E., & Michael, E. (2009). Modeling the effects of weather and climate change on malaria transmission. *Environmental health perspectives*, *118*(5), 620-626.

[25] Parham, P. E., Pople, D., Christiansen-Jucht, C., Lindsay, S., Hinsley, W., & Michael, E. (2012). Modeling the role of environmental variables on the population dynamics of the malaria vector Anopheles gambiae sensu stricto. *Malaria Journal*, *11*(1), 271.

[26] Van Lieshout, M., Kovats, R. S., Livermore, M. T. J., & Martens, P. (2004). Climate change and malaria: analysis of the SRES climate and socio-economic scenarios. *Global Environmental Change*, *14*(1), 87-99.

[27] Walker, K., & Lynch, M. (2007). Contributions of Anopheles larval control to malaria suppression in tropical Africa: review of achievements and potential. *Medical and veterinary entomology*, *21*(1), 2-21.

[28] Fillinger, U., Ndenga, B., Githeko, A., & Lindsay, S. W. (2009). Integrated malaria vector control with microbial larvicides and insecticide-treated nets in western Kenya: a controlled trial. *Bulletin of the World Health Organization*, *87*, 655-665.

[29] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[30] Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional*

data analyst. John Wiley & Sons.

[31] Dye, C. (2014). After 2015: infectious diseases in a new era of health and development. *Phil. Trans. R. Soc. B*, *369*(1645), 20130426.

[32] Buczak, A. L., Baugher, B., Guven, E., Ramac-Thomas, L. C., Elbert, Y., Babin, S. M., & Lewis, S. H. (2015). Fuzzy association rule mining and classification for the prediction of malaria in South Korea. *BMC medical informatics and decision making*, *15*(1), 47.

[33] Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Academic Press.

[34] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

[35] Larose, D. T., & Larose, C. D. (2015). *Data mining and predictive analytics*. John Wiley & Sons.

[36] Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Springer Science & Business Media.

[37] Pyle, D. (1999). *Data preparation for data mining*. morgan kaufmann.

[38] Lerman, K. (2007). Social information processing in news aggregation. *IEEE Internet Computing*, *11*(6).

[39] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007, October). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement* (pp. 29-42). ACM.

[40] Nazir, A., Raza, S., & Chuah, C. N. (2008, October). Unveiling facebook: a measurement study of social network based applications. In *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement* (pp. 43-56). ACM.

[41] Nazir, A., Raza, S., Gupta, D., Chuah, C. N., & Krishnamurthy, B. (2009, November). Network level footprints of facebook applications. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement* (pp. 63-75). ACM.

[42] Cha, M., Mislove, A., & Gummadi, K. P. (2009, April). A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web* (pp. 721-730). ACM.

[43] TREES, D., & Rokach, L. Department of Industrial Engineering. *Tel-Aviv University, liorr@ eng. tau. ac. il, Oded Maimon, Department of Industrial Engineering, Tel-Aviv University, maimon@ eng. tau. ac. il*.

[44] Willassen, S. (2003). Forensics and the GSM mobile telephone system. *International Journal of Digital Evidence*, *2*(1), 1-17.

[45] Casadei, F., Savoldi, A., & Gubian, P. (2006). Forensics and SIM cards: an Overview. *International Journal of Digital Evidence*, *5*(1), 1-21.

[46] Kim, K., Hong, D., Chung, K., & Ryou, J. C. (2007, December). Data acquisition from cell phone using logical approach. In *Proceedings of world academy of science, engineering and*

technology (Vol. 26).

[47] Jansen, W., & Moenner, L. (2008, January). Overcoming impediments to cell phone forensics. In *hicss* (p. 484). IEEE.

[48] Willassen, S. (2005, February). Forensic analysis of mobile phone internal memory. In *IFIP International Conference on Digital Forensics* (pp. 191-204). Springer, Boston, MA.

[49] Al-Zarouni, M. (2007). Introduction to mobile phone flasher devices and considerations for their use in mobile phone forensics.

[50] Thing, V. L., Ng, K. Y., & Chang, E. C. (2010). Live memory forensics of mobile phones. *digital investigation*, *7*, S74-S82.

[51] Hoog, A., & Strzempka, K. (2011). *iPhone and iOS forensics: Investigation, analysis and mobile security for Apple iPhone, iPad and iOS devices*. Elsevier.

[52] Aarthi, Malathi. (2015). *A Safety Data Migration of Argus Application. International Journal of Advanced Research in Biology, Ecology, Science and Technology (IJARBEST)*.

[53] Devore, J. (2007). Making sense of data: A practical guide to exploratory data analysis and data mining.

[54] Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT Press.

[55] Xiong, H., Pandey, G., Steinbach, M., & Kumar, V. (2006). Enhancing data analysis with noise removal. *IEEE Transactions on Knowledge and Data Engineering*, *18*(3), 304-319.

[56] Bowman, A. W., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations* (Vol. 18). OUP Oxford.

[57] Refaat, M. (2010). *Data preparation for data mining using SAS*. Elsevier.

[58] Swati, K., & Kumar, S. (2015). A comparative study of various data transformation techniques in data mining. *International Journal of Scientific Engineering and Technology*, *4*, 146-148.

[59] Lavrac, N., Keravnou, E., & Zupan, B. (2000). Intelligent data analysis in medicine. *Encyclopedia of computer science and technology*, *42*(9), 113-157.

[60] Shahar, Y., Tu, S. W., Das, A. K., & Musen, M. A. (1992, July). A problem-solving architecture for managing temporal data and their abstractions. In *Workshop on Implementing Temporal Reasoning, AAAI* (Vol. 92, pp. 141-52).

[61] Russ, T. A. (1990). Using hindsight in medical decision making. *Computer Methods and Programs in Biomedicine*, *32*(1), 81-90.

[62] Westphal, C., & Blaxton, T. (1998). Data mining solutions: methods and tools for solving real-world problems.

[63] Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge university press.

[64] Vijayarani, S., & Sylviaa, M. S. M. (2016). Dimensionality Reduction-A Study. *International Journal of Engineering Applied Sciences and Technology*, *1*, 163-170.

[65] Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. Elsevier.

[66] Kharal, R. (2006). *Semidefinite embedding for the dimensionality reduction of DNA microarray data* (Master's thesis, University of Waterloo).

[67] Yu, L., Ye, J., & Liu, H. (2007). Dimensionality Reduction for Data Mining-Techniques, Applications and Trends. In *Proc. SIAM Int. Conf. Data Min. Proc.*

[68] Padraig.(2007). *Dimension Reduction. Padraig Cunningham University College Dublin Technical Report UCD-CSI2007-7.*

[69] Fodor, I. K. (2002). A survey of dimension reduction techniques. *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory*, *9*, 1-18.

[70] Jin, R., Breitbart, Y., & Muoh, C. (2009). Data discretization unification. *Knowledge and Information Systems*, *19*(1), 1.

[71] Bay, S. D. (2001). Multivariate discretization for set mining. *Knowledge and Information Systems*, *3*(4), 491-512.

[72] Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Machine Learning Proceedings 1995* (pp. 194-202).

[73] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

[74] Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining (adaptive computation and machine learning)* (pp. 361-452). Cambridge, MA: MIT press.

[75] Ritschard, G. (2013). CHAID and earlier supervised tree methods. In *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 70-96). Routledge.

[76] Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, *58*(302), 415-434.

[77] Cellard, J. C., Labbe, B., & Savitsky,G. (1967). Le programme Elisée. Présentation et application. *Revue Metra*, 3(6), 511-519.

[78] Messenger, R., & Mandell, L. (1972). A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American statistical association*, *67*(340), 768-772.

[79] Morgan, J. N., & Messenger, R. C. (1973). THAID, a sequential analysis program for the analysis of nominal scale dependent variables.

[80] Gillo, M. W. (1972). MAID, a Honeywell 600 program for an automatized survey analysis. *Behavioral Science*, *17*(2), 251.

[81] Gillo, M. W., & Shelly, M. W. (1974). Predictive modeling of multivariable and multivariate data. *Journal of the American Statistical Association*, *69*(347), 646-653.

[82] Loh, Wei-Yin. (2014). *Fifty years of classification and regression trees. International Statistical Review 82.3:* 329-348.

[83] Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied statistics*, 119-127.

[84] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

[85] Quinlan, J. R. (1993). C4. 5: Programming for machine learning. *Morgan Kauffmann*, *38*, 48.

[86] Breiman, L., Friedman, J., Olshen, R., Stone, C., (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

[87] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

[88] Christensen, R. (2006). *Log-linear models and logistic regression*. Springer Science & Business Media.

[89] Priddy, K. L., & Keller, P. E. (2005). *Artificial neural networks: an introduction* (Vol. 68). SPIE press.

[90] Burke, H. B. (1996). Statistical analysis of complex systems in biomedicine. In *Learning from Data* (pp. 251-258). Springer, New York, NY.

[91] Bain, A. (1873). *Mind and Body the Theories of Their Relation by Alexander Bain*. Henry S. King & Company.

[92] James, W. (1890). The Principles of. *Psychology*, *2*, 94.

[93] Werbos, Paul John. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences.* Doctoral Dissertation, Applied Mathematics, Harvard University, MA.

[94] Werbos, P. J., & Werbos, P. (1994). Beyond regression: New tools for prediction and analysis in the behavioral sciences. The roots of backpropagation.

[95] Rumelhart, H., & Hinton, G. E. (1995). Williams, 1986.“. *Back propagation Training Algorithm”, Developed at MIT*.

[96] Parker, D. B. (1982). Learning logic. Invention report S81-64, File 1, Office of Technology Licensing. *October, Stanford University*.

[97] McCulloch, Warren S., and Walter Pitts. (1943). *A logical calculus of the ideas immanent in nervous activity.* The bulletin of mathematical biophysics 5.4: 115-133.

[98] Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

[99] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175-185.

[100] Mandal, J. K. (Ed.). (2016). *Handbook of research on natural computing for optimization problems*. IGI Global.

[101] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46). New York: IBM.

[102] Kazmierska, J., & Malicki, J. (2008). Application of the Naïve Bayesian Classifier to optimize treatment decisions. *Radiotherapy and Oncology*, *86*(2), 211-216.

[103] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[104] Kavitha, K., Kangaiammal, A., & Satheesh, K. (2015). Analysis on Classification Techniques in

Mammographic Mass Data Set. *International Journal of Engineering Research and Applications 5 (7)*, 32-35.

[105] Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine learning*, *29*(2-3), 103-130.

[106] Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, *73*(16), 5261-5267.

[107] Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998, July). A Bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 workshop* (Vol. 62, pp. 98-105).

[108] Weisberg, S. (2005). *Applied linear regression* (Vol. 528). John Wiley & Sons.

[109] Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (Vol. 4, p. 318). Chicago: Irwin.

[110] Reinartz, T. (1999). *Focusing solutions for data mining: analytical studies and experimental results in real-world domains*. Springer-Verlag.

[111] Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29-39). Citeseer.

[112] Marbán, Ó., Mariscal, G., & Segovia, J. (2009). A data mining & knowledge discovery process model. In *Data Mining and Knowledge Discovery in Real Life Applications*. InTech.

[113] Reifer, D. J. (2006). Software Management. Wiley-IEEE Computer Society Press, 7th. Edition.

[114] Snow, R. W., Hay, S. I., & Marsh, K. (2006). Malaria in Africa: sources, risks, drivers and disease burden 2005–2030. *Foresight Project, Office of Science and Technology*.

[115] Martens, P., Kovats, R. S., Nijhof, S., De Vries, P., Livermore, M. T. J., Bradley, D. J., ... & McMichael, A. J. (1999). Climate change and future populations at risk of malaria. *Global environmental change*, *9*, S89-S107.

[116] Martens, W. J. (1998). *Health and climate change: modelling the impacts of global warming and ozone depletion*. Earthscan, London.

[117] IPCC (Intergovernmental Panel on Climate Change). (2001). Special report on emissions scenarios. *A special report of working group III of the intergovernmental panel on climate change*.

[118] Caminade, C., Kovats, S., Rocklov, J., Tompkins, A. M., Morse, A. P., Colón-González, F. J., ... & Lloyd, S. J. (2014). Impact of climate change on global malaria distribution. *Proceedings of the National Academy of Sciences*, *111*(9), 3286-3291.

[119] Jones, A. E. (2007). *Seasonal ensemble prediction of malaria in Africa* (Doctoral dissertation, University of Liverpool).

[120] Craig, M. H., Snow, R. W., & le Sueur, D. (1999). A climate-based distribution model of malaria

transmission in sub-Saharan Africa. *Parasitology today*, *15*(3), 105-111.

[121] Van Lieshout, M., Kovats, R. S., Livermore, M. T. J., & Martens, P. (2004). Climate change and malaria: analysis of the SRES climate and socio-economic scenarios. *Global Environmental Change*, *14*(1), 87-99.

[122] Tompkins, A. M., & Ermert, V. (2013). A regional-scale, high resolution dynamical malaria model that accounts for population density, climate and surface hydrology. *Malaria journal*, *12*(1), 65.

[123] Béguin, A., Hales, S., Rocklöv, J., Åström, C., Louis, V. R., & Sauerborn, R. (2011). The opposing effects of climate change and socio-economic development on the global distribution of malaria. *Global Environmental Change*, *21*(4), 1209-1214.

[124] Lunde, T. M., Bayoh, M. N., & Lindtjørn, B. (2013). How malaria models relate temperature to malaria transmission. *Parasites & vectors*, *6*(1), 20.

[125] Ermert, V., Fink, A. H., Jones, A. E., & Morse, A. P. (2011). Development of a new version of the Liverpool Malaria Model. I. Refining the parameter settings and mathematical formulation of basic processes based on a literature review. *Malaria journal*, *10*(1), 35.

[126] Martens, W. J. M., Jetten, T. H., Rotmans, J., & Niessen, L. W. (1995). Climate change and vector-borne diseases: a global modelling perspective. *Global environmental change*, *5*(3), 195-209.

[127] Martens, W. J., Niessen, L. W., Rotmans, J., Jetten, T. H., & McMichael, A. J. (1995). Potential impact of global climate change on malaria risk. *Environmental health perspectives*, *103*(5), 458.

[128] Martens W. (1997): *Health impacts of climate change and ozone depletion: an eco-epidemiological modelling approach*(Doctoral dissertation, Maastricht University).

[129] Bayoh, M. N. (2001). *Studies on the development and survival of Anopheles gambiae sensu stricto at various temperatures and relative humidities* (Doctoral dissertation, Durham University).

[130] Parham, P. E., Pople, D., Christiansen-Jucht, C., Lindsay, S., Hinsley, W., & Michael, E. (2012). Modeling the role of environmental variables on the population dynamics of the malaria vector Anopheles gambiae sensu stricto. *Malaria Journal*, *11*(1), 271.

[131] Mordecai, E. A., Paaijmans, K. P., Johnson, L. R., Balzer, C., Ben-Horin, T., de Moor, E., ... & Lafferty, K. D. (2013). Optimal temperature for malaria transmission is dramatically lower than previously predicted. *Ecology letters*, *16*(1), 22-30.

[132] Lunde, T. M., Korecha, D., Loha, E., Sorteberg, A., & Lindtjørn, B. (2013). A dynamic model of some malaria-transmitting anopheline mosquitoes of the Afrotropical region. I. Model description and sensitivity analysis. *Malaria Journal*, *12*(1), 28.

[133] Afrane, Y. A., Zhou, G., Lawson, B. W., Githeko, A. K., & Yan, G. (2006). Effects of microclimatic changes caused by deforestation on the survivorship and reproductive fitness of

Anopheles gambiae in western Kenya highlands. *The American journal of tropical medicine and hygiene*, *74*(5), 772-778.

[134] Harrington, L. C., Jones, J. J., Kitthawee, S., Sithiprasasna, R., Edman, J. D., & Scott, T. W. (2008). Age-dependent survival of the dengue vector Aedes aegypti (Diptera: Culicidae) demonstrated by simultaneous release–recapture of different age cohorts. *Journal of medical entomology*, *45*(2), 307-313.

[135] Béguin, A., Hales, S., Rocklöv, J., Åström, C., Louis, V. R., & Sauerborn, R. (2011). The opposing effects of climate change and socio-economic development on the global distribution of malaria. *Global Environmental Change*, *21*(4), 1209-1214.

[136] Kleinschmidt, I., Omumbo, J., Briet, O., Van De Giesen, N., Sogoba, N., Mensah, N. K., ... & Teuscher, T. (2001). An empirical malaria distribution map for West Africa. *Tropical Medicine & International Health*, *6*(10), 779-786.

[137] Rahman, A., Kogan, F., Roytman, L., Goldberg, M., & Guo, W. (2011). Modelling and prediction of malaria vector distribution in Bangladesh from remote-sensing data. *International journal of remote sensing*, *32*(5), 1233-1251.

[138] Huang, F., Zhou, S., Zhang, S., Wang, H., & Tang, L. (2011). Temporal correlation analysis between malaria and meteorological factors in Motuo County, Tibet. *Malaria Journal*, *10*(1), 54.

[139] Tesfaye, S., Belyhun, Y., Teklu, T., Medhin, G., Mengesha, T., & Petros, B. (2012). Malaria pattern observed in the highland fringe of Butajira, Southern Ethiopia: a ten-year retrospective analysis from parasitological and metrological data. *Malaria World Journal*, *3*(5), 1-8.

[140] Ayele, D. G. (2013). *Use of Statistical Modelling and Analyses of Malaria Rapid Diagnostic Test Outcome in Ethiopia*(Doctoral dissertation, University of KwaZulu-Natal, Pietermaritzburg).

[141] Loha, E., & Lindtjørn, B. (2012). Predictors of Plasmodium falciparum malaria incidence in Chano Mille, South Ethiopia: a longitudinal study. *The American journal of tropical medicine and hygiene*, *87*(3), 450-459.

[142] Bizimana, J. P. Malaria hotspots in Rwanda-relative influence of climate variability and interventions.

[143] Feltelius, V., & Elleby, R. (2014). Habitat characterization for malaria vector mosquito larvae in Gamo Gofa, Ethiopia.

[144] Dessalegne, E., Suryabhagavan, K. V., & Balakrishnan, M. (2016). Malaria-risk assessment using geographical information system and remote sensing in Mecha district, West Gojjam, Ethiopia. *J Geomat*, *10*, 55-64.

[145] UNDP Emergency Unit, UNDP (2000). *Ethiopia: Region, Zone and Woreda Population and Relief Beneficiary data for 1999 and 2000*. Addis Ababa, Ethiopia.

[146] World Bank Country Office, World Bank. (2001) *World Population Prospects–the 2000 Revision*. Ethiopia

[147] Ceri, S., Fraternali, P., Bongio, A., Brambilla, M., Comai, S., & Matera, M. (2003). *Morgan*

Kaufmann series in data management systems: Designing data-intensive Web applications. Morgan Kaufmann.

[148] Goldberg, Y., & Elhadad, M. (2008, June). splitSVM: fast, space-efficient, non-heuristic, polynomial kernel computation for NLP applications. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers* (pp. 237-240). Association for Computational Linguistics.

[149] Shashua, A. (2009). Introduction to machine learning: Class notes 67577. *arXiv preprint arXiv:0904.3664*.

[150] Larose, D. T., & Larose, C. D. (2015). *Data mining and predictive analytics*. John Wiley & Sons.

Annex A: Pseudo Code of Weather Dataset Conversion

Input: weatherDataList WDL;
A value of (Region, Zone, Elevation, Longitude, Latitude, Element, year, January, February, March, April, May, June, July, August, September, October, November, December)
Output: newWeatherDataList NWDL;
A value of (Region, Zone, Elevation, Longitude, Latitude, Rainfall, Humidity, MinTemp, MaxTemp, Date (month/year))

```
locationYearList←select distinct region, zone, Year from WDL
yearList ← select distinct year from WDL
newWeatherDataObject ←null
FOR i=0, i<locationYearList.size, i++)
        Counter ←0
        initialize newWeatherDataObject

            FOR (k=0, k<WDL.size, K++)
               if (WDL(K).region=locationYearList(i).Region AND WDL(K).zone= locationYearList
               (i).zone   AND WDL(K).Year= locationYearList (i).Year)
               if(Counter==0) then

                        newWeatherDataObject.Region=WDL(K).Region
                        newWeatherDataObject.Zone=WDL(K).Zone
                        newWeatherDataObject.Longitude=WDL(K).Longitude
                        newWeatherDataObject.Latitude=WDL(K).Latitude
                        newWeatherDataObject.valueOf(WDL(K).Element )=WDL(K).January
                        newWeatherDataObject.Date=Date(January, WDL(K).Year)
                        NWDL.add(newWeatherDataObject)
                        newWeatherDataObject. valueOf(WDL(K).Element )=WDL(K).February
                        newWeatherDataObject.Date=Date(February, WDL(K).Year)
                        NWDL.add(newWeatherDataObject)
                        newWeatherDataObject. valueOf(WDL(K).Element )=WDL(K).March
                        newWeatherDataObject.Date=Date(March, WDL(K).Year)
                        NWDL.add(newWeatherDataObject)
                        newWeatherDataObject. valueOf(WDL(K).Element )=WDL(K).April
                        newWeatherDataObject.Date=Date(April, WDL(K).Year)
                        NWDL.add(newWeatherDataObject)
                        newWeatherDataObject. valueOf(WDL(K).Element )=WDL(K).May
                        newWeatherDataObject.Date=Date(May, WDL(K).Year)
                        NWDL.add(newWeatherDataObject)
                        newWeatherDataObject. valueOf(WDL(K).Element )=WDL(K).June
                        newWeatherDataObject.Date=Date(June, WDL(K).Year)
                        NWDL.add(newWeatherDataObject)
                        newWeatherDataObject. valueOf(WDL(K).Element )=WDL(K).July
                        newWeatherDataObject.Date=Date(July, WDL(K).Year)
                        NWDL.add(newWeatherDataObject)
                        newWeatherDataObject. valueOf(WDL(K).Element )=WDL(K).August
                        newWeatherDataObject.Date=Date(August, WDL(K).Year)
                        NWDL.add(newWeatherDataObject)
                        newWeatherDataObject. valueOf(WDL(K).Element )=WDL(K).September
                        newWeatherDataObject.Date=Date(September, WDL(K).Year)
                        NWDL.add(newWeatherDataObject)
                        newWeatherDataObject. valueOf(WDL(K).Element )=WDL(K).October
                        newWeatherDataObject.Date=Date(October, WDL(K).Year)
                        NWDL.add(newWeatherDataObject)
                        newWeatherDataObject. valueOf(WDL(K).Element )=WDL(K).November
```

```
                    newWeatherDataObject.Date=Date(November, WDL(K).Year)
                    NWDL.add(newWeatherDataObject)
                    newWeatherDataObject. valueOf(WDL(K).Element )=WDL(K).December
                    newWeatherDataObject.Date=Date(December, WDL(K).Year)
                    NWDL.add(newWeatherDataObject)
        if(else<>0) then
            FOR (n=0, n<NWDL.size, n++)
                    If(NWDL(n).Date== Date(January, WDL(K).Year)
                            NWDL(n). valueOf(WDL(K).Element )=WDL(K).January
                    else if(NWDL(n).Date== Date(February, WDL(K).Year)
                            NWDL(n). valueOf(WDL(K).Element )=WDL(K). February
                    else if(NWDL(n).Date== Date(March, WDL(K).Year)
                            NWDL(n). valueOf(WDL(K).Element )=WDL(K).March
                    else if(NWDL(n).Date== Date(April, WDL(K).Year)
                            NWDL(n). valueOf(WDL(K).Element )=WDL(K).April
                    else if(NWDL(n).Date== Date(May, WDL(K).Year)
                            NWDL(n). valueOf(WDL(K).Element )=WDL(K).May
                    else if(NWDL(n).Date== Date(June, WDL(K).Year)
                            NWDL(n). valueOf(WDL(K).Element )=WDL(K).June
                    else if(NWDL(n).Date== Date(July, WDL(K).Year)
                            NWDL(n). valueOf(WDL(K).Element )=WDL(K).July
                    else if(NWDL(n).Date== Date(August, WDL(K).Year)
                            NWDL(n). valueOf(WDL(K).Element )=WDL(K).August
                    else if(NWDL(n).Date== Date(September, WDL(K).Year)
                            NWDL(n). valueOf(WDL(K).Element )=WDL(K).September
                    else if(NWDL(n).Date== Date(October, WDL(K).Year)
                            NWDL(n). valueOf(WDL(K).Element )=WDL(K).October
                    else if(NWDL(n).Date== Date(November, WDL(K).Year)
                            NWDL(n). valueOf(WDL(K).Element )=WDL(K).November
                        END
                      END
                   END
                 END
              END
              FOR (i=0, J<NWDL.size, J++)
              save(NWDL(j)) //insert into the new weather data table
              END
              Counter ++
              NWDL <--null
        END
        END
```

Annex B: Pseudo Code of Malaria Dataset Conversion

Input: malariaDataList MDL;
A value of (Region, Zone, Woreda, Year, Month, EpidemicWeek, TotalMalariaConfirmedClinical)

Output: newMalariaDataList NMDL;
A value of (Region, Zone, Woreda, Year, Month, TotalMalariaConfirmedClinical)

```
locationDateList←select Region, Zone, Woreda, Year, Month from MDL
newWeatherDataObject NMDO←null
FOR i=0, i<locationDateList.size, i++)
        Counter ←0
        initialize NMDO;
```

```
                FOR (k=0, k<MDL.size, K++)
                    if (MDL(k).Region=locationDateList(i).Region AND MDL(k).Zone =
                    locationDateList(i).Zone   AND MDL(k).Woreda = locationYearList (i).Woreda AND
                    MDL(k).Year = locationDateList(i).YearAND MDL(k).Month =
                    locationDateList(i).Month)
                            if(Counter==0) then
                                    NMDO.Region =MDL(k).Region
                                    NMDO.Zone =MDL(k).Zone
                                    NMDO.Woreda =MDL(k).Woreda
                                    NMDO.Year =MDL(k).Year
                                    NMDO.Month =MDL(k).Month
                                    NMDO.TotalMalariaConfirmedClinical=NMDO.TotalMalariaConfirm
                                    edClinical +MDLO.TotalMalariaConfirmedClinical
                            elseif(Counter<>0) then
                                    NMDL(i).TotalMalariaConfirmedClinical=NMDL(i).TotalMalariaConf
                                    irmedClinical +MDL(K).TotalMalariaConfirmedClinical

                            END
                            NMDL.add(NMDO)

                    END
            END
        END
    savebatch(NMDL ) //insert into the new malaria data table
```

Annex C: Pseudo Code of Population Dataset Conversion

```
Input:      populationDataListPDL;
        A value of (Region, Zone, Woreda, TotalPopulation)
Output:  newPopulationDataList PMDL;
        A value of (Region, Zone, Woreda, Year, Month, TotalPopulation)
locationDateList←select Region, Zone, Woreda from PDL
newPopulationDataObjectNPDO←null
FOR i=0, i<locationDateList.size, i++)
        tempPopulation←0
        Counter ←0
        censusYear←1994


          FOR (k=0, k<PDL.size, K++)
                if (PDL(k).Region=locationDateList(i).Region AND PDL(k).Zone = locationDateList(i).Zone
                AND PDL(k).Woreda = locationYearList (i).Woreda
                  When censusYear<=2015 Do

            if(Counter2==0) then

                    FOR(int j=12;i<=1;j--)
                    initialize NPDO;
                    if(Counter2==0) then
                            if(Counter==0) then
                                    NPDO.Region =MDL(k).Region
                                    NPDO.Zone =MDL(k).Zone
                                    NPDO.Woreda =MDL(k).Woreda
```

```
                                    NPDO.Year =censusYear
                                    NPDO.Month =j
                                    NPDO.TotalPopulation =MDL(k).TotalPopulation
                                    tempPopulation=NPDO.TotalPopulation-
                                    NPDO.TotalPopulation*0.0025
                            elseif(Counter<>0) then
                                    NPDO.Region =MDL(k).Region
                                    NPDO.Zone =MDL(k).Zone
                                    NPDO.Woreda =MDL(k).Woreda
                                    NPDO.Year =censusYear
                                    NPDO.Month =j
                                    NPDO.TotalPopulation =tempPopulation
                                    tempPopulation=NPDO.TotalPopulation-
                                    NPDO.TotalPopulation*0.0025

                                END
                        else if(Counter2<>0)

                        FOR(int n=1;n<=12;n++)

                                    NPDO.Region =MDL(n).Region
                                    NPDO.Zone =MDL(n).Zone
                                    NPDO.Woreda =MDL(n).Woreda
                                    NPDO.Year =censusYear
                                    NPDO.Month =n
                                    NPDO.TotalPopulation =tempPopulation+tempPopulation*0.0025
                                    tempPopulation=NPDO.TotalPopulation

                                END
                        Counter++
                        NPDL.add(NPDO)
                     END
                    Counter2++
                censusYear++
                    tempPopulation=NPDO.TotalPopulation
                    END

END
savebatch(NPDL) //insert into the new malaria data table
```

Annex D: Pseudo Code to Merge Malaria, Weather, NDVI and Population Datasets

Input: newMalariaDataList NMDL;
A value of (Region, Zone, Woreda, Year, Month, TotalMalariaConfirmedClinical)
newWeatherDataList NWDL;
A value of (Region, Zone, Elevation, Longitude, Latitude, Rainfall, Humidity, MinTemp, MaxTemp, Date (month/year))
newPopulationDataListNPDL;
A value of (Region, Zone, Woreda, Year, Month, TotalPopulation)
ndviDataListNDL;
A value of (Region, Zone, Longitude, Latitude, Year, Month, ndvi)
Output: malariaDispersionDataListMDDL;

```
        A value of (Region, Zone, Woreda, Year, Month, TotalMalariaConfirmedClinical, Elevation,
        Longitude, Latitude, Rainfall, Humidity, MinTemp, MaxTemp, TotalPopulation,ndvi)
locationDateList←select Region, Zone, Woreda, Year, Month from NMDL
malariaDispersionDataObject MDDO←null
FOR i=0, i<NMDL.size, i++)
        initialize NMDO;

            FOR (k=0, k<NWDL.size, K++)
               if (NWDL(k).Region=NMDL(i).Region AND NWDL(k).Zone = NMDL(i).Zone AND
               NWDL(k).Year = NMDL(i).YearAND NWDL(k).Month = NMDL(i).Month)

                        FOR (n=0, n<NPDL.size, n++)
                                if (NWDL(k).Region=NPDL(n).Region AND NWDL(k).Zone =
                                NPDL(n).Zone AND NWDL(k).Year = NPDL(n).YearAND
                                NWDL(k).Month = NPDL(n).Month)

                                        FOR (m=0, m<NDL.size, m++)
                                                if (NDL(m).Region=NPDL(n).Region AND
                                                NDL(m).Zone = NPDL(n).Zone AND NDL(m).Year
                                                = NPDL(n).YearAND NDL(m).Month =
                                                NPDL(n).Month)
                                                        MDDO.Region=NPDL(n).Region
                                                        MDDO.Zone =NPDL(n).Zone
                                                        MDDO.Woreda = NPDL(n).Woreda
                                                        MDDO.Year=NPDL(n).Year
                                                        MDDO.Month=NPDL(n).Month
                                                        MDDO.TotalPopulation=NPDL(n).TotalPo
                                                pulation

                                                END
                                        END
                                MDDO.Elevation=NWDL(k).Elevation
                                MDDO.Longitude =NWDL(k).Longitude
                                MDDO.Latitude =NWDL(k).Latitude
                                MDDO.Rainfall=NWDL(k).Rainfall
                                MDDO.Humidity=NWDL(k).Humidity
                                MDDO.MinTemp=NWDL(k).MinTemp
                                MDDO.MaxTemp=NWDL(k).MaxTemp
                                MDDO.Year=Year(NWDL(k).Date)
                                MDDO.Month=Month(NWDL(k).Date)
                                END
                        END
                MDDO.TotalMalariaConfirmedClinical =NMDL(i).TotalMalariaConfirmedClinical
                    END
                MDDL.add(MDDO)
        END
        savebatch(MDDL) //insert into the malariadispersion data table
```

I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in any other university, and that all source of materials used for the thesis have been duly acknowledged.

Declared by:

Name: **Abreham Kassahun Shiferaw**

Signature: _____

Date: _____

Confirmed by advisor:

Name: **Dr. Solomon Atnafu**

Signature: _____

Date: _____