Africa's Malaria Epidemic Predictor: Application of Machine Learning on Malaria Incidence and Climate Data

Muthoni Masinde
Centre for Sustainable Smart Cities
Central University of Technology, Free State
Bloemfontein, South Africa
+27 501 3091
muthonimasinde@gmail.com

ABSTRACT

The 2019 World Malaria Report confirms that Africa continue to bear the burden of malaria morbidity. The continent accounted for over 93% of the global malaria incidence reported in 2018. Despite the numerous multi-level and consultative efforts to combat this epidemic, malaria continues to claim thousands of human lives, especially those of children under 5 years of age. Since malaria is preventable and treatable, one of the solutions towards reducing the number of deaths is by implementing an effective malaria outbreak early warning system that can forecast malaria incidence long before they occur. This way, policymakers can put mitigation measures in place. Tapping into the success of machine learning algorithms in predicting disease outbreaks, we present a malaria outbreak prediction system that is anchored on the well-established correlation between certain climatic conditions and breeding environment of the malaria carrying vector (mosquito). Historical datasets on climate and malaria incidence are used to train nine machine learning algorithms and four best performing ones identified based on classification accuracy and computation performance. Preceding the models' development, reliability and correlation analysis was carried out on the data; this was then followed by reduction of the dimensionality of the feature space of the two datasets. Given the power of deep learning in handling selectivity variance, the malaria predictor system was developed based on the deep learning algorithm. Further, the evaluation of the system was done using the Simulator function in RapidMiner and the accuracy of the predictions assessed using an independent dataset that was not used in the models' development. With prediction accuracy of up to 99%, this system has the potential in contributing to the fight against malaria epidemic in Africa and elsewhere in the world.

CCS Concepts

• Theory of computation→Theory and algorithms for application domains.• Applied computing→Health informatics.• Information Systems→Information systems applications→Data mining.

ICCDA 2020, March 9–12, 2020, Silicon Valley, CA, USA

ACM ISBN 978-1-4503-7644-0/20/03…$15.00
https://doi.org/10.1145/3388142.3388158

Keywords

Malaria incidence, machine learning, deep learning, artificial neural networks, disease decision support systems, Sub-Saharan Africa, RapidMiner.

1. INTRODUCTION

Machine learning is everywhere; in some cases, the users are unaware that they are using it. It entails data mining with view to extract pattens[1], [2]. Application of machine learning spans domains such as face recognition, speech recognition, disease predication, self-driving cars, web search and anomaly detection, among many others[1], [3]–[6]. Among the various machine learning techniques applicable in the medical domain, artificial neural networks (ANNs) are powerful due to the black box and versatile learner concepts. On the other hand, deep learning (based on neural networks) is driving the current exponential growth in machine learning. This is mainly due to its ability for intuitive decision-making ability. Deep learning entails learning high level abstractions in data. Deep learning addresses ANNs' inability to handle selectivity variance [7]. Other relevant machine learning techniques are decision trees, rough set theory and decision tables [1], [7]. Factors that determine the superiority of the techniques in their ability for knowledge discovery is accuracy, preciseness and reliability [6].

Disease decision support systems assist in diagnosis, prediction, classification and risk forecasting. One of the challenges in handling medical datasets is imbalance, conflict, incompleteness and vagueness [6]. The latter could explain why most of the existing solution tend to focus on assisting the diagnosis and classification [8] and only a few on the prediction of future occurrence [9]. Some of the tools available for supporting machine learning are MATLAB [10], RapidMiner [11] and SPSS [12]. In all these tools, some of the measures used to assess the performances of the models: Relative Error, Square Error, Root Mean Square Error, Correlation and Receiver Operating Characteristics.

Malaria is a vector-borne disease; it is transmitted through the bite of an infected Anopheles mosquito. The World Health Organization [13] classifies malaria under communicable diseases along the likes of cholera, influenzas and tuberculosis. Links between certain climatic conditions and malaria incidence have been established and are related to creation of conducive environment for the development (to maturity), biting capacity and survival periods of the Anopheles[14]. Three climatic factors associated with this environment are: temperature, humidity, rainfall. While temperatures ranging between 20 and 30^0C provide conducive environment to the vector, temperatures below 16^0C and higher than 30^0C curtails their ability to thrive. On the other

hand, the link between the vector and amount of rainfall is not straightforward. However, Rahman et al (2010) reported that over 50mm of monthly rainfall could increase the mosquitoes' development while a combination of very high rainfall and hot temperatures reduces the chances. This was later confirmed by [15]. Directly related to temperature and rainfall, is humidity - less than 60% is said to minimizes malaria incidences. There is also proven link between the Global Vegetation Index and Anopheles mosquito ability to pass on the malaria causing protozoa plasmodium parasite[16]. Figure 1 below, illustrates the connection between weather conditions, malaria, and transmission environment.

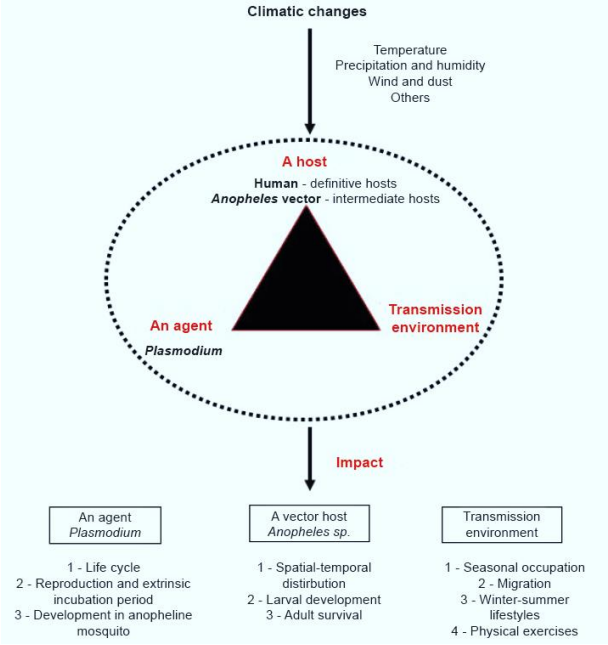

Figure 1. Effects of climate on malaria's transmission environment (Source: [14])

In the 2019 World Malaria Report [13], 93% (213 million of the 228 million) of the malaria incidences and 94% of deaths reported in 2018 occurred in Africa. Further, 19 countries in Sub-Saharan Africa and India accounted for the 85% of the cases. Besides, over 50% of the cases occurred in the following 6 countries: Nigeria (25%), Democratic Republic of Congo (12%), Uganda (5%), Cote d' Ivoire (4%), Mozambique (4%) and Niger (4%). These figures are mostly directly linked to the conducive (to the Anopheles mosquito) climate in the said countries [17]. Besides, the limited nature of medical resources in these countries make the prevalence and impacts of the malaria more profound [9] as illustrated in figure 2 below.

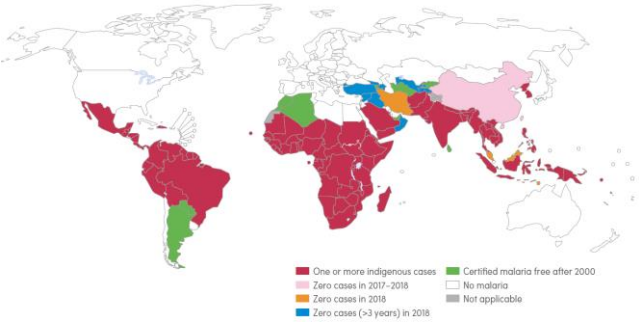

Figure 2. Incidence of malaria for 2018 (Source:[11])

The main outcome of the World Health Organization's Global Technical Strategy for Malaria 2016-2030 [17] is to reduce the global malaria burden by 90% by 2030. The strategy yielded impressive results between 2010 and 2013. However, the rate of reduction took a dive from 2014 and even showed some increases in some African countries[13]. On the other hand, it is now apparent that climate change will have significant impact on human infectious diseases such as malaria and this could partly explain the stagnation of the above Strategy. Research has shown that, by 2050, the climate change being witnessed currently has the potential of doubling the cause probability of malaria and expand geographical coverage (of areas vulnerable to malaria) to more areas beyond the Sub-Saharan Africa [14]. In [18], recommendations for employing an array of adaptation measures for countering this reality are made. These is anchored on the fact that human beings are active agents with the potential to reduce the resulting vulnerabilities. The three adaptation recommendations made (in [18]) are related to risk management strategies and surveillance enhancements proposals in [14].

There are numerous attempts towards the development of formal models for predicting malaria based on climatic conditions. A lot of this is taking place under the Intergovernmental Panel on Climate Change's (IPCC) findings[19]. A related (to this paper) example is an analysis of the connection between vegetation and weather patterns to incidences of malaria in Bangladesh [16]. In doing so, the authors considered, vegetation health and condition indices such as Advanced Very High Radiometer. In [20], Poisson regression was applied in establishing that the link between malaria and weather was due to the latter's influence on mosquito and parasite life cycle. A comprehensive list of such studies is presented in [14]. It is worth noting that other non-climate factors such as ecological and socioeconomic (such as the capacity of the healthcare system) play a role in determining the final impacts of malaria epidemic. Furthermore, most of these existing models are based on mathematical modeling [15], [20]–[22] and very few are based on machine learning.

Given the above gap analysis, the following research objectives (in the context of Africa) are pursued in this paper:

1) To make use of machine learning in developing models that go beyond empirical observations of the association between weather conditions and malaria incidence and develop more scientific explanations to this association.

2) To create a malaria early warning system that is effective and relevant to the Sub-Saharan African countries. The rationale for this is the fact that most Sub-Saharan African countries are unable to implement globally available adaptation measures due to high levels of poverty, and economic insatiability [9].

2. APPLICATION OF MACHINE LEARNING IN PREDICTING DISEASE OUTBREAKS

2.1 Related Literature

The underlying paradigm of machine learning is Artificial intelligence (AI). AI is a broader term that refers to the computing paradigm that aims to mimic human cognitive functions. It solves complex problems by developing computer systems with the capacity to perform tasks that require human intelligence such as learning and visual perception [2]. Different AI techniques are normally applied to analyze and model vast amounts of relevant

data which is then used to learn what conditions are most likely to result in disease outbreak [16].

Machine learning has been hailed for its power in predicting infectious disease. The technique has been deployed in health informatics for early detection, diagnosis and treatment of diseases [23]. For example, Vembandasamy et al [24], proposed Naive Bayes classifier for prediction of heart disease. On the other hand, Guo and Liu et al., [25] developed a machine learning model that took into consideration the climate factors to predict the outbreak of dengue in China.

Some of the popular techniques for machine learning algorithms include artificial neural networks (ANNs) and deep learning. ANNs simulate the structure and capabilities of the human brain which is made up of a network of neurons. ANNs mimic this feature of human brains by getting computers to work as the interconnected networks of neurons, learn and make decisions in the manner similar to humans. ANN, in its simplest form has three layers (input layer, hidden layer, and output layer). Rather than feeding the computers with data and teaching them how to do everything like it was in the initial AI approaches, machine learning provides the computers with data and allows them to learn by themselves from that data. Most learning machines use reinforcement learning and deep learning [2]. Deep learning is the new era of machine learning with the capabilities of adding more hidden layers to neural networking. This allows exploring complex non-linear patterns in the data [23]. This feature of deep learning makes computers not only capable of processing and learning from data but also to train themselves to process and learn from data [2]. Consequently, this makes deep learning suitable for prediction of certain disease outbreaks that are associated with climate factors.

2.2 Evaluating machine learning algorithms

Of most importance in machine learning is the accuracy of the algorithm – this generally refers to total number of correctly classified segments divided by the total number of test segments [26]. The most commonly used metrics for evaluating machine learning algorithms are classification accuracy and computation performance. Classification accuracy consists of three dimensions: accuracy, precision and recall [26]–[28]. In this case, accuracy refers to the percentage of correctly classified instances over the total number of instances while the number of class members is an attribute of precision. Other commonly used accuracy measures are sensitivity (true positive rate), specificity (true negative rate) and overall accuracy [26]. On the other hand, recall, also known as true positive rate, is the number of class members classified correctly. One way of ensuring assessing the relative performance and consistency of the algorithms is by averaging the measures under similar conditions [26].

Although in the era of super- and cloud computing, the computation performance metric could be relatively irrelevant, the reality of the elusive size of big data warrants consideration of machine algorithms' computation resources' requirements. This is amidst a mind-boggling figure of 2.5 quintillion bytes that has been estimated as the quantity of data being generated every day. Under such circumstances, the algorithm's performance in terms of building and classification speeds become important especially in live-threatening situations such as malaria outbreaks. Existing literature relating to the evaluation of the performance of machine learning algorithm is specific to application domain such as for urban patterns [26], IP traffic flow classification [28] and search problems [29]. No relevant application was found for Africa's malaria incidence.

2.3 Machine Learning Software Tools

Some of the most commonly used machine learning software tools include MATLAB, SPSS, Scikit learn, PyTorch, TensorFlow, Weka, KNIME, Colab, Apache Mohout, Accors.Net, Shogun, Keras.io and RapidMiner. Majority of these tools are freely accessible. An exhaustive catalogue of them can be found in [30]. Further, most of these tools provide interfaces for performing all the basic steps of machine learning, such as gathering data, preparation of the data, choosing a model, training, evaluation, hyper parameter tuning and prediction.

RapidMiner is both a commercial and free software tool. RapidMiner's maker provides a Community Edition of the software. Apart from providing a very user-friendly Graphical User Interface (GUI) for carrying out all the basic steps above, it provides very intuitive visualization tools and deployable prediction simulator for deploying ready-to-use applications[31].

MATLAB is an elaborate commercial software that provides high-performance language for technical computing. Among the many Toolboxes of MATLAB is the rich Deep Learning Toolbox that provides simple MATLAB commands and a GUI for creating and interconnecting the layers of a deep neural network. Some of the functions of this Toolbox are fetching and preprocessing data, feature extraction, training and evaluation. The well-established MATLAB environment allows this ToolBox to support advanced features such deep learning with big data on GPUs and in parallel [10].

The commercial Statistical Package Social Sciences (SPSS) comes with a few tool that can support machine learning [32]. Two of these are: (1) the neural networks that supports Multilayer Perceptron and Radial Basis Function; and (2) Classify function that supports seven classification algorithms.

3. DATA AND METHODS

3.1 Datasets

Two datasets were used in this research: one, related to historical incidences of malaria and two, historical weather data consisting of rainfall and temperature. Annual incidence of malaria (measured by the number of malaria cases per 1000 population at risk per year) reported between 2000 and 2017 was retrieved from the World Health Organization (WHO) data repository (http://apps.who.int/gho/data/node.gswcah). In this context, WHO defines population at risk as population living in areas where malaria transmission occurs. The data contained records for Malaria Incidence (MI) and Maternal Mortality Ratio (MR) from 107 countries for the 18 years under consideration. Apart from County Name (CN), the fourth data dimension was WHO Region (WR) (e.g. Africa, Europe and Eastern Mediterranean).The second is a global dataset consisting of total monthly rainfall and monthly mean temperatures for years 1901 to 2016. This was extracted from the World Bank's climate knowledge portal (https://climateknowledgeportal.worldbank.org/download-data).
In total, 272,832 rows of data (monthly Rainfall, monthly mean temperature, altitude and longitude) representing data for all countries was used.

3.2 Reduction of the dimensionality of the feature space

3.2.1 Conventional data exploration

Before selecting the most significant features to use for the machine learning process, conventional data exploration tools were applied. First, principle component analysis (PCA) [32], [33]

of the aspects constituting climate, geographical location and malaria incidence (and deaths) was performed. For each of the two datasets, data reliability was determined using Cronbach's Alpha test based on standardized items[32]. To further ascertain the coherencies of a dataset, standard deviations as well as inter-items covariance/correlation matrices were also considered. Finally, ANOVA with Friedman's[34] reliability tests were performed. While two factors (Mean Annual Temperature and Latitude) had significance of 1, geographical location (classified under WHO Region) had a significance of 0.371. All the other factors had a significance of 0. The results for Reliability Tests were as follows: Kaiser-Meyer-Olkin Measure of Sampling Adequacy = 0.695 and Bartlett's Test of Sphericity, measured by Approx. Chi-Square is 675.084. further, the Kendall's coefficient of concordance W was 0.883.

In order to combine the monthly-based climate with the annual-based malaria datasets, the annual total rainfall and mean annual temperatures were computed in SPSS. This resulted in the final 7-dimensional space dataset made up of: Year, WHO Region, Country Name, Maternal Mortality Ratio, Malaria Incidence, Latitude, Longitude, Annual Rainfall and Mean Annual Temperature. Further, the malaria incidence for years 2000 and 2005 were omitted from the dataset due the non-continuousness (data for the years in between was missing). Data for the two years was later used in the evaluation of the accuracy of the proposed malaria predictor.

3.2.2 Dimension Reduction

As stated above, Principal Component Analysis (PCA) was applied on six (excluding Maternal Mortality Ratio, Year and Country Name) using SPSS. This resulted in two components shown in table 1 and figure 2 below. Component 1 and 3 accounted for 30.16% and 29.35% (total of 61% for both) of the variance.

Table 1. Rotated Component matrix for datasets

	Component	
	1	**2**
WHOregion	**0,918**	-0,105
MalariaIncidence	-0,095	**0,624**
Latitude	-0,335	-0,647
Longitude	0,194	-0,545
AnnualRainfall	**0,884**	0,099
MeanAnnualTemperature	0,161	**0,797**

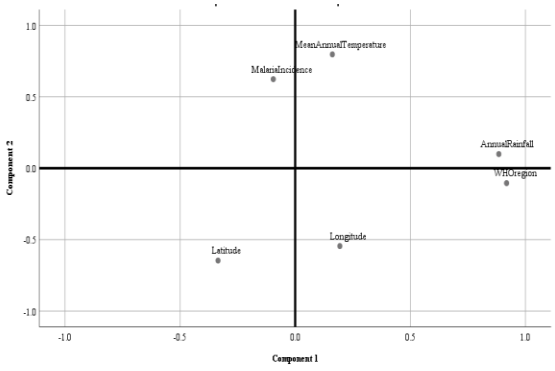

Figure 3. Datasets component plot in rotated space

Further dimension reduction exercise was carried out using the *k-means* clustering function of RapidMiner. Similar to the results of PCA, two clusters were identified and strongly correlated to the four dimensions similar to the ones in table 1 above.

3.3 Machine Learning

From the results of data pre-processing and feature reduction processes presented above, and in line with the statistics presented in the 2019 World Malaria Report [13], only data for the African countries was extracted and used in developing the machine learning algorithms. While RapidMiner was used in developing all the nine models, MATLAB's Deep Learning Toolbox was used to develop the Deep Learning model. The first run of RapidMiner provided the following weights

Table 2. Preliminary machine learning results - weighting of dataset dimensions

Attribute	Weight
WHOregion = Africa	0,6681
MaternalMortalityRatio	0,6127
MeanAnnualTemperature	0,3303
WHOregion = Americas	0,2990
WHOregion = Eastern Mediterranean	0,2256
Latitude	0,2018
WHOregion = South-East Asia	0,1898
WHOregion = Western Pacific	0,1236
Longitude	0,0632
Year	0,0409
AnnualRainfall	0,0261

The results in table 2 indicate overwhelming inclination towards the African. This further strengthened the decision to omit other regions in the subsequent machine learning modelling. Various plots (see figures 4 and 6) on the climate dataset for Africa confirmed the trend that mean annual temperatures are rising and that the rainfall patterns were more erratic.

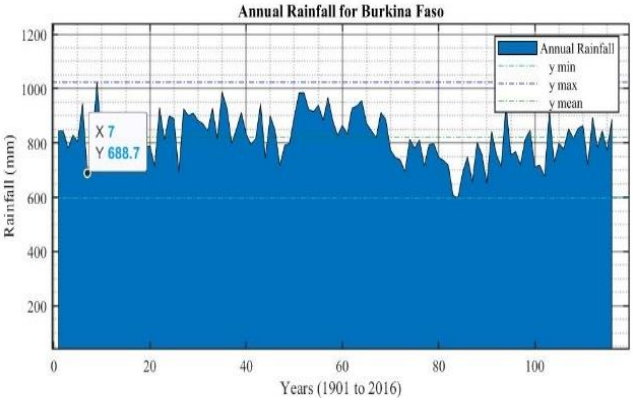

Figure 4. Mean annual temperature for Burkina Faso

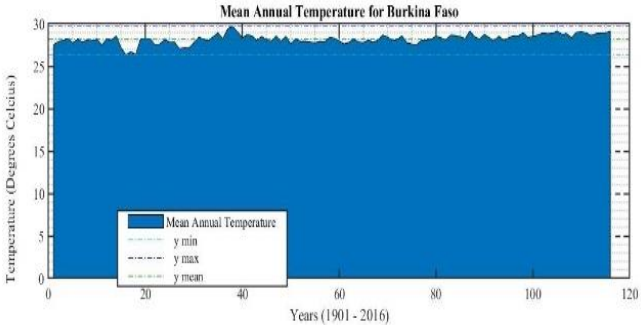

Figure 5. Total annual rainfall for Burkina Faso

3.3.1 Classification Performance

Using the Auto Model of RapidMiner, the six dimensions of the consolidated dataset were processed and resulted in the performances shown in table 3 below.

Table 3. Machine learning algorithm classification performance matrix

Model	Accuracy	AUC	Classification Error	F Measure	Precision	Recall	Sensitivity	Specificity
Decision Tree	75,10%	80,79%	24,90%	79,50%	89,29%	72,29%	72,29%	81,79%
Deep Learning	70,26%	68,81%	29,74%	79,86%	73,70%	87,67%	87,67%	29,05%
Fast Large Margin	48,82%	37,74%	51,18%	63,54%	63,02%	64,31%	64,31%	9,05%
Generalized Linear Model	66,99%	63,41%	33,01%	80,14%	66,99%	100,00%	100,00%	0,00%
Gradient Boosted Trees	84,05%	88,15%	15,95%	87,04%	91,74%	83,38%	83,38%	88,45%
Logistic Regression	66,99%	63,44%	33,01%	80,14%	66,99%	100,00%	100,00%	0,00%
Naive Bayes	66,99%	68,37%	33,01%	80,14%	66,99%	100,00%	100,00%	0,00%
Random Forest	82,88%	88,91%	17,12%	85,87%	91,52%	81,16%	81,16%	88,45%
Support Vector Machine	64,77%	74,12%	35,23%	78,49%	66,21%	96,67%	96,67%	0,00%

Table 4 below shows the results of subjecting performances in table 3 above into the respective ranking of machine learning algorithms. The best 4 performing algorithms are Gradient Boosted Trees, Random Forest, Decision Tree and Deep Learning respectively. The two worst performing ones are Fast Large Margin and Support Vector Machine respectively. For Recall and Sensitivity measures (table 5), individual algorithms had same ranking for both metrics.

Table 4. Machine learning algorithm classification performance ranking

Model	Accuracy	AUC	Classification Error	F-measure	Precision	Specifity	Overall Classification Rank
Gradient Boosted Trees	1	2	1	1	1	1	**(1) 6**
Random Forest	2	1	2	2	2	2	**(2) 9**
Decision Tree	3	3	3	7	3	3	**(3) 19**
Deep Learning	4	5	4	6	4	4	**(4) 23**
Generalized Linear Model	7	8	5	3	5	6	**(5) 28**
Logistic Regression	6	7	6	4	6	7	**(6) 29**
Naive Bayes	5	6	7	5	7	8	**(7) 30**
Support Vector Machine	8	4	8	8	8	9	**(8) 36**
Fast Large Margin	9	9	9	9	9	5	**(9) 45**

Table 5. Machine learning algorithm computation time (in milliseconds) performance ranking

Model	Total Time	Training Time (1,000 Rows)	Scoring Time (1,000 Rows)	Rank (Total)	Rank (Training)	Rank (Scoring)	Overall Rank
Decision Tree	1755,0	110,4	113,8	3	1	6	**3**
Deep Learning	2575,0	1642,9	105,7	6	9	4	**6**
Fast Large Margin	2183,0	149,4	89,4	4	2	2	**2**
Generalized Linear Model	1370,0	220,8	97,6	2	6	3	**4**
Gradient Boosted Trees	36884,0	996,8	122,0	9	8	7	**9**
Logistic Regression	1122,0	194,8	73,2	1	3	1	**1**
Naive Bayes	2465,0	194,8	105,7	5	4	5	**5**
Random Forest	17892,0	207,8	1040,7	7	5	9	**7**
Support Vector Machine	19811,0	496,8	642,3	8	7	8	**8**

3.3.2 *Computation performance*

A shown in table 5, algorithms have different rankings for the total, training, scoring, computation times. For instance, while Logistic Regression ranking in terms of training time (194.8 ms) was 3rd, the algorithm performed the best for the for both total time (1122.0 ms) and scoring time (73.2 ms). Gradient Boosted Trees, Deep Learning and Random Forest took the longest for total time, training time and scoring time respectively. In order to objectively rank the algorithms, the sum of the scores for the three dimensions of computation times (total, training and scoring) was obtained. In the final ranking (column Overall Rank in table 6), Logistic Regression was the best performing while Gradient Boosted Trees was the worst performer.

3.3.3 *Overall Performance*

Putting all the ranks together resulted in the ranking shown in table 6 below. The results show that combining both classification and computation time performances puts Gradient Boosted Trees, Random Forest, Decision Tree and Deep Learning as the 4 top performing algorithms respectively. Fast Large Margin remains the worst performing algorithm.

Table 6. Machine learning algorithm overall performance ranking

Model	Computation Rank	Classification Rank	Combined Rank
Gradient Boosted Trees	9	6	15
Random Forest	7	9	16
Decision Tree	3	19	22
Deep Learning	6	23	29
Logistic Regression	1	29	30
Generalized Linear Model	4	28	32
Naive Bayes	5	30	35
Support Vector Machine	8	36	44
Fast Large Margin	2	45	47

3.3.4 *Malaria Prediction System Development*

Given the advantages of deep learning presented earlier in this paper, this algorithm was selected for the development of the Malaria Predictor System. RapidMiner Prediction function was run to predict malaria incidence using four dimensions identified earlier. Each of these dimensions contributed the following weights to the prediction: Mean Annual Rainfall = 0.723, Annual Rainfall = 0.0603, Latitude = 0.271 and Longitude = 0.098. The ranges of each of these dimensions are illustrated in figures 6-8 below: Values in figures 6 and 7 confirm some of the findings documented in malaria literature – that is, the vector ability to thrive and sustain maturity of the plasmodia is between 20 and 29.5^{0}C and total annual rainfall between 50mm and 2,000mm. In the case of African Continent, malaria tend to be recorded along latitudes 0^0 to 15^0. The Malaria Predictor build using RapiderMiner Simulator is shown in figure 9 below. To predict an incidence, three inputs (rainfall, temperature and latitude) are input in the respective input fields. The simulator automatically displays the predicted incidence on the right side of the Simulator.

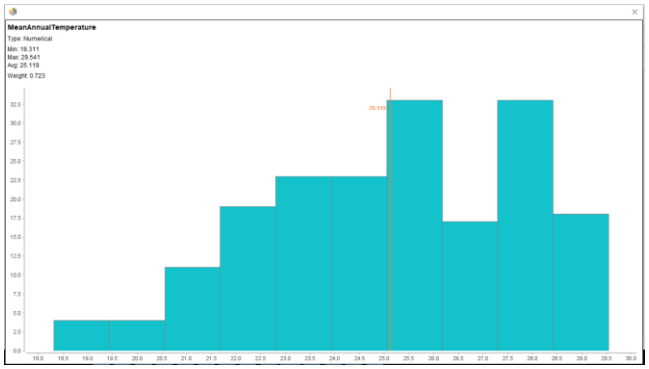

Figure 6. Effects of mean annual temperature on malaria incidence

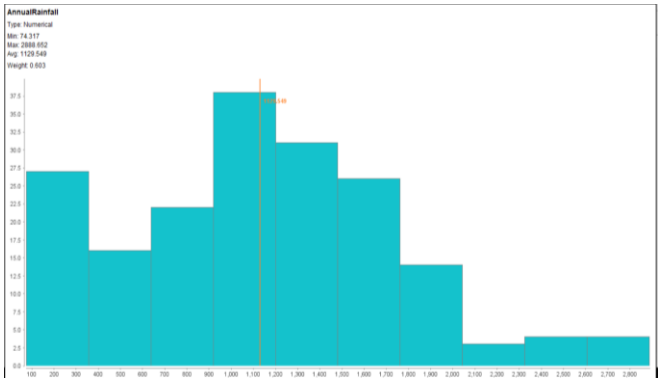

Figure 7. Effects of annual rainfall on malaria incidence

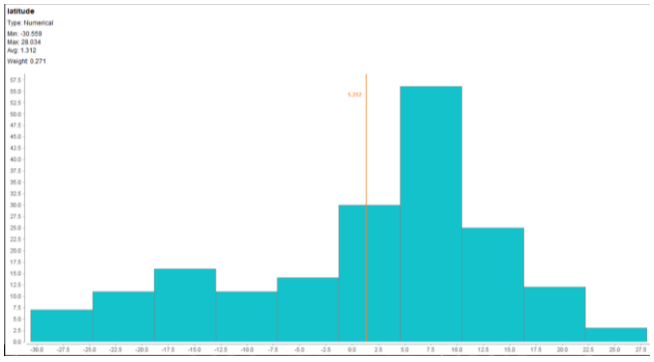

Figure 8. Contribution of latitude to malaria incidence

.

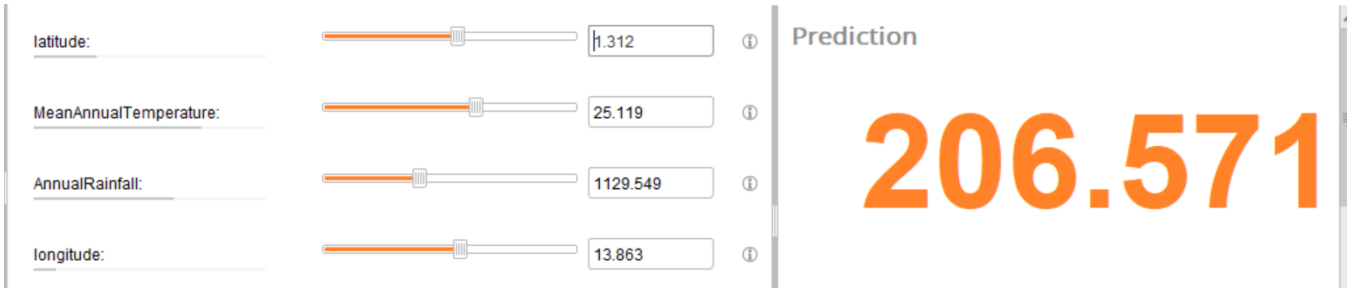

Figure 9. Deep Learning Simulator for Malaria Incidence

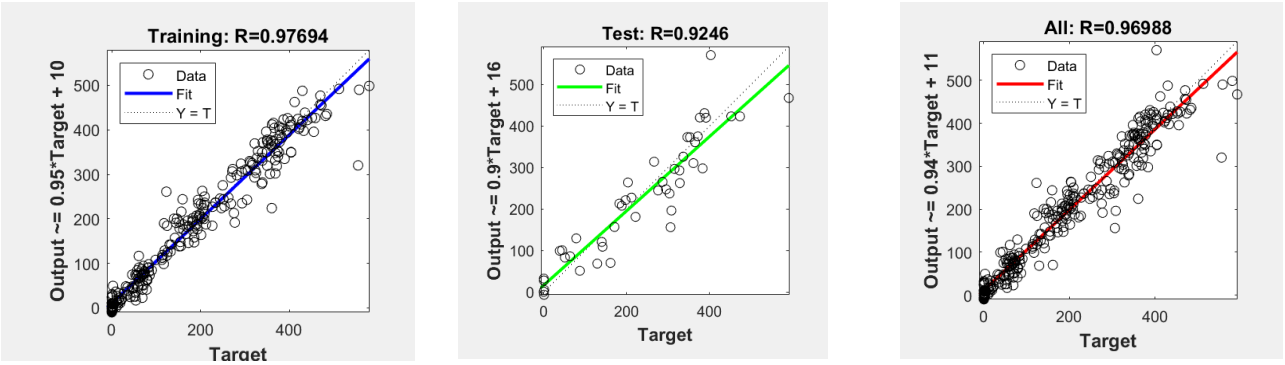

Figure 10. ANNs in MATLAB

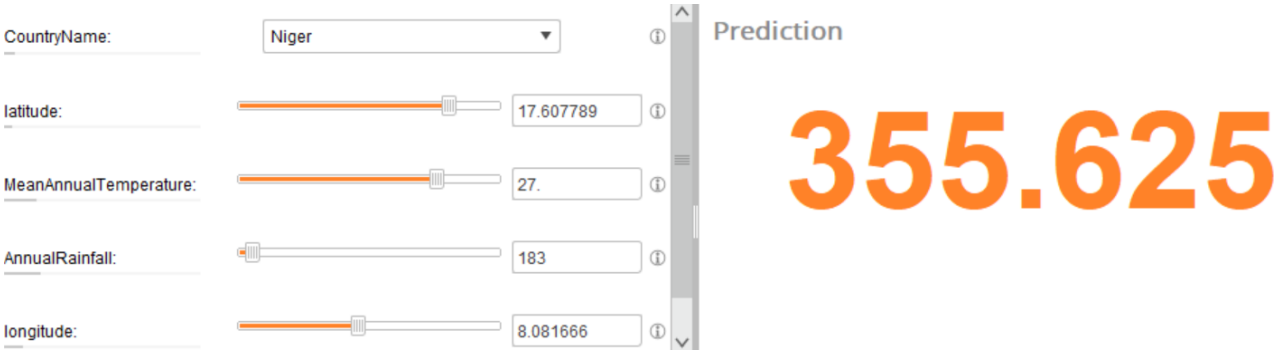

Figure 11. Malaria predictor system – illustration using Niger

MATALAB Deep Learning Toolkit (*nntraintool*) was used with random (*dividerand*) data division (70:15:15 for training, validation and testing respectively) and Bayesian Regularization (*trainbr*) training algorithm. A total of 21 hidden neurons were used for the 4 inputs (mean annual temperature, annual rainfall, latitude and longitude) and one output/target (malaria incidence). The limited size of the dataset made running of the Deep Learning function inappropriate – in its place, a larger number of hidden layers within the neural network function was used. As shown in figure 10 above, the forecasting model generated results that have correlations that are very close to 1.

4. EVALUATION, DISCUSSION AND CONCLUSION

4.1 Evaluation

In order to objectively evaluate the prediction model, malaria incidence dataset (for 2000 and 2005) not used in model development was run through the simulator. In order to predict future occurrence of malaria based on the three factors (mean annual temperature, total annual rainfall and latitude), the Simulator tool in RapidMiner was used to predict malaria incidence for 5 (selected based on the higher chance of incidence). The results of this evaluation are depicted in table 7 below.

Table 7. Malaria incidence predictor – evaluation results

Country	Predicted	Actual	Relative Error
Burkina Faso	504.378	545.89	7,60%
DRC	519.976	515.61	0,85%
Kenya	141.304	129.34	9,25%
Niger	355.625	355.40	0,06%
Nigeria	351.774	420.93	16,43%

The results should be interpreted based on the reality that the values for dataset dimension (rainfall, temperature, latitude, longitude and malaria incidence) are averages. The error rates therefore will reflect the diversity of the respective countries.

Take Nigeria for example which has up to 4 climate types: (1) topical monsoon characterized by warm temperatures (between 26^0C and 28^0C), high humidity and copious rainfall (ranging from 2,000 to 4,000mm per year); (2) Tropical savanna climate (also referred to as tropical wet and dry climate) which experiences temperatures ranging from 18^0C to 36^0C and annual rainfall of about 1,500mm; (3) Sahel climate or tropical dry climate which receives much less rainfall and temperatures can be as high as 40^0C; and (4) Alpine climate or highland climate or mountain climate. On the other hand, Kenya has up to seven very diverse climatic zones and the same can be said of other countries included in the evaluation set (https://cpdb.wmo.int/).

4.2 Discussion and Conclusion

Four well-documented findings about malaria guided the research presented in this paper. These are: (1) malaria is communicable disease that is transmitted through mosquito bites; (2) there is a strong correlation between malaria incidence and particular climatic conditions; (3) malaria is preventable and treatable. Based on this, very elaborate strategies (such as Global Technical Strategy for Malaria 2016-2030 [17]) have been implemented towards eradicate malaria from the face of the earth; and (4) climate change is slowly reversing gains from malaria prevention/eradication efforts and the situation in Africa (especially Sub-Saharan African countries) is still very serious. Given the above circumstances, the proposed Malaria Predictor System will go a long in contributing towards the WHO and country-level agenda to combat malaria epidemic.

As per our research objective 1, the results presented in this paper provide beyond empirical observations of the association between weather conditions and malaria incidence reported in Africa. More scientific explanations to this association has also been provided. The second objective is realized in form of a Malaria Predictor System implemented in RapidMiner. Although the relative error of the system could be as high as 40%, much lower (than 1%) are achievable from localized (to a district in a country for instance) data. Larger errors are recorded where the averaged data represents diverse climate types. The authors also acknowledge that errors could be introduced from the nature of most health-related datasets – they tend to be incomplete and vague. With such factors in mind, the proposed Malaria Predictor should only be used as guiding tool for planning purposes.

5. ACKNOWLEDGMENT

The project reported in this paper is funded by the South Africa's Research Foundation (NRF) grant for 2019: Thuthuka Funding Instrument (Unique Grant No: 117800).

6. REFERENCES

[1] B. Naqvi, A. Ali, M. A. Hashmi, and M. Atif, 'Prediction Techniques for Diagnosis of Diabetic Disease : A Comparative Study', vol. 18, no. 8, pp. 118–124, 2018.

[2] B. Marr, 'What Is The Difference Between Artificial Intelligence And Machine Learning?', *Forbes.com*, 2016.

[3] R. Deepak *et al.*, 'Optimizing neural networks for medical data sets : A case study on neonatal apnea prediction', *Artif. Intell. Med.*, vol. 98, no. January 2018, pp. 59–76, 2019.

[4] M. Masinde, M. Mwagha, and T. Tadesse, 'Downscaling africa's drought forecasts through integration of indigenous and scientific drought forecasts using fuzzy cognitive maps', *Geosci.*, 2018.

[5] M. Masinde, 'Survivability to sustainability of biodiversity', 2013.

[6] B. K. Sarkar, 'An e-healthcare system for disease prediction using hybrid data mining technique', pp. 628–661, 2019.

[7] J. Schmidhuber, 'Deep Learning in neural networks: An overview', *Neural Networks*, vol. 61, pp. 85–117, 2015.

[8] H. Chiroma *et al.*, 'Malaria Severity Classification Through Jordan-Elman Neural Network Based on Features Extracted From Thick Blood Smear', *NEURAL Netw. WORLD*, vol. 25, no. 5, pp. 565–584, 2015.

[9] B. Modu, N. Polovina, Y. Lan, S. Konur, A. T. Asyhari, and Y. Peng, 'Towards a Predictive Analytics-Based Intelligent Malaria Outbreak Warning System', *Appl. Sci.*, vol. 7, no. 8, Aug. 2017.

[10] M. H. Beale, M. T. Hagan, and H. B. Demuth, 'Deep Learning Toolbox ™ User ' s Guide How to Contact MathWorks', 2019.

[11] RapidMiner, 'RapidMiner Studi o Manual', 2014.

[12] IBM Corp, 'SPSS Statistics for Macintosh', *IBM Corp. Released 2019*. 2019.

[13] WHO, 'World Malaria Report 2019', Geneva:, 2019.

[14] W. Cella *et al.*, 'Do climate changes alter the distribution and transmission of malaria? Evidence assessment and recommendations for future studies', *Rev. Soc. Bras. Med. Trop.*, vol. 52, 2019.

[15] A. Midekisa, B. Beyene, A. Mihretie, E. Bayabil, and M. C. Wimberly, 'Seasonal associations of climatic drivers and malaria in the highlands of Ethiopia', *Parasit. Vectors*, vol. 8, Jun. 2015.

[16] A. Rahman, L. Roytman, M. Goldberg, and F. Kogan, 'Comparative analysis on applicability of satellite and meteorological data for prediction of malaria in endemic area in Bangladesh', *J. Trop. Med.*, vol. 2010, 2010.

[17] WHO, 'Global technical strategy for malaria 2016-2030', 2015.

[18] X. Wu, Y. Lu, S. Zhou, L. Chen, and B. Xu, 'Impact of climate change on human infectious diseases: Empirical evidence and human adaptation', *Environ. Int.*, vol. 86, pp. 14–23, Jan. 2016.

[19] IPCC, 'Proposed outline of the special report in 2018 on the impacts of global warming of 1 . 5 ° C above pre-industrial levels and related global greenhouse gas emission pathways , in the context of strengthening the global response to the threat of climate cha', *Ipcc - Sr15*, vol. 2, no. October, pp. 17–20, 2019.

[20] H. D. Teklehaimanot, M. Lipsitch, A. Teklehaimanot, and J. Schwartz, 'Weather-based prediction of Plasmodium falciparum malaria in epidemic-prone regions of Ethiopia I. Patterns of lagged weather effects reflect biological mechanisms', *Malar. J.*, vol. 3, pp. 1–11, 2004.

[21] M. Walker *et al.*, 'Temporal and micro-spatial heterogeneity in the distribution of Anopheles vectors of malaria along the Kenyan coast', *Parasit. Vectors*, vol. 6,

Oct. 2013.

[22] P. Goswami, U. S. Murty, S. R. Mutheneni, and S. T. Krishnan, 'Relative Roles of Weather Variables and Change in Human Population in Malaria: Comparison over Different States of India', *PLoS One*, vol. 9, no. 6, Jun. 2014.

[23] F. Jiang *et al.*, 'Artificial intelligence in healthcare: Past, present and future', *Stroke and Vascular Neurology*. 2017.

[24] K. Vembandasamy, R. Sasipriya, and E. Deepa, 'Heart Diseases Detection Using Naive Bayes Algorithm', *Int. J. Innov. Sci. Eng. Technol.*, 2015.

[25] P. Guo *et al.*, 'Developing a dengue forecast model using machine learning: A case study in China', *PLoS Negl. Trop. Dis.*, 2017.

[26] M. Wieland and M. Pittore, 'Performance evaluation of machine learning algorithms for urban pattern recognition from multi-spectral satellite images', *Remote Sens.*, vol. 6, no. 4, pp. 2912–2939, 2014.

[27] L. R. Hope and K. B. Korb, 'A Bayesian metric for evaluating machine learning algorithms', *Lect. Notes Artif. Intell. (Subseries Lect. Notes Comput. Sci.*, vol. 3339, pp. 991–997, 2004.

[28] N. Williams, S. Zander, and G. Armitage, 'A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification', *Comput. Commun. Rev.*, vol. 36, no. 5, pp. 7–15, 2006.

[29] L. Kotthoff, I. P. Gent, and I. Miguel, 'An evaluation of machine learning in algorithm selection for search problems', *AI Commun.*, vol. 25, no. 3, pp. 257–270, 2012.

[30] Maikel, 'Free and Open Machine Learning Documentation', 2019.

[31] G. Ertek, D. Tapucu, and I. Arin, *Text mining with rapidminer*. 2013.

[32] S. B. Green, *Using SPSS for Windows and Macintosh: Analyzing and Understanding Data:Using SPSS for Windows and Macintosh: Analyzing and Understanding Data*. 2016.

[33] R. Gonzalez, 'Applied Multivariate Statistics for the Social Sciences', *Am. Stat.*, 2003.

[34] T. Curran and W. J. Friedman, 'ERP old/new effects at different retention intervals in recency discrimination tasks.', *Brain Res. Cogn. Brain Res.*, 2004.