Leveraging Neural Networks in Malaria Control

Joseph Livesey
University of Liverpool
Liverpool, UK
Joseph.Livesey@Liverpool.ac.uk

Dominik Wojtczak
University of Liverpool
Liverpool, UK
d.wojtczak@liverpool.ac.uk

***Abstract*—In this paper we build a neural network model to predict prevalence of malaria for a given geographic location and year. We report on our experience of building the most suitable neural network architecture for this problem. We show that both utilizing dropout and Adam optimizer in the network training process is very effective and can lead to a precise model without overfitting issues. Incorporating rainfall data leads to a significant improvement in the precision of the model, highlighting the fact that this is an important factor in the spread of malaria. We then utilize the selected best neural network to predict the outcome of eradicating malaria at given locations. This can help to decide where to use limited resources, like vaccines or insecticides, for the largest possible impact in malaria control.**

***Index Terms*—malaria prevalence prediction, neural networks, malaria control**

I. Introduction

APPROXIMATELY 15 million deaths a year are caused directly by infectious disease. We focus in this paper on malaria which is one of the most deadliest disease despite being preventable and treatable. It is estimated that more than 400 thousands people died in 2019 due to malaria and most of them (60%) were children under 5. Furthermore, the African region accounts for 94% of malaria cases and deaths [1].

Reliable models of how infectious diseases spread and how they can be controlled are critical to reducing their impact and saving lives as a result. Generally such models are hand-crafted and then the parameters are adjusted to fit the data. Here, instead of building the transmission model, we train a neural network to predict the infection numbers in a given year based on the specific geographic location and predicted rainfall pattern for that year.

Many different approaches to forecasting the prevalence of infectious disease have been taken (see, e.g., [2] for a survey on this subject). Based on [2], in the case of malaria, the most common approach taken so far was the use of linear models and the main focus of such studies was mainly on China. According to [1], 90% of malaria deaths in 2010 were in Africa, which makes forecasting the prevalence of malaria in Africa of great importance. We take as a baseline a model that simply returns the average malaria prevalence across Africa (namely, it returns prevalence of 0.239 for every point). This was calculated based on every data point in the training set. We then build a linear regression model and then shift focus to building neural network of various depths for the same task to achieve better performance.

Related work. Several papers so far used very simple neural networks for predicting malaria prevalence for four different countries: China [3], [4], Brazil [5], India [6], [7] and Thailand [8]. Most of these studies used fewer than 100 data points in total. In [9] a neural network approach was used to predict malaria outbreaks in Canta in Brazil. That study found that neural networks can work better than a logistic regression model. This was only a small study, however it gives an indication that using neural networks on a larger scale to predict outbreaks of malaria can be beneficial.

Other approaches for malaria prevalence or outbreak prediction were using fuzzy systems [10], [11], extreme gradient boosting [12], [13], KNN [13], or Naive Bayes [13].

II. Dataset and Data Preprocessing

Data was collected from the Malaria Atlas Project (MAP) [14], where survey records are collated for the spread of malaria all across the world. Each piece of data has its own citation, which can be found in [15].

There are multiple different forms of malaria which affect humans. In this paper we focus on Plasmodium Falciparum, which is the most severe form in humans [1].

As we are focusing on predicting prevalence of malaria in Africa, we started by eliminating all irrelevant data points in this survey. It was also necessary to clean the data so that only complete data was used, with both coordinates of where the data was collected from, and the information on positive malaria cases. This left us with 9026 data points with which to create a model. Each of these data points had coordinates, information on the year the data was collected, and the number of people surveyed compared with the total number of people positive for malaria. All the models in this paper were trained on a training set of 5000 points, 2000 data points being kept for validation to tune the hyperparameters and 1901 data points for the test data to estimate the ultimate performance of the chosen model. A representation of these points can be seen in Figure 1. Coordinates were used as the independent variable, and proportion of positive malaria results as the dependent variable.

Rainfall data was gathered from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) [16] for all the locations in our dataset. The CHIRPS data was downloaded in PNG format, meaning it had to be matched up with the coordinates of our data points. It was noticed that some data points were seen to be on the boundaries between countries,

meaning no rainfall data could be extracted. For these points the average of the 8 points surrounding them were taken as their values. This process was run for all rainfall data between 1981 and 2014.

Finally, we normalized all values to the [0,1] interval.

III. Performance metric

We evaluated the performance of all the constructed models using the root mean squared error (RMSE), which is defined as

$$RMSE = \sqrt{\sum_{i=1}^{N}(x_i - x_i')^2/N}$$

where x_i-s are the actual observations, x_i' are the values predicted by the model, and N is the number of points in the dataset (training or validation). As prevalence values are in the range $[0, 1]$, RMSE also has to belong to this range.

IV. Models without Rainfall Data

We first tried to build models that do not take advantage of the rainfall data collected. This was done to see how important such climate data is for malaria models.

A. Linear Regression

The first aim was to create a model for the spread of malaria in Africa using the standard linear regression [17] approach. We built two models: one ignoring the year as an input variable and another considering it as one of the input variables, but their performance was essentially the same. We found that linear regression model is better than the baseline model (the one that always returns a constant prevalence value) and gave a RMSE of 0.228. It is as significant improvement over the constant prevalence model (whose RMSE is 0.263), but not as much as we have hoped for.

B. Neural Networks

The most important choice when building a neural network is deciding how many hidden layers are needed, and how many nodes will be in each hidden layer. Another is choosing the type of optimizer that will be used to train the network. We initially used a gradient descent optimizer and a neural network with 3 fully-connected hidden layers with 10 nodes each.

After much experimenting with learning rate, steps, batch sizes and hidden units we managed to train a model to have RMSE error of just 0.16, a significant improvement over the baseline. However, the RMSE on the validation set was actually worse than the 0.263 benchmark. This was a clear sign that the model was overfitting.

One way to improve the model and prevent overfitting was to limit the number of hidden layers and hidden nodes. This led to just a slight improvement over to the baseline model, on both training and validation. It was found that, in general, the smaller the neural network, the smaller the difference in RMSE values between the training and validation data, but this was at a trade off with their magnitude, which was higher.

We also tried to put coordinates in to bucketized sections, however again, this either led to no improvement, or overfitting. Another way was limiting the data to just East Africa (due to the majority of data points being focused here), but again this gave no significant improvement.

Finally, we decided to try a different optimizer. We used the Adam optimizer, as it is known to be simple and effective [18]. This indeed gave a better result than the gradient descent method. Using a learning rate of 0.005, 50000 steps, a 1000 batch size and three hidden layers, each of 7 nodes, the RMSE error was 0.188 for the training data. However, it was 0.227 for the validation data. This again is better than the benchmark, so is a promising result, however, is also a way away from the training set error, meaning some overfitting may well be occurring. Progression of training for this model can be seen in Figure 2.

In Figure 3, we can see the plot of actual malaria proportions on the y-axis and predicted proportions on the x-axis. As can be seen, the general trend seems correct, however there is a lot of room for improvement.

V. Models with Rainfall Data

At this point, we incorporated the rainfall data collected from [16] as clearly just collecting the coordinates and the year was insufficient in order to build a successful model.

A. Linear Regression

We started off by building a linear model to give the baseline for results. This gave a training RMSE of 0.240 and validation RMSE of 0.241. This is a clear but not hugely significant improvement over the constant prevalence model (whose RMSE is 0.263).

B. Neural Networks

Many trials of neural network were then attempted, trying to balance model performance with overfitting problems. As recommended in [19], it is best to limit hidden neurons to start with (bearing in mind this may not be optimal), so this is where we began.

A single hidden layer with just 3 nodes was created. Dropout [20] was set to 20 percent on all models in order to restrict overfitting where possible. This immediately gave a promising result, RMSE of 0.228 for the training set and RMSE of 0.232 for the validation set, much better than the linear regression model. This alone shows the promise of using neural network in order to predict the spread of disease as it shows it can be much more effective than linear regression models.

This error later improved to 0.222 (training) and 0.226 (validation) when using 5 nodes, and for 10 nodes, 0.218 and 0.222 respectively. Increasing the number of nodes from here improved both the training and validation error, but the gap between these grew.

We experimented with many different 2 and 3 hidden layer neural networks architectures. We tried varying the number of hidden nodes at each level, however it was found that this had

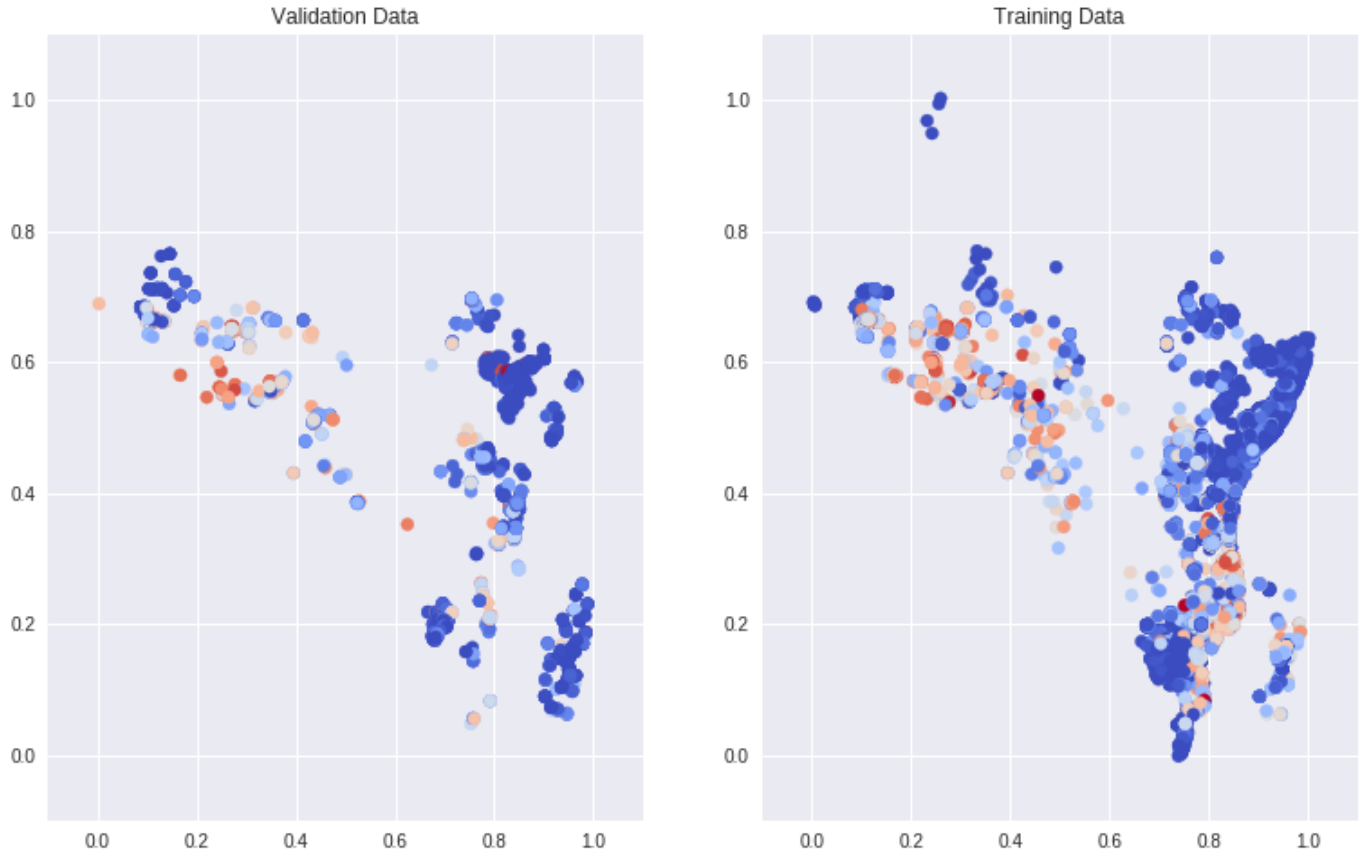

Fig. 1. Data points, plotted by coordinates. The color represents the proportion of people who tested positive for malaria.

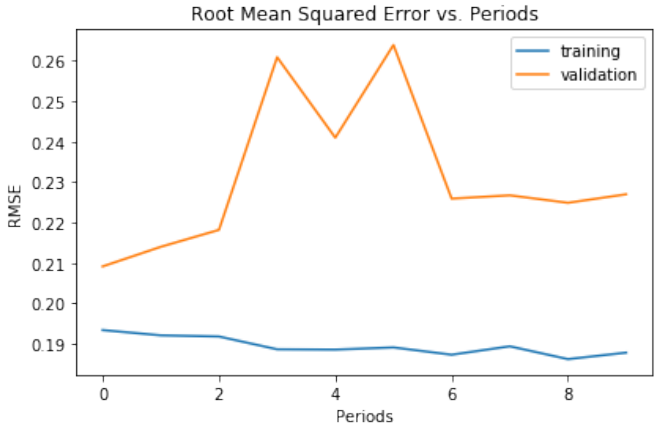

Fig. 2. The training of the neural network with 3 hidden layers each with 7 nodes

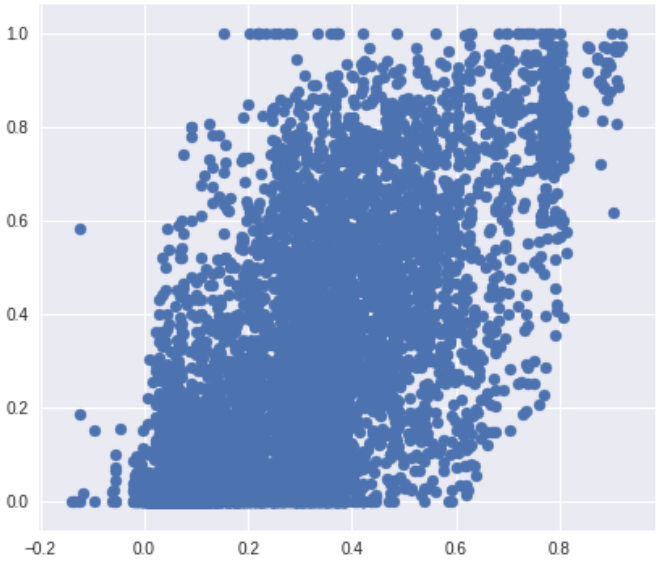

Fig. 3. Plot of actual malaria proportions on the y-axis and predicted proportions on the x-axis

little effect, so most of the testing from this point was with the same number of nodes on each level. Again we started with a small number, and worked up as the experiments continued.

A similar pattern emerged, the smaller the network the less difference between the RMSE on the training and validation sets there is, but also the higher is their magnitude. Interestingly, it was found that a structure of three hidden layers with

5 nodes each was actually significantly worse than a single hidden layer with 5 nodes (0.232 RMSE on training and 0.239 RMSE on validation). Three hidden layers with 10 nodes each results in RMSE 0.217 (training) and 0.220 (validation). The dropout is a big cause of this; with no dropout two hidden layers with 5 nodes each gives us RMSE 0.207 (training) and 0.209 (validation). Clearly this is because a dropout of 0.2 is very high when we have three hidden layers of 5 nodes each, however this seems a very solid model, much lower error and not much overfitting due to its relatively small size.

However this dropout is critical for larger networks. Three hidden layers with 30 nodes each give RMSE of 0.147 on training and RMSE of 0.198 on validation, a huge disparity and a clear sign of overfitting. With dropout included though we get RMSE of 0.174 on training and RMSE of 0.187 on validation, which is promising, but still a disparity.

Images of malaria spread predictions were also created using various different models. This was done to get a visualization of what the model actually thinks and predicts. As the malaria average for the database was roughly 0.25, the simple models largely under predicted malaria throughout the whole of Africa, and only with a more complicated network did the models start to capture how prevalence of malaria really varies across Africa.

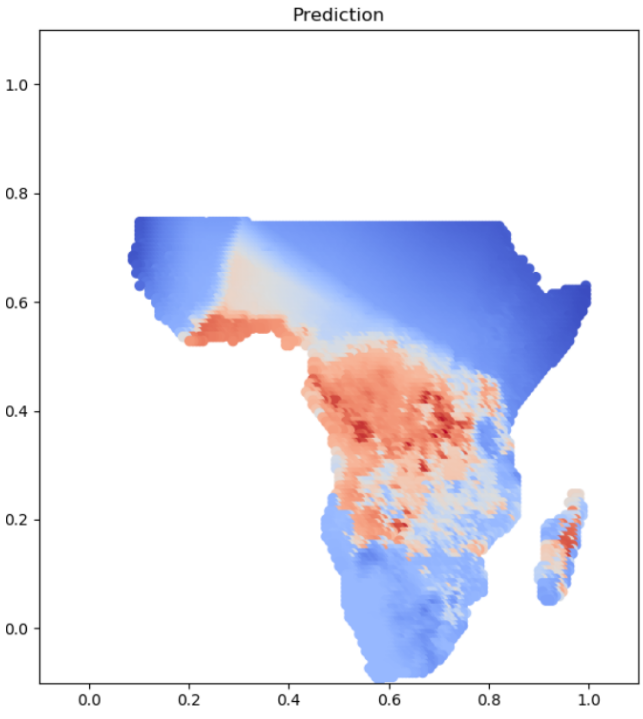

Fig. 4. The prediction made for the prevalence of malaria in Africa March 2014, using a neural network with two hidden layers of 30 nodes each. The average prevalence was 0.236.

Due to the nature of neural networks, each time the network is trained it gives a slightly different result, but using a network with two hidden layers with 30 nodes each and 0.2 dropout gave a good training error to validation error trade off, whilst having low errors and beginning to capture how instances of malaria really vary. (Example malaria prevalence prediction of this model can be seen in Figure 4.) This was the model that was selected in the end as the best one and evaluated on the test data. Its RMSE on training data was 0.187, on validation data was 0.193, and in the end we used the test data to check its performance and got RMSE of 0.194. As we can see, the model is very consistent even on totally unseen data and our model selection process was successful in the end.

VI. Malaria Control

The next step was to look in to where resources may best be used to combat malaria.

Points were artificially entered into the training data, to simulate areas where malaria had been eradicated to see which had the biggest effect on Africa as a whole. Four points were added, at the smallest distance interval, to simulate this area having had malaria eradicated for the particular month. This was attempted for different locations in Africa. We then retrained our best neural network model and measured the drop in the average prevalence rate as predicted by our model for the whole of Africa.

The findings were interesting: adding these points to the centre of Africa had a much bigger impact than on the coasts, even to areas which previously had a high proportion of malaria. Adding these points to Congo gave an average prevalence of 0.198, compared with 0.215 when adding these to Mozambique, however it needs to be noted that the original data set had far fewer points in the centre of Africa than on the coasts. Figures 5, 6 and 7 show examples of this.

VII. Conclusions

In this paper we showed that even with limited data a neural network is a powerful tool for predicting prevalence of malaria in Africa. We noted that overfitting is a common problem when training a neural network. The problem therefore came down to balancing the loss with the overfitting. Perhaps unsurprisingly, the bigger models with low dropout would have impressive RMSE on the training set, however a very poor result on the test set. This was an expected result and confirmed these models were overfitting. Adding parameters is a way to decrease RMSE, but we may get to the point where each instance in our dataset could have its own parameter, which clearly is not a sensible model for trying to make predictions on unseen data.

However by both utilizing dropout and Adam optimizer for learning is very effective and can lead to a precise model. Rainfall data lead to a significant improvement in the model, highlighting the fact that this is an important factor in the spread of malaria. In the end, by artificially inputting zero malaria data points into the model gives us a good idea where limited resources, e.g., vaccines or insecticides, could best be used to minimize its spread.

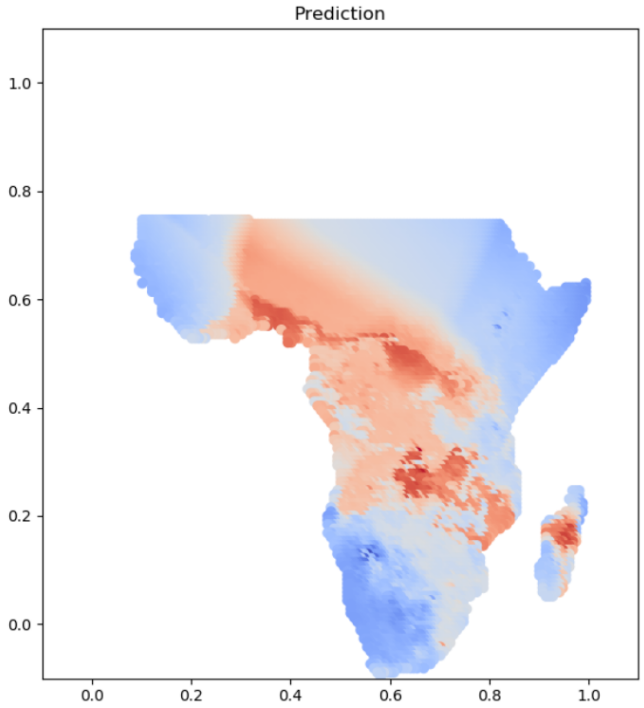

Fig. 5. The prediction made for the prevalence of malaria in Africa using a 30,30 neural network, after adding 4 malaria free data points to the west of Africa (Sierre Leone). The average prevalence was 0.217.

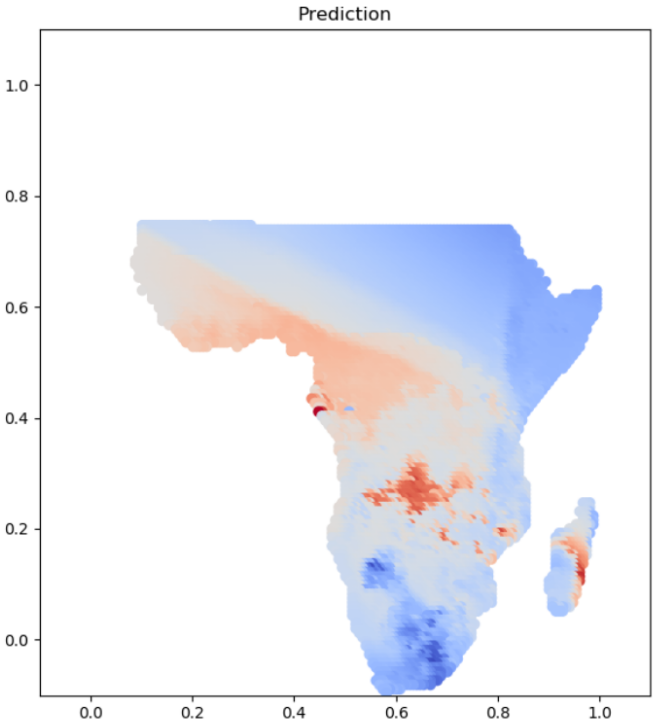

Fig. 6. The prediction made for the prevalence of malaria in Africa using a neural network with two hidden layers of 30 nodes each, after adding 4 malaria free data points to the east of Africa (Mozambique). The average prevalence was 0.215.

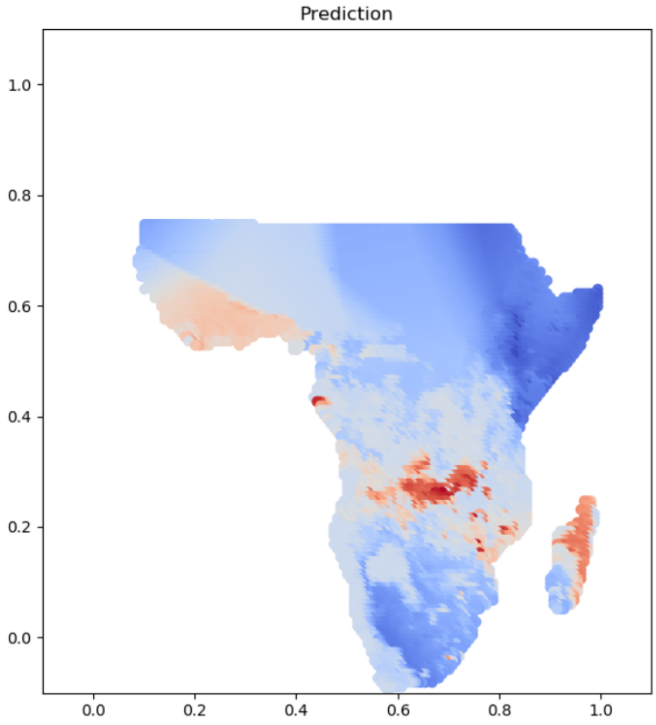

Fig. 7. The prediction made for the prevalence of malaria in Africa using a neural network with two hidden layers of 30 nodes each, after adding 4 malaria free data points to the centre of Africa (Congo). The average prevalence was 0.198.

REFERENCES

[1] "Fact sheet about Malaria." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/malaria

[2] K. Zinszer, A. D. Verma, K. Charland, T. F. Brewer, J. S. Brownstein, Z. Sun, and D. L. Buckeridge, "A scoping review of malaria forecasting: past work and future directions," *BMJ open*, vol. 2, no. 6, 2012.

[3] L. Zhou, F. Zhang, Y. Wang, L. Yu, Y. Shi, H. Jiang, L. Liu, and S. Nie, "A neural network model for the prediction of malaria in jingmen city, china," *International Journal of Infectious Diseases*, vol. 21, p. 270, 2014.

[4] M. Wang, H. Wang, J. Wang, H. Liu, R. Lu, T. Duan, X. Gong, S. Feng, Y. Liu, Z. Cui *et al.*, "A novel model for malaria prediction based on ensemble algorithms," *PloS one*, vol. 14, no. 12, p. e0226910, 2019.

[5] M. V. M. d. Lima and G. Z. Laporta, "Evaluation of prediction models for the occurrence of malaria in the state of amapá, brazil, 1997-2016: an ecological study," *Epidemiologia e Serviços de Saúde*, vol. 30, p. e2020080, 2021.

[6] P. Mohapatra, N. Tripathi, I. Pal, and S. Shrestha, "Determining suitable machine learning classifier technique for prediction of malaria incidents attributed to climate of odisha," *International Journal of Environmental Health Research*, pp. 1–17, 2021.

[7] V. Sharma, A. Kumar, L. Panat, G. Karajkhede, and A. Lele, "Malaria outbreak prediction model using machine learning," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 4, no. 12, 2015.

[8] R. Kiang, F. Adimi, V. Soika, J. Nigro, P. Singhasivanon, J. Sirichaisinthop, S. Leemingsawat, C. Apiwathnasorn, and S. Looareesuwan, "Meteorological, environmental remote sensing and neural network analysis of the epidemiology of malaria transmission in thailand," *Geospatial health*, pp. 71–84, 2006.

[9] G. B. da Cunha, J. F. Luitgards-Moura, E. L. M. Naves, A. O. Andrade, A. A. Pereira, and S. T. Milagre, "Use of an artificial neural network to predict the incidence of malaria in the city of Canta, state of Roraima," *Revista da Sociedade Brasileira de Medicina Tropical*, vol. 43, no. 5, 2010.

[10] A. L. Buczak, B. Baugher, E. Guven, L. C. Ramac-Thomas, Y. Elbert, S. M. Babin, and S. H. Lewis, "Fuzzy association rule mining and classification for the prediction of malaria in south korea," *BMC medical informatics and decision making*, vol. 15, no. 1, pp. 1–17, 2015.

[11] B. E. Chekol and H. Hagras, "Employing machine learning techniques for the malaria epidemic prediction in ethiopia," in *2018 10th Computer Science and Electronic Engineering (CEEC)*. IEEE, 2018, pp. 89–94.

[12] O. Nkiruka, R. Prasad, and O. Clement, "Prediction of malaria incidence using climate variability and machine learning," *Informatics in Medicine Unlocked*, vol. 22, p. 100508, 2021.

[13] G. Kalipe, V. Gautham, and R. K. Behera, "Predicting malarial outbreak using machine learning and deep learning approach: a review and analysis," in *2018 International Conference on Information Technology (ICIT)*. IEEE, 2018, pp. 33–38.

[14] "The Malaria Atlas Project." [Online]. Available: https://www.tki-dev.malariaatlas.org/

[15] "PR Survey Data." [Online]. Available: https://www.tki-dev.malariaatlas.org/pr-survey-data/

[16] "CHIRPS: Rainfall Estimates from Rain Gauge and Satellite Observations | Climate Hazards Center - UC Santa Barbara." [Online]. Available: https://www.chc.ucsb.edu/data/chirps

[17] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[18] J. Brownlee, "Gentle Introduction to the Adam Optimization Algorithm for Deep Learning," Jul. 2017. [Online]. Available: https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/

[19] S. Karsoliya, "Approximating number of hidden layer neurons in multiple hidden layer bpnn architecture," *International Journal of Engineering Trends and Technology*, vol. 3, no. 6, pp. 714–717, 2012.

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.