HIGH SCHOOL EDITION
Journal of Student Research
Volume 14 Issue 1 (2025)

MalariVis: A Bi-Modal Machine Learning System for Malaria Case Forecasting and Outbreak Prediction

Dylan Ma

Academy of Science and Technology, USA

ABSTRACT

Every 50 seconds, malaria claims another life, an annual cost of over 600,000 people. Eradication efforts have been widespread, yet cases have increased over time, and life-saving care often remains out of reach for those who need it most. While malaria is entirely treatable, hundreds of thousands continue to pass away each and every year; change is necessary. This study aimed to develop a new approach to the problem, creating a dual-focused system for both malaria incidence forecasting (regression) and outbreak prediction (classification), one to four months in advance. Built on clinical data from the National Institute for Pharmaceutical Research and Development and climatic datasets from the POWER Project, fourteen robust models were tested in Abuja, Nigeria: nine for regression and five for classification. The Seasonal Autoregressive Integrated Moving Average Model proved best for the first task, with an average R2 score of 0.92 across four months of forecasting and statistical significance ($p<0.05$). The Long Short-Term Memory Model performed just as well for the second task, with an average accuracy of 92.75% ($p<0.05$). Furthermore, in-depth feature analysis was conducted, showing wind speed, rainfall, and humidity as key driving factors of malaria incidence. The finalized models were implemented into *MalariVis*, a free, innovative, and easily-accessible application for malaria forecasting, enabling health officials and civilians alike to prepare for upcoming outbreaks before they happen, minimizing upcoming casualties. *MalariVis* could be the future of malaria prediction and prevention, saving lives on a pathway to eradication.

Introduction

Malaria is a disease transmitted through the bites of infected female *Anopheles* mosquitoes and is endemic in nearly one hundred countries around the world (World Health Organization [WHO], 2023a). Almost half of the world's population is at risk for malaria, and there were an estimated 247 million cases in 2022, with children under five being the most at risk, accounting for 80% of deaths in Africa (WHO, 2023a). The disease also leads to malnutrition, indirectly causing the deaths of half of all children under the age of five worldwide (National Institute of Allergy and Infectious Diseases [NIAID], 2021). Cases are disproportionately clustered in Africa, a region home to 95% of recorded malaria incidence and 96% of deaths (WHO, 2023a).

Transmission of malaria is directly correlated with multiple climatic factors, including temperature and rainfall. Warmer temperatures allow the parasite within infected mosquitoes to develop faster, leading to higher incidence rates (Fernando, n.d.). Thus, malaria is often tied with hot seasons in many regions around the world. Heavy rainfall in dry climates can also provide good breeding conditions for mosquitoes, once again, increasing incidence (Fernando, n.d.). As climate change becomes an ever-pressing issue, malaria will evolve into an even greater threat, reemerging worldwide and increasing opportunities for transmission.

Symptoms of malaria tend to begin days after infection and include fever and flu-like illness with chills, headaches, muscle aches, and tiredness (Centers for Disease Control and Prevention [CDC], 2023). These signs can progress if not treated to anemia, jaundice, kidney failure, seizures, comas, and in many cases, death (CDC, 2023). Some types of malaria can relapse, occurring again, while others lay dormant for years in hibernation before finally striking their hosts (CDC, 2023). Malaria's death rate varies from region to region based on medical infrastructure

and other factors, with Nigeria having the highest of over 33%, indicating a strong need for preventive measures (WHO, 2023a).

Currently, malaria prevention strategies include vaccines such as the R21 vaccine (only the second ever licensed vaccine against parasitic diseases and approved in October of 2023), antimalarial drugs for better treatment, improved diagnosis techniques, and vector management tools including bed nets, insecticides, and environmental modification (WHO, 2023b; NIAID, 2011). However, vaccines remain scarce and expensive, especially for developing countries, antimalarial drugs are declining in effectiveness due to the emergence and spread of resistant parasites, current diagnosis remains labor-intensive and slow, and vector management has faced setbacks in nearly every area, from insecticide-resistant mosquitoes to an increase in lack of funding across the globe (NIAID, 2011). Subsequently, another approach for malaria control is critically needed, without a reliance on advanced infrastructure and drugs, and with adaptability to the emergence and reemergence of new, diverse types of parasites and mosquitoes.

This study aimed to provide that approach through the usage of mathematical modeling through machine learning and deep learning techniques, creating an accurate and robust predictive model for malaria incidence and outbreak detection, one to four months in advance. It also looked to determine which models had the highest predictive capability, as well as the climatic features most important in malaria transmission through feature importance, creating a base for future work whether it be with malaria or any infectious disease in general, from influenza to dengue virus.

For the regression task of forecasting malaria incidence, the following highly reputable models were tested: Random Forest Regressor (RFR), Ridge Regression, Least Absolute Shrinkage and Selection Operator (LASSO), AutoRegressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA), Long Short-Term Memory (LSTM), Support Vector Regression (SVR), Negative Binomial Regression, and k-Nearest Neighbors (KNN). RFR is based on the Random Forest algorithm, in which an ensemble of decision trees is created to make accurate predictions for continuous numerical values. Each tree is essentially trained individually to predict a target value, and predictions are averaged for a final prediction. Ridge Regression, also known as L2 regularization, is a regression technique that is similar to Ordinary Least Squares (OLS), but with one extra caveat: an introduced regularization term that looks to prevent overfitting by adding a penalty on the size of coefficients. LASSO works similarly to Ridge Regression, but adds a regularization term by penalizing the value of coefficients, sometimes removing features completely, whereas Ridge Regression typically only decreases the coefficients of certain coefficients. ARIMA is a time-series forecasting method that combines three key components: autoregressive or AR, which represents the influence of past values on current values, differencing or I, which looks into how to make the data stationary, and moving average or MA, which accounts for the impact of past error values and noise on current values. SARIMA is simply an extension of ARIMA, which accounts for seasonal patterns in time series data. LSTM is a type of recurrent neural network (RNN) that performs exceptionally with predicting sequential data, with memory cells and gates to retain information over long sequences, allowing complex patterns to be recognized. It falls under the category of deep learning as opposed to machine learning, which focuses on neural networks with many layers, replicating the human brain in structure and function. SVR is typically used in regression tasks, finding a hyperplane that best fits training data while maintaining a specific margin of error defined by support vectors, hence the name. Negative Binomial Regression is a type of generalized linear model used for count data that shows overdispersion (variance is higher than normal), estimating mean size rather than a specific value. Finally, KNN is a simple yet versatile algorithm, making predictions based on the average value of the k-nearest data points.

For the classification task of predicting malaria outbreaks and their intensity from low to high, several reputable models were once again tested: Random Forest Classifier (RFC), Support Vector Machine (SVM), LSTM, Extreme Gradient Boosting (XGBoost), and KNN. RFC is similar to RFR due to its structure of decision trees, but it uses a majority vote system rather than an average to determine the final classification. SVM is similar to SVR in its use of a hyperplane that best separates data, but this time the margins are used to separate classes, finding a decision boundary. LSTM is used in a similar way for classification as it is used for regression, with the main difference being the output node structures with one for each class, rather than only one for the predicted numerical value. XGBoost works with decision trees, where each new tree corrects the errors of the previous one, garnering accuracy over time.

Finally, KNN makes predictions based on the majority class of a data point's k-nearest neighbors, relying on the idea that similar data points tend to belong in the same class.

Prior to creating the models, a Seasonal Trend decomposition using Loess (STL) was used, separating the seasonal data of malaria into three categories: seasonal, trend, and residual. Seasonal represents the regular patterns in the data, trend represents the long-term direction of the data, and residual represents the error and noise in the data, including outliers and inconsistencies. To clean the data, the residual was removed, and "seasonal" and "trend" were used as features for the models, allowing them to take seasonality into account, increasing accuracy. When needed, the data will also be lagged to allow the models to take past values into consideration when predicting future ones, increasing robustness.

Each model was compared to each other through accuracy, precision, and recall for classification models, and mean absolute error, root mean squared error, and R2 score for regression models. Statistical tests were also conducted to further determine which model performed best for the regression and classification tasks ($p < 0.05$). Climatic data sourced from the POWER Project by NASA including rainfall, wind speed, humidity, maximum temperature, and minimum temperature were used as features in the models (National Aeronautics and Space Administration [NASA], 2023). Malaria incidence was sourced from the diagnostic unit of the National Institute for Pharmaceutical Research and Development at Abuja, ranges from January 2000 to December 2015, and includes confirmed monthly cases of *P. falciparum* and *P. malariae* via blood smear testing, the most accurate diagnostic test for malaria (National Institute for Pharmaceutical Research and Development [NIPRD], 2023). To determine which climatic factors were most important in malaria transmission, feature analysis was conducted on the most accurate two models for regression and classification respectively.

The combination of machine learning and disease prediction in the researcher's project is nothing new, but the underlying features are different. For example, in 2021, a project looked to predict malaria cases in six countries across sub-Saharan Africa, but it only did so at the country scale, with no real way to implement the predicted results (Nkiruka et al., 2021). A previously mentioned study conducted in Abuja found that humidity was the greatest factor at lag zero and temperature the greatest at lag four, but it was limited to the three factors of humidity, temperature, and rainfall, and used only statistical modeling without machine learning, thus it was unable to predict malaria incidence in advance (Segun et al., 2020). In Burkina Faso, machine learning models such as Random Forest were used to predict outbreaks weeks in advance, but the precision lacks at only 30% for minor outbreaks, clearly with much room for improvement (Harvey et al., 2021). In 2022, a group of researchers were able to accurately predict malaria outbreaks using sea surface temperature as an important feature, a novelty in the field (Martineau et al., 2022). However, accuracy varies widely, from 30% to 90%, and is highly dependent on the season. The study is more aimed at long-term forecasting, up to nine months in advance, rather than monthly reporting.

The researcher's project shares similarities with many past journals, but there are no studies with high accuracy and great future forecasting capability, nor is there research into bi-model systems for the simultaneous detection and prevention of malaria. This project also predicts incidence and outbreaks anywhere from one to four months in advance, unprecedented for past studies. A large portion of research tends to use yearly cases on the country scale rather than monthly, thus implementation is hard and results do not typically capture all the complex dynamics and intricacies within malaria incidence. Accuracy also tends to be lacking in other models, and as more research is done, more features are found, and more models are tested, improvement is always possible. Finally, intensive feature analysis from one to four months has never been done before, shedding light on the long-term impact of various climatic factors on malaria trends.

This research project was implemented in an online interface to provide easy access to those in need, healthcare workers and civilians alike. By simply inputting past climatic data points, the selected model will automatically generate a predicted incidence and outbreak severity, acting as an early warning system for times to come, unprecedented in current-day malaria prevention strategies. This allows local health authorities to concentrate resources and preventive measures, rather than spreading them throughout the region, increasing effectiveness, and thus, saving lives. Early detection will also greatly reduce healthcare costs, as accurate prediction leads to early treatment,

tackling the disease before it progresses further. The research done over the course of the project will also help in determining which type of mathematical model is most suitable for epidemic detection, not just malaria, aiding in future endeavors regarding countless different infectious diseases. Lastly, robust feature analysis conducted will prove critical to further research, once again providing insight for future projects in the field, helping researchers understand the relationship between the climate and infectious disease transmission.

As climate change progresses, this research will become more and more vital in understanding future trends, as it takes into account changes over time, recognizing changing transmission patterns and expanding transmission seasons beneath the surface. This project is easily adaptable in the future by simply adding more data, expanding its scope while also increasing robustness and accuracy. Looking broader, with just the addition of larger datasets, this research can be expanded to regions across Sub-Saharan Africa, increasing effectiveness and saving lives across the board. It can also be applied to other similar diseases, creating a future where early detection using machine learning is commonplace.

Methods

Safety

This project was computer-based and thus safety was not applicable. The monthly malaria incidence dataset was not shared to ensure the anonymity of patients and prevent improper usage of health information.

Materials

- Computer
- Google Colab
- Microsoft Excel
- Monthly malaria incidence dataset from 2000 - 2015
- Monthly climatic values dataset from 2000 - 2015 with temperature, wind speed, humidity, and rainfall

Procedure

Study Area

The study area for this project was the capital city of Nigeria: Abuja. With a current metro area population of nearly four million, it lies in one of the hardest-hit regions by malaria in Sub-Saharan Africa. The city's climate is tropical, with a noticeable wet and dry season, accounting for notable fluctuations in malaria incidence every year. Abuja is located in central Nigeria in the Federal Capital Territory, near the Gulf of Guinea.

Data Management

A dataset was created by combining monthly malaria incidence values from 2000 to 2015, sourced from the NIPRD, with monthly climatic values from 2000 to 2015 consisting of temperature, wind speed, humidity, and rainfall, sourced from the NASA POWER Project (NIPRD, 2023; NASA, 2023). A column titled "Months" was added, which indicates the months since the beginning of the dataset, allowing models to capture long-term trends in the data. This was done in Microsoft Excel, and columns were labeled "Cases", "Months", "Temp", "Humidity", "Wind Speed", and "Rainfall". There were 192 total rows (not counting the initial row of column labels) and six columns. This dataset was titled "malaria_dataset.csv" and saved in the .csv format. A separate data set titled "malaria_for_class_dataset.csv" was created and saved in the .csv format for classification models, with the target values being binary. A zero

represented no outbreak, while one represented an observed outbreak. These outbreak classifications were derived from the original dataset, where case indices above the 50th percentile were considered an outbreak.

Data Preprocessing

Separate Google Colaboratory notebooks were created for each model to be tested: RFR, Ridge Regression, LASSO, SARIMA, ARIMA, LSTM, SVR, KNN, and Negative Binomial Regression for regression, and KNN, LSTM, SVM, RFC, and XGBoost for classification, yielding fourteen total notebooks. In each notebook, the dataset titled "malaria_dataset.csv" was set as a dataframe and read. The required libraries were imported for each model including pandas, numpy, sklearn.metrics, statsmodels.tsa.seasonal. Some machine learning models also had model specific libraries that were mandatory, such as statsmodels.tsa.arima_model for the ARIMA regression model. Others come from pre-installed libraries in Google Colab, such as Random Forest coming from within scikit-learn itself.

In every notebook, an STL decomposition was done and the dataset was split into its seasonal, trend, and residual components. The residual component represented the outliers and variance in the dataset that did not fit the seasonal nature of malaria in Abuja and Nigeria, and thus it was smoothed out. Then, the decomposed data was "lagged", which adds new features by allowing the model to take previous data into account (such as the temperature two months prior). Different amounts of lags were tested to determine the most effective one for accuracy (whether the data should be shifted back one month, two months, or more), with a lag of five months yielding the highest accuracy. Thus, the remaining dataset included the target value for regression models and the target class for classification models, along with features ranging from one month in the past to five due to the introduction of lags.

Model Building and Optimization

In all fourteen machine learning models, the target labels and features were set accordingly and the data was split into a training and testing set, with the former being 80% of the original data (2000 to roughly 2013), and the latter being 20% (2013 to 2015). Each model was built accordingly, and the hyperparameters were fine-tuned and tested on the testing set in a variety of ways, from modifying the architecture and adding regularization to reduce overfitting in the two LSTM models, to testing various parameters in the two Random Forest models. This was done both manually and autonomously using grid and random search, as well as brute force. Some models such as SARIMA and ARIMA required the dataset to be in the date-time format and were adjusted for. A final model was created in each notebook to be compared against one another. This procedure was done for prediction from one to four months in advance, with the dataset features being shifted "down" for each successive month.

Model Comparison and Analysis

To determine a model's effectiveness, its accuracy, precision, and recall were checked for classification models, and its mean absolute error (MAE), root mean squared error (RMSE), and R2 score for regression models. These metrics were conducted over multiple cross-validations, allowing the model to train and test on different "folds" of the dataset, leading to a more accurate analysis. These values were used to compare models to determine the optimal one for practical use for both regression and classification. Unpaired t-tests were conducted to confirm statistical significance ($p < 0.05$) of the top models. Various plots were generated through matplotlib to further investigate each model's performance, such as a confusion matrix for classification models. Feature importance was conducted for both regression and classification, the results were scaled and averaged, and the values were ranked, allowing for a comprehensive analysis of the leading climatic factors for malaria transmission.

Implementation

The highest-performing models for the regression and classification task were implemented into the *MalariVis* website, where users could learn malaria prevention strategies, treatment options, forecast future malaria incidence from one to four months in advance, and predict outbreaks on a similar timescale. To create this user interface, Wix was used, while the prediction application, implemented into the Wix website, was created and coded using Streamlit. To

design the app towards civilian use, separate models were trained and tested without the lagged features of "seasonal", "trend", and "cases", instead relying on the remaining climatic and easily-accessible features. For clinicians, those features were not removed, and an additional ability to add current case counts to train future models was created, increasing accuracy and predictive capability over time.

Results

Initially, the dataset was analyzed. Figure 1 depicts monthly malaria incidence by month and Figures 2 to 5 show the features by month. A seasonal pattern was noted, thus outbreaks for the classification dataset were determined by splitting the data into four main sections (January to March, April to June, July to September, and October to December) and classifying an outbreak as incidence in the upper 50th percentile in each section.

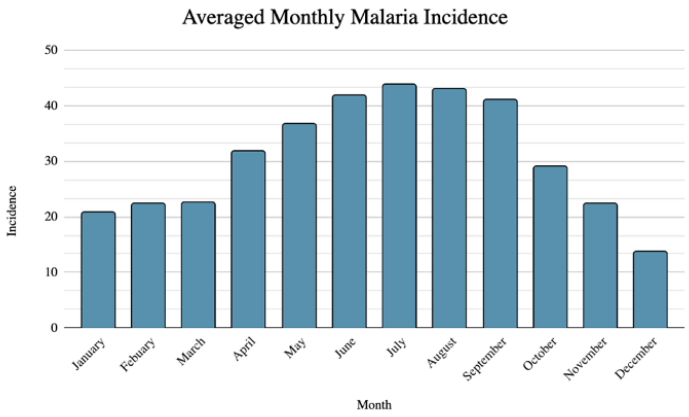

Figure 1. Malaria incidence by month, averaged. Month is the horizontal axis; malaria incidence is the vertical axis.

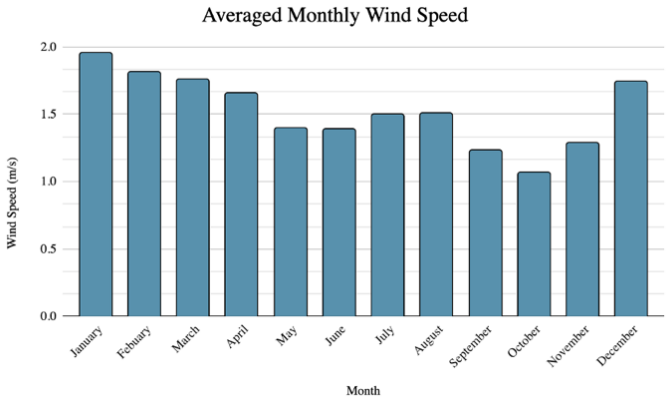

Figure 2. Wind speed by month, averaged. Month is the horizontal axis; wind speed in meters per second is the vertical axis.

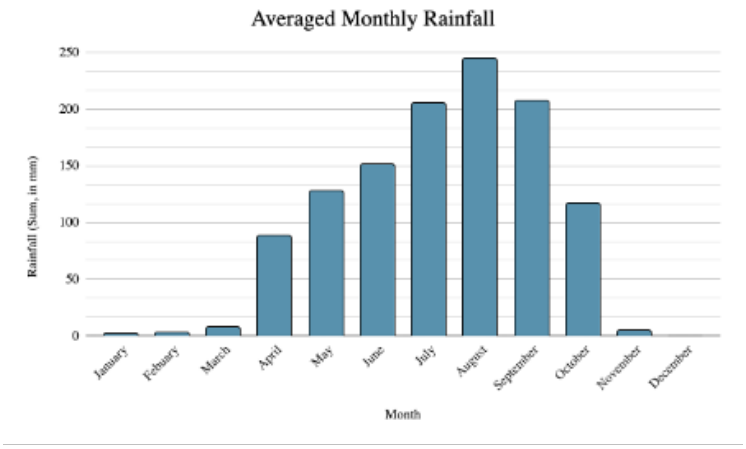

Figure 3. Rainfall by month, averaged. Month is the horizontal axis; rainfall in millimeters is the vertical axis.

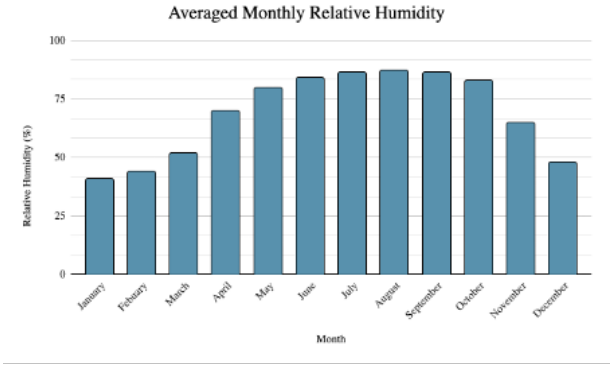

Figure 4. Relative humidity by month, averaged. Month is the horizontal axis; humidity (%) is the vertical axis.

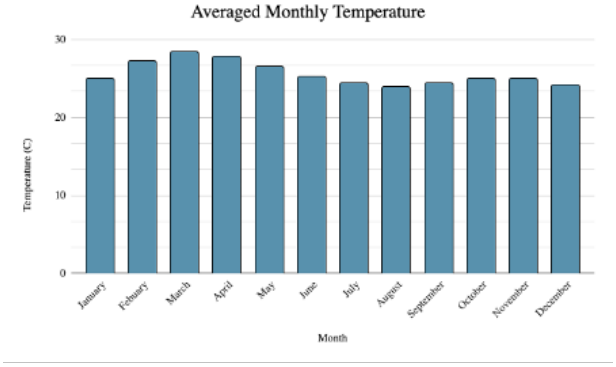

Figure 5. Temperature by month, averaged. Month is the horizontal axis; temperature in Celsius is the vertical axis.

The results from the fourteen models can be seen below. Table 1 depicts the MAE, RMSE, and R2 score of the nine regression models tested, on the time scale of one month in advance. Table 2 shows the same metrics but when forecasting two months, Table 3 three months in advance, and Table 4 four months in advance.

Table 1. MAE, RMSE, and R2 score for selected regression models: one month in advance.

Model Name	MAE	RMSE	R2
RFR	3.31 cases	4.21 cases	0.91
LASSO	4.97 cases	6.92 cases	0.83
Ridge Regression	4.93 cases	6.46 cases	0.84
SARIMAX	2.77 cases	3.86 cases	0.94
ARIMAX	6.45 cases	8.78 cases	0.62
Negative Binomial Regression	5.32 cases	7.21 cases	0.75
LSTM	4.31 cases	5.25 cases	0.87
SVR	4.75 cases	6.18 cases	0.85
KNN	5.26 cases	7.14 cases	0.77

Table 2. MAE, RMSE, and R2 score for selected regression models: two months in advance.

Model Name	MAE	RMSE	R2
RFR	3.71 cases	4.65 cases	0.90
LASSO	5.01 cases	6.89 cases	0.83
Ridge Regression	4.99 cases	6.70 cases	0.83
SARIMAX	3.19 cases	3.93 cases	0.93
ARIMAX	6.22 cases	8.55 cases	0.63
Negative Binomial Regression	6.02 cases	7.33 cases	0.71
LSTM	4.21 cases	4.96 cases	0.89
SVR	4.69 cases	6.11 cases	0.85
KNN	5.77 cases	7.82 cases	0.74

Table 3. MAE, RMSE, and R2 score for selected regression models: three months in advance.

Model Name	MAE	RMSE	R2
RFR	4.22 cases	5.11 cases	0.87
LASSO	5.12 cases	7.01 cases	0.83
Ridge Regression	5.21 cases	7.39 cases	0.82
SARIMAX	3.55 cases	4.17 cases	0.91
ARIMAX	6.85 cases	9.12 cases	0.60
Negative Binomial Regression	6.11 cases	7.54 cases	0.71
LSTM	4.31 cases	5.34 cases	0.88
SVR	5.22 cases	6.53 cases	0.83
KNN	5.89 cases	8.18 cases	0.73

Table 4. MAE, RMSE, and R2 score for selected regression models: four months in advance.

Model Name	MAE	RMSE	R2
RFR	4.44 cases	5.28 cases	0.87
LASSO	5.33 cases	7.55cases	0.81
Ridge Regression	5.47 cases	7.41 cases	0.81
SARIMAX	4.11 cases	4.77 cases	0.89

ARIMAX	6.84 cases	9.47 cases	0.61
Negative Binomial Regression	5.99 cases	7.65 cases	0.70
LSTM	4.29 cases	4.91 cases	0.87
SVR	5.57 cases	7.88 cases	0.80
KNN	5.93 cases	8.07 cases	0.73

Figure 6 below depicts the MAE for the selected regression models, Figure 7 shows the RMSE, and Figure 8, the R2 score, all from one to four months in advance in a bar graph for easy comparison.

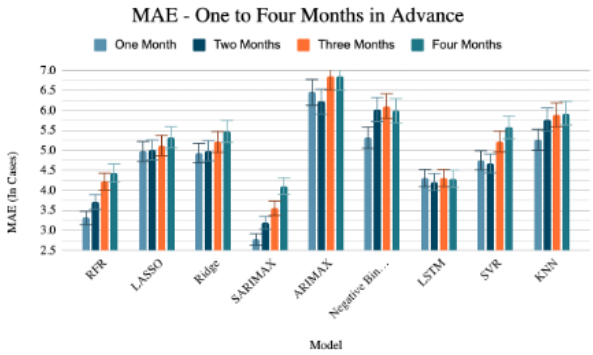

Figure 6. MAE for selected regression models, one to four months in advance. Model is the horizontal axis; MAE (in cases) is the vertical axis.

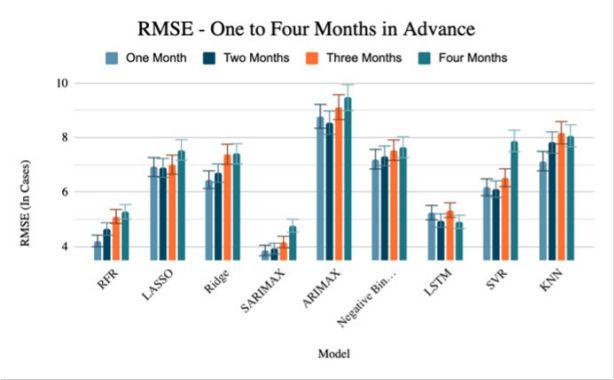

Figure 7. RMSE for selected regression models, one to four months in advance. Model is the horizontal axis; RMSE (in cases) is the vertical axis.

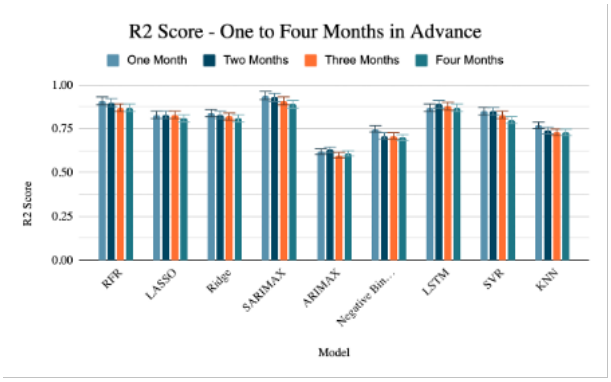

Figure 8. R2 score for selected regression models, one to four months in advance. Model is the horizontal axis; R2 score is the vertical axis.

Similar tables can be seen for the five classification models in Tables 5, 6, 7, and 8, showing accuracy, precision, and recall from one to four months in advance.

Table 5. Accuracy, precision, and recall for selected classification models: one month in advance.

Model Name	Accuracy	Precision	Recall
RFC	79%	0.81	0.78
SVM	88%	0.89	0.89
XGBoost	82%	0.82	0.83
LSTM	95%	0.95	0.95
KNN	89%	0.88	0.91

Table 6. Accuracy, precision, and recall for selected classification models: two months in advance.

Model Name	Accuracy	Precision	Recall
RFC	78%	0.79	0.77
SVM	87%	0.88	0.86
XGBoost	79%	0.82	0.77
LSTM	94%	0.95	0.94
KNN	87%	0.85	0.88

Table 7. Accuracy, precision, and recall for selected classification models: three months in advance.

Model Name	Accuracy	Precision	Recall
RFC	79%	0.79	0.81
SVM	84%	0.85	0.83
XGBoost	79%	0.82	0.76
LSTM	92%	0.92	0.93
KNN	86%	0.86	0.85

Table 8. Accuracy, precision, and recall for selected classification models: four months in advance.

Model Name	Accuracy	Precision	Recall
RFC	78%	0.77	0.79
SVM	82%	0.84	0.80
XGBoost	79%	0.81	0.80
LSTM	90%	0.89	0.91
KNN	82%	0.83	0.80

Figure 9 below depicts the accuracy for the selected regression models, Figure 10 shows the precision, and Figure 11, the recall, all from one to four months in advance in a bar graph for easy comparison.

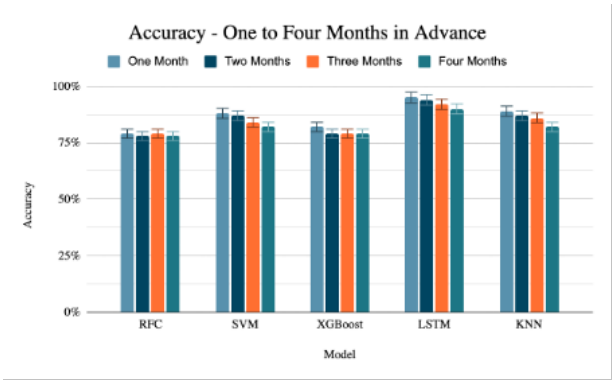

Figure 9. Accuracy for selected classification models, one to four months in advance. Model is the horizontal axis; accuracy is the vertical axis.

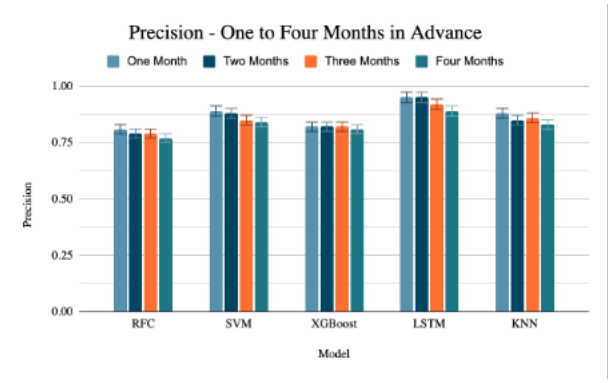

Figure 10. Precision for selected classification models, one to four months in advance. Model is the horizontal axis; precision is the vertical axis.

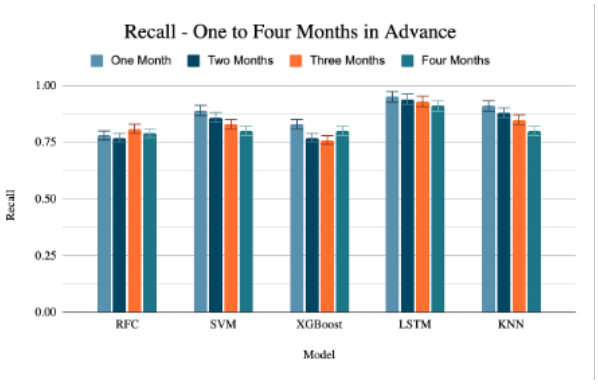

Figure 11. Recall for selected classification models, one to four months in advance. Model is the horizontal axis; recall is the vertical axis.

To accurately model the decrease in R2 score, and the increase in MAE and RMSE over time, the three metrics were graphed in the figures below (Figure 12, Figure 13, Figure 14).

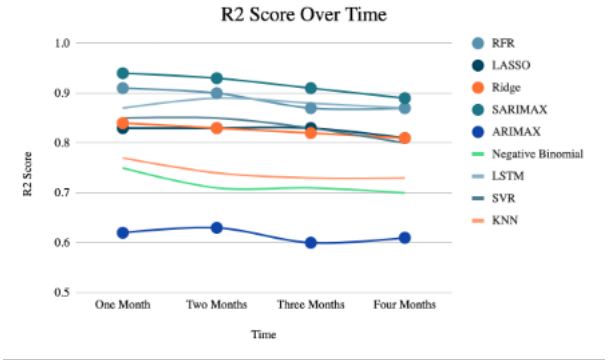

Figure 12. Decrease in R2 score for selected regression models over time. Time is the horizontal axis; R2 score is the vertical axis.

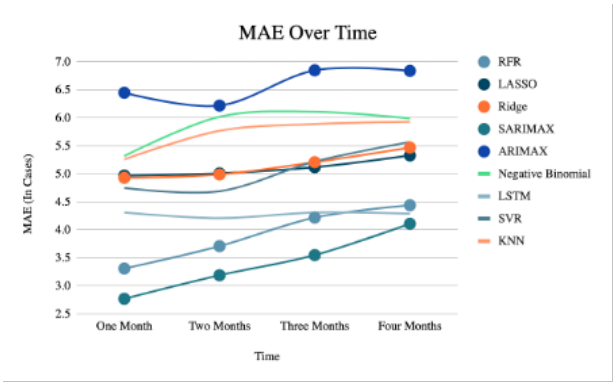

Figure 13. Decrease in MAE for selected regression models over time. Time is the horizontal axis; MAE (in cases) is the vertical axis.

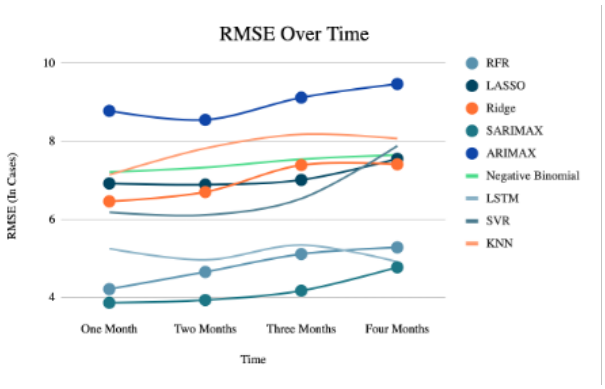

Figure 14. Decrease in RMSE for selected regression models over time. Time is the horizontal axis; RMSE (in cases) is the vertical axis.

To depict the decrease in accuracy, precision, and recall over time, the three metrics were graphed in the figures below (Figure 15, Figure 16, Figure 17).

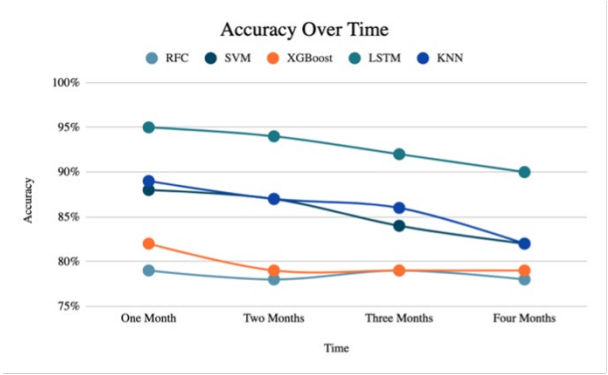

Figure 15. Decrease in accuracy for selected classification models over time. Time is the horizontal axis; accuracy is the vertical axis.

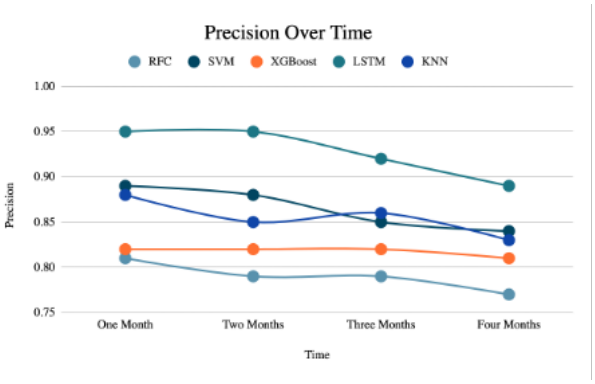

Figure 16. Decrease in precision for selected classification models over time. Time is the horizontal axis; precision is the vertical axis.

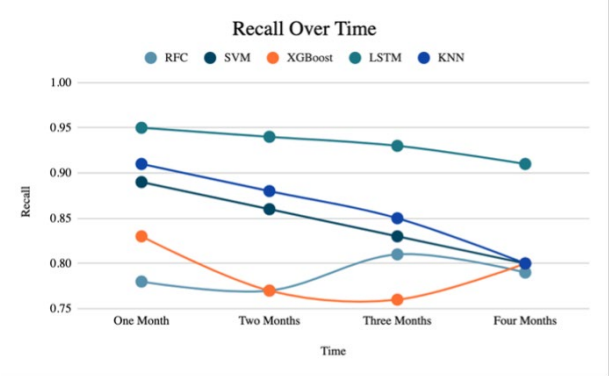

Figure 17. Decrease in recall for selected classification models over time. Time is the horizontal axis; recall is the vertical axis.

To analyze the most successful regression model, a line graph was generated, comparing predicted to actual incidence counts one to four months in advance, as seen in Figures 18 - 21.

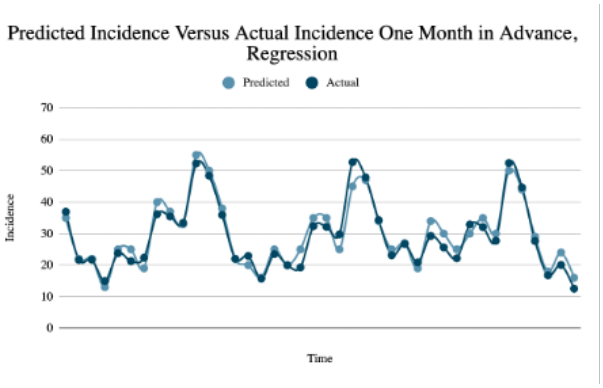

Figure 18. Predicted incidence versus actual incidence, one month in advance, regression. Time is the horizontal axis; incidence is the vertical axis.

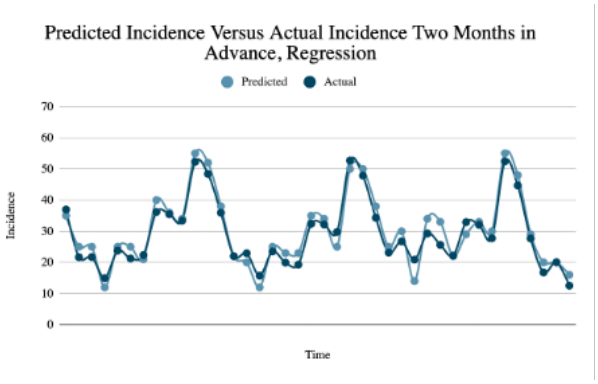

Figure 19. Predicted incidence versus actual incidence, two months in advance, regression. Time is the horizontal axis; incidence is the vertical axis.

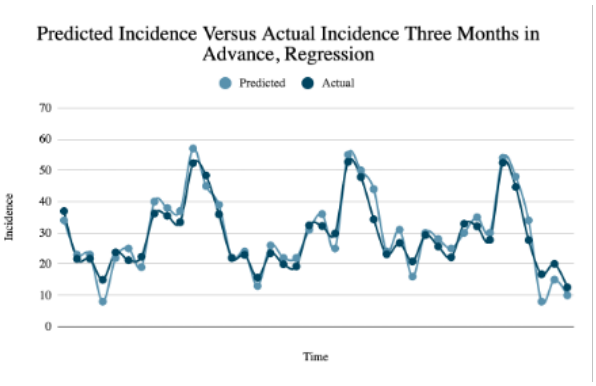

Figure 20. Predicted incidence versus actual incidence, three months in advance, regression. Time is the horizontal axis; incidence is the vertical axis.

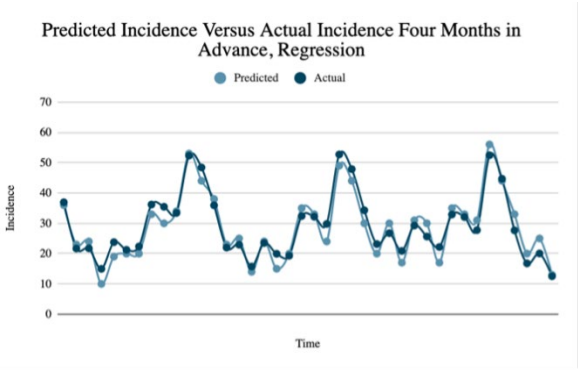

Figure 21. Predicted incidence versus actual incidence, four months in advance, regression. Time is the horizontal axis; incidence is the vertical axis.

To further analyze the most successful classification model, a confusion matrix was generated (prediction one to four months in advance), as seen in Figures 22 - 25.

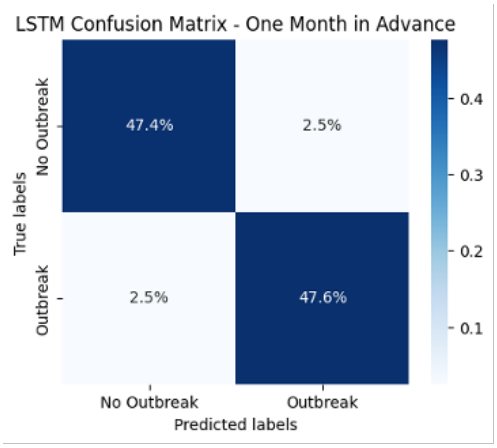

Figure 22. Generated confusion matrix, one month in advance, classification. The values in the confusion matrix are averaged percentages from the cross-fold validation tests and are rounded to the nearest tenth for simplicity.

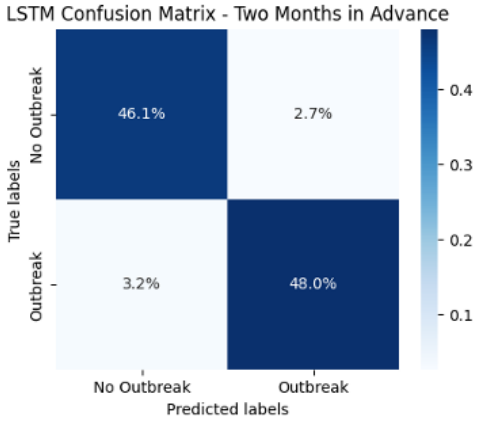

Figure 23. Generated confusion matrix, two months in advance, classification. The values in the confusion matrix are averaged percentages from the cross-fold validation tests and are rounded to the nearest tenth for simplicity.

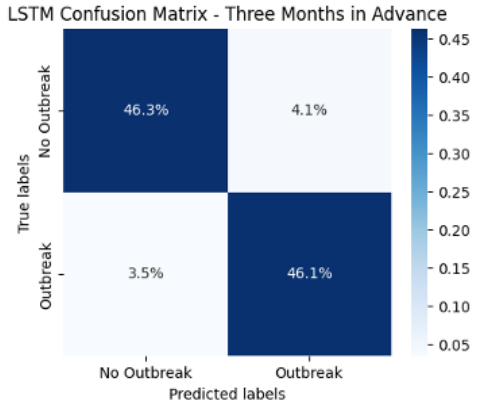

Figure 24. Generated confusion matrix, three months in advance, classification. The values in the confusion matrix are averaged percentages from the cross-fold validation tests and are rounded to the nearest tenth for simplicity.

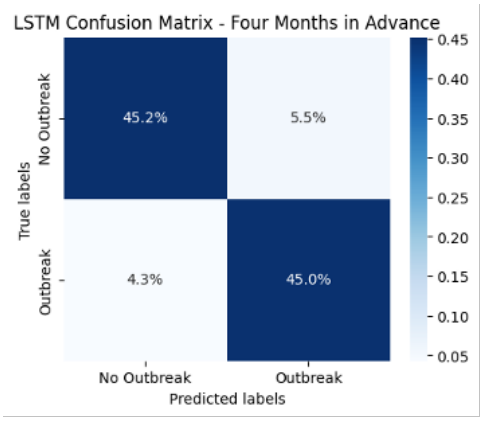

Figure 25. Generated confusion matrix, four months in advance, classification. The values in the confusion matrix are averaged percentages from the cross-fold validation tests and are rounded to the nearest tenth for simplicity.

Unpaired T-tests were used to reinforce the status of the best model: the highest performing models for the regression and classification tasks were compared against the rest of the models in each category from one to four months using the recorded metrics after each cross-validation fold. Statistical significance was determined with a recorded p-value less than 0.05 and observed for both SARIMAX and LSTM (classification).

Feature analysis was conducted on the two best performing models, using the coefficients for SARIMAX and permutation importance for the LSTM classification model. The former allowed the researcher to accurately determine the impact of each climatic variable in the overall equation to calculate incidence, and the latter observed the decrease in accuracy after mixing up each feature, thus indicating the importance. Both methods were scaled from zero to a hundred and then averaged to ensure an accurate feature analysis. The results can be seen in the bar charts below ranked by category (Figure 26) and overall (Figure 27, top eight features).

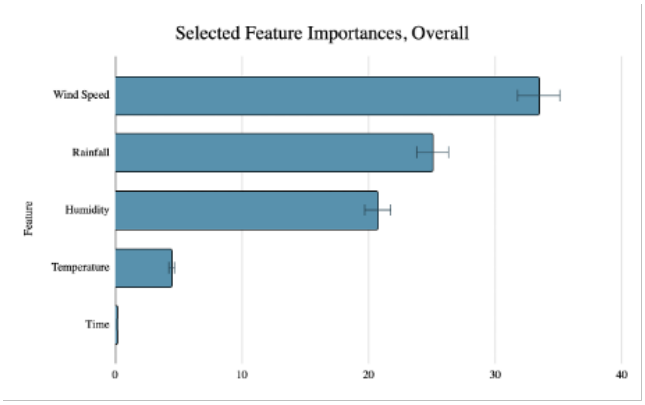

Figure 26. Selected feature importances ranked by category, overall. A scale from 0 to 100 is the horizontal axis; feature is the vertical axis.

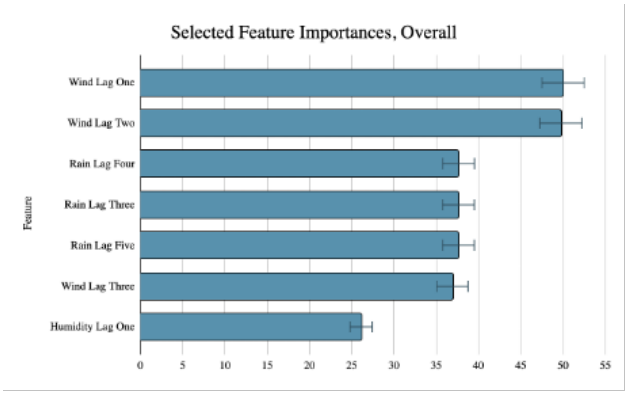

Figure 27. Selected feature importances ranked by category, overall. A scale from 0 to 100 is the horizontal axis; feature is the vertical axis.

Discussion

The results from the initial data analysis point towards the seasonality of malaria incidence and its features, as seen in Figures 1 to 5. Thus, an STL decomposition was used on the dataset, removing the residual (noise) components, allowing the machine learning models to capture the underlying seasonal patterns of malaria. This seasonality causes a variation in cases month by month, thus when classifying outbreaks, incidence counts were split into four distinct

sections (January to March, April to June, July to September, October to December) and values above the 50th percentile were classified as an outbreak. This led to a more holistic modeling, rather than a clustering of perceived "outbreaks" in the summer.

To increase accuracy, optimization techniques such as random search and grid search were used, finding the best hyperparameters for each model, both regression and classification. For classification, a SMOTE-ENN was used, balancing the dataset and reducing bias in predictions. Furthermore, the dataset for all fourteen models was lagged, allowing the models to train on past features to predict future incidence and outbreaks. Lagging the climatic features five times was determined optimal (creating twenty total climatic features), leading to the best performance. Past case incidence and the seasonal and trend components of the STL decomposition were also used as features and lagged accordingly, leading to thirty-six total features (including the "months" feature which tracks the number of months since the beginning of the dataset, allowing the models to discover changes in malarial patterns over time).

Some techniques to increase performance were tested but proved unreliable in overall accuracy. For example, a voting system was tried where a collection of models would predict either malaria incidence or classify potential outbreaks, but it proved ineffective. The features themselves were also refined as many initial variables proved unreliable at prediction and forecasting, including the percentage of bed nets, population, and poverty statistics.

For a more holistic analysis of both regression and classification models, cross-validation was used (ten folds) and the results were averaged. For the regression models, the SARIMAX proved most effective in all tested metrics from one to four months in advance, as seen in Tables 1 to 4. This could be attributed to the fact that the model is able to capture temporal patterns in malaria incidence, effectively adapting to the seasonality of the disease. A small decrease in performance over time was noted and some models proved similar in the long run to the SARIMAX, including the Random Forest Regressor and the LSTM. For the classification models, the LSTM had the highest accuracy, precision, and recall from one to four months, as seen in Tables 5 to 8. This could be due to the fact that LSTM models are well-suited for sequential data, picking up dependencies and patterns over long-periods of time in the dataset. In general, deep learning models like LSTM are quite effective at recognizing complex non-linear relationships in data, adapting to irregular patterns and the dynamics of malaria transmission. Once again, a decrease in accuracy over time was noted (5%), but unlike the regression models, no classification model seemed to perform close to the LSTM.

To further analyze results, for regression models, the predicted versus actual incidence was plotted, as seen in Figures 18, 19, 20, and 21. The SARIMAX was able to capture the trends in seasonality in malaria incidence, accounting for the rise in cases in the summer and the dip in cases later in the year. For the LSTM classification model, four confusion matrices were generated as seen in Figures 22 to 25, showing notable accuracy and a lack of false positives and negatives. A lack of bias by the model was also noted, showing the effectiveness of the SMOTE-ENN and the high performance in general of the LSTM.

Unpaired t-tests were used to further validate that the SARIMAX model was best for the regression task of forecasting malaria incidence and the LSTM model best for the classification task of prediction outbreaks in the future. Statistical significance ($p < 0.05$) was observed for both models from one to four months in advance, confirming their optimal performance.

Feature analysis was conducted to determine which climatic variables affected malaria transmission and incidence the most. When ranked by category (Figure 26), wind speed proved most important, potentially due to the movement patterns of the *Anopheles* mosquito, which primarily relies on wind to aid in travel. Wind can also disperse mosquito larvae to new water bodies, spreading transmission over a wider range of distance. Rainfall and humidity placed second and third respectively, potentially due to the fact that those two variables can lead to an increase in stagnant water, creating optimal breeding sites, allowing more mosquito larvae to grow and develop, leading to more vectors for transmission and thus, an increase in cases. Higher humidity levels have also been shown to extend the lifetime of mosquitoes over time, increasing the duration during which they can transmit the malaria parasite. Temperature seemed to have less of an impact, possibly because it typically remains stable in malaria endemic countries, especially in Abuja, diminishing its effect on predicting incidence, although it definitely has an impact on cases. Finally, the time feature, which tracked the number of months since the beginning of the dataset, had little to no impact

at all on cases, hinting at the lack of change in malaria prevention strategies over the past two decades as cases have stagnated, or even increased, over time. When ranked overall (Figure 27), wind speed and rainfall took six of the top seven features, with humidity (at lag one) being the final factor.

The SARIMAX model for regression and LSTM model for classification were implemented into *MalariVis*, an easily-accessible website and app for real-time prediction. The interface is designed for both clinician and civilian use, allowing users to forecast from one to four months in advance after inputting four basic climatic features. Clinicians have two additional features: the ability to input current malaria incidence to be used in future models for training and validation, and the opportunity to use past case counts as a feature, thus improving performance. The website itself includes critical information regarding malaria control and prevention, acting as a useful tool for millions of people living in malaria-endemic countries worldwide. Governments and health officials themselves can use *MalariVis* to properly allocate resources and prepare for future outbreaks, minimizing incidence counts and fatalities.

This project has filled the gap in the lack of malaria prevention strategies, shedding light on a new approach to the age-old issue. The accuracies of the SARIMAX and LSTM models have both surpassed past researches regardless of location, and the knowledge gained from the techniques used, models tested, and features employed will be critical in future studies. Further models can be tested, and new socioeconomic variables added, but in general, *MalariVis* lays the groundwork for the future of malaria prediction, forecasting, and prevention worldwide. Its applicability is unprecedented in the new era of data driven models, and the simple addition of pre-existing health records will quickly expand both geographic reach (new endemic regions including South Asia and South America, which share climatic characteristics with the study area) and predictive capability. *MalariVis* could be the future of malaria prediction, slowly turning the dream of eradication into a reality.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Centers for Disease Control and Prevention. (2023, June 28). *CDC - Malaria* . Centers for Disease Control and Prevention (CDC). https://www.cdc.gov/malaria/about/faqs.html#

E. Mbunge, R. C. Millham, M. N. Sibiya and S. Takavarasha, "Application of machine learning models to predict malaria using malaria cases and environmental risk factors," 2022 Conference on Information Communications Technology and Society (ICTAS), Durban, South Africa, 2022, pp. 1-5, doi: 10.1109/ICTAS53252.2022.9744657.

Fernando, S. D. (n.d.). *Climate Change and Malaria - A Complex Relationship*. United Nations. https://www.un.org/en/chronicle/article/climate-change-and-malaria-complex-relationship#:~:text=An%20increase%20in%20temperature%2C%20rainfall,it%20was%20not%20reported%20earlier

Harvey, D., Valkenburg, W., & Amara, A. (2021, June 18). Predicting malaria epidemics in Burkina Faso with machine learning. *PLOS ONE*, *16*(6). https://doi.org/10.1371/journal.pone.0253302

G. Kalipe, V. Gautham and R. K. Behera, "Predicting Malarial Outbreak using Machine Learning and Deep Learning Approach: A Review and Analysis," 2018 International Conference on Information Technology (ICIT), Bhubaneswar, India, 2018, pp. 33-38, doi: 10.1109/ICIT.2018.00019.

Lee, Y. W., Choi, J. W., & Shin, E.-H. (2021, February). Machine learning model for predicting malaria using clinical information. Computers in Biology and Medicine, 129. https://doi.org/10.1016/j.compbiomed.2020.104151

Marsh, K., & Snow, R. W. (1999). Malaria transmission and morbidity. Parassitologia, 41(1-3), 241-246. https://pubmed.ncbi.nlm.nih.gov/10697862/

Martineau, P., Behera, S. K., Nonaka, M., Jayanthi, R., Ikeda, T., Minakawa, N., Kruger, P., & Mabunda, Q. E. (2022). Predicting malaria outbreaks from sea surface temperature variability up to 9 months ahead in Limpopo, South Africa, using machine learning. *Frontiers in Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.962377

National Aeronautics and Space Administration. (2023). *POWER | DAVe*. NASA POWER. https://power.larc.nasa.gov/beta/data-access-viewer/

National Institute for Pharmaceutical Research and Development. (2023). National Institute for Pharmaceutical Research and Development (NIPRD). https://www.niprd.gov.ng/

Nkiruka, O., Prasad, R., & Clement, O. (2021). Prediction of malaria incidence using climate variability and machine learning. *Informatics in Medicine Unlocked*, *22*. https://doi.org/10.1016/j.imu.2020.100508

Segun, O. E., Shohaimi, S., Nallapan, M., Lamidi-Sarumoh, A. A., & Salari, N. (2020, May 16). Statistical Modelling of the Effects of Weather Factors on Malaria Occurrence in Abuja, Nigeria. *International Journal of Environmental Research and Public Health*, *17*(10), 3474. https://doi.org/10.3390/ijerph17103474

S. W. Lindsay & M. H. Birley (1996) Climate change and malaria transmission, Annals Tropical Medicine & Parasitology, 90:5, 573-588, DOI: 10.1080/00034983.1996.11813087

Tai, K.Y., Dhaliwal, J. Machine learning model for malaria risk prediction based on mutation location of large-scale genetic variation data. J Big Data 9, 85 (2022). https://doi.org/10.1186/s40537-022-00635-x

National Institute of Allergy and Infectious Diseases (2011, March 8). *Malaria Prevention, Treatment, and Control Strategies*. National Institute of Allergy and Infectious Diseases (NIAID). https://www.niaid.nih.gov/diseases-conditions/malaria-strategies

World Health Organization. (2023a). *Malaria*. World Health Organization (WHO). https://www.who.int/health-topics/malaria#tab=tab

World Health Organization. (2023b, October 2). *WHO recommends R21/matrix-M vaccine for malaria prevention in updated advice on immunization.* World Health Organization (WHO). https://www.who.int/news/item/02-10-2023-who-recommends-r21-matrix-m-vaccine-for-malaria-prevention-in-updated-advice-on-immunization