Machine Learning based Predictive Modelling for Disease Outbreak Threshold Estimation

Sristi Lakshmi Lalitha[1*], Kondameedi Ganesh[2], Mathangi Raj Kumar[2], Mirza Feroz Baig[2], Mohammad Sajid[2]

[1]Associate Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering (Data Science), Vaagdevi College of Engineering, Warangal, Telangana

*Corresponding Email: lalithayagnam03@gmail.com

Abstract

Malaria diagnosis relied heavily on manual microscopy, where a skilled technician examines blood smears under a microscope to identify and count malaria parasites. Traditional manual methods of malaria outbreak detection rely heavily on clinical reporting, manual surveillance, and reactive interventions after the spread has already begun. These systems are often delayed, data-deficient, and incapable of providing timely warnings. Such limitations result in increased transmission rates, delayed resource mobilization, and poor risk management. The objective of this work is to leverage machine learning models to develop a robust and proactive malaria outbreak prediction system based on mosquito species data and environmental features. The motivation stems from the need for faster, more accurate, and data-driven decision-making tools to predict outbreaks before they escalate. By automating the analysis using models like Decision Tree Regressor (DTR) and Multi-Layer Perceptron Regressor (MLPR), the proposed system offers high precision in predicting outbreak probabilities, as demonstrated by the MLPR model's R^2 score of 0.9494. This system not only forecasts potential outbreaks but also helps public health officials implement preventive measures proactively. The graphical analysis and predictive outputs provide actionable insights into species-specific risks and environmental triggers, enabling smarter allocation of healthcare resources. The proposed system addresses the shortcomings of manual methods by introducing an intelligent, scalable, and data-driven framework, ultimately contributing to improved public health response, reduced disease burden, and enhanced disease surveillance capabilities in malaria-endemic regions.

Keywords: Malaria diagnosis, Blood smear image analysis, Predictive modelling, Machine learning, Multilayer perceptron.

1. INTRODUCTION

Malaria remains one of the most significant public health challenges globally, particularly in regions like sub-Saharan Africa and South Asia. In India, malaria has been a persistent threat, with millions of cases reported annually, primarily in states like Odisha, Chhattisgarh, and Jharkhand. According to the World Health Organization (WHO), India accounted for nearly 4% of the world's malaria cases in 2020. The traditional method of diagnosing malaria involves microscopic examination of stained blood smears to identify and count the presence of *Plasmodium* parasites. This method, while effective, is highly dependent on the expertise of trained technicians and the quality of the equipment used, often leading to inconsistencies in diagnosis. The manual process is labor-intensive and time-consuming, making it less feasible in resource-limited settings where malaria is most prevalent. The motivation for this research is driven by the urgent need to overcome the limitations of traditional malaria diagnostics. With the global burden of malaria remaining high, particularly in low-income countries, there is a critical need for diagnostic tools that are not only accurate but also accessible and scalable. The goal is

to harness the power of ML learning to create a diagnostic tool that can operate efficiently in various settings, providing rapid and reliable results that can enhance early detection and treatment of malaria.

Adithya Rajnarayanan et al. [1] (2024) they proposed an Artificial Neural Network (ANN) is applied to predict the compartmental trajectories following mathematical analysis. SIR (Susceptible-Infected-Recovered) and SEIR (Susceptible-Exposed-Infected-Recovered), are the existing systems they used to develop their model. For that they got 60% of accuracy to their existing method and 71% for their proposed method. But they failed to introduce adoptability to sudden changes in their proposed model. F Grignaffini et al. [2] (2024) Proposes a CAD system enhanced with deep learning architectures, specifically CNNs, for automatic detection and classification of malaria parasites with 84.5% accuracy whereas coming to their existing system they have used image processing techniques combined with basic machine learning classifiers like KNN etc… with 54% of accuracy. Even though they proposed an accurate model but they have failed to incorporate the nonlinear factors. Khodadadi et al. [3] (2023) they existing system was manual reporting and statistical methods for monitoring disease patterns but those systems took lot of time to give an outbreak analysis for that resoan they developed a model by using SVM (Support Vector Machine) which is a machine learning algorithm and this overcome the drawbacks they faced with their existing systems and got the best accurate results.

Ileperuma et al. [4] (2023) their predictive model offered early warning signals for malaria outbreaks, which allowed authorities to implement preventive measures and allocate resources in advance. Coming to their traditional methods for predicting malaria prevalence are often rely on statistical techniques like linear regression or time series analysis. These models used historical malaria incidence data and climate factors like humidity and rainfall etc… to estimate future prevalence. Which gave around 15% less accuracy than their proposed system. Their proposed system even suffered to give the output for Additional Environmental and Socioeconomic Factors. Francisca Chibugo Udegbe et al. [5] (2023) Traditional diagnostic approaches used by them are often constrained by delayed data processing and analysis, fall short in rapidly evolving scenarios like infectious disease outbreaks. Recent advancements in real-time data integration are addressing these limitations, offering significant improvements in predictive accuracy and timely response. They got 55% of accuracy for their proposed system, but it limits to a small datasets. They unable to handle the large dataset.

TD Keno et al. [6] (2022) in this study they have taken vector-borne diseases (VBDs) as their existing system, but they result in lack of precision, for that reason they have proposed a model based on VBD which will take the real-time data as input and produced the high accurate output than their existing with the accuracy of around 33%. But their proposed system cannot integrate the Socioeconomic Factors. Elliot m bunge et al. [7] (2022) they applied logistic regression, decision trees classifier, support vector machine, and random forest classifier to predict malaria in Buhera district. The study shows that logistic regression performs better, with 43% accuracy, 52% precision and 90% F1-score than other machine learning classifiers when predicting malaria outbreaks using environmental risk factors. Mazni Baharom et al. [8] (2021) their existing models generally consider malaria as endemic to certain regions and track incidence through hospital records or community-based health data collection without advanced climate-driven analysis. They have used Fuzzy Logic Suitability (FLS) Model which can give 20% better accuracy than the traditional systems.

Johnson et al. [9] (2021) the model presented a detailed analysis of machine learning approaches like SVM(support vector machine) and decision tree in outbreak detecting area to give a lead to the researchers who want to work in this area. The traditional malaria surveillance their work is public health facilities to gather data on malaria cases. It uses a web-based National Health Management Information System (NHMIS) for data reporting. Even though they have developed a great model but they did not provide a data security system in their model.

doi: 10.48047/ijiee.2025.15.4.14

C. Giesen et al. [10] (2020) considering all articles indexed in PubMed, Scopus, Embase and CENTRAL. Search terms referring to MBD, CC and environmental factors were screened in title, abstract and keywords.Results A total of twenty-nine studies were included, most of them on malaria (61%), being Anopheles spp. (61%) the most commonly analyzed vector, mainly in Eastern Africa (48%). Seventy-nine percent of these studies were based on predictive models. Hernandez et al. [11] (2020) they collected secondary data in September 2017 from published articles and journals on this major issue to discuss the effects of climate change on VBDs in India through this article. This paper described briefly review the changing epidemiology of the most important vector-borne diseases in India. The proposed system integrates climate data and geographic information systems (GIS) into a predictive malaria surveillance. This system gives the advanced predictive modeling and real-time climate monitoring optimize malaria interventions.

Flores et al. [12] (2020) this study employed a mechanistic model to simulate mosquito population dynamics under various greenhouse gas emission and land-cover change scenarios based on climate data provided by a state-of-the-art regional climate model. Our results show a 12.6% decrease in the annual mosquito population in newly urbanized areas and a 5.9% increase in the annual mosquito population in existing urban areas. Furthermore, changing climate conditions are worked to cause a 15–17% reduction in the total annual mosquito population; however, the change will not be uniform throughout the year.

OJ Matthew et al. [13] (2020) they used multivariate regression analysis and lag correlation up to 4 months were performed to examine contributions of climate variation to the reported malaria cases. Results revealed that mosquitoes could survive all-year round with p values ranging between 0.40 and 0.96 under the prevailing mean climate. However, the climate suitability level for transmission of malaria was ‘moderate’ ($0.45 < p \leq 0.60$) in the dry season but ‘very high’ ($0.75 < p \leq 0.96$) in the wet. TI Visa et al. [14] (2020) the surveillance system was evaluated using the "2001 United States Centers for Disease Control's updated guidelines for Evaluating Public Health Surveillance Systems". Data were described using means, standard deviation, frequencies and proportions. Chi-squared for linear trends was used. They developed a model by using KNN and SVM with better accuracy also they have some gaps like still there is a risk of overfitiing in their work.

Kumar et al. [15] (2020) the Epidemic Prognosis Incorporating Disease and Environmental Monitoring for Integrated Assessment (EPIDEMIA) computer system was designed and implemented to integrate disease surveillance with environmental monitoring in support of operational malaria forecasting in the Amhara region of Ethiopia. For this they have used K mean algorithm to group similar malaria data points and detect outliers. The algorithm identifies anomalies by comparing new data with established clusters, flagging areas where malaria activity deviates from normal patterns. But earlier they used manual data collection which is unable to detect early-stage anomalies. The K-means algorithm also have a drawback that they will provide the false alarms. A, Khan et al. [16] (2020) the proposed future of this survey effects of rising temperatures on endemicity are at least one order of magnitude smaller than changes observed since about 1900 and up to two orders of magnitude smaller than those that can be achieved by the effective scale-up of key control measures by using XGboost algorithm. Predictions of an intensification of malaria in a warmer world, based on extrapolated empirical relationships or biological mechanisms, must be set against a context of a century of warming that has seen marked global declines in the disease and a substantial weakening of the global correlation between malaria endemicity and climate.

S Brugueras et al. [17] (2019) they used Rapid Diagnostic Tests (RDTs) as an existing method to detect malaria antigens in the blood via immunochromatographic tests with 34% of accuracy whereas ANN is the proposed method for their work with 52% of accuracy and also it can’t provide any security. Odu

Nkiruka et al. [18] (2019) this research proposes a machine learning-based model for the classification of malaria incidence using climate variability across six countries of Sub-Saharan Africa over a period of twenty-eight years. The work begins with a feature engineering process, which identifies the climate factors that affect the incidence of malaria, followed by the k means clustering process for outlier detection, and then, XGBoost algorithm for classification.

2. PROPOSED SYSTEM

ML learning-based approach for analyzing malaria infection data. It starts by preprocessing the dataset, including resampling and encoding categorical variables using LabelEncoder. The data is split into training and testing sets, and various machine learning models like Decision Tree and MLP (Multi-Layer Perceptron) are trained to predict malaria outbreak thresholds. Metrics like Mean Squared Error (MSE) and R^2 score are calculated to evaluate model performance. Finally, the trained models are used to make predictions on a test dataset, with results saved and displayed.

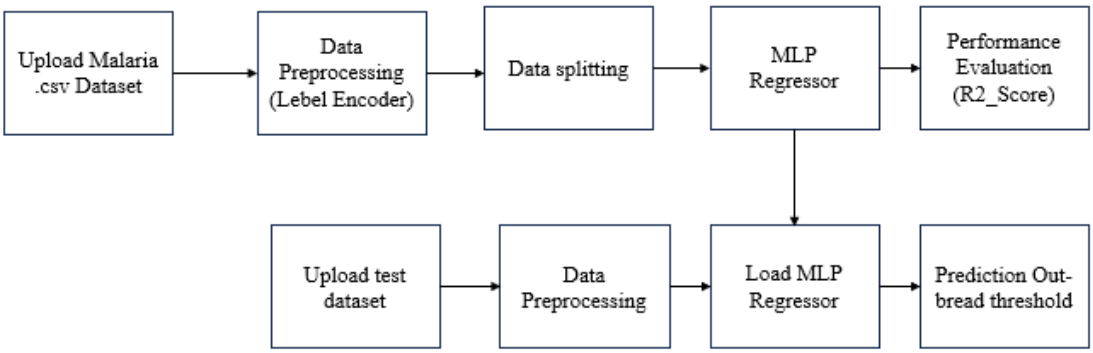

Figure 1: Block diagram of proposed malaria prediction system.

Step 1: Dataset

- The dataset "malaria.csv" is utilized, which contains information related to malaria infection, including features such as mosquito species and outbreak thresholds. This dataset is the foundation for training and testing machine learning models.

Step 2: Dataset Preprocessing

- **Null Value Removal:** The dataset is checked for any missing or null values, which are removed to ensure data quality.
- **Label Encoding:** The categorical features within the dataset are converted into numerical values using label encoding. This is essential for machine learning models to process the data effectively.

Step 3: Label Encoder

- A label encoder is applied to all categorical columns within the dataset to transform them into numerical values. This step ensures that the data can be fed into machine learning models without any issues.

Step 4: Data Splitting

- The dataset is split into training and testing sets using an 80-20 split. This is crucial for evaluating the model's performance on unseen data.

Step 5: Existing Model (Decision Tree Regressor)

- A Decision Tree Regressor model is trained on processed data. This model serves as a baseline to compare against the proposed model.
- The model is evaluated using metrics such as Mean Squared Error (MSE) and R^2 Score.

Step 6: Proposed Model (MLP Regressor)

- The proposed model is a Multi-Layer Perceptron (MLP) Regressor, which is a type of ML learning model.
- This model is trained on the same dataset and compared against the Decision Tree Regressor.
- The MLP Regressor is expected to provide better accuracy and performance due to its complex architecture.

Step 7: Performance Comparison

- The performance of the Decision Tree Regressor and MLP Regressor is compared using metrics such as Mean Squared Error (MSE) and R^2 Score.
- The comparison helps in determining the effectiveness of the proposed ML learning model over the traditional machine learning model.

Step 8: Prediction of Output from Test Data (Using MLP Model)

- The trained MLP model is used to predict outcomes on new test data.
- The predictions are compared with the actual values to evaluate the model's accuracy and reliability.
- The results are then added to the test dataset, which now includes the predicted values.

2.1 ML Model Building

Once the model is selected, it is trained using the training dataset. During training, the model learns the relationships between the input features and the output labels by adjusting its internal parameters. This process involves feeding the data into the model, computing the error (difference between predicted and actual outcomes), and optimizing the model parameters to minimize this error. Training is typically done iteratively, with the model improving its predictions over time. Techniques like cross-validation can be used during training to tune hyperparameters and prevent overfitting, ensuring the model generalizes well to new data.

A Multi-Layer Perceptron (MLP) is a class of feedforward artificial neural network (ANN) that consists of multiple layers of nodes. It is widely used for both regression and classification tasks. Each node (or neuron) in the network represents a mathematical function that computes a weighted sum of the inputs and applies an activation function to produce the output.

How It Works: An MLP operates by passing input data through multiple layers of neurons. Each layer in the network consists of several neurons, and each neuron in a layer is connected to every neuron in the subsequent layer, making the network "fully connected."

1. **Input Layer:** Receives the input data. The number of neurons in this layer corresponds to the number of features in the dataset.
2. **Hidden Layers:** One or more intermediate layers where the network performs most of its computations. Each neuron in these layers processes the inputs from the previous layer and

passes the output to the next layer. The weights of these connections are adjusted during the training process to minimize the error in predictions.

3. **Output Layer:** Produces the final output of the network. In regression tasks, this is typically a single neuron that provides the predicted continuous value.

Training an MLP involves adjusting the weights of the connections between neurons using a process called backpropagation. Backpropagation calculates the error at the output and propagates it backward through the network to update the weights, typically using an optimization technique like gradient descent.

Architecture:

- **Input Layer:** Neurons correspond to the input features.
- **Hidden Layers:** Consist of neurons connected with activation functions like ReLU (Rectified Linear Unit), Sigmoid, or Tanh that introduce non-linearity.
- **Output Layer:** A single neuron in the case of regression, which outputs the predicted value.

The architecture of an MLP can vary in terms of the number of hidden layers and the number of neurons per layer. A MLP network (with more hidden layers) can capture more complex patterns but require more computational resources and careful tuning to prevent overfitting.

3. RESULTS AND DISCUSSION

The dataset provided is designed to predict malaria outbreaks based on various environmental, biological, and healthcare-related factors. The outbreak_threshold is the output variable, representing the likelihood of a malaria outbreak in the area.

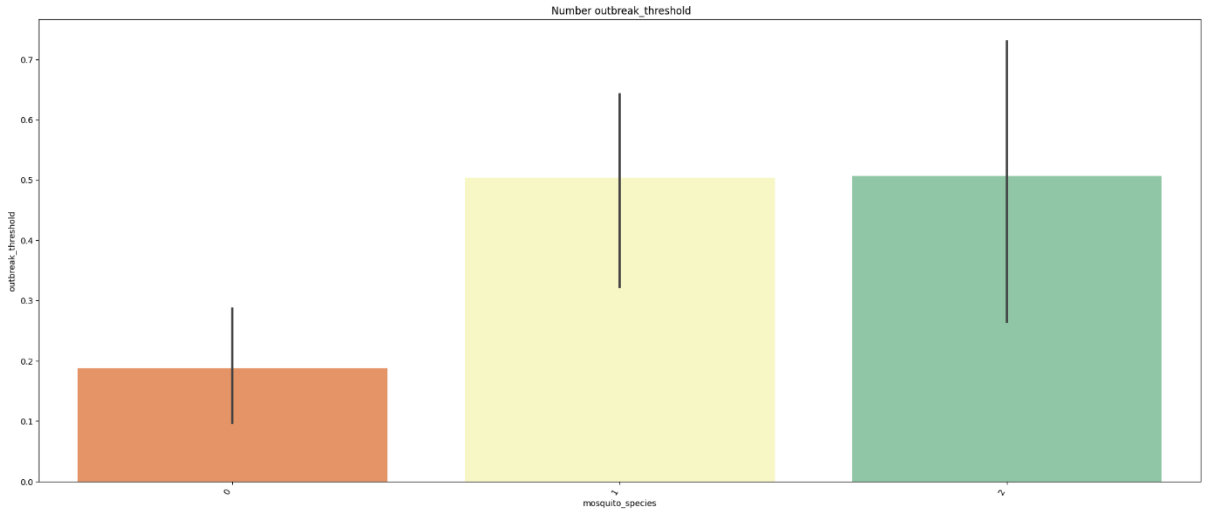

Figure 2: Mosquito species with outbreak.

This value ranges from 0 to 1, where a higher value indicates a higher probability of an outbreak. The dataset comprises seventeen categorical predictors and one continuous target. The predictors describe environmental, entomological, and infrastructural factors—temperature level (Low/High), humidity (Low/High), precipitation (Yes/No), historical outbreak frequency (High/Low), and human travel patterns (High/Low); dominant mosquito species (e.g., Aedes aegypti, Anopheles gambiae, Culex quinquefasciatus) alongside a binary indicator of whether those species are known malaria vectors; vector infection rates (Low/High); healthcare availability and accessibility (each Yes/No) and local malaria treatment success rates (Low/High); land use type (Urban/Rural); presence of mosquito breeding sites and stagnant waters (each Yes/No); IoT-measured air and water quality (Good/Bad); and presence of vegetation (Yes/No). The sole target variable, outbreak_threshold, is a real number between

0 and 1 reflecting the modeled probability of a malaria outbreak in the area. Figure 2 shows that Mosquito Sp ecies (Anopheles gambiae, Culex quinquefasciatus, Aedes aegypti) with outbreak threshold and Anop heles gambiae is very less threshold value as compare to both. Figure 3 shows that the scatter plot of D TR which tell very low accuracy and our predicated line is not good.

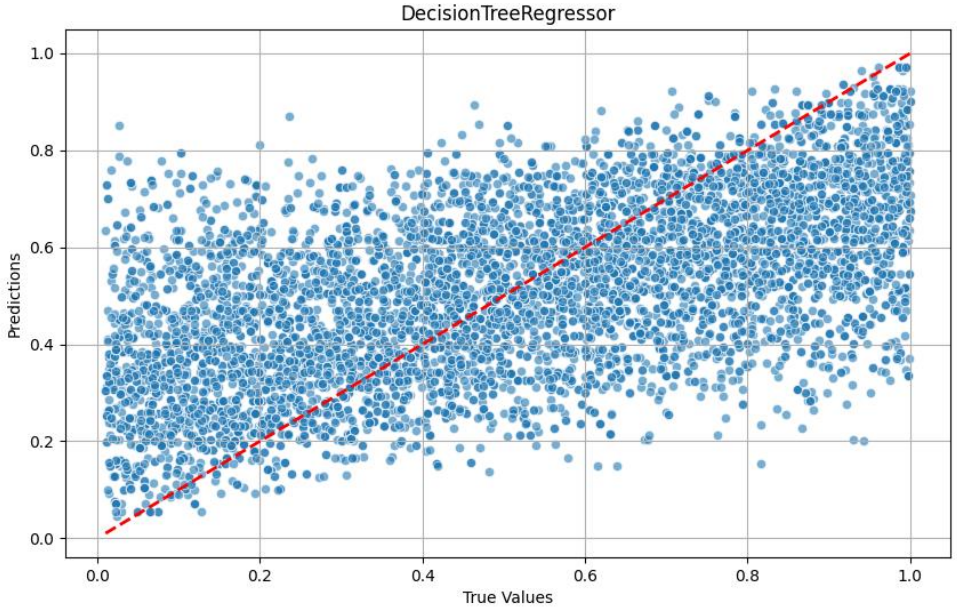

Figure 3: Scatter plot of DTR.

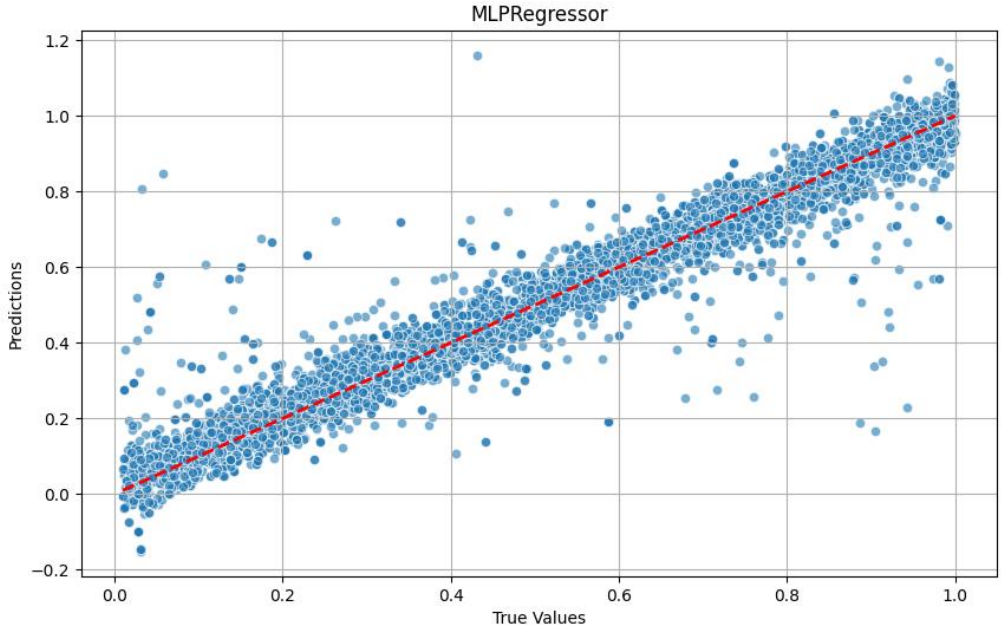

Figure 4: Scatter plot of MLPR.

Figure 4 is the scatter plot of the MLP Regressor which has the very best fit line as compared to the D TR. The MLP Regressor demonstrated significantly better performance than the Decision Tree Regres sor on the same test set of 7,800 records. It achieved a much lower Mean Absolute Error (MAE) of 0.0 400, indicating smaller average prediction errors. The Mean Squared Error (MSE) of 0.0042 and Root Mean Squared Error (RMSE) of 0.0649 are considerably smaller than those of the Decision Tree Regr essor, highlighting the model's higher precision and reduced variance in predictions. Moreover, the R-squared (R^2) value of 0.9494 signifies that the MLP Regressor explains 94.94% of the variability in th

doi: 10.48047/ijiee.2025.15.4.14

e data, showcasing its strong predictive accuracy and a marked improvement over the Decision Tree R egressor's 37.37% R^2. These metrics collectively establish the MLP Regressor as a much more effectiv e model for this task.

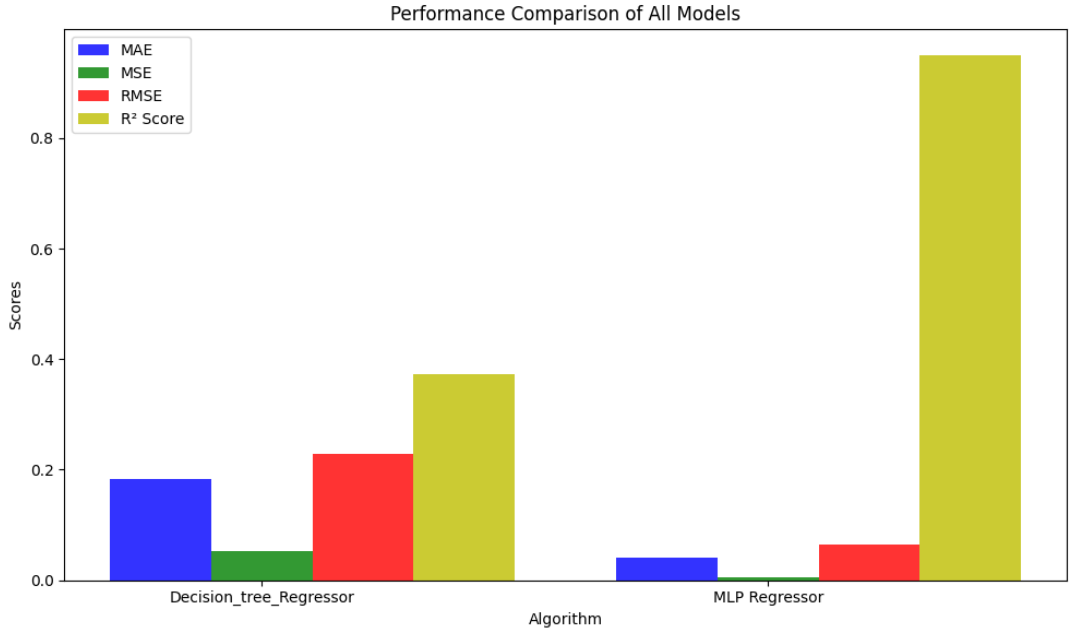

Figure 5: Comparison graph.

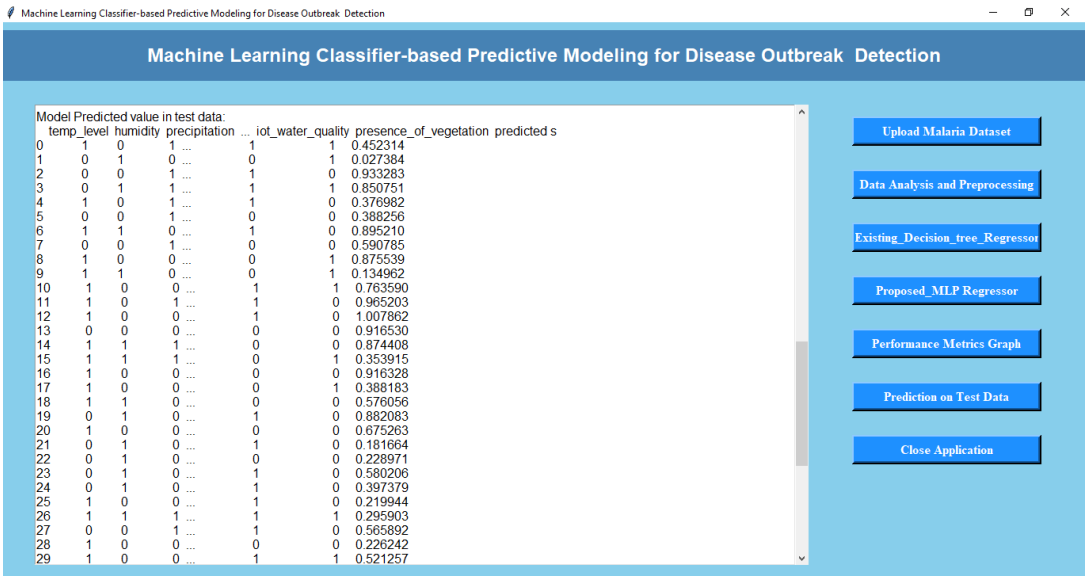

Figure 6: Predicted output.

Figure 6 shows that the predict column in your dataset represents a numerical value that corresponds to a prediction generated by a machine learning model. In the context of malaria infection diagnosis or risk prediction, this value could signify various outcomes depending on the model's purpose. Here are a few possible interpretations Outbreak threshold value The predicted value represents the predicted probability of a malaria outbreak occurring in the area under the given conditions. For example, a value of 0.452314 indicates a 45.23% likelihood of a malaria outbreak.

doi: 10.48047/ijiee.2025.15.4.14

Table 1: Comparative Analysis: Decision Tree Regressor vs. MLP Regressor.

Metric	Decision Tree Regressor	MLP Regressor
Mean Absolute Error (MAE)	0.1829	0.0400
Mean Squared Error (MSE)	0.0521	0.0042
Root Mean Squared Error	0.2283	0.0649
R-squared (R^2)	0.3737	0.9494

The comparative analysis between the existing Decision Tree Regressor (DTR) and the proposed MLP Regressor clearly demonstrates that the MLP Regressor significantly outperforms the DTR across all performance metrics. The Mean Absolute Error (MAE) and Mean Squared Error (MSE) for the MLP Regressor are notably lower (0.0400 and 0.0042, respectively) compared to those of the DTR (0.1829 and 0.0521), indicating much higher prediction accuracy. Similarly, the Root Mean Squared Error (RMSE) is substantially reduced from 0.2283 in the DTR to 0.0649 in the MLP Regressor, highlighting improved consistency in predictions. Most importantly, the R-squared (R^2) value for the MLP Regressor is 0.9494, showing a very strong fit to the data, in contrast to the DTR's R^2 of 0.3737, which indicates a weaker model. These results suggest that the proposed MLP Regressor is a far superior model in terms of accuracy, reliability, and overall performance.

4. CONCLUSION

This study has successfully demonstrated the application of machine learning techniques for predicting malaria outbreaks by analyzing mosquito species data and relevant environmental factors. A comparative analysis between the Decision Tree Regressor (DTR) and the Multi-Layer Perceptron Regressor (MLPR) was conducted to evaluate the efficiency and accuracy of predictive models. The Decision Tree Regressor, although straightforward and easy to interpret, showed limited performance with a Mean Absolute Error (MAE) of 0.1829 and a low R^2 score of 0.3737, indicating that it could only explain about 37.37% of the variability in the dataset. On the other hand, the MLPR significantly outperformed the DTR across all evaluation metrics. It achieved a remarkably low MAE of 0.0400 and an R^2 score of 0.9494, which suggests it could explain nearly 95% of the variance in the outcome variable. These results affirm that neural network-based models, due to their deep learning capabilities and ability to model non-linear relationships, are far more effective for complex prediction tasks such as disease outbreak forecasting. Furthermore, the graphical representations such as scatter plots and performance comparison charts reinforce the superior performance of MLPR in fitting and generalizing the patterns in the data. The visual analysis of species like *Anopheles gambiae*, *Culex quinquefasciatus*, and *Aedes aegypti* in relation to outbreak thresholds provides crucial insights into vector-specific risks. The predicted output values generated by the MLPR model serve as a numerical estimation of outbreak probabilities, which could serve as an early indicator for public health departments. Overall, this work demonstrates how artificial intelligence and machine learning can be vital tools in healthcare analytics, offering timely and data-driven decision support in the fight against vector-borne diseases such as malaria.

REFERENCES

[1] Adithya Rajnarayanan, Manoj Kumar, "Analysis of a mathematical model for malaria using data-driven approach", vol. 8 published 1 Sep 2024.

[2] F Grignaffini, P Simeoni, A Alisi 2024 "Computer-Aided Diagnosis Systems for Automatic Malaria Parasite Detection and Classification", Volume 19, published 11 August 2024, page 10-18.
[3] Khodadadi, Ehsaneh; Towfek, S. K, "Internet of Things Enabled Disease Outbreak Detection: A Predictive Modeling System.", Vol 10, published 2023, p84.
[4] Ileperuma, Kaveesha Jampani, "Predicting malaria prevalence with machine learning models using satellite-based climate information". First published 2023.
[5] Francisca Chibugo Udegbe, Ejike Innocent Nwankwo, "Real-Time data integration in diagnostic devices for predictive modeling of infectious disease outbreaks", Volume 4, published December 2023, p.525-545.
[6] TD Keno, LB Dano, GA Ganati, "The effect of climate variability on malaria transmission". vol. 6, published 29 June 2022, page no. 298-302.
[7] Elliot m bunge, Richard c milham, "application of machine learning to predict malaria using malaria cases and environmental risk factors", vol. 2, published 2022, pp.100 168.
[8] Mazni Baharom, Norfazilah Ahmad, "The Impact of Meteorological Factors on Communicable Disease Incidence and Its Workion: A Systematic Review", Vol 18, Published: 22 October 2021.
[9] Johnson, A, "The study Using machine learning for early detection of malaria outbreaks", Volume: 14, published: 2021, p.355 – 366.
[10] C Giesen, J Roche, "The impact of climate change on mosquito-borne diseases in Africa.", Vol. 3 Published online: 25 Jun 2020, Pages 287-301
[11] Hernandez, D, "Research on Climate change and its effects on vector-borne diseases", Vol 21, published 2020, pages479–483
[12] Flores, R, "impact of urbanization on malaria dynamics in tropical regions", vol. 46 , published 2020, p.4275–4282.
[13] OJ Matthew, "Investigating climate suitability conditions for malaria transmission and impacts of climate variability on mosquito survival in the humid tropical region",published 2020.
[14] TI Visa, O Ajumobi, "Evaluation of malaria surveillance system in Kano State", Volume 9, published 10 February 2020.
[15] Kumar, "The Integrating machine learning and remote sensing for malaria surveillance", published 2020.
[16] Khan, A, "Climate change and malaria - A global perspective", volume 465, published 20 2020, pages 342–345
[17] S. Brugueras, B. Fernández-Martínez, "Environmental drivers, climate change and emergent diseases transmitted by mosquitoes and their vectors in southern Europe: A systematic review", Volume 191, December 2020
[18] Odu Nkiruka, Rajesh Prasad, "Prediction of malaria incidence using climate variability and machine learning", Volume 22, published 2019, 100-201.