Journal for Educators, Teachers and Trainers
The PROFESIOLab electronic, peer-reviewed, open-access Magazine

ISSN 1989-9572

DOI:10.47750/jett.2024.15.05.43

ML REGRESSION-BASED PREDICTIVE MODELING FOR DISEASE OUTBREAK THRESHOLD ESTIMATION

G. Prabhakar1, Spurthi Vadlakonda 2, Vanam snehitha2, Sravani Siddam2

Journal for Educators, Teachers and Trainers, Vol.15(5)

https://jett.labosfor.com/

Date of Reception: 24 Oct 2024

Date of Revision: 20 Nov 2024

Date of Publication : 31 Dec 2024

G. Prabhakar1, Spurthi Vadlakonda 2, Vanam snehitha2, Sravani Siddam2 (2024). ML REGRESSION-BASED PREDICTIVE MODELING FOR DISEASE OUTBREAK THRESHOLD ESTIMATION. *Journal for Educators, Teachers and Trainers*,Vol.15(5).435-443

Journal for Educators, Teachers and Trainers, Vol. 15(5)

ISSN1989 –9572

ML REGRESSION-BASED PREDICTIVE MODELING FOR DISEASE OUTBREAK THRESHOLD ESTIMATION

G. Prabhakar[1], Spurthi Vadlakonda [2], Vanam snehitha[2], Sravani Siddam[2]

[1]Assistant Professor, [2]UG Student, [1,2]School of Computer Science and Engineering

[1,2]Malla Reddy Engineering College for Women (UGC-Autonomous), Maisammaguda, Hyderabad, 500100, Telangana.

ABSTRACT

Malaria diagnosis relied heavily on manual microscopy, where a skilled technician examines blood smears under a microscope to identify and count malaria parasites. This method, established in the early 1900s, has been the gold standard for malaria diagnosis but is labor-intensive, time-consuming, and requires significant expertise. In regions with limited healthcare resources, this has often led to misdiagnosis or delayed treatment. The objective of this study is to leverage ML learning techniques to develop an automated, accurate, and efficient diagnostic tool for detecting malaria infections from medical images, thereby improving diagnostic accuracy and reducing the time required for analysis. The title "ML Learning-Based Analysis for Malaria Infection Diagnosis" refers to the application of ML learning algorithms, a subset of machine learning, to analyze medical images and diagnose malaria infections. The approach aims to automate the detection process, making it faster and more reliable compared to traditional methods. Before the advent of machine learning or AI, the primary method for diagnosing malaria was manual microscopy, as mentioned earlier. This involved staining blood smears with special dyes, followed by meticulous examination under a microscope. The accuracy of this method largely depended on the technician's experience and the quality of the equipment, which could vary significantly, especially in low-resource settings. Traditional microscopy for malaria diagnosis, while effective, has several limitations, including the need for skilled personnel, the potential for human error, and the slow turnaround time for results. The motivation for this research stems from the need to address the shortcomings of traditional malaria diagnostic methods. With the global burden of malaria remaining high, especially in low-income countries, there is a pressing need for diagnostic tools that are not only accurate but also accessible and scalable. The proposed system involves the development of a ML learning model trained on a large dataset of labeled blood smear images. This model will automatically detect and classify malaria parasites in the images, offering a quick and accurate

diagnosis. By reducing reliance on human expertise, this system can provide consistent results across different settings, enhance early detection, and enable prompt treatment, ultimately contributing to better malaria control and eradication efforts.

KEYWORDS : Blood Smears, Diagnostic Accuracy, Parasite detection, Automated diagnosis

1. INTRODUCTION

The integration of ML learning into medical diagnostics offers a revolutionary approach to disease detection. By automating the analysis of medical images, ML learning can significantly improve the speed and accuracy of malaria diagnosis. This technology is particularly vital in regions with limited access to skilled healthcare professionals, where rapid and reliable diagnostics can save lives.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.

In real-world scenarios, especially in remote or underdeveloped regions, access to trained medical personnel and advanced diagnostic equipment is often limited. This leads to delays in diagnosis and treatment, contributing to higher morbidity and mortality rates. An automated, ML learning-based diagnostic tool can bridge this gap by providing a reliable, fast, and accessible means of diagnosing malaria. It can be deployed in clinics, hospitals, and even mobile health units, ensuring that patients receive accurate diagnoses and timely treatment, regardless of their location.The applications of this project extend beyond just malaria diagnosis. The developed ML learning model can be adapted to diagnose other parasitic infections by training it on relevant medical images. In addition, this technology can be integrated into telemedicine platforms, allowing remote diagnostics in areas with limited healthcare infrastructure. Healthcare providers can use this tool to screen large populations quickly, improving public health outcomes through early detection and treatment. Furthermore, the system can be used in research settings to analyze large datasets of blood smears, contributing to epidemiological studies and the development of new treatment strategies. Ultimately, this project has the potential to revolutionize infectious disease diagnostics, making it a vital tool in the global fight against malaria and other parasitic diseases.

2. LITERATURE SURVEY

David H. et al. [1] provided a comprehensive overview of the principles of data mining, focusing on its application across various domains, including healthcare. Koh H.C. et al. [2] emphasized the potential of data mining techniques in enhancing healthcare decision support systems, demonstrating how these methods can improve patient care and optimize resource management. Building on these concepts, Tomar D. et al. [3] conducted a survey on data mining approaches specifically within the healthcare sector, detailing various methodologies and their effectiveness in clinical settings. The World Health Organization [4] highlighted the global burden of malaria, particularly in Africa, and the need for advanced diagnostic tools.

Roca-Feltrer A. et al. [5] estimated the malaria morbidity in African children under five years, stressing the importance of effective disease management and prevention strategies, which could be informed by

data mining and analysis. Ibrahim et al. [6] in their study compared different classification techniques using WEKA for breast cancer. The aim of the study is to investigate the performance of different classification methods for a set of large datasets. The algorithms tested are Bayes Network, Radial Basis Function, Pruned Tree, Single Conjunctive Rule learner and Nearest Neighbours. The best algorithm on the breast cancer data sets is Bayes network classifier with the highest accuracy and lowest average error. Boris and Milan [7] performed prediction and decision making in healthcare using data mining. They analysed the usefulness of data mining in the healthcare sector and some of the obstacles that disable the effective and efficient prediction. Sharma et al. [8] in their study presented malaria outbreak prediction model using machine learning. In this study, they stated that the early prediction of malaria outbreak is the key to the control of malaria morbidity. This prediction can help as an early warning tool to identify potential outbreaks of malaria. The machine learning used for the data mining was classification algorithms based on support vector machine (SVM) and artificial neural network (ANN). Also, the total number of Plasmodium falciparum cases and an outbreak occurs in binary values yes or no. Root mean square error (RMSE) and receivers operating characteristics (ROC) were used to measure the performance of the models. Kapor and Rani [9] employed an efficient decision tree algorithm using J48 and reduced error pruning. In the study, decision trees were utilized to delineate decision-making process. The decision tree builds classification or regression models in the form of the tree structure, which divides the datasets into tinier and tinier subsets. Some of the benefits and limitations of the decision trees where highlighted. The paper introduces a new decision tree algorithm based on J48 and reduced error pruning. The tree obtained is fast decision tree learning and will be based on the information gain or reducing the variance.

Leopard et al. [10] worked on survey and analysis on classification and regression data mining techniques for disease outbreak prediction in datasets. In this study, the need to develop a strong model for the prediction of disease outbreak in various countries using data mining algorithms was discussed. The advantages and disadvantages of the different classification techniques were highlighted and also the accuracy measures in decision trees from previous publications from the year 2001 to 2014 were presented. Bbosa et al. [11] studied clinical malaria diagnosis: ruled-based classification statistical prototype. In the study, they were able to identify the predictors of malaria, developed data mining, statistically enhanced rule-based classification to diagnose malaria and developed an automated system to incorporate the rules and the statistical models. The prototype was evaluated for efficacy showing a sensitivity value of 70% across the age groups. They also presented tables for malaria prevalence, signs and symptoms of both hospital and diagnosis. Witten I.H. et al. [12, 15] discussed the practical aspects of data mining and machine learning, introducing tools and techniques essential for extracting meaningful patterns from large datasets, with a particular focus on the WEKA workbench [20]. Cao X. et al. [13] applied data mining techniques to analyze cancer vaccine trials, offering a high-level overview of the data's impact on immunology research. Cios K.J. et al. [14] highlighted the unique challenges of medical data mining, emphasizing its complexity and the need for specialized methods. Han J. et al. [15] explored advanced concepts and techniques in data mining, providing a comprehensive guide to handling large datasets, while Quinlan R.

3. 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.

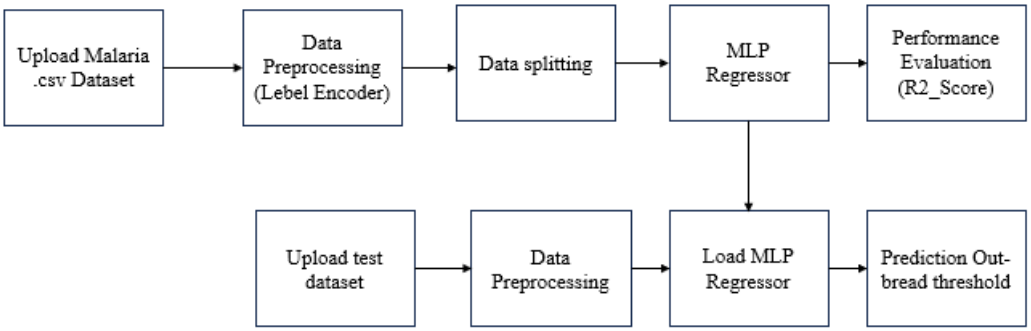

Figure 1:Proposed Block Diagram of Disease Prediction

The dataset is divided into two subsets: a **training set** and a **testing set**. Typically, this split is done using an 80-20 ratio, where 80% of the data is used to train the machine learning models, and the remaining 20% is reserved for testing and evaluating the model's performance on unseen data. This ensures that the model's accuracy is not just due to memorization of the training data but can generalize well to new data.

4. RESULTS AND DISCUSSION

The dataset is first loaded into a Pandas DataFrame, which contains various features relevant to malaria diagnosis. The data is then resampled to balance the dataset, ensuring that the model is trained on a representative sample of the data. Resampling is crucial for handling imbalanced datasets, where certain classes may be underrepresented. The next step involves preprocessing the data, which includes handling categorical variables through Label Encoding. This process converts categorical data into numerical values, allowing machine learning models to process the data effectively. The data is then split into features (X) and the target variable (y), where X represents the input features, and y represents the output or labels that the model aims to predict.

The dataset is split into training and testing sets using an 80-20 split. The training set is used to train the machine learning models, while the testing set is used to evaluate their performance. Splitting the data ensures that the model is tested on unseen data, providing an accurate assessment of its performance in real-world scenarios. To gain insights into the data, exploratory data analysis is conducted. This involves visualizing the relationship between various features, such as mosquito_species and outbreak_threshold. Visualization helps in understanding the data distribution, identifying patterns, and detecting any anomalies that might affect model performance.Two models are trained in this implementation: a Decision Tree Regressor and a Multi-Layer Perceptron (MLP) Regressor. Both models are trained on the training dataset, where the Decision Tree Regressor builds a tree-like structure to make predictions, and the MLP Regressor uses a ML learning approach with multiple hidden layers to learn complex patterns in the data.After training, the models are evaluated on the testing dataset using metrics such as Mean Squared Error (MSE) and R^2 Score. These metrics help assess the model's accuracy and its ability to generalize to new data. If the models perform well, they are saved for future use, ensuring that the training Finally, the trained models are used to make predictions on new, unseen test data. This step simulates the model's real-world application, where it predicts malaria infection status based on new patient data. The predictions are then stored and can be used for further analysis or integrated into a diagnostic system.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. This value ranges from 0 to 1, where a higher value indicates a higher probability of an outbreak.

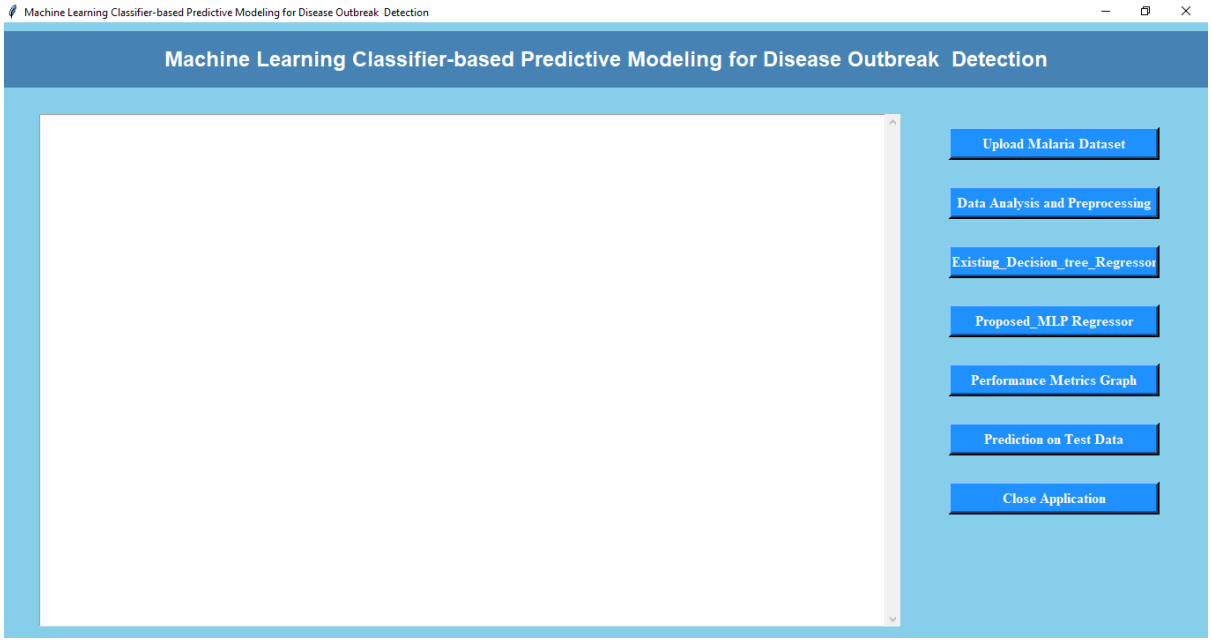

Figure 2: GUI

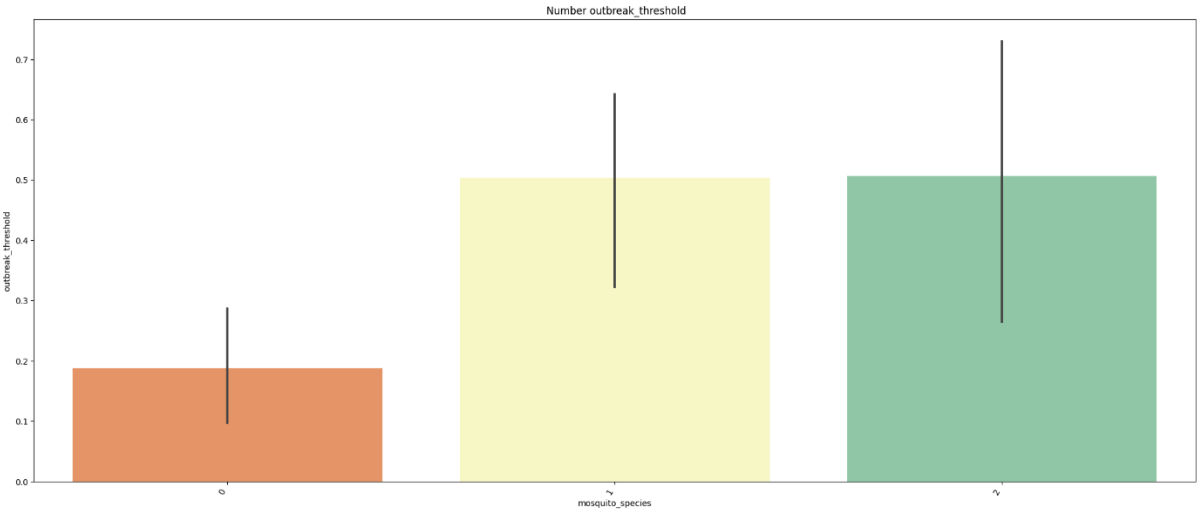

Figure 3: Mosquito Species with outbreak

Figure 3 shows that Mosquito Species (Anopheles gambiae, Culex quinquefasciatus, Aedes aegypti) with outbreak threshold and Anopheles gambiae is very less threshold value as compare to both.

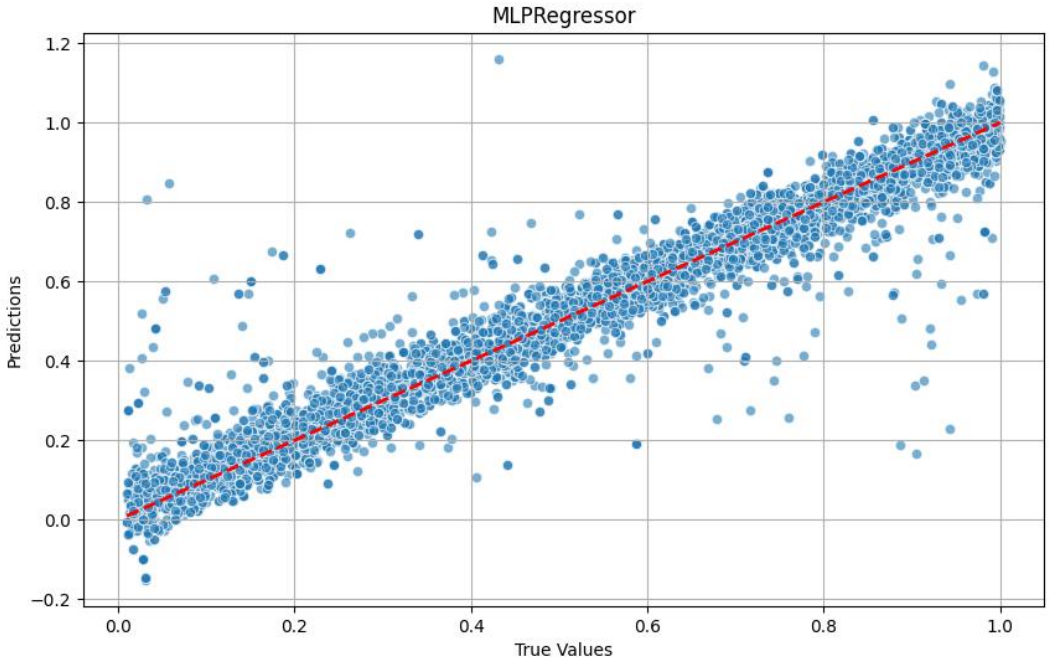

Figure 4: Metrics of MLP Regressor

The MLPRegressor demonstrated significantly better performance than the DecisionTreeRegressor on the same test set of 7,800 records. It achieved a much lower Mean Absolute Error (MAE) of 0.0400, in dicating 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 DecisionTreeRegressor, h ighlighting the model's higher precision and reduced variance in predictions. Moreover, the R-squared (R^2) value of 0.9494 signifies that the MLPRegressor explains 94.94% of the variability in the data, sh owcasing its strong predictive accuracy and a marked improvement over the DecisionTreeRegressor's 37.37% R^2. These metrics collectively establish the MLPRegressor as a much more effective model fo r this task.

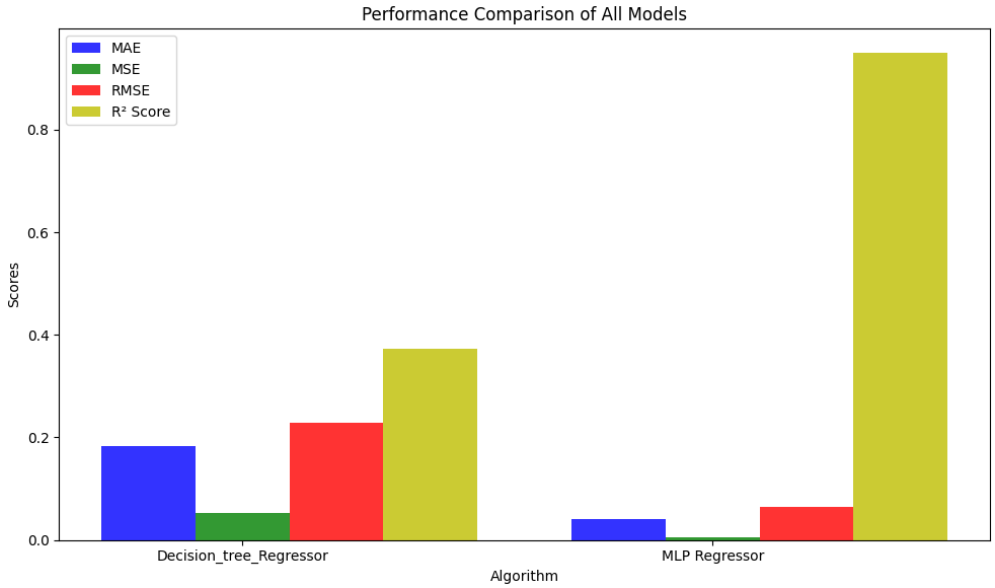

Figure 5: Comparison Graph

Figure 5 shows that the comparison graph in that one MLP Regressor is best metrics.

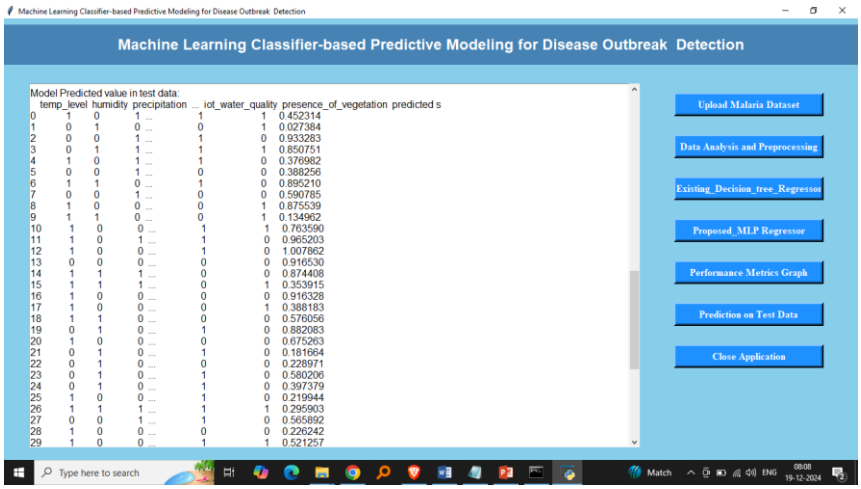

Figure 6: Predicted Output

Figure 6 is shows that the predict column in your dataset represents a numerical value that likely 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 predict value might represent the predicted probability of a malaria outbreak occurring in the area under the given conditions. For example, a value of 0.452314 might indicate a 45.23% likelihood of a malaria outbreak.

5. CONCLUSION

The research on ML learning-based analysis for malaria infection diagnosis represents a significant advancement in the field of medical diagnostics, particularly for infectious diseases that have a profound impact on global public health. The traditional methods of malaria diagnosis, while effective, are fraught with limitations, including dependency on skilled personnel, susceptibility to human error, and the time-consuming nature of the process. These challenges are particularly pronounced in low-resource settings, where the shortage of skilled technicians and inadequate healthcare infrastructure often lead to delayed or inaccurate diagnoses, ultimately contributing to higher morbidity and mortality rates.The integration of ML learning into malaria diagnosis offers a promising solution to these challenges. By training ML learning models on large datasets of labeled blood smear images, the proposed system can automate the detection and classification of malaria parasites. This automation significantly reduces the time required for diagnosis, minimizes human error, and provides consistent and accurate results regardless of the setting. The ML learning model can process and analyze medical images rapidly, making it possible to diagnose malaria with high accuracy even in remote or resource-limited areas where traditional microscopy may not be feasible. The scalability of ML learning-based systems makes them suitable for deployment in large-scale screening programs, potentially reaching a broader population and improving overall malaria control efforts. The use of AI in this context not only enhances the speed and accuracy of diagnosis but also democratizes access to high-quality diagnostic tools, making them available to regions that have traditionally been underserved.

REFERENCES

[1] David H, Mannila H, Smyth PC. Principles of data mining. MIT Press, Cambridge; 2001.

[2] Koh HC, Tan G. Data mining application in healthcare. Journal of Healthcare Information Management. 2005;19(2).

[3] Tomar D, Agarwal S. A survey on data mining approaches for healthcare. International Journal of Bio-Science and Bio-Technology. 2013;5(5):241-266.

[4] World Health Organization. World Malaria Report: Geneva; 2012.

[5] Roca-Feltrer A, Carneiro I, Armstrong Schellenberg JR. Estimates of the burden of malaria morbidity in Africa in children under the age of 5 years. Trop Med Int Health. 2008;13:771–83.

[6] Ibrahim F, Abu N, Osman A, Usman J, Kadri NA. Comparison of different classification techniques using weka for breast cancer. Biomed 06, IFMBE Proceedings. Springer-Verlag Berlin Heidelberg Publisher. 2007;15:520-523.

[7] Boris M, Milan M. Prediction and decision making in health care using data mining. Kuwait Chapter of Arabian Journal of Business and Management Review. 2012;1(12).

[8] Sharma V, Ajai K, Lakshmi P, Ganesh K, Anuradha L. Malaria outbreak prediction model using machine learning. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET). 2015;4(12).

[9] Kapor P, Rani R. Efficient decision tree algorithm using j48 and reduced error pruning. International Journal of Engineering Research and General Science. 2015;3(3): 2091-2730. [10] Leopard H, Cheruiyot KW, Kimani S. A survey and analysis of classification and regression data mining techniques for diseases outbreak prediction in datasets. The International Journal of Engineering and Science (IJES). 2016;5(9):01-11.

[11] Bbosa F, Ronald W, Peter J. Clinical malaria diagnosis: Ruled-based classification statistical prototype. Publisher: SpringerPlus. 2016;5:939.

[12] Witten HI, Frank E, Hall MA, Pal CJ. Data mining: Practical machine learning tools and techniques. Fourth Edition Morgan Kaufmann (Elsevier); 2017.

[13] Cao X, Maloney KB, Brusic V. Data mining of cancer vaccine trials: A bird's-eye view. Immunome Research. 2008;4:7. DOI: 10.1186/1745-7580-4-7 [14] Cios KJ, Moore GW. The uniqueness of medical data mining. To Appear in Artificial Intelligence in Medicine Journal; 2002.

[15] Witten IH, Eibe F, Christopher JP, Mark AH. The weka workbench “data mining: Practical machine learning tools and techniques”. Morgan Kaufmann, Fourth Edition. 2016;7. [16] Han J, Kamber M, Pei J. Data mining concepts and techniques third edition by Elsevier Inc. 2012;18- 19:622-624.