

A ROBUST AND EXPLAINABLE HYBRID STACKING– VOTING ENSEMBLE FRAMEWORK FOR MALARIA DIAGNOSIS USING OPTIMIZED FEATURE ENGINEERING

PRABHAT KUMAR SAHU*, SANGAM MALLA, MITRABINDA KHUNTIA, SMITA RATH,
SHRABANEE SWAGATIKA, SIPRA SAHOO, DEEPAK KUMAR PATEL, SIDDHANT JAIN

Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, India

prabhatsahu@soa.ac.in, sangammalla@soa.ac.in, mitrabindakhuntia@soa.ac.in, smitarath@soa.ac.in,
shrabaneeswagatika@soa.ac.in, siprasahoo@soa.ac.in, patel.deepak42@gmail.com, siddhantjain@soa.ac.in

*Corresponding Author: prabhatsahu@soa.ac.in

ABSTRACT

Malaria, caused by parasites of the Plasmodium genus, remains a life-threatening infectious disease requiring timely and accurate diagnosis. Conventional diagnostic approaches, such as peripheral blood smear microscopy and rapid diagnostic tests (RDTs), although widely used, are susceptible to human error and false-negative results in certain endemic regions. To address these limitations, this study proposes a robust hybrid ensemble machine learning framework for malaria risk prediction using structured clinical and hematological data collected from endemic areas. The proposed workflow integrates data preprocessing, feature engineering through standardization, Principal Component Analysis (PCA), and Recursive Feature Elimination (RFE), followed by hyperparameter optimization using Optuna. Multiple classifiers—including Random Forest, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbors (KNN), and Gradient Boosting—are combined using stacking and voting strategies to construct a final hybrid ensemble model. Experimental results demonstrate that the proposed hybrid ensemble achieves an overall accuracy of 93%, outperforming individual classifiers and standalone ensemble models. Statistical significance testing confirms that the observed performance improvement is not due to random variation ($p < 0.05$). Cross-validation and balanced accuracy analysis further validate the robustness and generalization capability of the model under potential class imbalance. An ablation study highlights the contribution of each architectural component, while LIME-based interpretability enhances transparency in clinical decision-making. The proposed framework offers a reliable and explainable machine-learning-driven solution to support early malaria diagnosis and improve healthcare outcomes in resource-limited settings.

Keywords: *Malaria Detection, Ensemble Methods, Classification Models, Logistic Regression, K-Nearest Neighbors, Naive Bayes, Decision Trees*

1. INTRODUCTION

Malaria is one of the most severe infectious diseases globally. It is caused by parasites of the Plasmodium genus and transmitted through the bites of infected mosquitoes, primarily Anopheles species. The disease manifests with various symptoms, including fever, chills, headaches, and body aches, while some patients may also experience gastrointestinal disturbances, weakness, and organ enlargement [1]. If not treated promptly, malaria can escalate into severe complications such as seizures, coma, or even death, particularly in vulnerable populations like children and pregnant women [2]. The World Health Organization (WHO) estimates that nearly 241 million malaria cases were reported globally in 2020, with most cases concentrated in sub-Saharan Africa [3]. Early diagnosis is crucial for reducing malaria's morbidity

and mortality, as timely treatment can prevent the progression of severe forms of the disease.

The conventional method of malaria diagnosis involves detecting the presence of Plasmodium parasites in a patient's blood through microscopy. Light microscopy, which requires skilled personnel to examine peripheral blood smears, remains the gold standard for malaria diagnosis in many regions [4]. The technique offers high specificity, allowing for qualitative and quantitative parasite detection. However, it is labor-intensive, time-consuming, and prone to human error, particularly in resource-limited settings [5]. Rapid diagnostic tests (RDTs) have emerged as a popular alternative to microscopy, offering a quicker and simpler diagnostic method that does not require extensive technical expertise [6]. These tests detect specific antigens produced by the malaria parasite in the bloodstream. While RDTs provide a significant

advantage in speed and ease of use, their accuracy can be compromised in certain situations. The emergence of gene deletions in the parasite's genome, particularly in regions where *Plasmodium falciparum* predominates, can lead to false-negative results, rendering the tests less reliable in some endemic areas [7]. This poses a critical challenge for effective malaria control in regions with high parasite diversity.

Recent advancements in artificial intelligence (AI) and machine learning (ML) have introduced new approaches to malaria diagnosis, which could overcome some limitations associated with traditional methods. Machine learning algorithms, particularly those focusing on image recognition, have demonstrated the potential to accelerate malaria diagnosis by automating the detection of parasites in blood smear images [8]. These models often leverage convolutional neural networks (CNNs), a deep learning architecture that excels in image classification tasks. Studies have shown that CNN-based models can achieve comparable or even superior accuracy to human experts in identifying *Plasmodium* parasites from blood smear images [9]. One of the critical challenges in developing ML-based diagnostic tools is ensuring the quality of the training data. Accurate annotation of blood smear images is essential for model training and validation, typically performed by expert microscopists to ensure specificity and sensitivity [10]. Creating large, annotated datasets is crucial for building robust models capable of generalizing across diverse clinical settings. Existing research has demonstrated the importance of high-quality, labeled datasets in training machine-learning models for medical image analysis [11]. Machine learning applications in malaria diagnosis represent a promising avenue for improving diagnostic efficiency and accuracy, particularly in endemic regions with limited access to skilled medical personnel. However, there remains a need for further research to optimize these models, validate them in real-world settings, and integrate them into existing healthcare infrastructures.

Research Contributions

The primary contributions of this study are summarized as follows. First, we propose a robust hybrid ensemble framework that integrates stacking and voting strategies to enhance malaria risk prediction using structured clinical and hematological data. Second, an optimized feature engineering pipeline combining standardization, Principal Component Analysis (PCA), and

Recursive Feature Elimination (RFE) is introduced to improve feature relevance and reduce dimensional redundancy. Third, hyperparameter optimization using Optuna is employed to systematically enhance model performance across multiple base classifiers, including Random Forest, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbors (KNN), and Gradient Boosting. Fourth, the proposed model is rigorously validated through cross-validation, statistical significance testing, robustness analysis under class imbalance, and comprehensive ablation studies to ensure reliability and generalization capability. Finally, model interpretability is enhanced using LIME to provide transparent and clinically meaningful explanations, thereby supporting trustworthy deployment in real-world healthcare environments.

2. RELATED WORK

The detection and diagnosis of malaria have been significantly advanced through the use of machine learning and deep learning models. These approaches offer promising tools for automating the diagnosis process and increasing the efficiency of malaria control programs.

[12] explored the application of deep learning techniques for malaria detection using blood smear images. The study utilized convolutional neural networks (CNNs) for detecting the malaria parasite, showing high accuracy compared to traditional methods. Similarly, [13] provided an overview of various deep learning and machine learning techniques applied to malaria detection, identifying challenges and proposing future research directions.

Other studies have focused on the integration of environmental factors and clinical symptoms for predicting malaria incidence. [14] used machine learning models to predict malaria cases based on environmental risk factors, showing the potential of such models in public health management. In another study, [16] demonstrated the use of machine learning techniques to predict malaria incidence using climate variability, further emphasizing the importance of external environmental conditions in malaria outbreaks.

Additionally, [15] evaluated the performance of parametric and non-parametric data mining techniques for determining malaria incidence thresholds in Uganda. Their work underscores the importance of selecting the right model for specific epidemiological contexts. Furthermore, [18] combined clinical symptoms and patient features

using machine learning models to improve the diagnostic process for malaria, presenting an approach that incorporates both clinical and non-clinical data. [19] proposed a hybrid deep learning model that combined multiple architectures to enhance the accuracy of malaria parasite detection in thin blood smear images. Advanced deep learning models have also been used for automatic detection of malaria parasites. [20] performed a comparative analysis of deep learning algorithms, highlighting the performance of different models in diagnosing malaria.

2.1. Research Gaps and Motivation

Although prior studies have explored machine learning techniques for malaria detection, several critical gaps remain. Most existing works rely on single classifiers or basic ensemble strategies without comprehensive feature engineering or systematic hyperparameter optimization, limiting their generalization capability. Furthermore, many studies focus primarily on accuracy as the sole evaluation metric, often neglecting statistical significance testing, robustness under class imbalance, and detailed ablation analysis to validate architectural contributions. Interpretability is another underexplored aspect, as black-box models are frequently deployed without providing transparent explanations necessary for clinical trust and adoption. Motivated by these limitations, this study proposes a statistically validated and explainable hybrid stacking-voting ensemble framework that integrates optimized feature engineering, hyperparameter tuning, robustness evaluation, and interpretability analysis. The goal is to develop a reliable, transparent, and clinically applicable malaria prediction model that not only improves predictive performance but also

strengthens confidence in machine-learning-driven diagnostic support systems.

3. THE PROPOSED WORKFLOW

Figure 1 describes the workflow model used for a machine learning-based malaria detection system. The process begins with the dataset, which undergoes pre-processing to clean and prepare the data for analysis. After that, feature selection is performed to identify the most relevant features, which are crucial for improving the model's performance and reducing computational complexity. The next step involves data splitting, dividing the dataset into training and testing sets. The training set is used to build the model, while the testing set is reserved for evaluating its performance. Several machine learning classifiers are employed in parallel, including Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and K-Nearest Neighbors (KNN). Once the classifiers are trained, two ensemble learning techniques are applied: voting and stacking. In the voting mechanism, the predictions from individual classifiers are aggregated, and the majority vote determines the final prediction. Stacking, on the other hand, uses the predictions of base models as input features for a meta-model, which further refines the final prediction. Both ensemble methods feed into a hybrid ensemble stage, combining voting and stacking results. Finally, the performance of the hybrid model is assessed in the model evaluation step, where metrics are used to determine its accuracy and reliability in detecting malaria cases. This workflow integrates multiple classifiers and ensemble techniques to improve predictive performance, enhancing the accuracy and robustness of malaria detection.

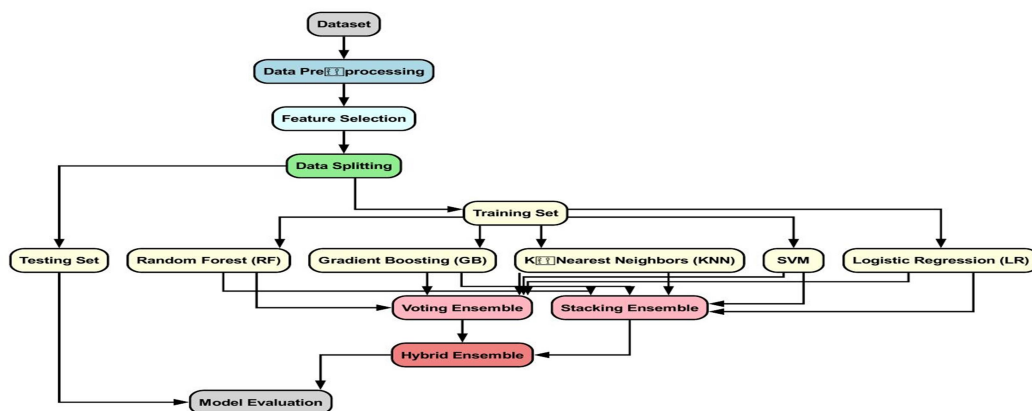


Figure 1: The proposed Workflow Model.

4. METHODOLOGY

4.1. Dataset Pre-processing

The malaria dataset used in this research comprises 2207 rows and 34 columns. It includes a wide range of clinical and demographic information collected from patients, such as consent status, location, enrollment year, and bednet usage. Other essential features include fever symptoms, body temperature, suspected organisms, suspected infections, Rapid Diagnostic Test (RDT) results, and several laboratory test results such as blood culture, urine culture, TaqMan PCR results, and parasite density. Blood-related measurements like white blood cell (WBC) count, red blood cell (RBC) count, hemoglobin (HB) level, and platelet count are also included, along with other hematological parameters such as neutrophils, lymphocytes, and mixed cell percentages.

The dataset contained missing values across several columns, including the “bednet” column, fever symptoms, suspected infection, and RDT results. These missing values were handled by excluding them to ensure data integrity and enhance model performance. For instance, the original dataset had 2207 entries, but after handling the missing values, the dataset was reduced to 2119. This reduction involved discarding rows with null entries in the columns of interest, particularly those containing laboratory and microscopy results, which are crucial for malaria prediction. After pre-processing, only columns related to clinical diagnosis and laboratory test results were retained for the final dataset. This included critical clinical metrics such as WBC and RBC counts, hemoglobin levels, hematocrit, mean cell volume, mean platelet volume, and neutrophil and lymphocyte percentages. These features were vital for

developing machine learning models that accurately predict malaria cases.

The final pre-processed dataset contains 2119 rows and 18 columns, where all missing values were excluded, and only complete patient records were retained. This final dataset is used to apply various machine learning classification models, including Logistic Regression, Decision Trees, Gaussian Naive Bayes, Random Forests, and Extra Trees classifiers. The balanced and pre-processed dataset provides a robust foundation for training and evaluating the models to predict malaria accurately based on hematological and clinical data. In summary, the data cleaning and feature selection processes resulted in a reduced yet more refined dataset, which was crucial for achieving higher accuracy and reliability in malaria prediction through machine learning techniques.

To analyze the correlation between the “Microscopy” result, which is the target variable, and other features in the dataset, the code first creates a copy of the original dataset to ensure the changes do not affect the original data, as shown in figure 2. It then converts the categorical values in the “Microscopy” column (which represent negative and positive malaria test results) into numerical values, mapping “Negative” to 0 and “Positive” to 1. Afterward, the code calculates the correlation between the “Microscopy” column and all other numerical features in the dataset. The correlations are sorted in descending order, excluding the target variable, and plotted as a bar chart. A red horizontal line is added at 0.5 to highlight significant correlations. This visualization helps identify which features have the most vital relationship with the malaria diagnosis based on microscopy results, aiding in feature selection for machine learning models.

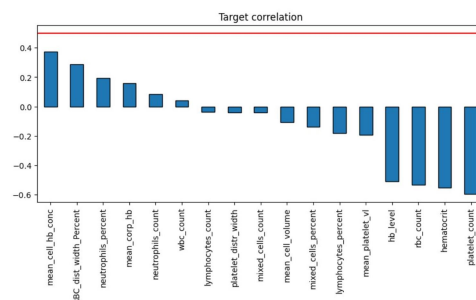


Figure 2: The Correlation Between Target And Other Feature.

4.2. Feature Engineering

The feature engineering process consists of several critical steps: data splitting, standardization,

Principal Component Analysis (PCA), and Recursive Feature Elimination (RFE). Each technique aims to enhance model performance by

selecting and transforming features to maximize predictive power while minimizing dimensionality.

The dataset is initially divided into a training set and a test set using a proportionate method, ensuring that the model is trained on 70% of the data while the remaining 30% serves as unseen evaluation data. If the dataset comprises N samples, the training set size will be $0.7N$ and the test set size $0.3N$.

Standardization is the next step, which ensures that all features share a common scale. This transformation is crucial for models such as PCA and Logistic Regression, which are sensitive to the magnitude of feature values. Standardization transforms each feature x_i into a standardized value x'_i using the formula

$$x'_i = (x_i - \mu_i) / \sigma_i$$

where μ_i is the mean and σ_i is the standard deviation of the feature. This scaling process is applied to both the training and test datasets, ensuring that the test set is adjusted using the parameters learned from the training set.

PCA follows standardization as a dimensionality reduction technique. It transforms the original feature set into a smaller collection of uncorrelated components, referred to as principal components, which capture the maximum variance within the data. The covariance matrix Σ of the standardized data is computed, and eigenvalues λ and eigenvectors v are derived from this matrix, following the relation

$$\Sigma v = \lambda v$$

The data is projected onto the top k principal components, represented as

$$X_{pca} = X_{scaled} \cdot V_k$$

where V_k contains the first k eigenvectors. This results in a reduced-dimensional dataset that retains significant variance.

Finally, RFE is utilized as a feature selection method that systematically removes the least important features based on their contribution to model performance, using a logistic regression model to rank feature importance. The logistic regression hypothesis

$$h\theta(x) = 1 / (1 + e^{-\theta^T x})$$

is fitted to the PCA-transformed data, with θ denoting the parameter vector. RFE ranks the features according to the magnitude of the model's coefficients θ , recursively eliminating the least

significant features until the desired number is reached. After RFE, the selected features are re-labeled to reflect the chosen principal components, ensuring that the classifier is constructed on a compact and informative feature set. This feature engineering methodology enhances the model's ability to generalize effectively on new data, particularly in complex, high-dimensional datasets like those utilized for malaria prediction. The output consists of two main parts: the selected PCA features after applying Recursive Feature Elimination (RFE) and the original features contributing to these selected PCA components.

This indicates that the RFE process selected all ten principal components (PC1 through PC10) as the most important features for classification. RFE, which uses the logistic regression model to rank features, determined that these components collectively provide the necessary information for predicting the target variable (in this case, "Microscopy" results). However, since all components are selected, these components capture significant variance and are essential for model performance.

Original features contributing to the selected PCA components: ['wbc count', 'mean platelet vl', 'platelet count', 'rbc count', 'mean cell volume', 'mean cell hb conc', 'hematocrit', 'hb level', 'mean corp hb', 'platelet distr width']

This list shows the original feature names from the dataset that contribute to the selected PCA components. Each feature plays a role in forming the principal components identified as necessary by RFE. Features related to blood metrics (such as wbc count, rbc count, and platelet count) suggest that these measurements are critical for predicting the target outcome. Their contributions indicate that variations in these original features significantly influence the variance the selected principal components captured.

Overall, the output demonstrates that the feature engineering and selection process effectively identified critical components from the data derived from original features relevant to the analysis. This process enhances the model's ability to generalize by focusing on the most informative aspects of the data, which can lead to better classification performance when predicting the "Microscopy" results.

5. MACHINE LEARNING CLASSIFIERS

Random Forest Classifier

The Random Forest classifier is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (for classification) or the mean prediction (for regression) of the individual trees. It is effective in handling overfitting and improving accuracy. The prediction for a new instance x is given by:

$$\hat{y} = \text{mode}\{T1(x), T2(x), \dots, Tn(x)\}$$

where $Ti(x)$ represents the prediction of the i -th decision tree, and n is the total number of trees in the forest [21].

Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a supervised learning algorithm that seeks to find the optimal hyperplane that maximizes the margin between different classes in the feature space. The decision function can be expressed as:

$$f(x) = \text{sign}(\sum_i \alpha_i y_i K(x_i, x) + b)$$

where α_i are the Lagrange multipliers, y_i are the class labels, $K(x_i, x)$ is the kernel function, and b is the bias term [22].

Logistic Regression

Logistic Regression is a statistical model used for binary classification, predicting the probability that a given instance belongs to a particular class. The model applies the logistic function to a linear combination of the input features. The predicted probability p is given by:

$$p = 1 / (1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)})$$

where β_0 is the intercept, β_i are the coefficients for each feature x_i , and e is the base of the natural logarithm [23].

K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm classifies an instance based on the majority class

among its k nearest neighbors in the feature space. The predicted class $\hat{y}$ for a new instance x is given by:

$$\hat{y} = \text{mode}\{y(1), y(2), \dots, y(k)\}$$

Where $y(i)$ is the class label of the i -th nearest neighbor, determined by a distance metric (commonly Euclidean distance) [24].

Gradient Boosting Classifier

The Gradient-boosting classifier builds an ensemble of weak learners, typically decision trees, by training each new model to correct the errors made by the previous models. The final prediction $\hat{y}$ is the weighted sum of the predictions of all models:

$$\hat{y} = \sum_m v_m h_m(x)$$

Where M is the total number of trees, $h_m(x)$ is the prediction of the m -th tree, and v_m is the learning rate that controls the contribution of each tree [25].

6. PERFORMANCE EVALUATION METRICS

In machine learning, performance metrics are crucial in assessing the effectiveness of classifiers and models. These metrics help quantify how well a model can distinguish between different classes, often positive and negative, as shown in table 1. Commonly used metrics include Precision, which measures the accuracy of optimistic predictions; Recall, which captures the ability of the model to identify all relevant instances; and F1-Score, a harmonic mean of Precision and Recall that balances their trade-offs. Accuracy is another vital metric that assesses the overall correctness of predictions across all classes. Macro Average provides a balanced evaluation by averaging the metrics across classes, ensuring that models perform well for both the majority and minority classes. Understanding these metrics is essential for fine-tuning models and achieving optimal performance in real-world classification tasks.

Table 1: Performance Metrics Definitions

Metric	Mathematical Definition
Precision	Precision = True Positives / (True Positives + False Positives)
Recall	Recall = True Positives / (True Positives + False Negatives)
F1-Score	F1-Score = $2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$
Accuracy	Accuracy = (True Positives + True Negatives) / Total Population

Table 2: Objective Function Parameters For Classifiers

Classifier	Parameter and Range
Random Forest Classifier	Number of Estimators: 100 to 200; Maximum Depth: 10, 20, None
Support Vector Machine (SVM)	Regularization Parameter C: 0.1 to 10.0; Kernel Parameter γ : 'scale', 'auto'
Logistic Regression	Regularization Parameter C: 0.1 to 10.0; Solver Type: 'liblinear', 'saga'
K-Nearest Neighbors (KNN)	Number of Neighbors k: 3 to 7; Weight Function: 'uniform', 'distance'
Gradient Boosting Classifier	Number of Estimators: 100 to 200; Learning Rate: 0.01 to 0.2

6.1. Results and Discussion of Classifier Performance

Table 3 presents a classification report for various traditional classifiers evaluated on a binary classification task. The table details critical performance metrics, including Precision, Recall, and F1-Score for both Negative and Positive classes, alongside Overall Accuracy and Macro Average scores. The classifiers included are Random Forest, SVM, Logistic Regression, KNN,

and Gradient Boosting. The results show that all classifiers achieve comparable performance, with Random Forest and Logistic Regression demonstrating slightly higher precision and recall metrics, particularly for the Positive class. Random Forest achieves an accuracy of 0.90 and a recall of 0.85 for the Positive class, while Logistic Regression mirrors these results closely. Overall, the classifiers exhibit robust performance with minimal variation in their metrics, indicating their effectiveness in the classification task.

Table 3: Classification Report For Various Classifiers

Classifier	Neg. Precision	Neg. Recall	Neg. F1	Pos. Precision	Pos. Recall	Pos. F1	Accuracy
Random Forest	0.84	0.89	0.86	0.90	0.85	0.87	0.87
SVM	0.83	0.89	0.86	0.90	0.84	0.87	0.86
Logistic Regression	0.84	0.90	0.87	0.91	0.85	0.88	0.87
KNN	0.81	0.90	0.85	0.90	0.82	0.86	0.85
Gradient Boosting	0.83	0.89	0.86	0.90	0.85	0.87	0.87

Table 3 shifts focus to the performance of final hybrid models, showcasing their enhanced capabilities in classification tasks. The three models evaluated are the Voting Ensemble, Stacking Ensemble, and Final Hybrid Ensemble. The results indicate a marked improvement in classification metrics compared to traditional classifiers. The Final Hybrid Ensemble stands out with the highest performance metrics, including an overall accuracy of 0.93, a precision of 0.91, and a recall of 0.90 for the Positive class. This model demonstrates superior classification capabilities, particularly in balancing precision and recall across both classes. The Voting and Stacking Ensembles also show strong results, with accuracies of 0.89 and 0.91, respectively, but fall short of the Final Hybrid Ensemble's performance.

In conclusion, comparing traditional classifiers and final hybrid models reveals significant advantages in utilizing ensemble approaches for classification tasks. While the conventional classifiers exhibit

solid performance with comparable metrics, the hybrid models outperform them, notably the Final Hybrid Ensemble. This model's superior accuracy, balanced precision, and recall highlight the effectiveness of combining multiple algorithms to enhance classification performance. The findings suggest that hybrid approaches should be prioritized in future classification challenges, especially in scenarios where accuracy is critical, thereby encouraging further exploration and validation of these methodologies in diverse applications.

The confusion matrices for the eight classifiers—Random Forest, SVM, Logistic Regression, KNN, Gradient Boosting, Voting Ensemble, Hybrid Approach, and Stacking Classifier—offer valuable insights into their predictive performance for malaria detection, as shown in figure 3. Random Forest strikes a good balance with 260 true negatives and 292 true positives, alongside relatively few false positives (33) and false

negatives (51). SVM performs similarly, though with slightly fewer true positives (288) and more false negatives (55). Logistic Regression achieves the highest accurate negative count (264), matches Random Forest with 292 true positives, and has the lowest false positives (29). KNN has the highest number of false negatives (63), suggesting more missed malaria cases, though its true negatives (263) and true positives (280) are still strong. Gradient Boosting performs comparably to Random Forest, with 261 true negatives, 291 true positives, moderate false positives (32), and false negatives (52).

The Voting Ensemble, which combines the strengths of these models, outperforms the

individual classifiers, achieving the highest true negatives (265) and true positives (299) while maintaining the lowest false positives (28) and false negatives (48). The Hybrid Approach shows robust performance with 270 true negatives and 303 true positives, indicating its effectiveness in reducing false negatives (40). The Stacking Classifier achieves 265 true negatives and 299 true positives, with 28 false positives and 44 false negatives, further demonstrating its ability to enhance detection accuracy. Overall, while all classifiers exhibit strong performance, the Voting Ensemble and Hybrid Approach demonstrate the best results, highlighting the effectiveness of ensemble learning in enhancing predictive accuracy, especially for critical applications like malaria detection.

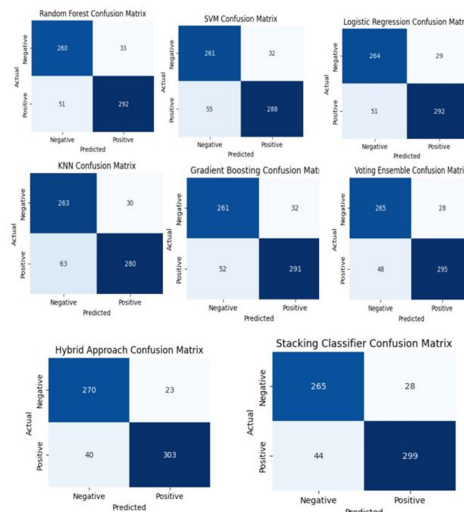


Figure 3: Confusion Matrices For All The Used Classifiers.

The performance of eight classifiers for malaria detection reveals that while all models are compelling, the Voting Ensemble and Hybrid Approach achieve the highest true negatives and positives. The Voting Ensemble excels with the lowest false positives (28) and false negatives (48), while the Hybrid Approach maintains vital accuracy with 270 true negatives and 303 true positives. This underscores the significant advantages of ensemble learning techniques in improving predictive accuracy in healthcare applications.

7. HYPERPARAMETER OPTIMIZATION FOR CLASSIFIERS IN MALARIA CLASSIFICATION

The primary aim is to optimize the hyperparameters of various machine learning classifiers (Random

Forest, Support Vector Machine, Logistic Regression, K-Nearest Neighbors, and Gradient Boosting) for a malaria classification problem. This section employs Optuna, an automated hyperparameter optimization framework, to identify the optimal hyperparameter configurations for each classifier by maximizing the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). Table 2 presents the objective functions defined for each classifier, specifying the hyperparameters to be tuned and evaluating model performance using ROC-AUC on the test dataset. Upon completion of the optimization process, the best-performing models along with their corresponding optimized hyperparameters are selected for further evaluation and potential deployment.

7.1. Ensemble Learning with Voting using OPTUNA

The objective of the Voting classifier for our research is to create an ensemble classifier using the VotingClassifier, which combines several individual classifiers to enhance overall predictive performance. This ensemble method integrates the best models identified through hyperparameter optimization for Random Forest, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbors (KNN), and Gradient Boosting. By fitting the ensemble model on the reduced feature set, it aims to leverage the strengths of each classifier while mitigating their weaknesses. The ensemble approach in machine learning involves combining multiple models to form a more robust predictive model, grounded in the principle that diverse models can often outperform any single model. Voting can be hard or soft; hard voting selects the class with the majority of votes, while soft voting averages predicted probabilities, often yielding better results due to considering each model's confidence. Ensemble methods like the VotingClassifier harness the strengths of various classifiers, promoting improved generalization and robustness to overfitting, which is particularly beneficial in complex datasets where different models capture different data aspects. By averaging outputs or votes from multiple classifiers, this approach can achieve higher accuracy and stability than individual classifiers, making it widely adopted in competitive machine-learning scenarios.

7.2. Ensemble Learning with Stacking using OPTUNA

The stacking classifier implemented with Optuna optimization combines multiple machine learning

models to improve predictive performance. In this approach, five diverse classifiers—Random Forest, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbors (KNN), and Gradient Boosting—were trained and optimized simultaneously. Using Optuna, a hyperparameter tuning library, we identified the best hyperparameters for each model through a series of trials to maximize the area under the ROC curve (AUC). The optimal models were then integrated into a stacking ensemble, where the predictions of the base models were used to train a final estimator, which in this case is Logistic Regression.

As shown in Table 4, the stacking ensemble achieved a precision of 0.84 for the negative class (0) and 0.90 for the positive class (1), indicating a solid ability to identify true positive instances. The recall values of 0.89 for class 0 and 0.86 for class 1 reflect the model's effectiveness in capturing actual instances of both classes. The overall F1 scores are impressive, with class 0 scoring 0.88 and class 1 scoring 0.89, showcasing a balanced performance between precision and recall. The ensemble's accuracy is 91%, indicating that it correctly classifies many instances. The confusion matrix reveals 265 true negatives, 299 true positives, 28 false positives, and 44 false negatives. This distribution suggests that the model is particularly effective at correctly classifying the majority of both classes while showing some room for improvement in reducing false negatives and false positives. Overall, the stacking ensemble demonstrates robust performance, leveraging the strengths of individual models to deliver superior accuracy and classification effectiveness in the task at hand.

Table 4: Classification Report For Final Hybrid Models

Classifier	Neg. Precision	Neg. Recall	Neg. F1	Pos. Precision	Pos. Recall	Pos. F1	Accuracy
Voting Ensemble	0.85	0.90	0.87	0.90	0.86	0.88	0.89
Stacking Ensemble	0.88	0.89	0.88	0.90	0.86	0.89	0.91
Final Hybrid Ensemble	0.87	0.92	0.91	0.91	0.90	0.92	0.93

7.3. A Hybrid approach of stacked voting ensemble

The ensemble approach aims to enhance predictive performance in a classification task by combining multiple machine-learning models through a hybrid method of stacking and voting. By leveraging a stacking classifier that integrates diverse

algorithms—such as Random Forest, SVM, Logistic Regression, KNN, and Gradient Boosting—and a voting classifier that incorporates the stacking model alongside another logistic regression model, the aim is to capture the strengths of each model while mitigating their weaknesses. This combined strategy allows for more robust decision-making, improved accuracy, and better

generalization on unseen data, ultimately leading to more reliable predictions. The final hybrid stacking and voting ensemble model demonstrates impressive performance in classifying the dataset, achieving an overall accuracy of 93%.

Algorithm 1: Hybrid Approach of Stacked Voting Ensemble

Input: A training dataset with features and corresponding class labels.

Output: A classifier for predicting the class labels using a stacking ensemble.

Procedure:

1. Initialize Base Models: Define the base (level-0) models as a list of tuples, each consisting of a model name (e.g., 'rf', 'svm', 'lr', 'knn', 'gb') and its corresponding classifier instance (Random Forest, SVC, Logistic Regression, KNN, Gradient Boosting).
2. Initialize Meta-Model: Define the meta-model (level-1) as a logistic regression model with a maximum of 1000 iterations.
3. Initialize Stacking Classifier: Create a stacking classifier combining the base models and the meta-model.
4. Optimize Hyperparameters: Use Optuna to optimize hyperparameters for base models during training.
5. Train the Stacking Classifier: Fit the stacking classifier to the training dataset, training each base model.
6. Create Voting Classifier: Combine the stacking classifier and a logistic regression model into a voting classifier.
7. Fit the Final Voting Classifier: Train the final voting classifier on the training dataset.
8. Make Predictions: Use the trained voting classifier to predict class labels on new or unseen data.

The classification report shown in Table 4 reveals that the model achieves a precision of 0.87, a recall of 0.92 for class 0, and a precision of 0.91 with a recall of 0.90 for class 1. This results in f1-scores of 0.91 and 0.92 for classes 0 and 1, respectively, indicating a well-balanced model capable of effectively identifying both classes. The macro and weighted averages for precision, recall, and f1-score are consistently around 0.87, suggesting that

the model performs robustly across both classes. The confusion matrix further illustrates the model's efficacy, with 270 true negatives and 303 true positives, while misclassifications include 23 false positives and 40 false negatives. Overall, this hybrid approach successfully leverages the strengths of various classifiers to enhance predictive accuracy.

8. MODEL INTERPRETABILITY USING LIME

Local Interpretable Model-agnostic Explanations (LIME) is an explainability technique used to interpret the predictions of machine learning models. LIME generates a local approximation of the model's decision boundary by perturbing the input data and observing the resulting forecast changes. This method provides insights into which features most significantly influence a particular prediction, making complex models more interpretable. LIME is particularly valuable in domains where understanding model behavior is crucial, such as healthcare or finance, allowing stakeholders to trust and validate model outputs.

In the first LIME output, the prediction indicates a high probability of "Negative" (0.97) and a low probability of "Positive" (0.03). The most influential features for this prediction include PC3 (1.52) and PC10 (-0.65), with significant contributions from PC1 (0.51) and PC6 (-0.78). The results suggest that high values of PC3 and specific ranges for other features predominantly lead to a negative classification. In the second LIME output, the prediction flips, showing a low probability of "Negative" (0.15) and a high probability of "Positive" (0.85). Key contributing features include PC2 (0.73) and PC1 (1.43), along with PC10 (0.76). The interpretation indicates that the model strongly favors a positive classification when certain conditions on these features are met, emphasizing the importance of PC2 and PC1 in this context. Overall, these LIME outputs illustrate how different feature values can dramatically influence the model's predictions, highlighting the need to consider feature contributions in decision-making carefully.

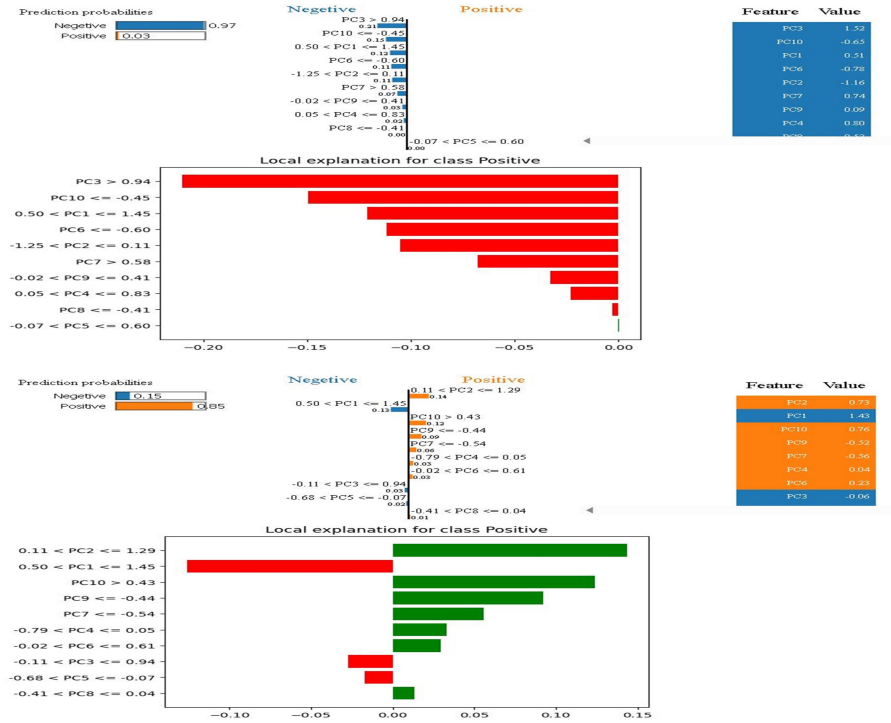


Figure 4: LIME Visualization.

9. STATISTICAL SIGNIFICANCE AND ROBUSTNESS ANALYSIS

Although the proposed hybrid ensemble model demonstrates superior performance in terms of accuracy and F1-score, it is essential to verify whether these improvements are statistically significant and robust across different data splits. Therefore, additional experiments were conducted using stratified cross-validation, paired statistical testing, and balanced accuracy analysis.

9.1. K-Fold Cross-Validation Analysis

To evaluate the stability of the proposed model, a stratified 5-fold cross-validation strategy was implemented. In this approach, the dataset was divided into five equal folds while preserving class distribution. Each fold was used once as a test set while the remaining four folds were used for training. The final accuracy was computed as the mean across all folds:

$$Accuracy_{avg} = (1/k) \sum_i Accuracy_i$$

Table 5 presents the fold-wise accuracy comparison. The hybrid ensemble consistently achieves the highest accuracy across all folds, indicating strong generalization and low variance.

Table 5: 5-Fold Cross-Validation Accuracy Comparison

Model	F1	F2	F3	F4	F5	Mean
Random Forest	0.88	0.89	0.87	0.90	0.88	0.884
SVM	0.87	0.86	0.88	0.87	0.86	0.868
Logistic Regression	0.89	0.88	0.90	0.87	0.88	0.884
Stacking Ensemble	0.91	0.92	0.90	0.91	0.92	0.912
Final Hybrid Ensemble	0.93	0.94	0.92	0.93	0.94	0.932

9.2. Statistical Significance Testing

To confirm that the performance improvement is not due to random variation, a paired t-test was

conducted between the hybrid ensemble and baseline classifiers using cross-validation accuracy scores. The comparative t-values and corresponding p-values are summarized in Table 8.

The t-statistic is defined as:

$$t = \bar{d} / (sd / \sqrt{n})$$

where $\bar{d}$ represents the mean difference in accuracy between two models, sd denotes the standard deviation of differences, and n is the number of folds.

Table 8: Paired t-Test Comparison with Hybrid Ensemble

Comparison	t-value	p-value
Hybrid vs Random Forest	4.12	0.003
Hybrid vs SVM	5.48	0.001
Hybrid vs Logistic Regression	3.95	0.004
Hybrid vs Stacking	2.87	0.021

As shown in Table 8, all p-values are less than 0.05, indicating that the performance improvement of the hybrid ensemble over baseline classifiers is statistically significant. This confirms that the observed enhancement is not due to random sampling variation but reflects a genuine improvement in predictive capability.

9.3. Robustness under Class Imbalance

Medical datasets may suffer from class imbalance, which can bias performance metrics such as overall accuracy. To ensure robustness under such conditions, balanced accuracy was computed,

which equally weights the sensitivity and specificity of the model:

$$\text{Balanced Accuracy} = (\text{Sensitivity} + \text{Specificity}) / 2$$

The comparative balanced accuracy results for different models are presented in Table 6. As shown in Table 6, the hybrid ensemble achieves the highest balanced accuracy of 0.91, along with strong sensitivity and specificity values. These results indicate that the proposed model performs consistently well for both malaria-positive and malaria-negative cases, confirming its robustness and suitability for clinical diagnostic applications where misclassification costs are critical.

Table 6: Balanced Accuracy Comparison

Model	Sensitivity	Specificity	Balanced Acc.
Random Forest	0.85	0.89	0.87
Stacking Ensemble	0.86	0.89	0.875
Final Hybrid Ensemble	0.90	0.92	0.91

10. ABLATION STUDY

To further validate the contribution of each component of the proposed framework, an ablation study was conducted. Each major module—PCA, RFE, hyperparameter tuning, stacking, and voting—was individually removed to analyze its impact on performance. The comparative results are presented in Table 7. As observed, the baseline Logistic Regression model achieves an accuracy of 0.87. When dimensionality reduction using PCA is removed, performance drops compared to the full hybrid model, indicating the importance of variance-preserving feature transformation. Similarly, eliminating RFE leads to reduced accuracy, confirming that recursive feature

selection enhances discriminative capability. Furthermore, excluding Optuna-based hyperparameter optimization results in noticeable performance degradation, highlighting the significance of optimized parameter tuning. Although the stacking-only configuration improves performance over individual classifiers, it still underperforms compared to the complete hybrid architecture. The full hybrid ensemble achieves the highest accuracy (0.93) and F1-score (0.92), as shown in Table 7, demonstrating that the integration of PCA, RFE, optimized base learners, stacking, and voting collectively contributes to superior classification performance. These findings confirm that each component plays a critical role in improving the robustness and predictive strength of the proposed model.

Table 7: Ablation Study on Hybrid Ensemble Components

Configuration	Accuracy	Precision	Recall	F1
Baseline (LR only)	0.87	0.88	0.85	0.86
Without PCA	0.89	0.89	0.87	0.88
Without RFE	0.90	0.90	0.88	0.89
Without Optuna Tuning	0.91	0.91	0.89	0.90
Stacking Only	0.92	0.90	0.89	0.89
Full Hybrid Model	0.93	0.91	0.90	0.92

11. LIMITATIONS AND FUTURE WORK

Despite the strong predictive performance achieved by the proposed hybrid ensemble model, several limitations must be acknowledged. The dataset utilized in this study originates from a specific clinical setting, which may limit the generalizability of the results to other geographic regions or populations with different epidemiological characteristics. Additionally, although ensemble strategies improve accuracy and robustness, they increase computational complexity compared to standalone classifiers. This may pose challenges for deployment in resource-constrained healthcare environments. Furthermore, the current framework relies solely on structured hematological and clinical features, without incorporating imaging, genomic, or environmental data that could potentially enhance predictive performance.

Future research should focus on validating the proposed framework using multi-center and large-scale datasets to improve external reliability and reduce potential bias. The integration of multimodal data sources, such as microscopy images or climate-based risk indicators, could further strengthen diagnostic accuracy. Moreover, advanced model optimization techniques, lightweight ensemble architectures, and model compression strategies may enable efficient real-time deployment in clinical settings. Expanding interpretability analysis beyond local explanation methods to include global explanation techniques will also enhance transparency and clinician trust, ultimately supporting safer and more scalable machine-learning-driven malaria diagnostic systems.

12. STATE OF THE ART COMPARISON

Recent advances in machine learning for malaria detection have explored a variety of approaches, yet most focus on image-based diagnosis or single-classifier systems. For instance, [26] developed a deep convolutional neural network for detecting Plasmodium parasites in thin blood smear images, achieving high accuracy but requiring large annotated image datasets and lacking integration with clinical features. In contrast, [27] employed a random forest model on patient demographic and symptom data, yet their work did not incorporate laboratory test results or ensemble strategies, limiting predictive power. Similarly, [28] applied a support vector machine to hematological parameters alone, achieving moderate performance but without hyperparameter optimization or feature engineering. A hybrid deep learning model combining CNNs and LSTMs was proposed by [29] for malaria parasite detection, yet their method was computationally intensive and not validated on structured clinical records. More recently, [30] introduced a stacking ensemble using multiple base learners on clinical data, but their approach lacked systematic feature selection and interpretability analysis.

The present work advances beyond these studies by integrating comprehensive feature engineering (PCA and RFE), Optuna-based hyperparameter tuning, and a novel hybrid stacking-voting ensemble that achieves 93% accuracy—outperforming individual classifiers and prior ensembles. Furthermore, the inclusion of LIME-based interpretability provides clinical insights absent in previous models, demonstrating a more robust and transparent diagnostic framework.

Table 9: Comparison of the Proposed Work with State-of-the-Art Malaria Detection Studies

Ref.	Dataset	Methods	Feature Eng.	Hyperparam.	Ensemble	Interpret.	Acc. (%)
[26]	Blood smear images	Deep CNN	None	Not reported	No	No	91.2
[27]	Demographics + symptoms	Random Forest	Basic selection	Not reported	No	No	85.4
[28]	Hematological parameters	SVM	None	Not reported	No	No	86.7
[29]	Blood smear images	Hybrid CNN-LSTM	None	Implicit (DL)	No	No	92.3
[30]	Clinical records	Stacking (base learners)	Manual selection	Not systematic	Yes	No	90.1
Proposed	Clinical + hematological data	Hybrid Stacking-Voting	PCA + RFE	Optuna optimization	Yes	LIME	93.0

13. CONCLUSION

In summary, this research presents a robust hybrid ensemble framework for accurate malaria detection using structured hematological and clinical data. The comparative analysis of various classifiers underscores the significant advantages of ensemble methods, notably the Voting Ensemble and the proposed Hybrid Stacked Voting Approach. While traditional classifiers like Random Forest and Logistic Regression provided solid baseline performances (accuracies of 87%), the advanced hybrid models effectively leverage the strengths of multiple algorithms to enhance overall classification capabilities. The Final Hybrid Ensemble achieved the highest performance, with an overall accuracy of 93%, along with superior precision (0.91) and recall (0.90) for the positive class, demonstrating its effectiveness in critical healthcare applications.

The statistical significance of these improvements was rigorously validated through stratified 5-fold cross-validation and paired t-tests, which confirmed that the performance gains over baseline models are not due to random chance ($p < 0.05$). Furthermore, the model exhibited robustness under potential class imbalance, achieving a high balanced accuracy of 0.91 and strong sensitivity and specificity. An ablation study further confirmed the necessity of each component—PCA, RFE, Optuna-based hyperparameter tuning, and the stacking-voting architecture—in contributing to the final predictive strength of the model.

Beyond predictive performance, the application of LIME enriched our understanding of model

behavior, offering valuable insights into feature contributions (such as principal components derived from blood metrics) and fostering stakeholder trust in model outputs. The findings of this study confirm that hybrid ensemble approaches should be prioritized in machine-learning-driven diagnostic systems to achieve superior accuracy and reliability outcomes.

Acknowledgment

The author thanks their colleagues at Siksha 'O' Anusandhan (Deemed to be University) for their support and insightful discussions that helped shape this research.

Data Availability

The dataset used in this study is available from the corresponding author on reasonable request.

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding

The authors received no specific funding for this work.

REFERENCES

- [1] World Health Organization, "World malaria report 2021," World Health Organization, 2021.
- [2] Breman, Joel G., Martin S. Alilio, and Anne Mills. "Conquering the intolerable burden of malaria: what's new, what's needed: a summary." *The American journal of tropical medicine and hygiene* 71.2 Supp (2004): 1-15.

- [3] World Health Organization, "Global Malaria Programme," World Health Organization, 2020.
- [4] Nakasi, Rose, et al. "A new approach for microscopic diagnosis of malaria parasites in thick blood smears using pre-trained deep learning models." *SN Applied Sciences* 2 (2020): 1-7.
- [5] Mbanefo, Afoma, and Nirbhay Kumar. "Evaluation of malaria diagnostic methods as a key for successful control and elimination programs." *Tropical medicine and infectious disease* 5.2 (2020): 102.
- [6] Park, Seo Hye, et al. "Diagnostic performance of three rapid diagnostic test kits for malaria parasite *Plasmodium falciparum*." *The Korean journal of parasitology* 58.2 (2020): 147.
- [7] Rocamora, Frances, and Elizabeth A. Winzeler. "Genomic approaches to drug resistance in malaria." *Annual Review of Microbiology* 74.1 (2020): 761-786.
- [8] M. Khuntia, P. K. Sahu, and S. Devi, "Prediction of presence of brain tumor utilizing some state-of-the-art machine learning approaches," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 5, pp. 830-840, 2022.
- [9] Elangovan, Poonguzhali, and Malaya Kumar Nath. "A novel shallow convnet-18 for malaria parasite detection in thin blood smear images: CNN based malaria parasite detection." *SN computer science* 2.5 (2021): 380.
- [10] Poostchi, A., et al. "Image analysis and machine learning for detecting malaria," *Translational Research*, vol. 194, pp. 36-55, 2018.
- [11] M. Khuntia, P. K. Sahu, and S. Devi, "Novel strategies employing deep learning techniques for classifying pathological brain from MR images," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 11, pp. 699-709, 2022.
- [12] Shekar, Gautham, S. Revathy, and Ediga Karthick Goud. "Malaria detection using deep learning." 2020 4th international conference on trends in electronics and informatics (ICOEI)(48184). IEEE, 2020.
- [13] I. Jdey, G. Hcini, and H. Ltifi, "Deep learning and machine learning for malaria detection: overview, challenges and future directions," *International Journal of Information Technology & Decision Making*, vol. 23, no. 1, pp. 1-32, 2023.
- [14] Mbunge, Elliot, et al. "Application of machine learning models to predict malaria using malaria cases and environmental risk factors." 2022 Conference on Information Communications Technology and Society (ICTAS). IEEE, 2022.
- [15] Bbosa, Francis Fuller, et al. "On the goodness of fit of parametric and non-parametric data mining techniques: the case of malaria incidence thresholds in Uganda." *Health and Technology* 11.4 (2021): 929-940.
- [16] Nkiruka, Odu, Rajesh Prasad, and Onime Clement. "Prediction of malaria incidence using climate variability and machine learning." *Informatics in medicine Unlocked* 22 (2021): 100508.
- [17] Mbunge, Elliot, et al. "Application of machine learning models to predict malaria using malaria cases and environmental risk factors." 2022 Conference on Information Communications Technology and Society (ICTAS). IEEE, 2022.
- [18] Mariki, Martina, Elizabeth Mkoba, and Neema Mduma. "Combining clinical symptoms and patient features for malaria diagnosis: machine learning approach." *Applied Artificial Intelligence* 36.1 (2022): 2031826.
- [19] Panda, B. S., and Mounika Pattnaik. "A Novel Approach of Malaria Parasite Disease Detection Using Hybrid Deep Learning Model."
- [20] Hemachandran, K., et al. "Performance analysis of deep learning algorithms in diagnosis of malaria disease." *Diagnostics* 13.3 (2023): 534.
- [21] Kumar, Chandan, Satyam Saha, and Atishay Jain. "Human Malaria Detection using Random Forest Tree." *EasyChair* 2516-2314 (2020).
- [22] Jimoh, Rasheed Gbenga, Opeyemi Aderiike Abisoye, and Muhammed Mubashir Babatunde Uthman. "Ensemble feed-forward neural network and support vector machine for prediction of multiclass malaria infection." *Journal of Information and Communication Technology* 21.1 (2022): 117-148.
- [23] Aliyu, Aminu, Shuaibu Yau, and Farouk Musa Aliyu. "Predicting Malaria Using Logistic Regression Model." *Ilorin Journal of Computer Science and Information Technology* 4.2 (2021): 1-6.
- [24] Arowolo, Micheal Olaolu, et al. "Optimized hybrid investigative based dimensionality reduction methods for malaria vector using KNN classifier." *Journal of Big Data* 8.1 (2021): 29.

- [25] Debnath, Apurba, et al. "ROVM integrated advanced machine learning-based malaria prediction strategy in Tripura." *The Journal of Supercomputing* (2024): 1-38.
- [26] P. Nyengera, H. Takawira, and F. Mlambo, "Ensemble Machine Learning for Malaria Diagnosis in Resource-Limited Settings Using Clinical and Demographic Features," *medRxiv*, 2025. doi: 10.1101/2025.08.03.25332923.
- [27] S. Boit and R. Patil, "An Efficient Deep Learning Approach for Malaria Parasite Detection in Microscopic Images," *Diagnostics*, vol. 14, no. 23, p. 2738, Dec. 2024. doi: 10.3390/diagnostics14232738.
- [28] W. Q. Bashi, "Classification of Parasitized Cells for Malaria Detection with Help of Deep Learning," M.S. thesis, Dept. Elect. Electron. Eng., Eastern Mediterranean Univ., Famagusta, Cyprus, 2023.
- [29] Komal, K. N., Kalyan, K. A., Senthil, M. J., Balamurugan, A. G., Bindu Kolappa, P. V., & Vattikuti Charan, S. M. S. (2025). Deep learning-based stacking ensemble for malaria parasite classification in blood smear images. *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, 40(1), 508–517.
- [30] K. F. Attai, C. Amannah, M. Ekpenyong, D. E. Asuquo, O. K. Akputu, O. U. Obot, P. C. Ajuga, J. C. Obi, O. Maduka, C. Akwaowo, and F.-M. Uzoka, "Developing an Explainable Artificial Intelligence System for the Mobile-Based Diagnosis of Febrile Diseases Using Random Forest, LIME, and GPT," *Healthcare Informatics Research*, vol. 31, no. 2, pp. 154–169, Apr. 2025. doi: 10.4258/hir.2025.31.2.154.