IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier

Artificial Intelligence Based Early Detection of Dengue Using CBC Data

NUSRAT JAHAN RIYA, MRITUNJOY CHAKRABORTY AND RIASAT KHAN
Electrical and Computer Engineering, North South University, Dhaka, Bangladesh
Corresponding author: Riasat Khan (riasat.khan@northsouth.edu)

ABSTRACT Dengue fever is a tropical mosquito-transmitted disease spread through the Aedes mosquito, where the human body works as the primary host. Each year, densely populated countries such as Bangladesh, Thailand, and India, particularly in the Southeast Asian region, experience the majority of dengue outbreaks worldwide. Notably, in 2023, Bangladesh endured an unprecedented dengue outbreak, registering the highest number of cases in over two decades since 2000. This research aims to facilitate early detection of dengue from patients' complete blood count (CBC) medical laboratory reports collected from two hospitals in Dhaka, Bangladesh. The custom-built dataset, comprising 320 samples and 14 hematology features, is used to evaluate diverse artificial intelligence techniques. This dataset documents suspected dengue cases in Bangladesh from May 2023 to October 2023, reflecting a significant outbreak period, including a gender distribution ratio of 5:3 male to female patients. Various preprocessing steps, handling missing values and outliers, one-hot encoding, synthetic oversampling, and removing redundant features, are applied to the employed dataset. Five feature selection methods and diverse machine learning algorithms, along with ensemble learning and transformer-based models, are implemented. The stacking ensemble classifier achieved the highest performance, with an accuracy of 96.88% and an F1 score of 0.9646. The stacking technique has been built using the LightGBM meta-classifier and XGBoost, Logistic Regression, and Multilayer Perceptron base learners. **The collected private CBC dengue dataset and the implementation codes will be available after the manuscript has been accepted.**

INDEX TERMS Complete Blood Count, Dengue Prediction, Explainable AI, Feature Selection, Machine Learning, Ensemble Learning, Transformer Model

I. INTRODUCTION

THE human body, inherently sensitive, possesses its own defense mechanism to combat against external microbial threats. Nevertheless, humans frequently fall victim to viral or bacterial infections, resulting in diseases that are significantly lethal. Dengue fever, for instance, is a viral disease primarily transmitted to humans by the Aedes mosquito. Every year, millions across the globe suffer from dengue fever, with thousands falling victim to its consequences [1]. According to the World Health Organization (WHO) and the European Union, in 2023, over six million people in nearly 92 countries were affected by dengue fever. Bangladesh alone recorded more than 0.31 million cases and over 1,600 deaths from this hemorrhagic fever [2].

Dengue is most prevalent in urban or peri-urban areas within the tropical and subtropical regions of the world, attributed mainly to insufficient sanitation, haphazard development and unplanned urbanization [3]. According to the latest review by the WHO, the countries in the African, Southeast Asian, and Western Pacific regions have the highest incidence of dengue fever. Among the countries in the Southeast Asian region, Bangladesh recorded the highest number of dengue cases between June and October. The number of affected patients and fatalities due to dengue in 2023 was the highest in recent decades [4]. Hence, early, efficient, and rapid detection and response measures for this arboviral disease are crucial. The escalating trend in the number of affected individuals underscores the imperative for implementing appropriate preventive future measures to avert surpassing previous records in terms of both affected patients and fatalities [5].

While dengue symptoms primarily arise from the bite of the Aedes mosquito, the virus typically remains dormant for a period before becoming apparent. Though not inherently fatal, dengue presents a range of debilitating symptoms similar to those of other diseases. Typically, individuals with dengue experience intense fever, excruciating bodily pain, nausea, loss of appetite, and various types of skin rashes. Despite the

absence of specific symptoms in the initial two weeks post-infection, patients often experience a sudden deterioration in health [6]. Pathological tests usually reveal a decrease in platelets in the blood, indicating a critical condition. Dengue virus has four serotypes. When someone is infected with one serotype, the body develops long-term immunity to it. However, the consequences can be severe if they are subsequently infected with a different serotype. While dengue symptoms may not be severe initially, upon a second infection, dengue can lead to severe conditions like shock syndrome, internal bleeding, or multiple organ failure.

Considering the recent surge in dengue infections, this study introduces artificial intelligence (AI) approaches that enable early detection of dengue by employing various critical hematologic features. In this work, a private dataset has been collected from two local hospitals in Dhaka, Bangladesh. The dataset comprises complete blood count (CBC) data for 320 individuals, classified as dengue 'positive' or 'negative,' and 14 attributes. The dataset has been preprocessed employing diverse techniques. Various machine learning models have been applied, i.e., Logistic Regression, Random Forest, SVM, LightGBM, XGBoost, and stacking classifier. Additionally, we deployed five deep-learning models – MLP, ANN, CNN, Bi-LSTM, and GRU and two advanced transformer models, TabPFN and TabTransformer. Hyperparameter tuning with GridSearchCV and Keras Tuner framework, and five feature selection methods have been employed to extract essential features. The pivotal role of various features in decision-making, mainly focusing on the interpretability of black-box models, is investigated using the LIME-based explainable AI approach. The study offers several significant contributions, which can be summarized as follows:

1) A major contribution of this work is to present a private CBC hematology report-based dengue dataset comprising 320 samples and 14 characteristic features collected from two local hospitals in Dhaka, Bangladesh.
2) Stacking ensemble model constructed from LightGBM meta-classifier and XGBoost, Logistic Regression, and Multilayer Perceptron base learners has been applied. TabPFN and TabTransformer-based transformers and advanced deep learning models are implemented.
3) GridSearchCV and Keras Tuner are applied to tune the best hyperparameters of the applied machine learning and deep learning models. Five feature selection methods have also been used to identify the most salient features.
4) Employing an explainable AI tool, LIME, this research shed light on the key features that significantly impact the most on predicting dengue positive and negative cases.
5) The novelty of this work is to apply explainable stacking ensemble and transformer-based AI models and investigate significant features employing a private blood test report-based dengue dataset.

II. LITERATURE REVIEW

In recent years, the advancement of AI has facilitated rapid and accurate diagnosis of various diseases through machine learning techniques. Machine learning enables the accurate identification of diseases such as diabetes, Parkinson's, Alzheimer's, cardiovascular diseases, ocular diseases, etc. Machine learning primarily involves training algorithms with new data and providing insights about patterns. As a result, with the assistance of this vast repository of data, precise disease identification becomes possible. To determine if a patient is suspected of having dengue, individuals typically visit the nearest hospital or clinic and undergo multiple tests, such as a CBC, IgM/IgG antibody test, NS1 antigen test, etc. Following blood collection, various pathological procedures are performed, and it generally takes many hours and costs a reasonable sum of money [7]. A specialized doctor then reviews the reports to assess the severity of dengue fever based on the results. These procedures can be intensive for the people of low-income countries like Bangladesh due to the scarcity of available specialized doctors and pathologists. As a result, many researchers are striving to make dengue prediction more efficient and cost-effective by presenting various approaches and ideas. This section below delivers a detailed overview of existing methods for automatically detecting dengue fever using CBC hematology samples, blood smear images, environmental factors, and other relevant parameters from recent articles.

Davi et al. [8] utilized the human genome data of 102 patients to predict dengue flavivirus using machine learning models. The authors investigated the patients at high risk of developing extreme phenotypes despite moderate symptoms. Among the applied machine learning algorithms, the ANN model demonstrated the best accuracy score of 86%, with a sensitivity of 98% by extracting features using SVM RFE. Sarma and other researchers [9] designed an automatic dengue prediction model using machine learning algorithms based on the recent outbreak in Bangladesh. The researchers collected raw data from the patients from Dhaka and Chittagong, Bangladesh's two largest and most densely populated cities. The decision tree algorithm achieved the highest accuracy of 79%.

Fernández et al. [10] applied a logistic regression model to diagnose dengue fever based on the features of approximately 550 patients with febrile illness. The applied logic regression model attained 69.2% accuracy for the positive cases with 86.2% sensitivity and 0.66 AUC score. Mayrose and researchers [11] demonstrated an automated dengue prediction model using several machine learning techniques and blood smear image samples based on the lymphocyte nucleus and platelets. The authors achieved the best performance using the SVM classifier with 95.74% accuracy and 0.96 F1 coefficient.

Mello et al. [12] presented predictive models for dengue based on real patient data admitted to Paraguay's health centers with dengue fever symptoms. The applied ANN attained the maximum accuracy of 96% with the highest sensitivity

and specificity of 96% and 97%, respectively. Dey et al. [13] initiated to predict dengue cases based on 11 states' data of Bangladesh. The authors empirically analyzed how environmental factors affect the rise and fall of dengue cases. The applied Support Vector Regression algorithm demonstrated the best results with an R2 score of 0.75. The Multiple Linear Regression algorithm illustrated an excellent performance with a 0.62 R2 coefficient.

Abdualgalil et al. [14] utilized clinical data from a local medical center of Yemen to predict dengue using efficient machine learning techniques. They implied five machine learning algorithms that performed efficiently on the utilized clinical data. The Extra Tree Classifier algorithm demonstrated the best performance with 99.12% accuracy and 0.99 F1 coefficient. Ong and his colleagues [15] depicted the transmission rate of dengue with meteorological data by comparing different machine learning algorithms. This study used multiple variables, algorithms, vector indices, and meteorological data. An ensemble machine learning algorithm, XGBoost, with the Boruta feature selection technique, achieved the highest accuracy (81%) and 0.815 AUC.

Chaw et al. [16] developed an AI-based automatic model that predicts if there is a chance of shock development among all dengue patients. They used physiological data from ill patients at the University of Malaya Medical Centre and trained the model based on these collected data. Among the applied machine learning models, the decision tree approach attained the maximum F1 score of 0.92 and 0.64 AUC. Sarwar and his colleagues [17] introduced a model that can accurately predict the number of dengue-affected patients. The authors considered various environmental factors in Dhaka, including humidity, temperature, and rainfall, as these variables critically influence dengue outbreaks. After implementing statistical algorithms, the SVM algorithm achieved the highest R-squared coefficient of determination of 0.92.

Akter et al. [18] conducted a comparative study to predict dengue fever in Dhaka city, evaluating the effectiveness of time series analysis and machine learning techniques. Based on the time series, the applied ARIMA model hypothesizes the forecasts of dengue outbreaks with a 15.29 mean absolute percentage error (MAPE). The neural network model demonstrated superior performance, achieving the lowest MAPE of 1.15. Majeed et al. [19] executed various hybrid AI models to predict cases of dengue viral fever in five regions of Malaysia. Various hybrid LSTM models have been applied by combining stacked, temporal and spatial attention approaches. The spatial stacked attention with the LSTM technique demonstrated the best performance with the lowest RMSE of 3.17.

It can be understood from the reviews of the related articles that significant works have been initiated on automatic dengue prediction employing advanced machine learning and deep learning techniques. However, most of these works did not investigate the dengue virus's significant clinical and environmental features. Few of these articles applied state-of-the-art explainable AI techniques to interpret the AI model's predictions.

III. METHODOLOGY

Figure 1 illustrates the working steps of the proposed automatic dengue prediction study. Initially, data collection is followed by data preprocessing, which includes categorical encoding and median imputation of null values. Exploratory data analysis and removal of outliers with Z-scores are conducted before splitting the data into training and testing sets. The training data undergoes SMOTE synthetic oversampling, feature selection using five methods, and classification algorithm application with hyperparameter optimization. The selected best parameter combination is then applied, and explainable AI techniques are used to enhance model interpretability. Finally, the model predicts dengue disease on the test data, classifying results as positive or negative. The detailed methodology of the working sequences of this research is discussed in the subsequent paragraphs.

A. DATA COLLECTION

A major contribution of this work is the presentation of a private CBC hematology report-based dengue dataset comprising 320 samples and 14 characteristic features collected from two hospitals in Dhaka, Bangladesh. Informed consent was obtained from the participants, and the study was approved by the Institutional Review Board of North South University (IRB approval number: 2023/OR-NSU/IRB/1001). We focus on identifying key factors in these reports and understanding the correlations among various features. The attributes of the employed dataset comprise Serial, Date, Gender, Age, Haemoglobin, ESR, WBC, Neutrophil, Lymphocyte, Monocyte, Eosinophil, Basophil, RBC, and Platelets. The reports span from May 2023 to October 2023, a period marked by a high rate of dengue outbreaks. This outbreak was highly correlated with rainfall, temperature, and an increase in the breeding rate of Aedes mosquitoes. The proposed dataset includes a diverse age range of patients, from children as young as eight months to adults up to 81 years old, with a distribution of 200 males and 120 females among the 320 patients. The distribution of dengue classes of the employed dataset is illustrated in Figure 2.

B. DATA PREPROCESSING

After collecting the data, the required preprocessing steps are completed which have been briefly described below.

Drop less essential columns: First, we have dropped the less required two attributes, i.e., Serial and Date, to make the working approach smoother. This approach is also known as a part of dataset cleaning that increases efficiency. As mentioned above, dropping the columns has helped to concentrate more on the critical hematologic features of the employed dataset.

Handling Null Values: It is observed that the highest number of null values constitutes the ESR attribute of the dataset, with 49 missing entries. In contrast, the Neutrophil, Lymphocyte, Monocyte, Eosinophil, and RBC features each

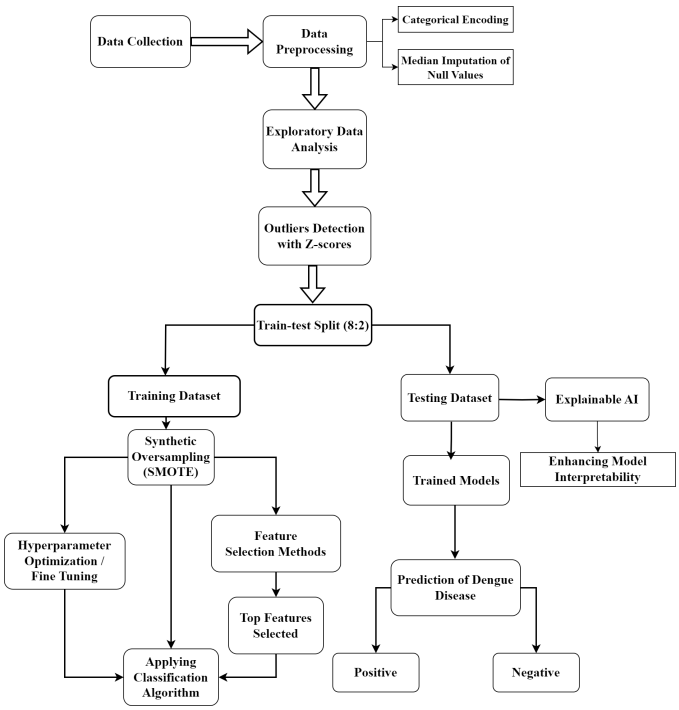

FIGURE 1. Working steps of the proposed automatic dengue prediction study

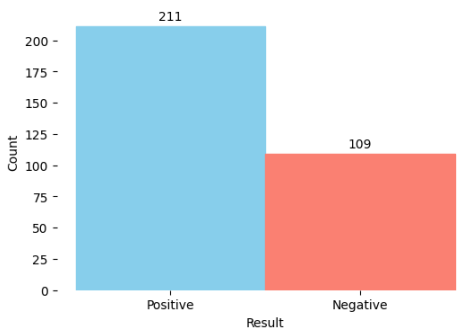

FIGURE 2. Distribution of dengue classes of the employed dataset

contain one null value. Additionally, two null values are identified in the Basophil feature. To address these missing values, the median imputation technique has been employed for the aforementioned features.

One-hot Encoding and Feature Scaling: In this work, the one-hot encoding technique has been used to intrinsically substitute the Gender and Dengue Class into '0' and '1.' The remaining numerical features are transformed to a comparable level employing a Gaussian distribution-based standard feature scaling framework.

Synthetic Oversampling: Synthetic Minority Oversampling Technique (SMOTE) has been used to address class imbalance by generating synthetic data points for minority class (negative dengue cases) training samples.

TABLE 1. Summary of various numerical features of the employed private dataset

Attribute	Min	Max	Mean	Standard Deviation
Age	0.600	81.000	33.895	16.522
Haemoglobin	1.700	19.200	12.614	2.566
ESR	1.000	130.000	27.764	27.323
WBC	0.540	32.600	7.590	4.901
Neutrophil	4.000	92.000	59.441	16.168
Lymphocyte	4.000	88.000	32.111	14.863
Monocyte	1.000	29.000	6.617	3.257
Eosinophil	0.000	11.000	1.996	1.614
Basophil	0.000	2.000	0.008	0.117
RBC	0.000	7.070	4.655	0.957
Platelets	5.000	540.000	110.316	102.110

C. EXPLORATORY DATA ANALYSIS

Figure 2 represents the ratio between positive and negative instances within the employed dengue dataset. Here, 'positive' and 'negative' denote the outcomes corresponding to 'Result.' In this dataset, 211 instances are labeled as positive and 109 as negative dengue cases. As a result, the CBC dataset encloses a total of 320 outputs.

Figure 3 portrays the gender distributions of the employed dataset. 62.5% and 37.5% of the instances correspond to males and females, respectively. The statistical descriptions of various numerical features of the employed private CBC dengue dataset are summarized in Table 1. In the case of Platelets, a vast difference between the minimum and the maximum value has been illustrated. It indicates a higher likelihood of positive compared to negative dengue cases. According to Figure 4, a variation of ages can be observed in the curated dengue dataset, ranging from eight months to

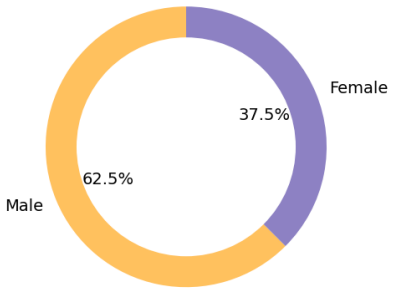

FIGURE 3. Gender distributions of the employed dataset

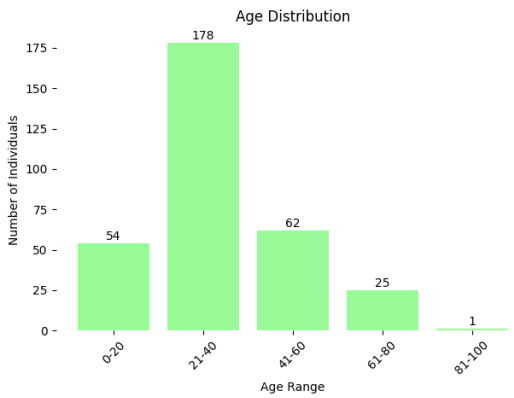

FIGURE 4. Age distributions of the employed dataset

81 years.

The pair plot of the selected features from the dengue CBC report dataset displays the relationships between various hematologic parameters such as age, hemoglobin, WBC, neutrophil, lymphocyte, RBC, and platelets, distinguishing between positive and negative dengue test results. Figure 5 reveals noticeable clustering patterns, particularly in WBC, platelets, and lymphocytes, where positive dengue cases (marked in purple) tend to have lower WBC and platelet counts, highlighting significant hematologic differences between positive and negative dengue patients.

D. DATA CLEANING AND REMOVING OUTLIERS

Figure 6 presents a heat map of various numerical features from the dataset, using Pearson's correlation method to visualize inter-feature correlations. A threshold value of 0.85 is set to determine significant correlations. Features exceeding this threshold are considered highly correlated. Neutrophils and lymphocytes show an inverse correlation with a value of -0.96, but neither feature is eliminated. Another correlation between hemoglobin and RBC yields a value of 0.83 below the threshold, resulting in no feature elimination. This analysis ensures that none of the features are excluded based on the established correlation criteria.

Violin plots of four selected features are depicted in Figure 7, which constitute versatile tools for visualizing numerical data distributions. The violin plots for Haemoglobin and Red Blood Cells (RBC) demonstrate symmetric distributions for both positive and negative situations. Low tails on Hemoglobin instances indicate few outliers in both groups. Similarly, neither the positive nor negative RBC distributions exhibit notable outliers and are consistent with low variability.

The distributions show greater variance for platelets and ESR (erythrocyte sedimentation rate). The ESR distribution of positive cases is skewed, with a lower median at about 25 and a noticeable tail that indicates outliers at higher values. The ESR distribution is wider in negative cases, with outliers at the higher end and a median close to 50. In positive situations, platelets have a rigid, symmetric distribution with a median of 150, indicating few outliers; in negative cases, the dispersion is larger, with a higher median of around 300 and more notable outliers, especially at higher values.

In this work, the Z score method has been used to address outliers, which can be expressed as:

$$Z < X - 2\sigma \tag{1}$$

$$Z > X + 2\sigma \tag{2}$$

where, X represents the feature value and σ denotes the standard deviation. For the hemoglobin feature, the range of values gets more concentrated, indicating fewer extreme values, but both positive and negative cases stay symmetric and centered around the same medians. With a marginally smaller range, RBC distributions simultaneously maintain their symmetry and center medians around 5. Although the tails of the ESR distributions for both groups have shrunk, the median for negative cases remains higher, indicating that excessively high values have been eliminated. With positive instances remaining concentrated around 150 and negative cases around 300, platelet distributions become more uniform with a narrower spread, indicating a more consistent distribution free of notable outliers.

E. ALGORITHMS

1) Logistic regression

Logistic regression [20] is a statistical technique that links a set of discrete or continuous independent variables to a binary dependent variable. It is a powerful tool that produces robust models. It predicts dependent data by examining the correlation between one or more already present independent variables.

2) SVM

Support vector machine [21] is an effective supervised learning technique for outliers identification, regression, and classification. In order to enable the prediction of labels from one or more feature vectors, it seeks to establish a decision boundary between two classes.

3) Random Forest

Random forest [22] is a classifier that uses multiple decision trees on different subsets of the input dataset and averages

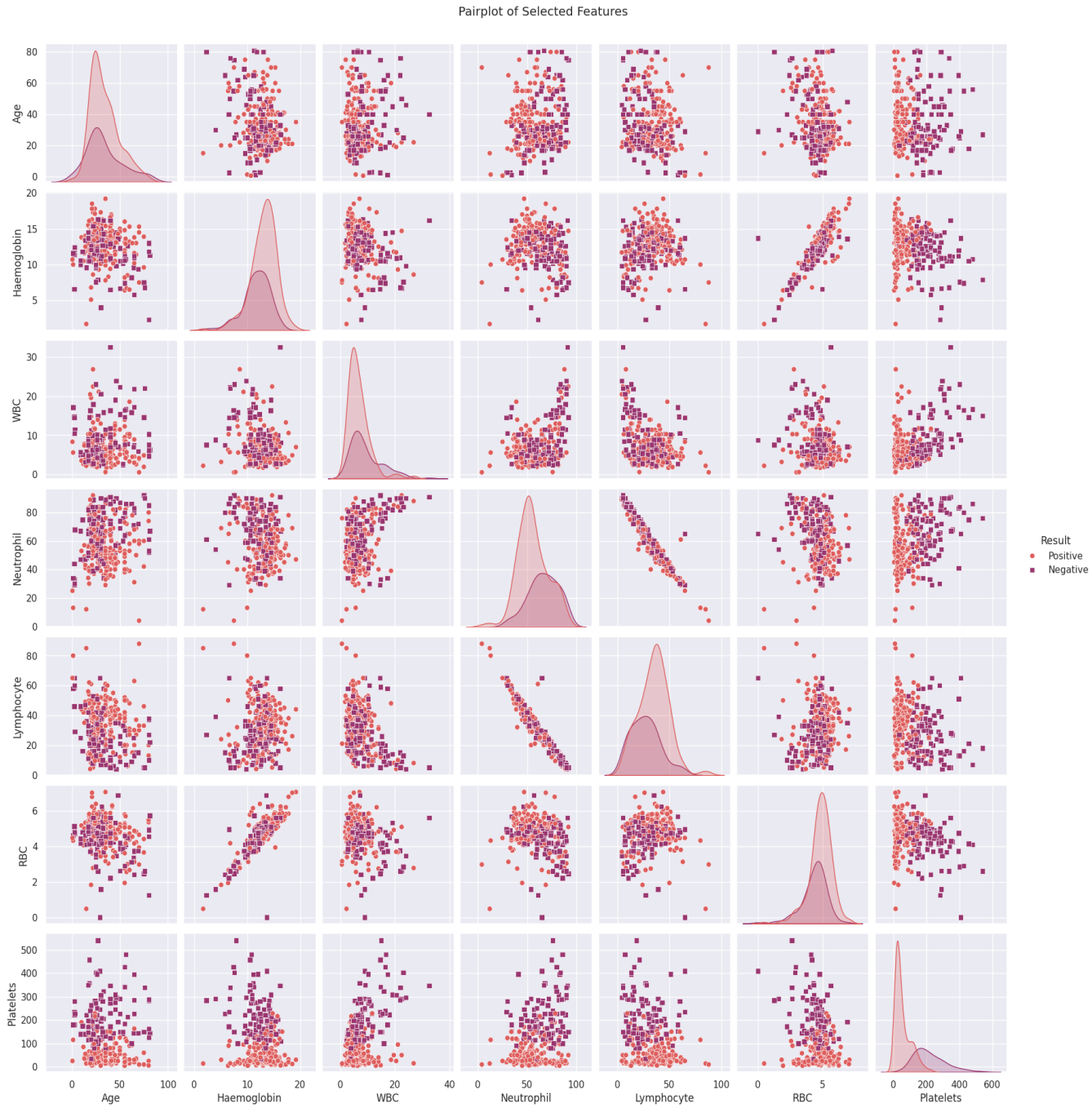

FIGURE 5. Pair plot of various features of the employed dataset

the results to increase the dataset's predicted accuracy. It's possible that some decision trees may anticipate the right output while others may not since random forest uses numerous trees to forecast the class of the dataset.

4) Naive Bayes

The naive Bayes algorithm [23] relies on Bayes' theorem, and it surmises that any single feature of a dataset is conditionally independent of the given class output. Gaussian Naive Bayes algorithm expresses the probability of a hypothesis. Though this algorithm is mostly useful for continuous data, it is also very efficacious for classification data. This model is quite simple yet effective.

5) AdaBoost

Adaptive Boosting (AdaBoost) [24] is a technique for ensemble learning that was first developed to boost the performance of binary classifiers. The ensemble approaches AdaBoost trains and deploys trees one after the other. By connecting a series of weak classifiers in AdaBoost, boosting is imple-

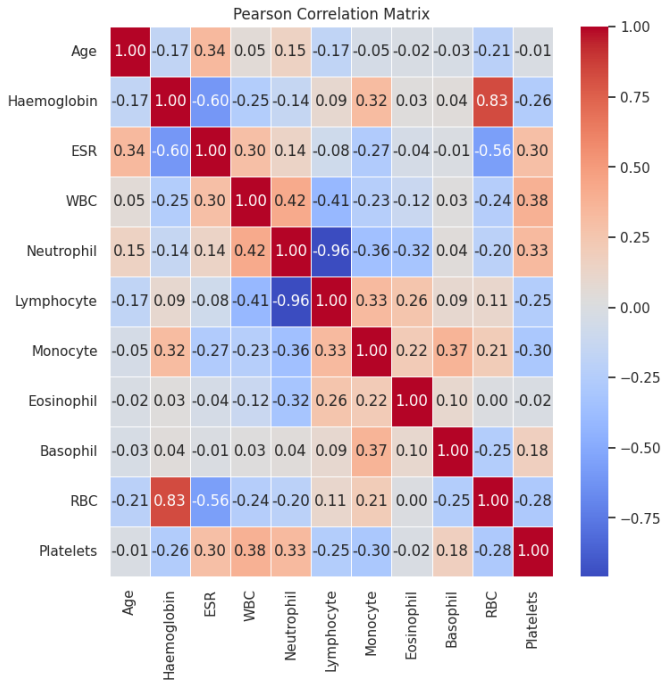

FIGURE 6. Heat map of various numerical features of the employed dataset

mented. Each weak classifier attempts to correct samples that were incorrectly classified by the weak classifier before it.

6) XGBoost

Extreme Gradient Boosting (XGBoost) algorithm is known as the XGBoost algorithm. It holds a prominent place among all the machine learning algorithms because of its performance and speed. Its portability is also a remarkable feature as it can be run on any platform and this algorithm is also integrable with multiple systems out there. Among the other boosting algorithms, it tends to perform faster [25] due to its followed concept of parallelization.

7) MLP

A multilayer perception (MLP) is a type of artificial neural network made up of several node levels, such as an input layer, an output layer, and one or more dense layers. It's a powerful model capable of learning complex patterns in data and also handles various types of data, including structured and unstructured data.

8) LightGBM

Light Gradient Boosting Machine (LightGBM) [26] is an extended version of the gradient boosting algorithm but with higher scalability. This algorithm requires less computational duration than other algorithms and gives more accurate predictions with its leaf-wise tree growth approach. Its confined depth makes the algorithm more robust as well.

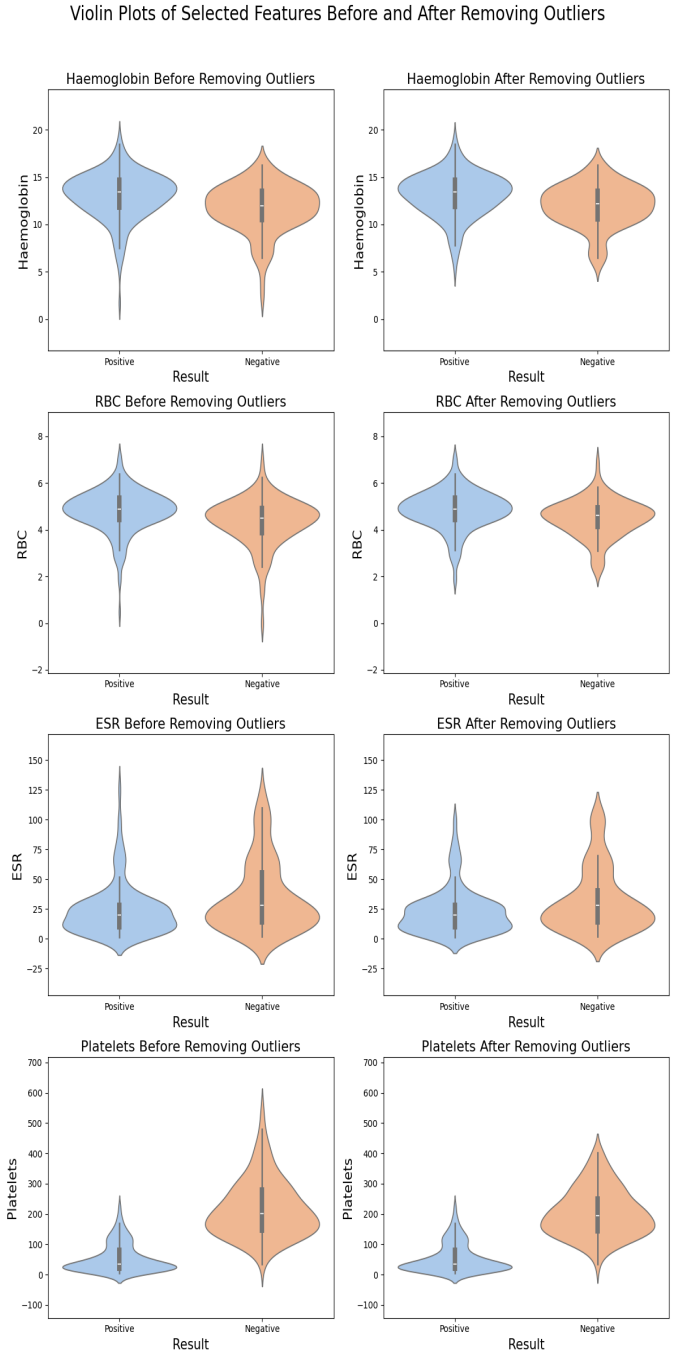

FIGURE 7. Violin plot of four selected features

9) Stacking Classifier

A stacking classifier is an ensemble machine learning approach applied within our model, specifically designed for addressing ad hoc circumstances. Multiple top-performed base algorithms are trained as the base models and implemented on the split data. Then, on the next level, the base models' primary outputs are considered the new features for the meta-classifier to get the highest final prediction. In this research, we have utilized the top-performed three statistical models (MLP, XGBoost, and Logistic Regression) as the

base models for their considerably better performance and scalability. On the other hand, the LightGBM has been used as the meta-classifier. The algorithm of the proposed stacking ensemble model applied in this automatic dengue prediction system is summarized in Algorithm 1.

Algorithm 1 Algorithm of the proposed stacking ensemble model

1: **Start**
2: **Step 1:** Split the dataset into training, validation, and test sets.
3: **Step 2:** Evaluate all models using cross-validation to obtain their accuracies.
4: **Step 3:** Select the best-performing models based on accuracy.
5: **Step 4:** Assign LightGBM (LGBM) as the meta-classifier and XGBoost (XGB), Logistic Regression (LR), and Multilayer Perceptron (MLP) as Base models.
6: **Step 5:** Train the base models (XGB, LR, MLP) with the training samples.
7: **Step 6:** Generate a new dataset using the predictions from the base models.
8: **Step 7:** Train the meta-classifier (LGBM) with the validation samples.
9: **Step 8:** Perform inference on the test samples using the trained meta-classifier (LGBM).
10: **End**

10) ANN

Artificial Neural Network (ANN) [27] algorithm is known as its working mechanism is quite similar to the human brain. Its work process includes an activation function that is a vital component facilitating the generation of output layers through the summation of input products and corresponding weights. Moreover, the inputs of the dataset are processed in the forward direction. Last but not least, ANN is simpler than other neural networks because of its feed-forward characteristics.

11) CNN

Convolutional Neural Network (CNN) indicates a class of deep learning models whose core architecture is the convolutional layer that processes the structured arrays of data. CNN model, often referred to as ConvNet, is specifically designed to automatically and adaptively learn spatial hierarchies of features, primarily focusing on image and video-based data.

12) GRU

In this research, a lightweight variant of the LSTM (Long Short-Term Memory) model, GRU (Gated Recurrent Unit) has been implemented. It is distinct in its integration of both long-term and short-term memory within its hidden state. This modified algorithm features two essential gates: the update gate and the reset gate, each designed with a clear understanding of memory mechanisms. The update gate is responsible for retaining memory information, while the reset gate facilitates memory-forgetting processes. The equations for both gates are almost similar, but the weights are distinct for both cases.

13) Bi-LSTM

For sequence-based classification problems, the Bi-LSTM algorithm performs conspicuously better than the LSTM model. This model is regarded as the extended version of the LSTM model. In this algorithm, the encoding is performed in both the forward and backward directions. Finally, the result is concatenated from both ends.

14) TabPFN

Tabular data Prior-data Fitted Network (TabPFN) [28] works significantly well on tabular classification datasets. The best feature of this model is that without any prior training, feature selection, and hyperparameter tuning, this model brings out the highest possible accuracy within a few seconds of application. It utilizes the in-context learning method that enables the model to learn the sequences from the given input.

15) TabTransformer

TabTransformer is a transformer-based model that specializes in training for tabular data. This model is quite prevalent for its efficiency and scalability in terms of handling tabular data without that many intensive preprocessing tasks. For our study, we have used the TensorFlow data pipeline with a sigmoid activation function, and this activation function has been applied to the outputs of all the layers to bring out the values within the range of 0 to 1. However, this model's self-attention mechanism is remarkably handy in intricating [29] and finding relationships among features of the employed dataset.

F. HYPERPRAMETERS OPTIMIZATION

Hyperparameter optimization is a fundamental procedure focused on identifying the optimal values for a machine learning model's parameters. Usually, prior to the application of this technique, conventional machine learning models are typically implied on the dataset, often resulting in suboptimal performance metrics. However, upon integrating hyperparameter optimization methodologies such as GridSearchCV and RandomizedSearchCV, statistical algorithms exhibit notable improvements in predictive performance compared to their pre-implementations. Overall, this process plays a vital role in achieving the ultimate optimal performance metrics such as precision, recall, accuracy, and F1 score. Conversely, the Keras Tuner optimization method is deployed to optimize deep learning algorithms in this work. This approach alleviates the need for extensive manual experimentation, enhancing model performance and improved accuracy rates. Conversely, the Keras Tuner optimization method is deployed to optimize deep learning algorithms. This approach alleviates the necessity for extensive manual experimentation, resulting in enhanced model performance and improved accu-

racy rates. In this research, GridSearchCV has been utilized for all the applied machine learning techniques, MLP, and LightGBM models. Conversely, we employed Keras Tuner to optimize deep learning models, including ANN, CNN, Bi-LSTM, and GRU.

G. FEATURE SELECTION METHOD

After implementing multiple machine learning and deep learning algorithms on the dataset, five feature selection methods have been applied that increase interpretability and model accuracy. These five methods are as follows: Pearson correlation, Recursive Feature Elimination, SelectKBest with ANOVA F-value, Chi-Square Test, and Extra Trees Classifier.

1) Pearson Correlation

The Pearson Correlation-based feature selection method is essential to finding the relationship between two features. Based on the input variables of the utilized CBC dengue dataset, we have decided to iterate this method via multiple machine-learning models. As this method determines the optimal accuracy among features by plotting points relative to the line, it enhances the clarity of visualizing the method's impact.

2) RFE

The Recursive Feature Elimination (RFE) feature selection method works on removing the most unimportant feature from the dataset. This process keeps repeating until it eliminates all the unimportant features from the dataset and selects the most important one.

3) SelectKBest

SelectKBest is a filter-based or univariate feature selection method in machine learning. It utilizes metrics like the chi-square test, ANOVA F-value, or both statistical tests to evaluate each feature. After scoring features, it selects only the most salient k features based on the score.

4) Chi-Square Test

A chi-square test is a statistical testing method for categorical data that is also considered a hypothesis testing method. This testing method determines the notable difference between the observed and expected data.

5) Extra Trees Classifier

The Extra Tree Classifier, a type of ensemble learning, is a powerful tool for classification tasks. It adds randomness to feature selection and combines results from multiple uncorrelated decision trees in a forest to predict outcomes.

IV. RESULTS AND DISCUSSIONS

This section presents the results of the applied AI models for the proposed automatic dengue detection system. Precision, recall, macro F1, and accuracy scores are some important performance measurement metrics demonstrated in this section that are determined using (3) to (6), respectively.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP+FP+TN+FN}} \tag{6}$$

Table 2 presents the performance metrics of various machine learning models evaluated for dengue prediction without hyperparameter optimization. The TabPFN classifier achieved the highest accuracy of 94.79%, with a precision of 0.9387 and a macro F1 score of 0.9419. Other notable performances include the stacking classifier model with an accuracy of 93.75%, and the XGBoost model, which demonstrated high recall (0.9297) and precision (0.9122), achieving an accuracy of 92.71%.

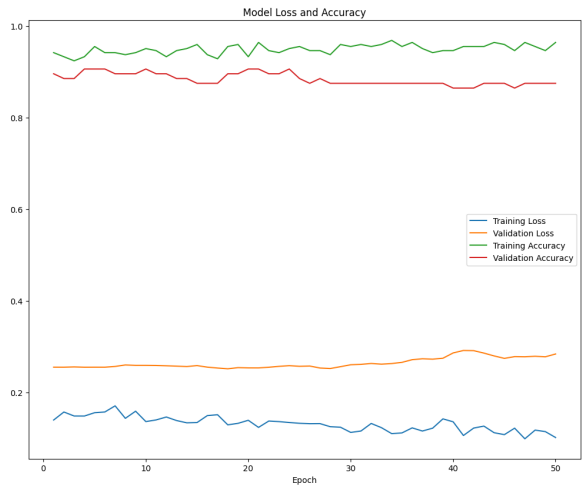

FIGURE 8. Training and validation accuracy and loss vs. epochs for the TabPFN model

The training and validation accuracy and loss of the TabPFN model with the change of epochs are illustrated in Figure 8. The training and validation accuracy remains high and stable at 94% to 95% throughout the epochs. In contrast, the training and validation loss show minimal fluctuations, indicating a consistently performing model with good generalization capabilities.

A. FEATURE SELECTIONS TECHNIQUES RESULTS

Table 3 displays the dengue prediction accuracies of various algorithms after applying different feature selection methods, including Pearson Correlation, RFE, SelectKBest, Chi-Square Test, and ExtraTree Classifier. The Random Forest algorithm consistently performs well across all methods, with a maximum accuracy of 93.75% using both the Pearson Correlation and Chi-Square Test methods, demonstrating the

TABLE 2. Performance metrics for various models (without hyperparameter optimization)

Algorithms	Precision	Recall	Accuracy (%)	Macro F1 Score
Logistic Regression	0.8885	0.9141	90.62	0.8982
SVM	0.8885	0.9141	90.62	0.8982
Naive Bayes	0.7884	0.7969	81.25	0.7922
Random Forest	0.8925	0.8984	90.62	0.8953
AdaBoost	0.8732	0.8672	88.54	0.8701
XGBoost	0.9122	0.9297	92.71	0.9198
MLP	0.8550	0.8207	83.33	0.8333
LightGBM	0.9175	0.9080	92.19	0.9124
Stacking Classifier	0.9364	0.9219	93.75	0.9285
ANN	0.7723	0.7723	79.17	0.7723
CNN	0.7969	0.7884	81.25	0.7922
GRU	0.8046	0.7751	81.25	0.7856
Bi-LSTM	0.8125	0.8416	81.25	0.8084
TabPFN	0.9387	0.9453	94.79	0.9419
TabTransformer	0.8828	0.8828	89.58	0.8828

TABLE 3. Accuracy (%) for various algorithms after applying feature selection methods

Algorithms	Pearson Correlation	RFE	SelectKBest	Chi-Square Test	ExtraTree Classifier
Logistic Regression	91.667	91.667	91.667	93.750	91.667
SVM	88.542	88.542	88.542	90.625	88.542
Random Forest	93.750	92.708	91.667	93.750	86.458
Naive Bayes	81.250	81.250	81.250	80.208	91.667
AdaBoost	87.5	88.542	87.500	85.417	88.542
XGBoost	91.667	93.750	91.667	91.667	85.417
MLP	89.583	90.625	90.625	90.625	86.458
LightGBM	91.667	92.708	91.667	90.625	86.458
Stacking Classifier	92.19	93.75	93.75	87.50	90.63

robustness of this algorithm with feature selection. Similarly, the stacking classifier and XGB model both achieved the highest accuracy score of 93.75% for the RFE method, and the stacking classifier also matched this score for the SelectKBest method.

B. HYPERPARAMETER OPTIMIZATION RESULTS

Table 4 presents the accuracy of various machine learning algorithms after applying the GridSearchCV hyperparameter optimizer. The stacking classifier achieves the highest accuracy at 96.88%, indicating a significant improvement compared to its performance without optimization (93.75%). The accuracy of Logistic Regression improved from 90.62% to 92.95%, and XGBoost improved from 92.71% to 93.75%. Overall, most algorithms show improved accuracy post-optimization, highlighting the effectiveness of GridSearchCV in enhancing model performance.

The accuracy and best-optimized hyperparameters of various deep learning models after applying Keras Tuner for hyperparameter optimization are summarized in Table 5. The CNN model achieved the highest accuracy at 86.53%, while both the Bi-LSTM and GRU models reached an accuracy of 83.33%.

The confusion matrix for the best-performing stacking classifier model is illustrated in Figure 9, highlighting its efficient predictive performance. The model accurately classified 20 instances of class 0 and 42 cases of class 1, with only 2 misclassifications, demonstrating its high effectiveness in distinguishing between the two dengue classes for the employed CBC dataset.

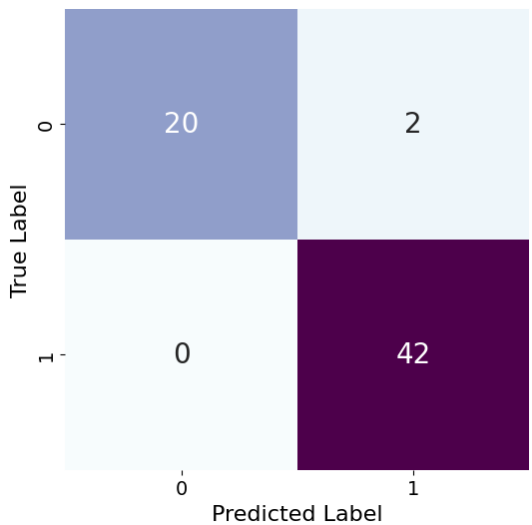

FIGURE 9. Confusion matrix for the best-performed stacking classifier model

Figure 10 displays the Receiver Operating Characteristic (ROC) curve for the stacking classifier, which achieved an impressive Area Under the Curve (AUC) score of 0.9919. This high AUC value indicates the applied stacking ensemble model's excellent ability to distinguish between the positive and negative dengue classes, confirming its robust predictive performance.

In this work, Local Interpretable Model-agnostic Explanation (LIME)-based eXplainable AI (XAI) has been utilized

TABLE 4. Accuracy for various ML algorithms after applying GridSearchCV hyperparameter optimizer

Algorithms	Best Hyperparameters	Accuracy (%)
Logistic Regression	C = 10, penalty = $l2$	92.95
SVM	C = 10, kernel= linear	92.26
Random Forest	max_depth = 5, n_estimators = 150	92.6
AdaBoost	learning_rate = 0.1, n_estimators = 150	92.97
XGBoost	learning_rate = 0.1, n_estimators = 150	92.928
LightGBM	learning_rate = 0.05, n_estimators = 200, feature_fraction = 1.0	93.75
MLP	activation = ReLU, hidden_layer_sizes = (100,)	95.83
Stacking Classifier	base_estimators = 100, $alpha$ = 0.01, final_l.rate = 0.01, final_estimator = 100	**96.88**

TABLE 5. Accuracy for various deep learning models after applying Keras Tuner hyperparameter optimizer

Models	Best Hyperparameters	Accuracy (%)
CNN	filters = 64, units = 480, activation = ReLU, optimizer = RMSProp, learning_rate = 0.0001	86.53
Bi-LSTM	unit 1 = 96,unit 2 = 64, unit 3 = 32, activation = ReLU, optimizer = Adam, learning_rate = 0.001,	83.33
GRU	units = 64, activation = sigmoid, optimizer = Adam, learning_rate = 0.001, dropout = 0.2	83.33

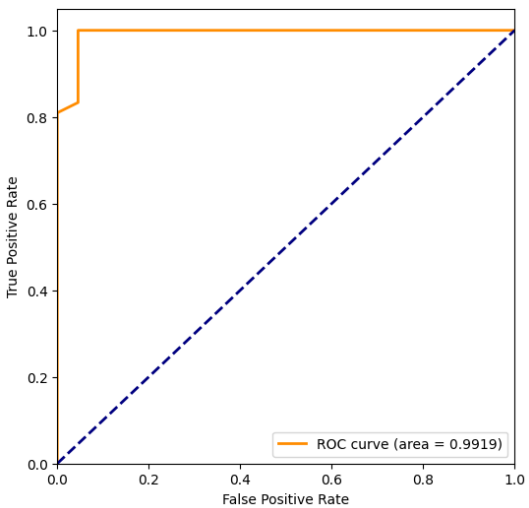

FIGURE 10. ROC-AUC curve for the stacking classifier

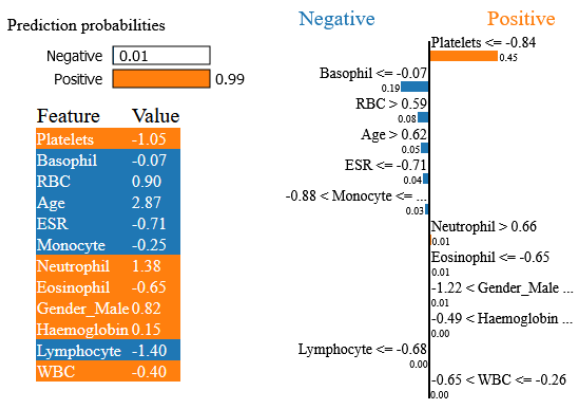

FIGURE 11. Dengue prediction interpretation of a positive case instance using LIME explainable AI

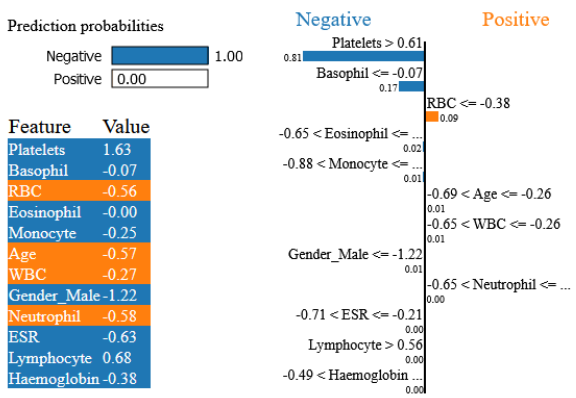

FIGURE 12. Dengue prediction interpretation of a negative case instance using LIME explainable AI

to explain how the black box stacking ensemble machine learning model predicts an outcome. This framework works efficiently by approximating the model's outcomes locally, for instance, training an interpretable linear model on various perturbed versions of the instance. This explanation provides insight into why the model classifies instances as either 'Positive' or 'Negative' in the employed dengue CBC report-based dataset. Figure 11 shows a confidence score of 0.99 for the positive case. Platelets, Neutrophil, Eosinophil, Gender, Haemoglobin, and WBC are the six most impactful features that played a significant role in determining the corresponding positive class. On the other hand, according to Figure 12, Plateles, Basophil, Lymphocyte, Monocyte, and Eosinophil act as the most prominent factors for predicting negative dengue class.

Figure 13 presents a radar chart comparing the performance of the top five models, stacking, MLP, LightGBM, TabPFN, and AdaBoost, across different performance met-

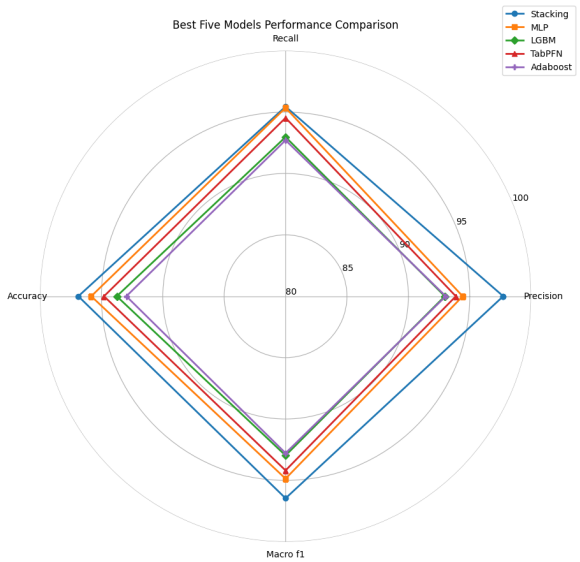

FIGURE 13. Radar chart of the best-performing models

rics. The stacking ensemble model demonstrates superior performance across all metrics, particularly in accuracy and recall. In contrast, the other models show competitive but slightly lower scores for the proposed dengue prediction system.

A comparative analysis of the proposed dengue prediction system with various other studies is presented in Table 6. The proposed system, using a LightGBM meta-classifier-based stacking ensemble technique and data from CBC reports in Dhaka, Bangladesh, achieved an accuracy of 96.88% and an F1 score of 0.9646, which is competitive with other high-performing models like the ANN (96% accuracy) and ETC (99.12% accuracy). Notably, the proposed system demonstrated strong precision and recall metrics (97.73% and 95.45%, respectively), highlighting its effectiveness in accurately predicting dengue cases compared to other articles.

V. CONCLUSIONS

This research introduces various AI techniques to predict the dengue virus employing a private CBC report dataset. The dataset comprises 320 samples and 14 hematology features collected from local hospitals in Dhaka, Bangladesh. Diverse dataset preprocessing steps are implemented to the dataset, i.e., handling missing values and outliers, one-hot encoding, feature standardization, synthetic oversampling, and removing redundant features. Various machine learning, deep learning and transformer-based models are applied to predict positive and negative dengue cases. The hyperparameters of the applied models are optimized by employing the GridSearchCV and Keras Tuner frameworks. A stacking ensemble approach constructed with LightGBM meta-classifier and XGBoost, Logistic Regression, and MLP base learners accomplishes the best performance among the machine learning models. The MLP neural network model performs best among the deep learning models. Finally, the LIME XAI approach has been applied to investigate the salient features and interpret the predictions provided by the stacking classifier. In the future, the employed dataset can be expanded by adding new data from a larger cohort of patients. Multimodal architecture can be applied using blood smear images for the same patient data. A multiclass problem can be defined using mild, moderate, severe positive, and negative dengue case samples.

REFERENCES

[1] M. B. Khan, Z.-S. Yang, C.-Y. Lin, M.-C. Hsu, A. N. Urbina, W. Assavalapsakul, W.-H. Wang, Y.-H. Chen, and S.-F. Wang, "Dengue overview: An updated systemic review," Journal of Infection and Public Health, vol. 16, pp. 1625–1642, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1876034123002587

[2] N. Ali, "The recent burden of dengue infection in Bangladesh: A serious public health issue," Journal of Infection and Public Health, vol. 17, pp. 226–228, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1876034123004392

[3] S. Roy, A. Biswas, M. T. A. Shawon et al., "Land use and meteorological influences on dengue transmission dynamics in Dhaka city, Bangladesh," Bulletin of the National Research Centre, vol. 48, 2024. [Online]. Available: https://doi.org/10.1186/s42269-024-01188-0

[4] N. Sharif, N. Sharif, A. Khan, and S. K. Dey, "The epidemiologic and clinical characteristics of the 2023 dengue outbreak in Bangladesh," Open Forum Infectious Diseases, vol. 11, 2024. [Online]. Available: https://doi.org/10.1093/ofid/ofae066

[5] M. E. H. Kayesh, I. Khalil, M. Kohara, and K. Tsukiyama-Kohara, "Increasing dengue burden and severe dengue risk in Bangladesh: An overview," Tropical Medicine and Infectious Disease, vol. 8, 2023. [Online]. Available: https://www.mdpi.com/2414-6366/8/1/32

[6] D. C. Kajeguka, F. M. Mponela, E. Mkumbo, A. N. Kaaya, D. Lasway, R. D. Kaaya, M. Alifrangis, E. Elanga-Ndille, B. T. Mmbaga, and R. Kavishe, "Prevalence and associated factors of dengue virus circulation in the rural community, handeni district in Tanga, Tanzania," Journal of Tropical Medicine, vol. 2023, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2023/5576300

[7] M. A. Kabir, H. Zilouchian, M. A. Younas, and W. Asghar, "Dengue detection: advances in diagnostic tools from conventional technology to point of care," Biosensors, vol. 11, no. 7, p. 206, 2021.

[8] C. Davi, A. Pastor, T. Oliveira, F. B. de Lima Neto, U. Braga-Neto, A. W. Bigham, M. Bamshad, E. T. Marques, and B. Acioli-Santos, "Severe dengue prognosis using human genome data and machine learning," IEEE Transactions on Biomedical Engineering, vol. 66, pp. 2861–2868, 2019.

[9] D. Sarma, S. Hossain, T. Mittra, M. A. M. Bhuiya, I. Saha, and R. Chakma, "Dengue prediction using machine learning algorithms," in Humanitarian Technology Conference. IEEE, 2020, pp. 1–6.

[10] E. Fernández, M. Smieja, S. D. Walter, and M. Loeb, "A predictive model to differentiate dengue from other febrile illness," BMC Infectious Diseases, vol. 16, pp. 1–7, 2016.

[11] H. Mayrose, G. M. Bairy, N. Sampathila, S. Belurkar, and K. Saravu, "Machine learning-based detection of dengue from blood smear images utilizing platelet and lymphocyte characteristics," Diagnostics, vol. 13, 2023. [Online]. Available: https://www.mdpi.com/2075-4418/13/2/220

[12] J. D. Mello-Román, J. C. Mello-Román, S. Gomez-Guerrero, and M. García-Torres, "Predictive models for the medical diagnosis of dengue: a case study in paraguay," Computational and Mathematical Methods in Medicine, vol. 2019, 2019.

[13] S. K. Dey, M. M. Rahman, A. Howlader, U. R. Siddiqi, K. M. M. Uddin, R. Borhan, and E. U. Rahman, "Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data in bangladesh: A machine learning approach," PLoS One, vol. 17, p. e0270933, 2022.

[14] B. Abdualgalil, S. Abraham, and W. M. Ismael, "Early diagnosis for dengue disease prediction using efficient machine learning techniques based on clinical data," Journal of Robotics and Control (JRC), vol. 3, pp. 257–268, 2022.

[15] S. Q. Ong, P. Isawasan, A. M. M. Ngesom, H. Shahar, A. M. M. Lasim, and G. Nair, "Predicting dengue transmission rates by comparing different

TABLE 6. Comparison of the proposed system with similar dengue prediction studies

References	Data Source	Sample Size	Classifier	Accuracy (%)	F1 score	Other Metrics
[8]	Data from three hospitals of Recife, Brazil	102	ANN	86%	N/A	Sensitivity (Recall): 98%, Specificity: 51%
[9]	Real-time patient dataset from hospitals in Bangladesh	209	Decision Tree	79%	0.79	Precision: 79%, Recall: 79%
[10]	Data from Tegucigalpa and San Pedro Sula, Honduras	548	Logistic Regression	69.2%	N/A	Sensitivity (Recall): 86.2%
[11]	Hematology Laboratory, Kasturba Hospital, Manipal, India	94 blood smear images	SVM	95.74%	0.9434	Sensitivity (Recall): 98.15%, Specificity: 92.50%, Precision: 94.64%
[12]	Public health system of Paraguay	4332	ANN	96%	N/A	Sensitivity (Recall): 96%, Specificity: 97%
[14]	Epidemiological Monitoring Center of Yemen	6694	ETC	99.12%	0.9913	Precision: 99.08%, Recall: 99.18%
[15]	Health Department of Federal Territory of Malaysia	N/A	XGBoost	81%	N/A	N/A
[16]	University of Malaya Medical Centre	170	Adaptive Logistic Regression	N/A	0.93	Precision: 65%, Recall: 100%
This work	CBC reports from local hospitals in Dhaka, Bangladesh	320	Stacking Ensemble	96.88%	0.9646	Precision: 97.73%, Recall: 95.45%, AUC: 0.99

machine learning models with vector indices and meteorological data," Scientific Reports, vol. 13, 2023.

[16] J. K. Chaw, S. H. Chaw, C. H. Quah, S. Sahrani, M. C. Ang, Y. Zhao, and T. T. Ting, "A predictive analytics model using machine learning algorithms to estimate the risk of shock development among dengue patients," Healthcare Analytics, vol. 5, 2024.

[17] M. T. Sarwar and M. Al Mamun, "Prediction of dengue using machine learning algorithms: Case study Dhaka," in International Conference on Electrical, Computer & Telecommunication Engineering. IEEE, 2022, pp. 1–6.

[18] T. Akter, M. T. Islam, M. F. Hossain, and M. S. Ullah, "A comparative study between time series and machine learning technique to predict dengue fever in dhaka city," Discrete Dynamics in Nature and Society, vol. 2024, no. 1, p. 2757381, 2024.

[19] M. A. Majeed, H. Z. M. Shafri, Z. Zulkafli, and A. Wayayok, "A deep learning approach for dengue fever prediction in Malaysia using LSTM with spatial attention," International Journal of Environmental Research and Public Health, vol. 20, 2023. [Online]. Available: https://www.mdpi.com/1660-4601/20/5/4130

[20] R. Real, A. M. Barbosa, and J. M. Vargas, "Obtaining environmental favourability functions from logistic regression," Environmental and Ecological Statistics, vol. 13, pp. 237–245, 2006.

[21] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of support vector machine (SVM) learning in cancer genomics," Cancer Genomics & Proteomics, vol. 15, pp. 41–51, 2018.

[22] A. Liaw, M. Wiener et al., "Classification and regression by randomforest," R news, vol. 2, pp. 18–22, 2002.

[23] D. J. Hand and K. Yu, "Idiot's bayes—not so stupid after all?" International Statistical Review, vol. 69, no. 3, pp. 385–398, 2001.

[24] A. Yulianto, P. Sukarno, and N. A. Suwastika, "Improving AdaBoost-based Intrusion Detection System (IDS) Performance on CIC IDS 2017 Dataset," in Journal of Physics: Conference Series, vol. 1192. IOP Publishing, 2019.

[25] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[26] J. Zhang, D. Mucs, U. Norinder, and F. Svensson, "LightGBM: An effective and scalable algorithm for prediction of chemical toxicity–application to the tox21 and mutagenicity data sets," Journal of Chemical Information and Modeling, vol. 59, pp. 4150–4158, 2019.

[27] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016, http://www.deeplearningbook.org.

[28] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "TabPFN: A transformer that solves small tabular classification problems in a second," arXiv preprint arXiv:2207.01848, 2022.

[29] T. K. Vyas, "Deep learning with tabular data: A self-supervised approach," arXiv preprint arXiv:2401.15238, 2024.

NUSRAT JAHAN RIYA received the B.Sc. degree in Computer Science and Engineering from the Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh. Her current research interests include artificial intelligence, machine learning, deep learning, and natural language processing.

MRITUNJOY CHAKRABORTY received the B.Sc. degree in Computer Science and Engineering from the Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh. He is currently working as a Research Assistant at the Department of ECE, North South University, Dhaka, Bangladesh. His research interests include computer vision, natural language processing, and machine learning.

RIASAT KHAN earned his B.Sc. degree in Electrical and Electronic Engineering from the Islamic University of Technology, Bangladesh, in 2010. He further pursued his academic journey, completing both the M.Sc. and Ph.D. degrees in Electrical Engineering at New Mexico State University, Las Cruces, USA, in 2018. Presently, he holds the position of Associate Professor in the Department of Electrical and Computer Engineering at North South University, Dhaka, Bangladesh. His research interests include data science, machine learning, computational bioelectromagnetics, and power electronics.