Ensemble model-based arrhythmia classification with local interpretable model-agnostic explanations

Md. Rabiul Islam[1], Tapan Kumar Godder[2], Ahsan-Ul-Ambia[3], Ferdib-Al-Islam[4], Anindya Nag[4], Bulbul Ahamed[5], Nujhut Tanzim[4], Md. Estiak Ahmed[6]
[1]Department of Biomedical Engineering, Islamic University, Kushtia, Bangladesh
[2]Department of Information and Communication Technology, Islamic University, Kushtia, Bangladesh
[3]Department of Computer Science and Engineering, Islamic University, Kushtia, Bangladesh
[4]Department of Computer Science and Engineering, Northern University of Business and Technology Khulna, Khulna, Bangladesh
[5]Department of Computer Science and Engineering, Sonargaon University, Dhaka, Bangladesh
[6]Computer Science and Engineering Discipline, Khulna University, Khulna, Bangladesh

Article Info

Article history:

Received May 6, 2024
Revised Dec 22, 2024
Accepted Jan 27, 2025

Keywords:

Arrhythmia
Electrocardiogram
Extreme gradient boosting
LIME
Machine learning

ABSTRACT

Arrhythmia can lead to heart failure, stroke, and sudden cardiac arrest. Prompt diagnosis of arrhythmia is crucial for appropriate treatment. This analysis utilized four databases. We utilized seven machine learning (ML) algorithms in our work. These algorithms include logistic regression (LR), decision tree (DT), extreme gradient boosting (XGB), k-nearest neighbors (KNN), naïve Bayes (NB), multilayer perceptron (MLP), AdaBoost, and a bagging ensemble of these approaches. In addition, we conducted an analysis on a stacking ensemble consisting of XGB and bagging XGB. This study examines various arrhythmia detection techniques using both a single base dataset and a composite dataset. The objective is to identify the optimal model for the combined dataset. This study aims to evaluate the efficacy of these models in accurately categorizing normal (N) and abnormal (A) heartbeats as binary classes. The empirical findings demonstrated that the stacking ensemble approach exhibited superior accuracy when used with the combined dataset. Arrhythmia classification models rely on this as a crucial component. The binary classification achieved an accuracy of 98.61%, a recall of 97.66%, and a precision of 97.77%. Subsequently, the local interpretable model-agnostic explanations (LIME) technique is employed to assess the prediction capability of the model.

Corresponding Author:

Ferdib-Al-Islam
Department of Computer Science and Engineering, Northern University of Business and Technology Khulna
Khulna-9100, Bangladesh
Email: ferdib.bsmrstu@gmail.com

1. INTRODUCTION

Arrhythmias, which are irregular rates or rhythms of heartbeats, reveals faulty heart functions. Severe arrhythmias can result in insufficient blood circulation, leading to potential harm to the brain and heart, and in some circumstances, abrupt cardiac death. Hence, it is crucial to monitor cardiac activities and identify arrhythmias for the sake of individuals' welfare. Arrhythmia detection can be utilized to promptly identify the onset of heart disease, expedite the administration of initial medical assistance, and ultimately preserve human lives. Arrhythmias, characterized by abnormal heart rhythm, are a prevalent cardiac condition that impacts a significant number of individuals globally and has the potential to be life-threatening. Based on data provided by the World Health Organization (WHO), cardiovascular disorders such as stroke and heart failure, have been the primary causes of death worldwide in recent years.

Arrhythmias can manifest as a sign of diverse underlying illnesses, with cardiovascular disease being among them. To reduce the risk of potentially fatal consequences such as stroke, coronary artery disease, and sudden cardiac arrest and to avoid the need for future intrusive and demanding therapies, early diagnosis of arrhythmias is essential [1].

The electrocardiogram (ECG) examines the heart's electrical activity throughout time as repeating signals; this makes it an essential tool for monitoring cardiac functioning and detecting abnormal heart rhythms. Periods represent the changing patterns of rhythmic activity within a single heartbeat. Arrhythmia identification involves classifying cardiac cycles as either normal or abnormal. The former describes the heart in its normal functioning, whereas the latter describes abnormal heart cycles that may cause harm to the heart or even cardiac arrest. Rapid detection of abnormal cardiac cycles using ECG data is our primary goal.

Machine learning (ML) has demonstrated its efficacy in automating and improving the accuracy of arrhythmia diagnosis in the healthcare industry. Heartbeats can be classified using many approaches, with ML models being one of them. Classifiers train models to reliably divide heartbeats into five categories: normal (N), supraventricular ectopic (SVEB), ventricular ectopic (VEB), fusion (F), and unknown (Q). However, the objective of this project is to develop classifiers that may create models capable of categorizing heartbeats into two distinct classes: normal (0) and abnormal or arrhythmia (1). The basis dataset utilized in this work comprises a substantial proportion of the normal heartbeat class while exhibiting minimal proportions of SVEB, VEB, F, and Q classes. Combine the SVEB, VEB, and F classes into a single class representing aberrant heartbeat or arrhythmia. For the purpose of enhancing model performance, this work integrates four datasets. The dataset had a limited number of examples for the SVEB, VEB, F, and Q classes. The model faces greater difficulty in discerning the distinctive characteristics and patterns of a class when there are limited instances of it. This could impede the model's ability to generalize effectively and accurately classify novel examples of that class. This work contributes to the previous literature [2] on ECG classification by utilizing a ML method and ensemble learning model to enhance accuracy. To do this, information pertaining to the SVEB, VEB, and F categories was extracted from several sources and incorporated into the original foundational dataset. Upon merging fresh data into the dataset, the SVEB, VEB, and F multiple classes were transformed into an abnormal heartbeat class or arrhythmia class, while the normal class remained unaltered. The abnormal heartbeat class now has a greater number of instances.

Ensembling multiple models instead of relying on a single ML model offers several benefits. Firstly, it reduces the risk of overfitting by combining predictions from different models, leading to more robust and generalizable results. Secondly, ensembles can capture a wider range of patterns and relationships in the data, enhancing predictive performance. Lastly, ensembling can help mitigate the weaknesses of individual models, resulting in better overall performance. Using a combined dataset composed of five datasets can also improve model performance. By merging multiple datasets, we can increase the diversity and richness of the data, enabling the model to learn more comprehensive patterns and relationships. This approach can lead to better generalization to unseen data and improved model performance compared to using a single dataset alone. In summary, ensembling models and combining datasets are effective strategies to enhance predictive performance, increase model robustness, and improve generalization ability. Models can readily identify the distinctive characteristics and patterns of a certain class, enabling them to precisely classify new instances of that class. Next, the synthetic minority oversampling technique (SMOTE) will be employed to generate more synthetic data by oversampling the minority class. The integration of four databases and the implementation of SMOTE has resulted in enhanced overall accuracy, recall, and precision metrics.

Throughout the years, numerous researchers have employed various methodologies to identify the presence of arrhythmia disease. The finished works are succinctly stated in the following terms. A model for ECG heartbeat classification proposed by Al-Mousa *et al*. [1] improves recollection for categories F and Q while maintaining the same recall for the other kinds. This analysis relied on the Massachusetts Institute of Technology-Beth Israel Hospital arrhythmia database (MIT-BIH) supraventricular database as its basic dataset. To improve the recall for the F and Q classes, the authors combined data from other datasets and added it to the fundamental dataset. For this combined dataset, they used SMOTE to achieve balance. The random forest (RF) algorithm outperformed all others with an accuracy rate of 97% and recall values of 93, 95, 95, and 30% for N, SVEB, VEB, F, and Q, respectively. Sakib *et al.* [2] discussed the difficulty of incorporating artificial intelligence (AI) into advanced internet of things (IoT) sensors to detect irregular heart rhythms using ECG data. The authors laid forth a method for arrhythmia classification via a lightweight deep learning (DL) strategy. This technique classified four distinct kinds of heartbeats using a 1D convolutional neural network (CNN) design. Four separate PhysioNet datasets were used for this evaluation. Compared to well-established approaches like RF, k-nearest neighbors (KNN), and delay nonlinear equation-based optimization, the proposed DL strategy performed better in the experiments. When employed on virtualized microcontrollers connected to IoT sensors, the DL model shows outstanding performance in terms of processing time and memory use. When tested on the MIT-BIH supraventricular dataset (94.12% accuracy), the MIT-BIH arrhythmia dataset (94.97% accuracy), the Institute of Cardiological Technics

(INCART) 12-lead arrhythmia dataset (94.97% accuracy), and the Sudden Cardiac Death Holter (SCDH) (96.67% accuracy), the proposed model performs well. A new method for identifying irregular heartbeats was presented by Wang *et al*. [3] that makes use of the EasyEnsemble algorithm in conjunction with global heartbeat data obtained from the MIT-BIH arrhythmia database. In addition to outperforming the competition overall, their suggested strategy boosts minority category performance while keeping majority category performance high. On average, the suggested model achieved a recall rate of 55.4% for type F, an accuracy of 91.7% for type N, 89.9% for type VEB, and 87.8% for type SVEB.

To classify ECG data, Khan *et al*. [4] revealed a tailored CNN model. The authors made use of PhysioNet's MIT-BIH arrhythmia database. With an average recall of 95.40% and a total accuracy of 95.2%, their proposed model successfully categorizes arrhythmia. A model that chooses the best subsets of features to distinguish one class from another was presented in [5]. A support vector machine (SVM) binary classifier is used to accomplish this task by comparing two sets of data one after the other. utilizing the MIT-BIH arrhythmia database, the proposed feature selection method was assessed. Classification accuracy is 86.66% on average when using the proposed feature selection method. For the classes N, SVEB, VEB, and F, it achieved recall rates of 88.94, 79.06, 85.48, and 93.81%, respectively. In their investigation, Alarsan and Younes [6] used a combined database of the MIT-BIH supraventricular arrhythmia and the MIT-BIH arrhythmia databases, and they used three models to it: decision tree (DT), RF, and gradient boosting trees. Using RF for multiclassification, the authors achieved a high accuracy of 98.03%. A diagnostic model developed for the purpose of detecting cardiac arrhythmia was published by Singh and Singh [7]. To find the most important characteristics, this study used three filter-based methods for selecting features. To test how well the feature selection method worked, the writers used three separate models: JRip, linear SVM, and RF. Using a gain ratio selected feature strategy with a selected group of 30 features and a RF classifier, the research attained a maximum accuracy of 85.58%.

Abdelmoneem *et al.* [8] presented a highly effective algorithm for detecting cardiac arrhythmia. They explored different oversampling techniques to address the issue of imbalanced datasets. The ensemble classifier, SVM, and RF with random sampling obtained a remarkable accuracy of 98.18%. This research also introduced a mobile system design that incorporates an algorithm for diagnosing and categorizing cardiac arrhythmia illnesses. Manju and Nair [9] established a model that classifies arrhythmias into ten categories, with one category representing normal conditions and the others representing distinct forms of arrhythmias. The authors derive characteristics from 12-lead ECG data. This study employed the extreme gradient boosting (XGB) algorithm to do feature reduction and balanced the dataset using the SMOTE edited nearest neighbors (ENN) technique. They employed multiple supervised ML methods for classification. The results demonstrated that the proposed model effectively categorizes different forms of arrhythmia with a remarkable level of precision, reaching an accuracy rate of 97.48%. This work presented a practical approach to accurately and efficiently classify arrhythmias by utilizing sophisticated data preprocessing techniques and ML algorithms. Peimankar *et al.* [10] introduced an ensemble learning method to automatically classify common cardiac arrhythmia. The study employed three classification algorithms, namely RF, AdaBoost, and artificial neural network (ANN), utilizing twenty-six features collected from ECG data. Upon evaluating 44 recordings from the MIT-BIH arrhythmia database, it was discovered that the RF, AdaBoost, and ANN algorithms had high individual accuracy rates of 96.16, 96.16, and 94.49%, respectively. Remarkably, the ensemble model achieves a remarkable 96.18% increase in total accuracy. The study indicated that the classification of arrhythmias using an ensemble learning approach is both reliable and user-friendly.

Sraitih *et al.* [11] introduced a novel ECG arrhythmia classification system that utilizes a large ECG database with an inter-patient paradigm. The objective is to improve the identification of less common arrhythmia categories without utilizing feature extraction. Four supervised ML models, namely SVM, KNN, RF, and an ensemble of these three models, were employed. The models underwent testing using actual inter-patient ECG records from MIT-database (MIT-DB). Prior to testing, the data was segmented and normalized. The focus of the testing was on classifying the following types of beats: normal beat (NOR), left bundle branch block beat (LBBB), right bundle branch block beat (RBBB), and premature atrial contraction (PAC). The results demonstrate that SVM surpassed other approaches in all criteria, obtaining an accuracy of 0.83. Furthermore, the SVM model shown efficacy in terms of computational expenditure, a pivotal aspect in the implementation of ECG arrhythmia classification algorithms.

Guo and Lin [12] introduced an AI framework to precisely identify atrial fibrillation by analyzing ECG signals. By employing feature extraction and ensemble learning techniques, the system attained an impressive accuracy rate of 92%. For model training and to demonstrate the efficiency of this parameter combination, the scientists used an arrangement of P wave morphology and heart rate variability characteristics. AI ensemble learning methods like Bagging, AdaBoost, and stacking were used by the writers of this study. When combined with many models, the stacking ensemble learning method produced the most accurate predictions. Along with an F1 score of 92.31% and an area under the curve (AUC) value of 91.10%, the findings include a sensitivity rate of 88% and a specificity rate of 96%. By combining a bidirectional long short-term memory with a CNN, Lu *et al*. [13] presented a novel DL architecture for autonomous arrhythmia categorization. The MIT-BIH and St-Petersburg datasets are used for training and evaluation of the model. The MIT-BIH dataset yielded a training accuracy of 100%, a validation accuracy of 98%, and a testing accuracy of 98%. The training, validation, and testing accuracy scores for the St-Petersburg dataset were 98, 95, and 95%, respectively. Getting very accurate results, particularly when dealing with the MIT-BIH data collection. By comparing the models' performance to that of preexisting models, we can see that this one performs better on the MIT-BIH dataset. Hassan *et al.* [14] introduced a ML algorithm designed to automatically detect heart disease. In this inquiry, unbalanced ECG samples are used to train SVM, logistic regression (LR), and AdaBoost classifiers. AdaBoost and LR are ranked as the top-performing models and are combined together to enhance their performance through ensembling. The ensemble model demonstrates superior HD identification ability on the PTB-ECG and MIT-BIH datasets, achieving high accuracy, F1-score, and AUC values. The ensemble model achieved accuracy scores of 94.60, 94.90, and 95.10% for the PTB-ECG dataset, and 92.10, 92.60, and 95% for the MIT-BIH dataset, in terms of accuracy, F1-score, and AUC, respectively. The following are the key contributions to this research.

– By combining data from four distinct databases, our study increased the diversity and richness of the dataset, leading to better generalization and more accurate classification of arrhythmias.
– We also use ensemble learning techniques, specifically a stacking ensemble of XGBoost and bagging XGBoost (EBXGB), which significantly improved the model's performance and compared model performance with traditional ML algorithms.
– Implementing the SMOTE helped in generating synthetic data to balance the minority classes, further enhancing the model's accuracy, recall, and precision.
– We utilized local interpretable model-agnostic explanations (LIME) to determine the impact of the features on the model's outcome.

The remaining portions of the paper are organized in the following manner: section 2 outlines the strategy used in this research. Section 3 presents the outcomes of our research and compares our work with earlier studies. Lastly, section 4 provides the conclusion.

2. METHOD

In this section, the proposed workflow of our research method has been discussed in detail. Figure 1 shows the proposed workflow. In Figure 1, data preparation and model training workflow begins with acquiring and cleaning four distinct datasets to remove inconsistencies and irrelevant information. These cleaned datasets are then merged into a single dataset, which undergoes further cleaning to ensure quality and consistency. The output column is transformed using a label encoder, converting categorical data into numerical format and simplifying the multi-class classification problem into a binary classification. Feature variables are standardized to ensure equal contribution to the model. The dataset is split into training and test sets, typically in an 80:20 ratio, to evaluate the model's performance realistically. To address class imbalance in the training set, the synthetic SMOTE is applied, generating synthetic examples for the minority class. Various ML models are then trained on this balanced and preprocessed data. Finally, the models are evaluated using metrics such as accuracy, precision, recall, and F1-score, providing a comprehensive assessment of their performance. This systematic approach ensures the development of robust and reliable ML models, particularly valuable in research settings where data quality and model accuracy are paramount. The algorithm for preparing data and training our proposed model is shown in Algorithm 1.

Algorithm 1. Data preparation and model training

```
Input: Dataset_1, Dataset_2, Dataset_3, Dataset_4
Output: Model evaluation results
Step 1: Load datasets
    Dataset_1 ← load_data("path/to/dataset1")
    Dataset_2 ← load_data("path/to/dataset2")
    Dataset_3 ← load_data("path/to/dataset3")
    Dataset_4 ← load_data("path/to/dataset4")
Step 2: Clean individual datasets
    Dataset1_clean ← clean_data(Dataset_1)
    Dataset2_clean ← clean_data(Dataset_2)
    Dataset3_clean ← clean_data(Dataset_3)
    Dataset4_clean ← clean_data(Dataset_4)
Step 3: Merge cleaned datasets
    Merged_Datasets ← merge_datasets([Dataset1_clean, Dataset2_clean, Dataset3_clean,
    Dataset4_clean])
```

Step 4: Use Label Encoder on output column
Merged_Datasets['type'] ← label_encode(*Merged_Datasets*['type'])
Step 5: Convert multi-classification to binary classification
Merged_Datasets['binary_class'] ← convert_to_binary(*Merged_Datasets*['type'])
Step 6: Split features and labels
X ← *Merged_Datasets*.drop(columns=['binary_class'])
y ← *Merged_Datasets*['binary_class']
Step 7: Standardize feature variables
X_standardized ← standardize(*X*)
Step 8: Split the dataset into training and test sets
(*X_train*, *X_test*, *y_train*, *y_test*) ← train_test_split(*X_standardized*, *y*, *test_size*=0.2, *random_state*=42)
Step 9: Apply SMOTE on training data
(*X_train_resampled*, *y_train_resampled*) ← apply_smote(*X_train*, *y_train*)
Step 10: Train models
models ← train_models(*X_train_resampled*, *y_train_resampled*)
Step 11: Evaluate models on test data
valuation_results ← evaluate_models(models, *X_test*, *y_test*)
Step 12: Return evaluation results

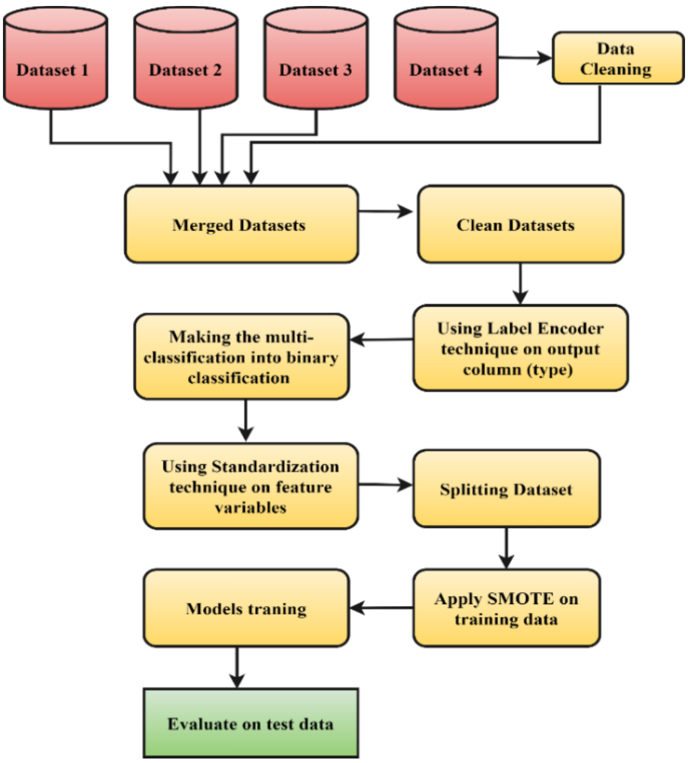


Figure 1. The workflow of the proposed work

2.1. Datasets collection

In this study, the MIT-BIH supraventricular arrhythmia database [15] is used as the main dataset or foundation dataset. Some of the other datasets used include the INCART 2-lead arrhythmia database, the SCDH database, and the MIT-BIH arrhythmia database. All of the data in the collection came from Kaggle. The 78 ECG recordings that make up the MIT-BIH supraventricular arrhythmia dataset have a duration of around 30 minutes apiece. A single pulse is represented by each of the 184,428 occurrences in the collection. Across all of these datasets, there are a total of 34 characteristics, which encompass a patient's record and the classification type of their heartbeat (label). The remaining 32 characteristics are partitioned into two groups, each including 16 features. One set corresponds to the lead II signal, while the other set corresponds to the lead V5 features [14]. Figure 2 provides clear evidence of a significant imbalance in the classes of the dataset.

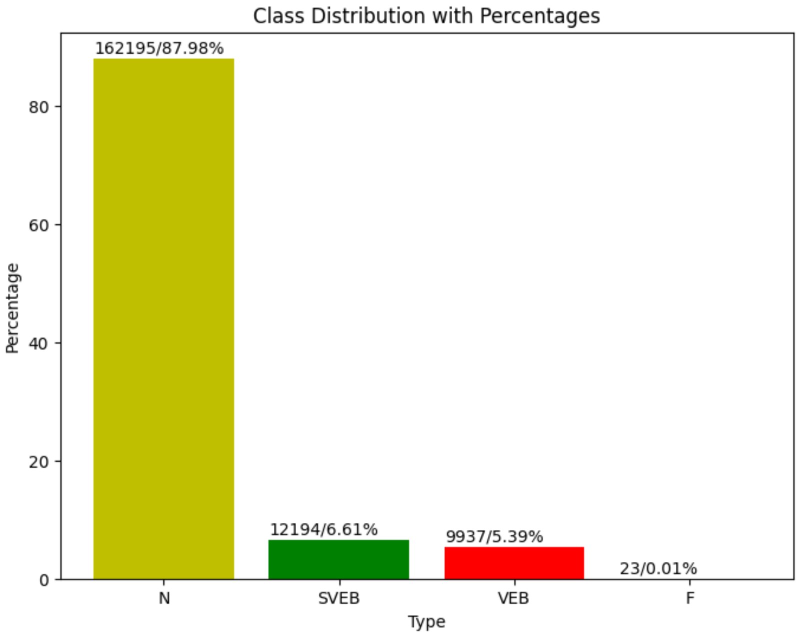


Figure 2. Class distribution of the base dataset

2.2. Data pre processing

Applying the model to an imbalanced dataset can lead to misleading accuracy and ultimately result in subpar performance. The SCDH dataset has numerous occurrences that have missing values. Consequently, it is necessary for us to purify this dataset. In this study, we address the issue of missing values by imputing them with the column-wise mean values. In addition, we exclude any records that contain missing values explicitly in the type column. Afterwards, data regarding the SVEB, VEB, and F categories was obtained from other databases and added to the original dataset. The 'record' part was removed as it just pertains to the patient number and does not contribute to the prediction of the heartbeat type. The 'type' column in the dataset is classified as the object data type. Hence, it is necessary to transform the object type into a numerical type using the label encoder technique. Next, using the StandardScaler technique to standardize the feature variable, ensuring that they have a similar range and distribution. This has the potential to improve both performance and accuracy. Using StandardScaler is beneficial when the dataset deviates from a normal distribution. The formula for standardization is as (1).

$$X' = \frac{X-\mu}{\sigma}, \tag{1}$$

To address the imbalance in this dataset, we employed the SMOTE to correct this unevenness. The objective of SMOTE is to achieve class distribution parity by synthesizing false samples for the minority class. Afterward, the dataset is partitioned into a training dataset and a test dataset. The models are trained using a dataset specifically designated for training purposes, and the ML algorithms go through multiple rounds to optimize the hyperparameters. Evaluation of the models is conducted using the test dataset.

2.3. Applied machine learning algorithms and ensemble models

This study investigates multiple ML algorithms and ensemble models to assess their effectiveness in classifying arrhythmias. The ML algorithms used are multilayer perceptron (MLP), AdaBoost, LR, DT, KNN, naïve Bayes (NB), and XGB. Creating a bagging ensemble consisting of LR, DT, KNN, NB, and XGB models. Ensemble combining XGB and Bagging XGB.

2.3.1. Logistic regression

The primary objective of the supervised ML technique called LR is to forecast the probability of different classes based on particular independent variables. LR differs from linear regression in that it use the sigmoid function to calculate the probability of an instance belonging to a specific class, rather than producing continuous output values. This algorithm demonstrates proficiency in binary classification tasks by applying a sigmoid function to the output of the linear regression function, generating probabilities [16].

2.3.2. Bagging logistic regression
In the process of bagging using LR, a technique called random sampling with replacement is employed. This involves training many instances of LR separately on different subsets of the dataset. Each LR or basic model in this set learns to identify distinct patterns in the data as a result of the variability introduced by the subsets. When creating predictions, each LR model produces its output. The final predictions for a classification job are determined by taking a majority vote among the outputs of all the LR models.

2.3.3. Decision tree
DT is a popular supervised ML approach. You may use this tool for both regression and classification problems. It's quite flexible. The dataset's properties serve as the inner nodes, the outcomes as the leaf nodes, and the decision rules as the branches in a tree structure. The two main kinds of nodes in DT are decision nodes and leaf nodes. Decision nodes are utilized for making determinations and can possess several branches, while leaf nodes are employed to symbolize the ultimate outcomes of such determinations and do not possess any extra branches. The test or decisions are determined based on the characteristics of the given dataset, and a visual depiction of potential solutions is offered depending on stated criteria [17]. The process involves the development of a hierarchical structure, similar to that of a tree, starting with the root node and expanding with additional branches.

2.3.4. Bagging decision tree
Bagging is an effective ensemble ML technique that pairs well with DT. The bagging technique involves training multiple DT separately on different subsets of the dataset using random sampling with replacement [18]. This variability enables each tree to capture distinct patterns in the data. Each decision tree in bagging yields its own output. The final predictions for a classification task are determined by a majority vote among all of the DT outputs.

2.3.5. Bagging extreme gradient boosting
When employing XGB with bagging, random sampling with replacement is utilized to train multiple XGB models separately on different subsets of the dataset. Each of these XGB or basic models is trained to detect distinct patterns in the data as a result of the variety introduced by the subsets. When creating predictions, each individual XGB model produces its own output. The final predictions for a classification job are determined by taking a majority vote among the outputs of all the XGB models [19].

2.3.6. Bagging k-nearest neighbor
In the bagging technique with KNN, random sampling with replacement is used to train several KNN models independently on different subsets of the dataset. Each of these KNN or base models learns to capture unique patterns in the data caused by variability. During the prediction process, each KNN algorithm produces its output. In classification tasks, the final prediction is determined by a majority vote [20].

2.3.7. Bagging naïve Bayes
When employing NB for bagging, multiple NB classifiers are trained independently on different subsets of the dataset, which are generated through random sampling with replacement. Due to the heterogeneity caused by the subsets, each NB or base model acquires the capacity to identify distinct patterns and correlations within the data. During the prediction process, each unique NB model generates its output [21].

2.3.8. Stacking ensemble of extreme gradient boosting and bagging extreme gradient boosting
Stacking is an effective ensemble learning technique in ML, where the predictions of multiple base models are combined to achieve a final prediction that demonstrates improved performance. It is alternatively referred to as a stacked ensemble or stacked generalization. A stacking ensemble can be likened to a collection of experts led by a leader. The leader takes into account the outputs of each expert before making the final decision. Applying a stacking ensemble to a big and diverse dataset is advantageous. This diversity allows the model to efficiently learn the correlation between the predictions of the base models and the target variable. The study utilizes XGB and bagging XGB as base estimators, with LR serving as the meta-model [22]. To enhance performance of the classifiers, the optimization of several hyperparameters has been performed. Table 1 displays the optimized hyperparameters.

2.4. Performance evaluation metrics
To assess the performance of arrhythmia classification from EEG, many evaluation metrics are utilized. Each targeting a specific aspect of its performance using precision, recall, F1-score, and accuracy metrics. These metrics offer a numerical evaluation of the model's ability to precisely classify arrhythmia [23].

Table 1. Optimized model performance used fine-tuned

Model	Hyperparameters Tuned
LR	C, penalty, solver, max-iter
XGB	n-estimators, max-depth, learning rate, gamma
DT	criterion, max-depth, min-samples-split
KNN	n_neighbors
NB	var_smoothing
AdaBoost	base-estimator, n_estimators, learning_rate
MLP	hidden_layer_sizes

2.5. Model explanation using LIME

LIME is a versatile tool that enhances our comprehension of the decision-making process of intricate models. It achieves this by constructing a clear and comprehensible framework based on a specific scenario, offering insights into the behavior of the black-box model in that particular context. An advantageous aspect of LIME is its compatibility with several ML models. Therefore, LIME serves as a valuable tool for improving our understanding of models in other fields. For this study, the authors employing LIME to ascertain the influence of various variables on the outcomes and uncover the underlying rationale behind the model's decision-making process. LIME is a robust technique developed to improve the comprehensibility of intricate ML models at a specific level. The aim is to enhance the lucidity and comprehension of forecasts by incorporating the notion of localized explanations. The emphasis lies on the interpretability of individual data instances rather than the complete model. This approach functions by creating modified samples in the vicinity of the specific event of interest, causing random fluctuations in the characteristic values. LIME is a technique that effectively estimates the intricate decision border of a model near a specific instance, without being constrained to a single model. This is achieved by developing a locally interpretable model, usually in the form of a linear model. LIME utilizes kernelized weights to guarantee that the perturbed samples have a significant impact on the local model [24].

The weights assign higher priority to samples that are closer to the original instance. The significance of the feature is evaluated by analyzing the coefficients of this specific model, which measures the impact of each characteristic on the decision-making process. The most influential traits, as assessed by their highest relevance scores, provide a localized explanation that identifies the variables with the greatest impact on the specific prognosis. LIME is valuable in several industries, especially in sensitive sectors like healthcare or finance, where comprehending the fundamental rationale behind specific predictions is vital [25]. LIME plays a crucial role in building trust and enabling the implementation of ML models in real-life situations by offering clear and understandable explanations of the model's decision-making process for each particular case [26]. The formula for LIME is as (2).

$$L\ (x, f, \pi)\ = \sum_{i=1}^{d} \pi_i\ (x)\ f_i(x) + C\ |\pi|_1 \quad (2)$$

The variables in the (2) are defined as follows: x represents the instance, f represents the approximation model, π represents the feature importance weights, d represents the feature count, and C represents the regularization parameter. The equation consists of a weighted combination of features and a regularization factor that encourages sparsity in the weights assigned to each feature. The optimization issue entails finding weights Π that minimize the discrepancy between predictions made by a black-box model and forecasts made by an approximation model for a given instance x.

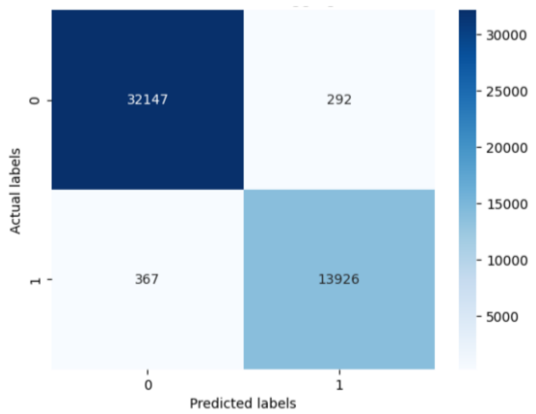
3. RESULTS AND DISCUSSION

Table 2 illustrates the performance comparison of ML models and ensemble technique models when utilizing a single base database versus when using combined databases. Combining datasets often leads to an increase in accuracy in most cases. Table 2 demonstrates that in the combined dataset, the stacking EBXGB and BXGB outperforms all other models utilized in this study in terms of accuracy. The BXGB model achieves an accuracy of 98.59%, a recall of 97.43%, and a precision of 97.95%. The EBXGB model achieves an accuracy of 98.61%, a recall of 97.66%, and a precision of 97.77%.

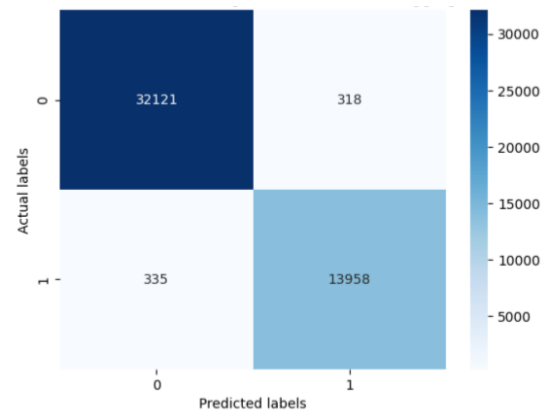
The confusion matrix of the best model (EBXGB) is shown in Figure 3. Figure 3(a) represents the confusion matrix for the EBXGB model using a single database. Figure 3(b) represents the confusion matrix for the EBXGB model using the combined database. In Figure 4, The receiver operating characteristic (ROC) curve of the EBXGB model has been illustrated. It can be seen that the AUC is 1.00.

Table 2. The performance of all model single and combined databases

Model	Single database			Combined database		
	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)
Without Bagging						
LR	57.87	85.94	90.79	90.22	87.50	93.28
XGB	91.65	91.88	98.02	97.75	97.45	98.53
DT	82.24	90.02	96.46	94.86	95.89	97.16
KNN	84.47	93.99	97.20	96.29	97.56	98.11
NB	36.09	55.16	82.87	83.55	78.87	88.79
AdaBoost	88.02	87.90	97.11	96.93	96.15	97.89
MLP	86.86	92.76	97.44	97.39	97.19	98.34
With Bagging						
Bagging LR	52.69	86.12	89.04	90.51	85.83	92.91
BXGB	91.84	92.51	98.11	97.95	97.43	98.59
Bagging DT	88.86	93.95	97.86	97.52	97.52	98.48
Bagging KNN	85.69	94.67	97.62	97.09	97.60	98.37
Bagging NB	36.63	54.28	83.22	84.07	78.63	88.91
Proposed Model (EBXGB)	90.51	93.03	97.99	97.77	97.66	98.61



(a)



(b)

Figure 3. Confusion matrix of the best model (EBXGB) of using: (a) single database and (b) combined database

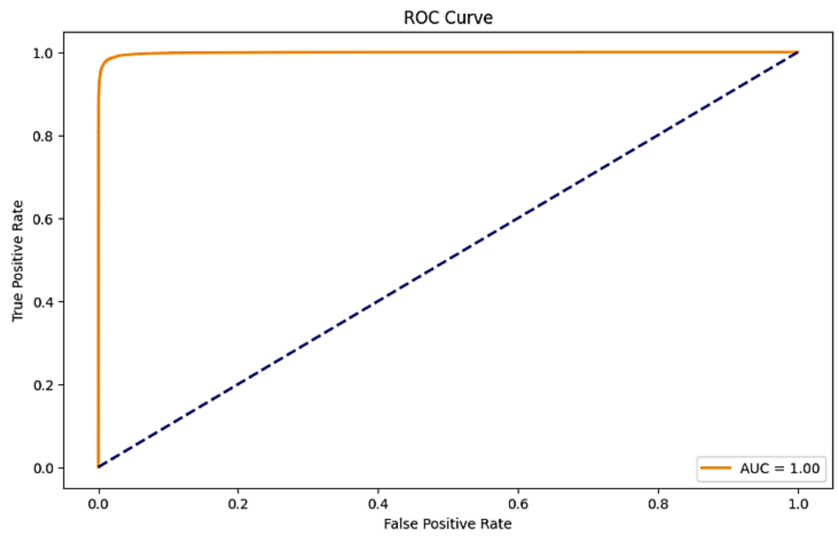


Figure 4. ROC curve of the EBXGB model

LIME is employed to comprehend the decision-making process of a stacking model. Figures 5 and 6 depict an application of the LIME-based XAI method to analyze the stacking model. In Figure 5, the stacking model predicts arrhythmia based on the following criteria: a QRS interval greater than 0.48, a post-RR interval greater than 0.95, a QT interval greater than 0.50, a QT interval greater than 0.47, an ST interval greater than 0.49, a qPeak greater than 0.47, a post-RR interval greater than 0.90, a qrs_morph1 greater than 0.40, and a qrs_morph0 greater than 0.47. In Figure 6, the stacking model predicts non-arrhythmia based on the following criteria: a QT interval of less than -0.53, a QT interval of less than -0.47, a pre_RR interval of more than 0.65, a QRS interval of less than -0.12, a QRS morph3 value of more than 0.62, and a sPeak value of less than 0.10.

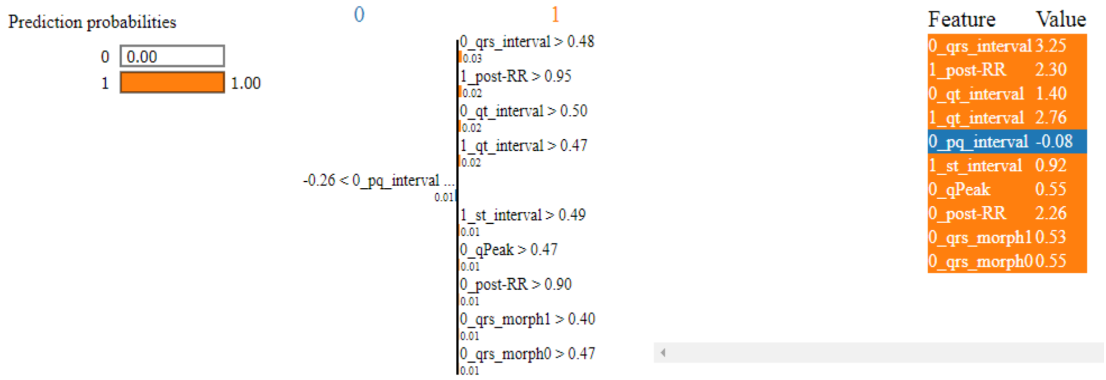


Figure 5. LIME explainable prediction interpretation (predict arrhythmia)

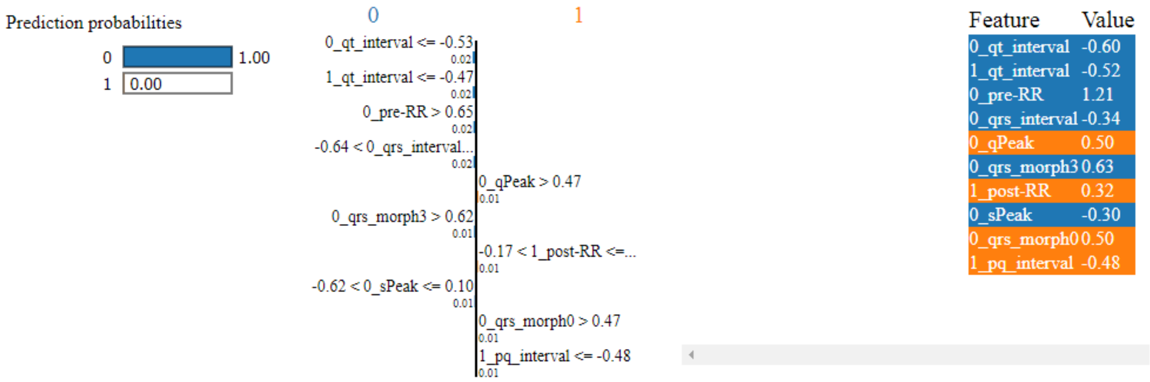


Figure 6. LIME explainable prediction interpretation (predict non-arrhythmia)

While previous studies have explored arrhythmia detection using single datasets and individual ML models, they have not explicitly addressed the potential improvements in detection accuracy and model interpretability that could be achieved by combining multiple datasets and employing ensemble learning techniques. The limited number of examples for certain arrhythmia classes in earlier research has also impeded the models' ability to generalize effectively. This study seeks to fill these gaps by integrating data from multiple sources and using ensemble methods to enhance model robustness and performance, thereby providing a more comprehensive and accurate approach to arrhythmia detection. This approach helps in better generalization and more accurate classification of arrhythmias. Additionally, we employed ensemble learning techniques, specifically a stacking EBXGB, which significantly improved model performance. By using the SMOTE to balance the dataset, we further enhanced the model's accuracy, recall, and precision. This comprehensive strategy of combining datasets and leveraging advanced ensemble methods provides a more robust and interpretable solution for arrhythmia detection.

The EBXGB model yields superior results in terms of precision, recall, and accuracy. Utilize the EBXGB model, the accuracy score for the single database was 97.99%, with precision and recall values of 90.51 and 93.03% respectively. The combined database earned accuracy, precision, and recall ratings of 98.61, 97.77, and 97.66% respectively. We utilize LIME analysis to identify essential features, as well as to classify data based on different classes. Table 3 presents a comparison between our suggested model and the models used in earlier studies. The proposed model, EBXGB outperforms existing state-of-the-art arrhythmia detection algorithms with an accuracy of 98.61%. The EBXGB model not only outperforms but also adds model explanation capabilities utilizing LIME, which improves interpretability and clinician confidence. This achievement highlights the possibility of incorporating advanced ensemble methods and interpretability tools into clinical practice, resulting in more accurate and dependable arrhythmia identification than previous models that lacked explanatory characteristics.

Table 3. A comparison between the models from previous research and our proposed model

Reference	Accuracy (%)	Model explanation
[1]	97	×
[2]	96.67	×
[3]	95.6	×
[4]	95.20	×
[5]	86.66	×
[10]	96.18	×
[11]	83	×
[13]	98	×
[14]	95.10	×
Proposed work (EBXGB model)	98.61	Yes

This research makes a substantial contribution to the field of arrhythmia detection by illustrating the ability of a stacking ensemble to accurately classify normal and aberrant heartbeats, as well as the efficacy of a variety of ML algorithms. The potential for these models to improve early diagnosis and opportune intervention, thereby reducing the risk of heart failure, stroke, and sudden cardiac arrest, is demonstrated by their high accuracy of 98.61%. The application of advanced techniques such as LIME and multiple datasets to enhance model interpretability fosters trust among clinicians by providing a more profound comprehension of model decisions. This research establishes a strong foundation for the integration of sophisticated ML models into clinical practice, which has the potential to enhance patient outcomes and advance personalized healthcare.

This study, which focuses on the identification of arrhythmia using different ML techniques, including a highly effective stacking ensemble, has significant constraints. The data sources are restricted to only four databases, which might not encompass all kinds of arrhythmia and could result in dataset bias. The binary classification method oversimplifies the complexity of arrhythmia, and the absence of external validation raises concerns about the ability of the model to be applied to different situations. Insufficient information is available regarding feature selection, computing resources, and clinical integration, which makes it difficult to replicate and apply in practice. In addition, the utilization of LIME for interpretability is subject to its constraints, and there is no evaluation of real-time implementation or longitudinal data analysis.

4. CONCLUSION

This study effectively improved the overall accuracy, precision, and recall of ML models, notably bagging and stacking ensemble models, by integrating three supplementary databases with a foundational database. The study highlights the significance of using varied datasets to get reliable outcomes by showcasing how combining datasets and selecting suitable hyperparameters can enhance model performance. The EBXGB shown exceptional performance in binary classification, obtaining an accuracy of 98.61%, recall of 97.66%, and precision of 97.77%. In contrast, the naïve Bayes model exhibited the lowest level of performance. By utilizing the LIME technique, the model's usefulness was improved as it provided valuable information about its decision-making process, thereby promoting confidence among clinicians. While the primary emphasis of this study is the detection of arrhythmia using ECG data, its approaches have wider ramifications and can be applied to other areas of medical diagnosis. To improve generalizability, future research should create multi-class classification, involve varied datasets, and carry out external validations. To guarantee practical application and therapeutic trust, emphasis should also be placed on real-time analysis, sophisticated interpretability techniques, and thorough evaluation metrics.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to the Department of Biomedical Engineering, Islamic University, Kushtia, Bangladesh, for their support and encouragement throughout this research.

FUNDING INFORMATION

The authors declare that this research was self-funded, with no financial support from any external funding agency, grant, or institution.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Md. Rabiul Islam	✓	✓	✓	✓			✓		✓	✓				✓
Tapan Kumar Godder	✓					✓	✓			✓		✓	✓	
Ahsan-Ul-Ambia		✓				✓				✓		✓	✓	
Ferdib-Al-Islam	✓		✓	✓	✓				✓	✓	✓			
Anindya Nag	✓		✓	✓	✓				✓	✓	✓			
Bulbul Ahamed		✓						✓	✓					
Nujhut Tanzim				✓	✓			✓	✓		✓			
Md. Estiak Ahmed		✓	✓	✓	✓			✓	✓		✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest regarding the research, authorship, and publication of this paper.

DATA AVAILABILITY

The data used in this study are publicly available from PhysioNet at https://doi.org/10.1161/01.cir.101.23.e215, reference number [15].

REFERENCES

[1] A. Al-Mousa, J. Baniissa, T. Hashem, and T. Ibraheem, "Enhanced electrocardiogram machine learning-based classification with emphasis on fusion and unknown heartbeat classes," *Digital Health*, vol. 9, 2023, doi: 10.1177/20552076231187608.

[2] S. Sakib, M. M. Fouda, Z. M. Fadlullah, N. Nasser, and W. Alasmary, "A proof-of-concept of ultra-edge smart IoT sensor: A continuous and lightweight arrhythmia monitoring approach," *IEEE Access*, vol. 9, pp. 26093–26106, 2021, doi: 10.1109/ACCESS.2021.3056509.

[3] T. Wang, C. Lu, W. Ju, and C. Liu, "Imbalanced heartbeat classification using EasyEnsemble technique and global heartbeat information," *Biomedical Signal Processing and Control*, vol. 71, 2022, doi: 10.1016/j.bspc.2021.103105.

[4] M. M. R. Khan, M. A. B. Siddique, S. Sakib, A. Aziz, A. K. Tanzeem, and Z. Hossain, "Electrocardiogram heartbeat classification using convolutional neural networks for the detection of cardiac arrhythmia," *4th International Conference on IoT in Social, Mobile, Analytics and Cloud, ISMAC 2020*, pp. 915–920, 2020, doi: 10.1109/I-SMAC49090.2020.9243474.

[5] Z. Zhang, J. Dong, X. Luo, K. S. Choi, and X. Wu, "Heartbeat classification using disease-specific feature selection," *Computers in Biology and Medicine*, vol. 46, no. 1, pp. 79–89, 2014, doi: 10.1016/j.compbiomed.2013.11.019.

[6] F. I. Alarsan and M. Younes, "Analysis and classification of heart diseases using heartbeat features and machine learning algorithms," *Journal of Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0244-x.

[7] N. Singh and P. Singh, "Cardiac arrhythmia classification using machine learning techniques," *Engineering Vibration, Communication and Information Processing*, pp. 469–480, 2019, doi: 10.1007/978-981-13-1642-5_42.

[8] S. S. Abdelmoneem, H. H. Said, and A. A. Saad, "Arrhythmia disease classification and mobile based system design," *Journal of Physics: Conference Series*, vol. 1447, no. 1, 2020, doi: 10.1088/1742-6596/1447/1/012014.

[9] B. R. Manju and A. R. Nair, "Classification of cardiac arrhythmia of 12 lead ECG using combination of SMOTEENN, XGBoost and machine learning algorithms," *2019 International Symposium on Embedded Computing and System Design, ISED 2019*, 2019, pp. 48–55, doi: 10.1109/ISED48680.2019.9096244.

[10] A. Peimankar, M. J. Jajroodi, and S. Puthusserypady, "Automatic detection of cardiac arrhythmias using ensemble learning," *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, pp. 383–388, 2019, doi: 10.1109/TENCON.2019.8929348.

[11] M. Sraitih, Y. Jabrane, and A. H. El Hassani, "An automated system for ECG arrhythmia detection using machine learning techniques," *Journal of Clinical Medicine*, vol. 10, no. 22, 2021, doi: 10.3390/jcm10225450.

[12] Z. Guo and L. Lin, "Cognitive physiological data analysis based on the XGBoost algorithm to realize positive perceptual sample classification," *Journal of Intelligent and Fuzzy Systems*, vol. 44, no. 4, pp. 6525–6543, 2023, doi: 10.3233/JIFS-222656.

[13] M. Lu *et al.*, "A stacking ensemble model of various machine learning models for daily runoff forecasting," *Water*, vol. 15, no. 7, 2023, doi: 10.3390/w15071265.

[14] S. U. Hassan, M. S. M. Zahid, T. A. A. Abdullah, and K. Husain, "Classification of cardiac arrhythmia using a convolutional neural network and bi-directional long short-term memory," *Digital Health*, vol. 8, pp. 1-13, 2022, doi: 10.1177/20552076221102766.

[15] A. L. Goldberger *et al.*, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, 2000, doi: 10.1161/01.cir.101.23.e215.

[16] J. Jeppesen, J. Christensen, P. Johansen, and S. Beniczky, "Personalized seizure detection using logistic regression machine learning based on wearable ECG-monitoring device," *Seizure: European Journal of Epilepsy*, vol. 107, pp. 155–161, 2023, doi: 10.1016/j.seizure.2023.04.012.

[17] Y. Zheng, J. Ding, F. Liu, and D. Wang, "Adaptive neural decision tree for EEG based emotion recognition," *Information Sciences*, vol. 643, 2023, doi: 10.1016/j.ins.2023.119160.

[18] K. M. Alalayah, E. M. Senan, H. F. Atlam, I. A. Ahmed, and H. S. A. Shatnawi, "Effective early detection of epileptic seizures through EEG signals using classification algorithms based on t-distributed stochastic neighbor embedding and k-means," *Diagnostics*, vol. 13, no. 11, 2023, doi: 10.3390/diagnostics13111957.

[19] F. J. Ramírez-Arias *et al.*, "Evaluation of machine learning algorithms for classification of EEG signals," *Technologies*, vol. 10, no. 4, 2022, doi: 10.3390/technologies10040079.

[20] C. Wu *et al.*, "Application of artificial intelligence ensemble learning model in early prediction of atrial fibrillation," *BMC Bioinformatics*, vol. 22, 2021, doi: 10.1186/s12859-021-04000-2.

[21] H. A. Siddiqa, M. Irfan, S. F. Abbasi, and W. Chen, "Electroencephalography (EEG) based neonatal sleep staging and detection using various classification algorithms," *Computers, Materials and Continua*, vol. 77, no. 2, pp. 1759–1778, 2023, doi: 10.32604/cmc.2023.041970.

[22] A. Rath, D. Mishra, and G. Panda, "Imbalanced ECG signal-based heart disease classification using ensemble machine learning technique," *Frontiers in Big Data*, vol. 5, 2022, doi: 10.3389/fdata.2022.1021518.

[23] M. M. Hassan *et al.*, "Machine learning-based rainfall prediction: unveiling insights and forecasting for improved preparedness," *IEEE Access*, vol. 11, pp. 132196–132222, 2023, doi: 10.1109/ACCESS.2023.3333876.

[24] I. Hussain *et al.*, "An explainable EEG-based human activity recognition model using machine-learning approach and LIME," *Sensors*, vol. 23, no. 17, 2023, doi: 10.3390/s23177452.

[25] M. S. Sanim, F. Al-Islam, and K. M. Hasan, "Identification of COVID-19 from other upper respiratory tract infections using random undersampling and LIME-based XAI model," *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*, 2023, doi: 10.1109/ICCCNT56998.2023.10307102.

[26] F. Al-Islam, A. Saha, E. J. Bristy, M. R. Islam, R. Afzal, and S. A. Ridita, "LIME-based explainable AI Models for predicting disease from patient's symptoms," *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*, 2023, doi: 10.1109/ICCCNT56998.2023.10307223.

BIOGRAPHIES OF AUTHORS

Md. Rabiul Islam graduated from the Department of Computer Science and Engineering, Islamic University in 2001 and obtained postgraduate degree from the same in 2002. He is pursuing his Ph.D. from the Department of Biomedical Engineering, Islamic University. He joined in the Department of Biomedical Engineering as a lecturer on July 2018. Before that, he was a faculty of Northern University Bangladesh. He has near about 15 publications both in national and international journals. He has also been supervising enormous number of students both in graduate and postgraduate level. His research interest in machine learning based bio-signal processing, medical imaging and prosthetics and orthotics. He can be contacted at email: rabinub@gmail.com.

Dr. Tapan Kumar Godder graduated from the Department of Applied Physics and Electronics Rajshahi University, Bangladesh and obtained his post-graduate degree from the same in 1994 and 1995 respectively. He was awarded his master of philosophy and doctor of philosophy from the Faculty of Applied Science and Technology, Islamic University, Bangladesh in 2007 and 2012 respectively. He joined in the Department of Information and Communication Technology, Islamic University as a Professor on 4th may, 2013. Since then, he has been professing, supervising and guiding the students in many ways. He has near about 40 publications in national and international journals and conferences. He supervised near about 20 students in M.Phil. and Ph.D. level. He can be contacted at email: tkict@yahoo.com.

Dr. Ahsan-Ul-Ambia graduated from the Department of Applied Physics and Electronics, Rajshahi University, Bangladesh and did his post-graduate degree from the same in 1993 and 1994 respectively. He obtained his doctor of philosophy from the graduate school of science and technology, Shizuoka University Japan in 2009. He joined in the Department of Computer Science and Engineering, Islamic University as a Professor on 2013. Since then, he has been professing, supervising and guiding the students in many ways. He has near about 35 publications in national and international journals and conferences. He is also the author of some renowned books. He supervised about 15 students in M.Phil. and Ph.D. level. He was the chairman of Department of Computer Science and Engineering, provost of residential hall and director ICT cell in Islamic University. He can be contacted at email: aambia@cse.iu.ac.bd.

Ferdib-Al-Islam completed B.Sc. engineering in computer science and engineering from Bangabandhu Sheikh Mujibur Rahman Science and Technology University (BSMRSTU), Gopalganj and M.Sc. engineering in the same subject from Khulna University of Engineering and Technology (KUET). He is currently working as an assistant professor in the Department of Computer Science and Engineering at Northern University of Business and Technology Khulna, Bangladesh. His research interests include machine learning, deep learning, explainable AI, and computer vision. He acted as a reviewer and a member of TPC in reputed journals and conferences. Apart from this, he has professional experience as a software engineer in IoT and machine learning. He can be contacted at email: ferdib.bsmrstu@gmail.com.

Anindya Nag (member, IEEE), obtained a B.Tech. degree from Adamas University in Kolkata, India, and an M.Sc. degree in computer science and engineering from Khulna University in Bangladesh. He presently holds the position of lecturer at the Department of Computer Science and Engineering at North Western University in Khulna, Bangladesh. His research focuses on wireless resource management in 6G and future generations, healthcare, the industrial internet of things (IIoT), medical internet of things (MIoT); natural language processing (NLP); machine learning: deep learning; artificial intelligence (AI); blockchain; cloud computing; and networking systems. He serves as a reviewer for various reputable journals and international conferences. He has written and collaborated on approximately 30 articles, which include peer-reviewed journals and conference papers published by IEEE/ACM. He can be contacted at email: anindyanag@ieee.org.

Bulbul Ahamed born in Munshiganj, Bangladesh, currently serves as a Professor in the Department of Computer Science and Engineering (CSE) at Sonargaon University (SU), Bangladesh. Previously, he served as assistant professor, senior lecturer, and lecturer in the Department of CSE at Northern University Bangladesh (NUB). His research interests encompass speech recognition, machine learning, artificial intelligence, cybersecurity, e-commerce, and social issues. He has authored articles in esteemed journals and conferences both domestically and internationally. He obtained his bachelor's degree in CSE from NUB and completed his MBA in MIS and marketing from the same university. Additionally, he earned his M.Sc. in CSE from United International University (UIU), Bangladesh. He has received several prestigious awards, including the Chancellor’s Gold Medal and Summa Cum Laude, in recognition of his outstanding academic achievements. Currently, he is pursuing his Ph.D. in computer science and engineering at Jahangirnagar University, Bangladesh. He can be contacted at email: bulbul_cse@su.edu.bd.

Nujhut Tanzim earned her B.Sc. degree in computer science and engineering (CSE) from Northern University of Business and Technology Khulna. Her research interests include artificial intelligence and machine learning. She can be contacted at: nujhattanzim@gmail.com.

Md. Estiak Ahmed earned his B.Sc. degree in computer science and engineering (CSE) from Northern University of Business and Technology Khulna. His research interests include artificial intelligence, machine learning, and signal processing. Currently, he is pursuing M.Sc. degree in computer science and engineering from Khulna University in Bangladesh. He can be contacted at email: estiakpial9@gmail.com.