An Outlier Detection and Feature Ranking based Ensemble Learning for ECG Analysis

Venkata Anuhya Ardeti[1]
Research Scholar
Dept. of ECM, Koneru Lakshmaiah Education Foundation
Vaddeswaram, Andhra Pradesh, India

Venkata Ratnam Kolluru[2]
Associate Professor
Dept. of ECM, Koneru Lakshmaiah Education Foundation
Vaddeswaram, Andhra Pradesh, India

George Tom Varghese[3]
Associate Professor
Dept. of EIE, St. Joseph's College of Engineering and Technology, Palai, Kottayam, Kerala, India

Rajesh Kumar Patjoshi[4]
Associate Professor
Dept. of ECE, National Institute of Science and Technology
Berhampur, Odisha, India

***Abstract*—Automated classification of each heartbeat class from the ECG signal is important to diagnose cardiovascular diseases (CVDs) more quickly. ECG data acquired from the real-time or clinical databases contains exceptional values or extreme values called outliers. The separation and removal of outliers is very much useful for improving the data quality. The presence of outliers will influence the results of machine learning (ML) methods such as classification and regression. Outlier identification and removal plays a significant role in this area of research and is a part of signal denoising. Also, most of the traditional ECG-signal processing methods are facing the difficulty in finding the essential key features of recorded signal. In this work, an extreme outlier detection technique known as improved inter quartile range (IIQR) filtering method is used to find the outliers of the signal for the feature ranking process. In addition, an optimized random forest (ORF) based heterogenous ensemble classification model is proposed to improve the true positive and runtime on the ECG data. The classification of each heartbeat type is classified with majority voting technique. Ensemble learning and majority voting rule is used to enhance the accuracy of heart disease prediction. The proposed feature ranking based ORF ensemble classification model (LR + SVM + ORF + XGBoost + KNN) is evaluated on the MITBIH arrhythmia database and produces an overall accuracy of 99.45% which significantly outperforms the state-of-the-art methods such as, (LR + SVM + RF + XGBoost + KNN) with 96.17% accuracy, ensemble deep learning accuracy of 95.81% and ensemble SVM accuracy of 94.47%.**

***Keywords*—Feature ranking; improved inter quartile range; majority voting; outlier detection; optimized random forest**

I. Introduction

World health organization (WHO) reported that globally cardiovascular diseases (CVDs) are the leading cause of death, having a significant impact on the nation's financial and health-care systems. The people residing in low-and middle-income countries are most affected with CVDs due to lack of access to effective and equitable healthcare services, resulting the increased mortality rate at a younger age [1]. An electrocardiogram (ECG) is a non-invasive tool for detecting CVDs that produces reliable findings with affordable cost. But the beat-by-beat analysis of ECG waveform data manually is tedious, inaccurate, and overwhelming [2]. So, efficient, and precise automated methods for beat classification have gotten significant interest recently. Due to the enormous data quantity and sparseness of medical data, obtaining an essential feature set for classification problems is becoming increasingly difficult. The performance of most classifiers is improved by eliminating the irrelevant or redundant features [3, 4]. Feature selection helps to avoid overfitting and high dimensionality problems in machine learning by reducing the number of features in the model and tries to optimize the model performance [5]. Most conventional classification methods are independent of dynamic feature selection due to large data size and high dimensionality. Feature selection through ranking tries to reduce the computational complexity of the model by compromising the classifier performance [6]. ECG "feature ranking and classification" are the significant tasks to medical and scientific researchers due to its higher-dimension feature space and small sample size. Existing techniques reviewed in the literature concentrates on using single base classifiers and independent of feature ranking process for feature selection. These models are limited to small data size and suffer with high dimensional feature space.

In this study, an extreme level outlier detection filtering approach is proposed for the detection of outliers from the ECG data taken from MITBIH Arrhythmia dataset [7]. The proposed outlier technique is an improvement to the traditional inter-quartile range outlier detection method [8, 9]. The extreme level outliers are removed to improve the error rate in the classification. After filtering, a hybrid kernel-based feature selection approach is developed to find the ranks of the features. The features with highest ranks having highest probability are considered for classification and the features with lowest ranks are neglected. These optimal set of features are used to predict the abnormality using classification model. In this research, we have developed an ensemble learning model involves five base classifiers with majority voting mechanism. The proposed ensemble method outperforms the state-of-the-art base classification techniques with an accuracy of 99.45%.

The rest of the paper is organized as follows: Section 2 gives a comprehensive review on various feature extraction and classification techniques reported in the literature. Section 3 discusses the proposed extreme level outlier detection-based hybrid random forest ensemble classification model. This section is divided into four subsections, where Section 3.1 deals with the ECG dataset used for training and testing the model. Section 3.2 explains the extreme outlier detection-based filtering approach used to find the outliers present in the data, while Section 3.3 describes the enhanced entropy-based feature ranking process, and Section 3.4 discusses the proposed optimized random forest ensemble learning model to classify the individual classes of ECG heartbeats. Section 4 report the results and discussions obtained for proposed ensemble classification model. Finally, the conclusions are drawn in Section 5.

II. Motivation towards Ensemble Learning

High dimensional biomedical data includes large amount of redundant and irrelevant features. If all the features are considered of equal importance, then the accuracy, time and spatial complexity of the model can be severely impacted. Hence, feature selection is considered as a significant step in the diagnosis of diseases based on high-dimensional biomedical data. The idea behind feature selection is to select an appropriate feature subset [10] which will act as a suitable foundation for future classification. It enhances the generalization capability of the prediction model, optimizes the homogeneity of the prediction algorithm, improves the computational performance, and avoid overfitting.

Feature selection algorithms are categorized into three types i.e., filters, wrappers, and embedded methods. Filter methods evaluate each feature independently of the classifier, rank the features according to some evaluation criterion and select the best ones [11]. This evaluation can be performed by using entropy for instance [12]. Wrappers methods evaluate the classifier's performance on various subsets of features and select the subset with maximum performance. These approaches are slower than filter methods and are dependent on the classifier used. Furthermore, feature subset selection is an NP-hard process that requires significant computation time and memory [13]. Genetic algorithm, Random search and greedy stepwise are some traditional algorithms used for feature subset selection. Embedded approaches on the other hand select the features during the learning process like artificial neural networks do [14]. Some studies on the other hand, applied various dimensionality reduction techniques on high dimensional database to diminish the size of the feature space. Most popular techniques such as principal component analysis (PCA) [15], singular value decomposition (SVD) [16] and linear discriminant analysis (LDA) [17] are used for biometric authentication applications.

The optimized feature set is given as input to the classifier to recognize the information about cardiac diseases from ECG. Abnormal classification has become a valuable and promising technique for early assessment of arrhythmia. Mohebbanazz et al., [18] proposed an optimized decision tree (DT) and adaptive boosted optimized decision tree method for classification of six types of ECG beats and evaluated the performance of the model on MITBIH arrhythmia database achieved an effective accuracy of 98.77% compared to state-of-the-art techniques. [19] introduced a time-efficient, reliable, and low-complexity resource-saving architecture with random forest (RF) classifier to classify two major types of arrhythmias such as supraventricular ectopic beats ventricular ectopic beats (VEB) and (SVEB). Classification performance of the model reaches to the f1 scores of 81.05% for SVEB and 97.07% for VEB. Another popular classifier known as, Support Vector Machine (SVM), [20] is a linear classifier that separates the classes linearly by creating a hyperplane from high-dimensional space. It captures the non-linear relationships of the ECG signal, detects the heartbeats, and classifies the data as normal/abnormal with high accuracy. Researchers proposed several SVM based classification techniques to detect arrhythmias in literature which involves Multi-class SVM [21], SVM with NN [22]. However, due to its high dimensionality space, it suffers with computational constraints. An efficient real-time time series cardiac disease prone weight (CDPW) Naive Bayes classification technique is implemented in [23] that estimate the posterior probability of different features using fuzzy rule and measures the CDPW value. The model performance was evaluated using 15000 records of real-time ECG data, and greater precision values were produced with less time complexity. The author in [24] employed K-Nearest Neighbour classifier on MITBIH arrhythmia database to classify five types of ECG beats and attained an accuracy of 98.40% for isolating the signals. A robust extreme gradient boosting technique is utilized in [25] to classify five ECG beat classes from both MITBIH data and self-collected single-lead wearable ECG dataset. The developed model outperforms the traditional models with an accuracy of 99.14% on MITDB and 98.68% on wearable ECG dataset. Ahmed et al., in [26] employed artificial neural network (ANN) to classify the ECG heartbeats from two imbalanced datasets MITBIH arrhythmia and PTB databases. They focus on penalising the loss value of ANN by assigning the class weights which outperforms the state-of-the techniques. However, researchers have demonstrated that the ensemble system can increase the performance of a base classifier.

Motivated by the development of several ML models, and a bid to improve accuracy, we propose a heterogeneous ensemble learning framework. Ensemble learning is the process of integrating various learners together to improve the stability and prediction ability of the classification model. It has been successfully applied in solving various machine learning problems includes feature selection, classification, and prediction. Fig. 1 shows a typical block diagram of an ensemble classification model, which includes three primary blocks: training datasets, base classifiers, and a combiner. In recent studies, researchers have proved that performance of the base classifier can be improved with ensemble classification method. Jose et al., [27] proposed random forest ensemble classification technique to diagnose cardiac arrhythmia. In this model the more informative features were selected using ranking criteria on training dataset. The performance of the learning model is evaluated on MITBIH arrhythmia database and obtained an accuracy of 96.14% and f1-score of 97.7%, 90.5% and 73% for normal, ventricular and supraventricular beats. An ensemble of random forest and support vector machine is implemented in [28], to classify five types of cardiac arrhythmia's and obtained an accuracy of 98.21%. Recent advances in technology proposed deep learning-based ensemble classification technique for improved cardiac diagnosis [29]. Experimentation is carried out on PTBDB and MITBIH database and found an accuracy, F1 score and area under curve (AUC) of 0.98, 0.93 and 0.92 for MITBIH datasets and 0.99, 0.986 and 0.995 for PTB dataset. A hybrid heterogeneous ensemble classification model for the prediction of heart disease is proposed in [30]. The performance of the model is evaluated on the Kaggle dataset and reports 98% accuracy which outperforms the weak learners. As many ensemble classification techniques reported in the literature, developing a robust ensemble learning model with lowest error and greater accuracy is still becoming a challenging task.

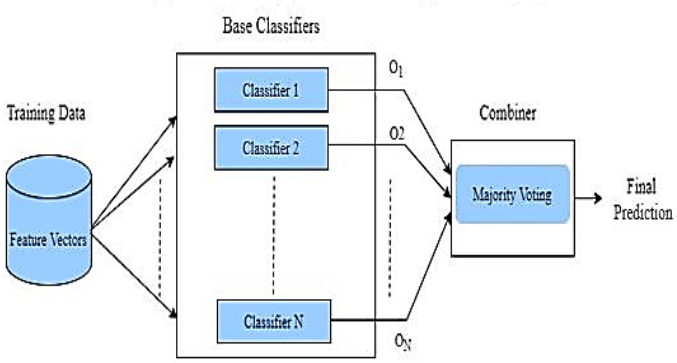

Fig. 1. Architecture of Ensemble Learning System.

III. Extreme Level Outlier Detection and ORF based Ensemble Classification

The outlier detection-based ensemble classification framework is shown in Fig. 2. It includes the following stages: 1) Data Acquisition, 2) Preprocessing, 3) Feature Ranking and 4) Classification.

A. Data Acquisition

The ECG data used for this proposed framework is taken from the clinical pre-recorded MITBIH Arrhythmia database. It consists of 48 ECG recordings, each spanning 30 minutes and captured at 360 Hz per channel with an 11-b resolution and a 10-mV range. In this work, we have evaluated the proposed model on two datasets i.e., dataset1 and dataset2. The dataset1 is the DWT processed training data taken from [24] having the specified features of amplitude, RR intervals, Speed, etc., and dataset2 consists of raw DWT processed coefficients of MITBIH arrhythmia dataset. Most recent studies concentrated on the evaluation of four classes such as N (Normal), S (Supraventricular), V (Ventricular) and F (Fusion) beats. In this work, we focus on detecting the N (Normal Synus Rhythm), and three arrhythmia's such as B (Ventricular Bigeminy), T (Ventricular Trigeminy) and VT (Ventricular Tachycardia). The waveform representation of ventricular bigeminy, ventricular tachycardia and ventricular trigemini is shown in Fig. 3(a), 3(b) and 3(c), respectively.

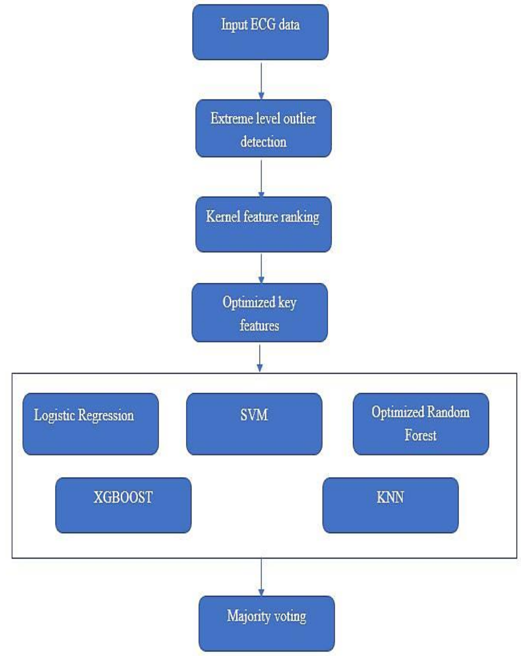

Fig. 2. Framework for Extreme Outlier Detection-and ORF based Heterogenous Ensemble Learning.

B. Data Preprocessing

Data preprocessing is the foremost step before implementing any machine learning technique. It improves the quality of data and makes it useful for modelling. The outlier technique used in this model is extreme value outlier detection. In this approach, the extreme level outliers are removed to improve the error rate in the classification problem. This approach is an extension to the traditional quartile-based filtering approach. The mathematical equation for finding the outliers is represented as follows:

At first, the data is sorted in ascending and split into three quartiles,

$$A[] = SortedAttIndices(); \tag{1}$$

The 25th, 50th and 75th percentile of data represented in three quartiles in the form of $\lambda_1, \lambda_2, \lambda_3$ and is represented as,

$$\lambda1 = V(F(|A|/4)); \tag{2}$$

$$\lambda2 = (V(F(|A|/2) + V(F(|A|/2+1)))/2; \tag{3}$$

$$\lambda3 = (V(F((|A| - |A|/4 - 1)) + V(F((|A| - |A|/4))))/2; \tag{4}$$

Inter Quartile Range (IQR) between the first and third quartiles can be calculated as,

$$\theta = \lambda3 - \lambda1; \tag{5}$$

The upper and lower extreme values of the outliers can be detected using,

$$UE[] = \lambda3 + \eta.\log(\Gamma\theta) \tag{6}$$

$$LE[] = \lambda1 - \eta.\log(\Gamma\theta) \tag{7}$$

Fig.3(a). Wave representation for ECG signal record 105 of MITDB representing Bigeminy condition.

Fig.3(b). Wave representation for ECG signal record 106 of MITDB representing ventricular tachycardia condition.

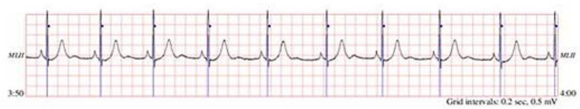

Fig.3(c). Wave representation for ECG signal record 106 of MITDB representing ventricular trigeminy condition.

Fig. 3. Waveform Representation of Bigeminy, Ventricular Tachycardia and Ventricular Trigeminy Conditions.

$$\Gamma(v/2, x/2) = \int_{x}^{inf} r^{v-1}.e^{-r}dr \tag{8}$$

$$UOutlier = \lambda 3 + \eta.\max(\Gamma(\lambda 3 + \lambda 1, 9)), \log(\Gamma\theta)) \tag{9}$$

$$LOutlier = \lambda 3 - \eta.\max(\Gamma(\lambda 3 + \lambda 1, 9)), \log(\Gamma\theta)) \tag{10}$$

C. Feature Ranking

After filtering approach, feature rankings are evaluated for better classification accuracy in the machine learning algorithms (e.g., support vector machines, logistic regression, Naive Bayes, random forest and artificial neural networks). For each feature, the usual technique of calculation is to estimate the distribution mean and standard deviation. Feature selection is a technique of eliminating redundant and nonessential data from the dataset discovering those features from those which have a significant effect on the outcome (e.g. higher accuracy in learning, lower computational cost, and better model interpretability).

In the proposed feature selection approach, an advanced kernel estimator is used to improve the hyper-parameters of the algorithm. The kernel estimator calculates the correlation values of each individual feature corresponding to the input dataset. In the proposed model, the hyperparameters were initialized using the kernel estimator and probabilistic based entropy measure. Here, each feature is ranked using the probability algorithm. The subset of top k features is selected as essential key features of the classification problem. Gaussian Estimator uses the kernel probability function to estimate the conditional variance of input data features.

$$B_f = uniqueCV(D); // \text{Unique column values} \tag{11}$$

$$HB_f = Histobins[] = histogrambin(D) \tag{12}$$

$$GaussianKernel : GK(f, \theta) = e^{-\theta^2} / (2 * f^2) \tag{13}$$

$$\psi = gkv = GK(\sum HB_f, \sum B_f); \tag{14}$$

$$KernelProbability = KP(D) = | HB_f / (\sum \psi * HB_f) | \tag{15}$$

$$GaussianEntropy : GE(d_i) = -GK(\sum_i d_i.\log(d_i), \mu_d) \tag{16}$$

The Conditional Gaussian Estimator (CGE) can be calculated as,

$$CGE : CGE(d_i) = -GK(\sum_i d_i.\log(d_i), \mu_d) - GE(\sum_i d_i) \tag{17}$$

$$KernelProbEstimation(kp) : IPSO(kp) = GE(\sum_i kp) - CGE(kp) \tag{18}$$

Gaussian entropy is used to find the entropy value of the feature based on the Gaussian Estimator. Conditional Gaussian Estimator is used to check the conditional probability of each feature value based on the Gaussian probability estimator. Finally, improved optimization hyperparameters are computed using the Gaussian Kernel Estimator and Conditional Gaussian Estimator.

D. Ensemble Classification

In this phase, an optimized random forest approach is proposed by using the key features. A heterogenous ensemble learning model is implemented by using a set of base classifiers. we have developed an ensemble learning model comprised of four base classifiers and one optimized random forest (ORF) classifier with majority voting mechanism. The standard random forest (RF) technique employs its own entropy measure for classification, but the proposed ORF uses an enhanced entropy measure for beat classification. Ensemble learning and majority voting rule is used to enhance the accuracy of heart disease prediction.

Algorithm Steps:

Step 1: Input ECG data
Step 2: Pre-process input data for missing values.
Step 3: Gradient filter is used to transform the data from unequal distribution
Step 4: For each randomized sample S_i
Do
Enhanced entropy:

$$Pr = -Prob(D_i).\log(Prob(D_i))$$

$$Ent(D) = \sum_i Pr$$

PE=Math.cbrt(entropy(data)*total*GHDSplitCriterion.computeHellinger(data)) *Pr/ (chiVal(data)).

for each sample in the test data check
If (PE>0)
Then
$S' = Classify((D_i, D_j));$
else
continue
end for

IV. Results and Discussion

On the heart ECG dataset, experiments are run in the python environment. On large, high-dimensional datasets, the extreme level outlier detection technique is used improve the true positive rate and accuracy. A hybrid heterogeneous ensemble classification framework is developed in this work to improve the overall classifier performance. Majority voting technique is employed to find the appropriate decision of individual base classifiers.

The proposed ensemble learning method compares the results using the entire training data set. As a result, each cross-validation model's prediction accuracy tends to be higher than base classification models. The proposed kernel-based feature selection-classification model outperforms conventional models, according to experimental results. A confusion matrix contains information about a classification model's actual and predicted classifications. The data in the matrix is commonly used to evaluate the performance of such a model. The classification results are displayed in the Confusion matrix shown in Table I. It depicts the connection between the actual and predicted classes. It also demonstrates how many true features were predicted as true as well as false.

TABLE I. Confusion Matrix

Actual Values		
Predicted Values	TN	FP
	FN	TP

True Negative (TN): The predicted values were correctly identified as true negatives.

True Positive (TP): The predicted values turned out to be true positives.

False Positive (FP): The predicted values were misinterpreted as true positives. i.e., the negative values were predicted to be positive.

False Negative (FN): The predicted values were incorrectly predicted as actual negatives, i.e., positive values were incorrectly predicted as negatives.

We can deduce the following statistical measures from the confusion matrix:

1) Precision: Refers to the percentage of correct positive cases.

$$\text{Precision} = \frac{TP}{(TP + FP)} \tag{19}$$

2) Recall or sensitivity: Represents the number of correctly identified positive cases.

$$\text{Recall} = \frac{TP}{(TP + FN)} \tag{20}$$

3) F1 score: Defined as the harmonic mean of precision and recall.

$$\text{F - measure} = 2 * \left[\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}\right] \tag{21}$$

4) Accuracy: It is the percentage of correct predictions out of a total number of predictions.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{22}$$

The following figures and tables show the experimental results of the MIT-ECG Data

Table II describes the sample ECG signal data with specified number of features such as RR, speed, age, sex, medicine, and class. This dataset is used to train the model using the proposed classification framework.

Fig. 4 illustrates the existing ensemble learning model on the input MITDB dataset. From the figure, it is observed that the experimentation is carried out on the dataset with ensemble of several base classifiers such as LR, SVM, RF, XGBoost and KNN. The learning model correctly classifies the instances of bigeminy, normal tachycardia and ventricular tachycardia beats with 96.17% accuracy.

Fig. 5 explains the proposed ensemble learning model on the input MITDB dataset. Here an optimized random forest (ORF) based heterogeneous ensemble learning model is developed and experiment is conducted on the dataset. From the figure, it is observed that the ORF ensemble learning model optimizes the relevant and redundant features based on entropy value and correctly classifies the instances with 99.455accuracy compared to existing ensemble learning models.

Table III lists out the comparison of proposed ensemble learning method with state-of-the-art classification techniques. The proposed optimized random forest-based ensemble model exhibits the superior performance among all the state-of-the-art methods stated.

Fig. 6 shows the comparative analysis of proposed ensemble heat-beat detection to the conventional models for accuracy metric. In this figure, as the number of samples increases along with features space, proposed model has better heat-beat detection accuracy than the previous models. Here, the cross validation is performed for 10 samples, 20 samples, 30 samples, 40 samples and 50 samples and accuracy performance for proposed model is observed.

TABLE II. DWT PROCESSED MITBIH ARRHYTHMIA DATA TAKEN FROM [24]

Amplitude	RR	Speed	Age	Sex		Medicine	Arrhythmia
0.915824	1.841667	0.49728	24	F		Yes	(B
0.794527	1.541667	0.515369	24	F		Yes	(B
0.764521	1.377778	0.554894	24	F		Yes	(B
1.039003	1.591667	0.652777	24	F		Yes	(B
2.003128	1.563889	1.280863	24	F		Yes	(B
1.688101	0.772222	2.18603	24	F		Yes	(B
1.668218	1.852778	0.900388	24	F		Yes	(B
1.995825	2.258333	0.88376	24	F		Yes	(B
0.976473	1.766667	0.55272	24	F		Yes	(B
1.191556	1.725	0.690757	24	F		Yes	(B
0.674095	1.847222	0.364924	24	F		Yes	(B
1.404964	2.263889	0.620598	24	F		Yes	(B
0.804747	1.841667	0.436967	24	F		Yes	(B
0.805688	1.880556	0.428431	24	F		Yes	(B
0.561085	1.711111	0.327907	24	F		Yes	(B
0.816213	1.702778	0.479342	24	F		Yes	(B
2.284514	1.680556	1.35938	24	F		Yes	(B
1.343283	1.85	0.726099	51	F		Yes	(B
1.346042	1.827778	0.736437	51	F		Yes	(B
1.295921	1.847222	0.701551	51	F		Yes	(B
1.171628	1.891667	0.619363	51	F		Yes	(B
1.205873	1.841667	0.654773	51	F		Yes	(B
1.198081	1.805556	0.663553	51	F		Yes	(B
1.147332	1.888889	0.607411	51	F		Yes	(B
1.354347	1.833333	0.738734	51	F		Yes	(B
1.469106	1.805556	0.813659	51	F		Yes	(B
1.350584	1.85	0.730046	51	F		Yes	(B
1.22675	1.838889	0.667115	51	F		Yes	(B
1.274882	1.811111	0.703922	51	F		Yes	(B
1.200385	1.819444	0.659753	51	F		Yes	(B
1.16124	1.894444	0.612971	51	F		Yes	(B
1.21954	1.897222	0.642803	51	F		Yes	(B
1.302295	1.886111	0.690466	51	F		Yes	(B
1.216297	1.886111	0.64487	51	F		Yes	(B
1.679835	1.522222	1.103542	51	F		Yes	(B
1.194743	1.852778	0.644839	51	F		Yes	(B
1.253372	1.791667	0.699556	51	F		Yes	(B
2.019494	1.683333	1.199699	51	F		Yes	(B
0.646879	1.736111	0.372602	51	F		Yes	(B
1.531183	1.930556	0.793131	51	F		Yes	(B
1.133171	1.847222	0.613446	51	F		Yes	(B
1.231869	1.830556	0.672948	51	F		Yes	(B

0.931352	0.813889	1.144323	69	M		Yes	(N
0.926876	0.813889	1.138824	69	M		Yes	(N
0.874316	0.813889	1.074244	69	M		Yes	(N
0.799794	0.788889	1.013823	69	M		Yes	(N
0.751938	0.788889	0.953161	69	M		Yes	(N
0.811479	0.788889	1.028636	69	M		Yes	(N
0.905821	0.816667	1.109169	69	M		Yes	(N
0.835362	0.652778	1.279703	69	M		Yes	(N
0.656436	0.991667	0.661952	69	M		Yes	(N
0.836695	0.841667	0.994093	69	M		Yes	(N
0.8435	0.808333	1.043506	69	M		Yes	(N
0.789682	0.794444	0.994005	69	M		Yes	(N
0.757089	0.769444	0.983943	69	M		Yes	(N
0.880419	0.838889	1.049507	69	M		Yes	(N
0.841678	0.855556	0.983779	69	M		Yes	(N
0.740295	0.822222	0.900359	69	M		Yes	(N
0.795441	0.830556	0.957721	69	M		Yes	(N
0.854628	0.819444	1.042936	69	M		Yes	(N
0.772212	0.794444	0.972015	69	M		Yes	(N
0.947551	0.8	1.184439	69	M		Yes	(N
0.838421	0.788889	1.062787	69	M		Yes	(N
0.945943	0.822222	1.150471	69	M		Yes	(N
0.837924	0.869444	0.963746	69	M		Yes	(N
0.808701	0.822222	0.983555	69	M		Yes	(N
0.866462	0.786111	1.102213	69	M		Yes	(N
1.041742	0.794444	1.311283	69	M		Yes	(N
0.806761	0.772222	1.044726	69	M		Yes	(N
0.846139	0.786111	1.07636	69	M		Yes	(N
0.898411	0.813889	1.10385	69	M		Yes	(N
0.79059	0.813889	0.971374	69	M		Yes	(N
0.685813	0.827778	0.828499	69	M		Yes	(N
0.777621	0.844444	0.920867	69	M		Yes	(N
1.035218	0.808333	1.280682	69	M		Yes	(N
0.80213	0.772222	1.03873	69	M		Yes	(N
0.724004	0.8	0.905005	69	M		Yes	(N
0.827432	0.788889	1.048857	69	M		Yes	(N
0.908186	0.858333	1.058081	69	M		Yes	(N
0.82578	0.841667	0.981125	69	M		Yes	(N
0.740944	0.825	0.898114	69	M		Yes	(N
0.839793	0.802778	1.046109	69	M		Yes	(N
1.005817	0.836111	1.202971	69	M		Yes	(N
0.756791	0.791667	0.955947	69	M		Yes	(N
0.799774	0.788889	1.013798	69	M		Yes	(N
0.780397	0.819444	0.952349	69	M		Yes	(N

0.886177	0.847222	1.045979	69	M		Yes	(N
0.774806	0.877778	0.882691	69	M		Yes	(N
0.751943	0.822222	0.914525	69	M		Yes	(N
0.864779	0.777778	1.111859	69	M		Yes	(N
0.865518	0.802778	1.078154	69	M		Yes	(N
0.707004	0.811111	0.871649	69	M		Yes	(N
0.709407	0.797222	0.889849	69	M		Yes	(N
0.828879	0.836111	0.99135	69	M		Yes	(N
0.81786	0.830556	0.984714	69	M		Yes	(N
0.74823	0.825	0.906946	69	M		Yes	(N
0.761216	0.811111	0.938485	69	M		Yes	(N
0.963256	0.788889	1.221029	69	M		Yes	(N
0.971116	0.783333	1.239722	69	M		Yes	(N
0.717338	0.805556	0.890489	69	M		Yes	(N
0.80705	0.841667	0.958871	69	M		Yes	(N
0.89964	0.833333	1.079568	69	M		Yes	(N
1.034261	0.830556	1.245265	69	M		Yes	(N
0.733918	0.805556	0.911071	69	M		Yes	(N
0.792855	0.777778	1.019385	69	M		Yes	(N
0.751469	0.797222	0.942609	69	M		Yes	(N
0.737086	0.783333	0.940961	69	M		Yes	(N

```
=== Classifier model (full training set) ===


 Exsiting Ensemble Classifier((LR+SVM+RF+XGBOOST+KNN) For ECG MIT Data
==========

Correctly Classified Instances        2819               96.1788 %
Incorrectly Classified Instances       112                3.8212 %
Kappa statistic                          0
Mean absolute error                      0.0377
Root mean squared error                  0.1364
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             2931

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.481     0.014     (B
                 1.000    1.000    0.962      1.000    0.981      ?        0.493     0.961     (N
                 0.000    0.000    ?          0.000    ?          ?        0.488     0.013     (T
                 0.000    0.000    ?          0.000    ?          ?        0.485     0.010     (VT
Weighted Avg.    0.962    0.962    ?          0.962    ?          ?        0.492     0.925

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
    0   42    0    0 |    a = (B
    0 2819    0    0 |    b = (N
    0   39    0    0 |    c = (T
    0   31    0    0 |    d = (VT
```

Fig. 4. Existing Ensemble Learning Result.

```
=== Classifier model (full training set) ===

 Proposed Ensemble Classifier For ECG MIT Data
==========
Correctly Classified Instances        2915               99.4541 %
Incorrectly Classified Instances        16                0.5459 %
Kappa statistic                          0.9267
Mean absolute error                      0.0034
Root mean squared error                  0.0412
Relative absolute error                  8.8962 %
Root relative squared error             30.1902 %
Total Number of Instances             2931

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.006    0.724      1.000    0.840      0.849    0.998     0.840     (B
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     (N
                 0.590    0.000    1.000      0.590    0.742      0.766    0.998     0.853     (T
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     (VT
Weighted Avg.    0.995    0.000    0.996      0.995    0.994      0.995    1.000     0.996

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
   42    0    0    0 |    a = (B
    0 2819    0    0 |    b = (N
   16    0   23    0 |    c = (T
    0    0    0   31 |    d = (VT
```

Fig. 5. Optimized Random Forest based Ensemble Learning Result.

TABLE III. COMPARISON OF PROPOSED ENSEMBLE MODEL WITH OTHER STATE-OF-THE-ART METHODS

Ref	Dataset	Classifier	Performance metrics
[18]	MITBIH	ODT+ Adaptive boosted ODT	98.77% ACC
[24]	MITBIH	KNN	98.40% ACC
[25]	MITBIH, Wearable dataset	XGBOOST	99.14% ACC| MITBIH, 98.68% ACC| Wearable dataset
[26]	MITBIH	KNN+DT	97.64% Acc
		SVM	97.58% Acc
		Ensemble Approach	97.78% Acc
		ANN with Class Weights	98.06% Acc
[27]	452 samples of sample data	RF Ensemble	90% Acc
[28]	MITBIH	RF+SVM	98.2% Acc
[29]	MITBIH	DL Ensemble	98% Acc 0.93 F1score 0.92 AUC
[30]	Kaggle	LR+SVM+DT+NB+KNN	98% Acc
Proposed Ensemble Method	MITBIH	LR+SVM+RF+XGBOOST+KNN	96.1% Acc
		LR+SVM+ORF+XGBOOST+KNN	**99.45% Acc**

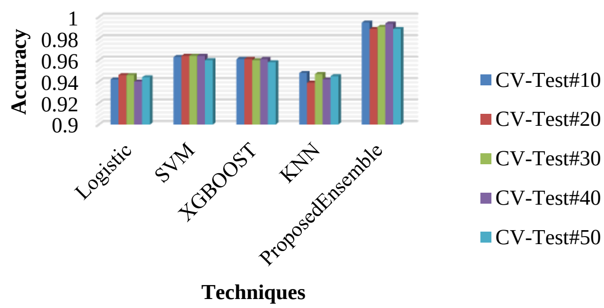

Fig. 6. Comparative Analysis of Proposed Framework to the Conventional Frameworks for ECG Heartbeat Detection for Accuracy Metric.

Fig. 7 depicts the comparative analysis of proposed ensemble heartbeat detection to the conventional models for recall metric. In this figure, as the number of samples increases along with features space, proposed model has better recall than the previous models. Here, the cross validation is performed for 10 samples, 20 samples, 30 samples, 40 samples and 50 samples and the recall for proposed model is observed.

Fig. 8 presents the comparative analysis of proposed ensemble heat-beat detection to the conventional models for F-measure metric. In this figure, as the number of samples increases along with features space, proposed model has better heat-beat detection F-measure than the previous models.

Table IV lists the comparative analysis of proposed ensemble heartbeat detection to the conventional models for AUC metric. Here, the cross-validation test of 10 samples to 50 samples is done and classification result is observed over various classifiers. In this table, as the number of samples increases along with features space, proposed model has better heartbeat detection AUC than the previous models.

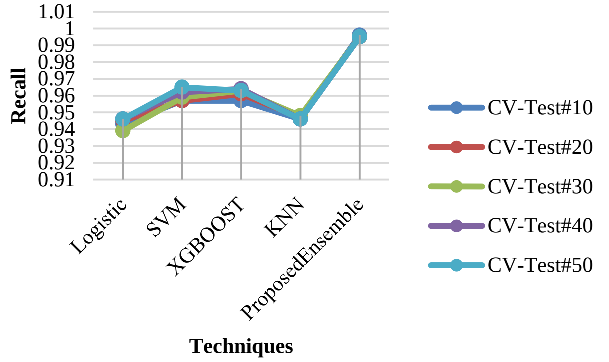

Fig. 7. Comparative Analysis of Proposed Framework to the Conventional Frameworks for ECG Heartbeat Detection for Recall Metric.

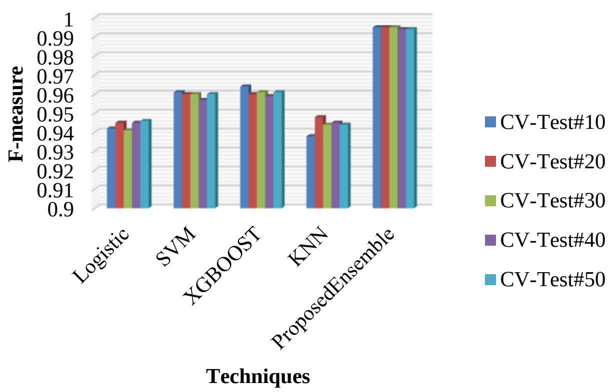

Fig. 8. Comparative Analysis of Proposed Framework to the Conventional Frameworks for ECG Heartbeat Detection for F1-Score Metric.

Table V lists the comparative analysis of proposed ensemble heartbeat detection to the conventional models for precision metric. In this table, as the number of samples increases along with features space, proposed model has better heartbeat detection precision than the previous models.

TABLE IV. COMPARISON OF PROPOSED FRAMEWORK IN TERMS OF AUC METRIC

CV-Test	Logistic	SVM	XGBOOST	KNN	HRF Ensemble Learning
CV-Test#10	0.943	0.963	0.961	0.94	0.995
CV-Test#20	0.942	0.964	0.964	0.942	0.975
CV-Test#30	0.941	0.965	0.965	0.942	0.974
CV-Test#40	0.939	0.963	0.965	0.942	0.985
CV-Test#50	0.941	0.961	0.963	0.942	0.971

TABLE V. COMPARISON OF PROPOSED FRAMEWORK IN TERMS OF AUC METRIC

CV-Test	Logistic	SVM	XGBOOST	KNN	Proposed Ensemble
CV-Test#10	0.943	0.958	0.959	0.939	0.995
CV-Test#20	0.945	0.961	0.963	0.941	0.985
CV-Test#30	0.945	0.96	0.959	0.944	0.974
CV-Test#40	0.939	0.962	0.962	0.94	0.985
CV-Test#50	0.94	0.961	0.963	0.939	0.978

V. CONCLUSION

In this paper, an optimized ensemble learning approach is implemented on the MITBIH arrhythmia dataset for better decision making. Since most of the base classifiers are independent of data size and outliers, the proposed improved inter quartile range outlier detection-based optimized random forest ensemble learning model has better efficiency in terms of outliers filtering and data classification problem. This outlier technique proposed in this work is an improvement to the traditional inter quartile range outlier detection method which removes the extreme level outliers and improves the accuracy in the classification process. After filtering, a kernel-based feature selection approach is implemented to find the ranks of the features. In addition, this paper proposed an enhanced entropy measure used in decision trees of random forest algorithm to get the optimal set of features. Finally, the ensemble learning classifies each class of heartbeat by majority voting principle and achieved an accuracy of 99.45% which outperforms the various state-of-the-methods.

ACKNOWLEDGMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] Shreya Bhattacharyya, Souvik Majumder, Arrhythmic Heartbeat Classification Using Ensemble of Random Forest and Support Vector Machine Algorithm, IEEE Transactions on Artificial Intelligence, Volume 2, Issue 3, June 2021.

[2] Aurore Lyon, Ana Minchole, Juan Pablo Martínez, Pablo Laguna, and Blanca Rodriguez, Computational techniques for ECG analysis and interpretation in light of their contribution to medical advances, Journal of the Royal Society Interface, vol. 15(138), 2018, https://dx.doi.org/10.1098%2Frsif.2017.0821.

[3] Asir Antony Gnana Singh Danasingh, Identifying Redundant Features using Unsupervised Learning for High-Dimensional Data, SN Applied Sciences, July 2020.

[4] Jafar Abdollahi & Babak Nouri-Moghaddam, A Hybrid Method for Heart Disease Diagnosis utilizing Feature Selection based Ensemble Classifier Model Generation, Iran Journal of Computer Science, May 2022, https://dx.doi.org/10.1007/s42044-022-00104-x.

[5] Utkarsh Mahadeokhaire, R. Dhanalakshmi, Stability of Feature Selection Algorithm: A review, Journal of King Saud University - Computer and Information Sciences, June 2019.

[6] Pierre Michel, Nicolas Ngo, Jean-Francois Pons, A Filter Approach for Feature Selection in Classification: Application to Automatic Atrial Fibrillation Detection in Electrocardiogram Recordings, BMC Medical Informatics and Decision Making, 21, Article Number 130, May 2021.

[7] A. Venkata Anuhya, Venkata Ratnam Kolluru and Rajesh Kumar Patjyoshi, A Deep Learning Approach for Detection and classification of QRS Contours using Single-lead ECG, International Journal of Pharmaceutical Sciences and Research, Vol. 12, No. 2, pp. 75-91, June 2020.

[8] Alex Barros, Paulo Resque, Joao Almeida, Renato Mota, Data Improvement Model based ECG Biometric for User Authentication and Identification, Sensors, Vol.20, Issue 10, May 2020.

[9] John Hart, Inter-quartile Analysis of Resting heart rate and heart rate Variability following Spinal Adjustment; A Case Study, Neuroscience Discovery, January 2019, http://dx.doi.org/10.7243/2052-6946-7-1.

[10] Bingtao Zhang, Peng Cao, Classification of High Dimensional Biomedical Data Based on Feature Selection using Redundant Removal, PLOS ONE, April 2019.

[11] Muhammed Rizwan, Bradley M Whitaker and David V Anderson, AF detection from ECG recordings using feature selection, sparse coding, and ensemble learning, Physiological Measurement, 39, 124007 (10pp), 2018, https://doi.org/10.1088/1361-6579/aaf35b.

[12] Yuwei Zhang, Yuan Zhang, Benny Lo, Wenyao Xu, Wearable ECG Signal Processing For Automated Cardiac Arrhythmia Classification using CFASE-Based Feature Selection, Expert Systems, pp 1-13, April 2019, https://doi.org/10.1111/exsy.12432.

[13] G. Angayarkanni, Dr. S. Hemalatha, Selection of Features Associated with Coronary Artery Diseases (CAD) using Feature Selection Techniques, Journal of Xi'an University of Architecture & Technology, Vol.12 (11), pp. 686-699, 2020.

[14] Vinay Varma K, Embedded Methods for Feature Selection in Neural Networks, October 2020, https://doi.org/10.48550/arXiv.2010.05834.

[15] Agnieszka Wosiak, Principal Component Analysis based on Data Characteristics for Dimensionality Reduction of ECG Recordings in Arrhythmia Classification, Open Physics, Vol. 17, Iss. 10, September 2019, https:// dx.doi.org/10.1515/phys-2019-0050.

[16] S. Wu, P. Chen, A. L. Swindlehurst, and P. Hung, Cancelable biometric recognition with ECGs: subspace-based approaches, IEEE Transactions Information Forensics and Security, Vol. 14, no. 5, pp. 1323-36, May 2019.

[17] Chulseung Yang, Gi Won Ku, Jeong-Gi Lee, Sang-Hyun Lee, Interval-Based LDA Algorithm for Electrocardiograms for Individual Verification, IEEE Transactions Information Forensics and Security, Vol. 10 (17), August 2020.

[18] Mohebbanaaz, L. V. Rajani Kumari and Y. Padma Sai, Classification of ECG beats using optimized decision tree and adaptive boosted optimized decision tree, Signal, Image and Video Processing, 16, pages 695-703, October 2021.

[19] Bo-Han Kung, Po-Yuan Hu, Chiu-Chang Huang, Cheng-Che Lee, Chia-Yu Yao, and Chieh-HsiungKuan, An Efficient ECG Classification System Using Resource-Saving Architecture and Random Forest, IEEE Journal of Biomedical Health Informatics, vol. 25(6), pp 1904-1914, 2021, DOI: 10.1109/JBHI.2020.3035191.

[20] Venkatesan, C., Karthigaikumar, P., Paul, A., Satheeskumaran, S., & Kumar, R, ECG Signal Preprocessing and SVM Classifier-Based Abnormality Detection in Remote Healthcare Applications, IEEE Access: Practical Innovations, Open Solutions, vol. 6, pp 9767-9773, 2018, https://doi.org/10.1109/access.2018.2794346.

[21] Jha, C. K., & Kolekar, M. H, Cardiac Arrhythmia Classification using Tunable Q-Wavelet Transform based Features and Support Vector Machine Classifier. Biomedical Signal Processing and Control, 59(101875), 2020, https://doi.org/10.1016/j.bspc.2020.101875.

[22] Sahoo, S., Kanungo, B., Behera, S., &Sabut, S, Multiresolution Wavelet Transform Based Feature Extraction and ECG Classification to Detect Cardiac Abnormalities. Measurement: journal of the International Measurement Confederation, Vol. 108, pp 55-66, 2017, https://doi.org/10.1016/j.measurement.2017.05.022.

[23] S. T. Aarthy and J. L. Mazher Iqbal, Time series real time naive bayes electrocardiogram signal classification for efficient disease prediction using fuzzy rules, Journal of Ambient Intelligence and Humanized Computing, Vol. 12, pp 5257-5267, 2021.

[24] VedavathiGauribidanurRangappa, Sahani Venkata AppalaVaraprasad Prasad, and Alok Agarwal, Classification of Cardiac Arrhythmia Stages using Hybrid Features Extraction with K-Nearest Neighbour Classifier of ECG Signals, International Journal of Intelligent Engineering and Systems, Vol.11, No.6, 2018.

[25] Huaiyu Zhu, Yisheng Zhao, Yun Pan, HanshuangXie, Fan Wu and Ruohong Huan, Robust Heartbeat Classification for Wearable Single-Lead ECG via Extreme Gradient Boosting, Sensors, Vol. 21, 5290, 2021, https://doi.org/10.3390/s21165290.

[26] Md. Atik Ahmed, Kazi Amit Hasan, Khan Fashee Monowar, Nowfel Mashnoor, ECG Heartbeat Classification Using Ensemble of Efficient Machine Learning Approaches on Imbalanced datasets, International Conference on Advanced Information and Communication Technology, 2020, https://doi.org/10.1109/ICAICT51780.2020.9333534.

[27] Jose Francisco Saenz-Cogollo and Maurizio Agelli, Investigating Feature Selection and RandomForests for Inter-Patient Heartbeat Classification, Vol. 13, Issue 4, March 2020.

[28] Shreya Bhattacharyya; Souvik Majumder; Papiya Debnath, Arrhythmic Heartbeat Classification Using Ensemble of Random Forest and Support Vector Machine Algorithm, IEEE Transactions on Artificial Intelligence, Vol. 2, Issue 3, June 2021.

[29] Adyasha Rath, Debahuti Mishra, Ganapati Panda, Suresh Chandra satapathy and Kaijian Xia, Improved heart disease detection from ECG signal using deep learning based ensemble model", Sustainable Computing: Informatics and Systems, Vol. 35, 100732, September 2022.

[30] Suresh Subramanian and Y Angeline Christobel, A Hybrid Machine Learning Model to Predict Heart Disease Accurately, Vol. 15, Iss. 12, pp. 527-534, March 2022, https://doi.org/ 10.17485/IJST/v15i12.104.