Feature Selection by ModifiedBoostARoota and Classification by CatBoost Model on High Dimensional Heart Disease Datasets

Anuradha. P and Vasantha Kalyani David

Abstract—**As heart disease is the leading cause of mortality worldwide, early detection and prevention of the disease would reduce the mortality rate. Various Machine Learning Algorithms are employed in the classification and prediction of diseases. For accurate prediction, Feature Selection algorithms are employed to choose features that have a significant association with the disease or target variable. This would reduce computing time and improve the prediction performance. In this paper, ModifiedBoostARoota (MBAR) algorithm was used for Feature Selection, and classifiers CatBoost, XGBoost, Decision Tree, Extra Trees Classifier, Support Vector Classifier, Logistic Regression, K Nearest Neighbors, Naive Bayes, and Random Forest were applied on UCI Arrhythmia dataset and UCI Z-Alizadeh Sani dataset. Synthetic Minority Over Sampling Technique (SMOTE) was used to balance the dataset. A comparison of the performance of the models on the imbalanced and balanced datasets shows that MBAR with CatBoost classifier gives better accuracy of 92.76% on the balanced Z-Alizadeh Sani dataset and 86.33% on the balanced Arrhythmia dataset.**

Index Terms—**Heart disease, feature selection, CatBoost, classification.**

I. Introduction

Heart disease is the leading cause of human deaths globally. According to the World Health Organization, Cardio Vascular Diseases were responsible for 38 percent of the 17 million premature deaths (below 70 years of age) caused by noncommunicable diseases in 2019. The four main types of heart diseases are i) Heart failure, ii) heart valve disease, iii) Cardiac Arrhythmia and iv) coronary artery disease.

i) If a valve in the heart is damaged or diseased, it leads to heart valve disease. ii) When the heart muscle becomes weak or when heart chambers are not filled with sufficient blood, the heart will not be able to pump the adequate blood required for the body. This condition is called heart failure.

iii) Cardiac Arrhythmia indicates an abnormality in the sequence of electrical impulses, causing the improper beating of the heart [1]. Arrhythmias may be harmless or life-threatening. The heart's electrical activity can be recorded using Electrocardiography (ECG or EKG), which can help diagnose Arrhythmias [1]. To predict Arrhythmia, analysis of each heartbeat of the ECG records might be done for long hours or days [2]. To detect abnormalities quickly and correctly like Arrhythmia in ECG, Machine Learning algorithms can be used, which would be a support for the medical practitioner. iv) Coronary Artery Disease (CAD) arises due to the accumulation of plaque inside the lining of the coronary arteries that would block blood flow to the heart [3].

High blood pressure, diabetes, low HDL cholesterol, family history, high LDL cholesterol, and smoking are the traditional risk factors for CAD [4]. Machine Learning Algorithms when applied much earlier in life on these risk factors can predict whether an individual is likely to get heart disease or not. In case, if the prediction is positive then, preventive measures to avoid CAD would be to adopt a healthy lifestyle, which includes good nutrition and physical activity [4].

A. *Machine Learning Algorithms*

Machine learning techniques are devised to predict the target/ output/ dependent variable, for the given input/ predictor variables [5]. Various Machine learning algorithms are available and focus of all research works would be to choose the right algorithm that would best suit for the specific dataset. For supervised learning where the output is known, various algorithms namely Linear Regression, Random Forest, Logistic Regression, CatBoost, Support Vector Machines, K Nearest Neighbors, Decision Trees, XGBoost etc., are widely used.

B. *Feature Selection*

In datasets, especially in high dimensional datasets, not all features contribute to the prediction of the target or outcome variable. So, selecting the features that are highly associated with the target/class variable would highly contribute to effective prediction as well as save computing time.

C. *Synthetic Minority Over-Sampling Technique (SMOTE)*

In an imbalanced dataset, all classes will not have an equal number of instances. The classifiers perform better on balanced datasets compared to imbalance datasets. N.V. Chawla *et al*., in their paper on SMOTE, showed that better classifier performance can be achieved by over-sampling the abnormal/ minority class and under-sampling the normal/ majority class. [6].

The objective of this work is to focus on the limitations mentioned by the authors in their previous research work in [7]. A feature selection technique called ModifiedBoostARoota algorithm (MBAR) was devised and

Manuscript received April 6, 2022; revised July 10, 2022.

Anuradha. P and Vasantha Kalyani David are with the Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, deemed to be University, Coimbatore, India (e-mail: anujith72@gmail.com).

DOI: 10.7763/IJCTE.2022.V14.1321

applied only on low dimensional heart disease datasets in [7]. The authors had earlier mentioned that MBAR was not applied on high-dimensional datasets due to time constraint.

Therefore, in this work, MBAR is applied on high dimensional heart disease datasets; namely, Arrhythmia dataset and Z-Alizadeh Sani dataset available in the UCI Machine Learning repository where the performance of the classifiers is compared when applied on these datasets with and without selected features. Also, the performances of the selected classifier on the imbalanced and balanced datasets are compared.

The following subsections consist of related work discussion and methodology in Section II, brief description of the datasets in Section III, results and discussion in Section IV, limitations in Section V and conclusion in Section VI.

II. Related Literature Survey and Methodology

On reviewing the feature selection and classification techniques used in earlier research works done on the Arrhythmia dataset, it is observed that A. Mustaqeem *et al.* [8] had accomplished Feature Selection by creating shadow features of each feature based on z-score feature importance by Random Forest Classifier. Those features with a z-score less than the maximum value of shadow features were eliminated [8]. On applying repeated ten-fold cross-validation, those authors found that Multi-layer perceptron gave higher accuracy of 78.2% compared to other classifiers [8].

A. Mustaqeem *et al*., in their work in [9], applied Support Vector Machine (SVM) based methods including one-against-all (OAA), one-against-one (OAO), and error-correction code (ECC). The OAO method when used with 80/20 data split, achieved an accuracy rate of 81.11% and on 90/10 data split, the accuracy obtained was 92.07%.

Khare *et al*., in [10], employed Spearman Rank Correlation for selecting features and Principal Component Analysis (PCA) was used for feature extraction. Then SVM was employed for classification, which gave an accuracy of 85.98%.

Fei Yang *et al.* [11] used an advanced approach for missing-value imputation called Robust Principle Component Analysis (RPCA) along with Zero, Mean, and PCA imputation methods. They modified KDF-WKNN by a correction factor. This modified kernel Difference-Weighted KNN (MKDF-WKNN) classification algorithm was used to manage the imbalance datasets problem and an accuracy of 73% was achieved [11].

Ersen Yılmaz had designed an expert system where feature selection was implemented by F-score and classification was done using Least Squares Support Vector Machines (LS-SVM), in which, Gaussian radial basis function was used as the kernel. The accuracy obtained was 82.09% [12].

Jadhav *et al.* used momentum learning rule with back-propagation algorithm which yielded 82.22% classification accuracy [13].

M. A. Khan and Y. Kim applied the hybrid model, principal components analysis (PCA) with LSTM for classification and attained a classification accuracy of 93.5% [14].

Mitra and Samanta employed correlation-based feature selection (CFS) with linear forward selection search [15]. On applying the Incremental back-propagation neural network (IBPLN) and Levenberg-Marquardt (LM) model, a classification accuracy of 87.71% was obtained [15].

Shimpi *et al.*, found that Support Vector Machine classifier yielded a better accuracy of 91.2% compared to other models considered in their work [16].

Ayar *et al*., applied the hybrid model, genetic algorithm along with Decision Tree~~,~~ for Classification. This hybrid model when applied on two-classes achieved an accuracy of 86.96% [17].

The review of the classification works done on Z-Alizadeh Sani dataset are as follows:

Kolukisa *et al.* applied the linear discriminant analysis and the SVM algorithm, which yielded an accuracy of 92.74% [18].

Kolukisa *et al.* in [19] devised an adaptive ensemble classifier consisting of Logistic Regression, k-Nearest Neighbor, Linear Discriminant Analysis, Support Vector Machine, Naïve Bayes classification algorithms and obtained 88.38% accuracy [19].

Gupta *et al*. designed a computational intelligent system, C-CADZ, using fixed analysis of mixed data (FAMD) and Binary Bat Algorithm (BBA) for feature extraction, after which an accuracy of 97.37% was achieved by applying an ensemble model of Random Forest and Extra Tree classifier [20].

Arabasadi *et al.* [21] focused on the concept that CAD occurs if one of the left circumflex (LCX) or left anterior descending (LAD) or right coronary (RCA) arteries is stenotic [22]. By using hybrid Neural Network-Genetic algorithms model, those authors achieved 93.85% accuracy. Dahal *et al*. compared five classifiers and observed that the SVM model’s prediction was more effective with an accuracy of 89.47% [23].

Cuvitoglu and Isik used Principal Component Analysis (PCA) t-test for feature selection, where five classifiers were compared and Artificial Neural Networks (ANN) yielded an Area-Under-the-Curve value of 93% [24].

Alizadeh Sani *et al.* in [25], used cost-sensitive algorithms along with base classifiers of Support Vector Machine (SVM), K-Nearest Neighbors (KNN), C4.5, Sequential Minimal Optimization (SMO), and Naïve Bayes with ten-fold cross-validation, and better accuracy of 92.09% was achieved by Sequential Minimal Optimization (SMO) [25].

The summarized form of related work on both datasets can be seen in Table II and Table V, where the proposed model is also compared with the related work. Table II shows that almost all the authors have used feature selection on Arrhythmia dataset. In the future, as an extension of these related works, tree-based models can be experimented on both datasets.

Fig. 1 depicts the methodology adopted in this work. The ModifiedBoostARoota (MBAR) algorithm is used for Feature Selection and Classifiers CatBoost, XGBoost, Logistic Regression, Decision Tree, Support Vector Classifier, K Nearest Neighbors, Extra Trees Classifier, Gaussian Naive Bayes and Random Forest were applied on

UCI Arrhythmia dataset and Z-Alizadeh Sani dataset. The stratified 10-fold cross validation accuracy score with three repeats of all the above-mentioned classifiers is compared. Also, on splitting the datasets as 70% train and 30% test sets, the precision, recall, f1 score and AUC score of all these classifiers are analyzed and the best performing classifier is selected. The selected model is applied on the two unbalanced feature selected datasets, SMOTE-balanced feature selected datasets and on the two datasets with no feature selection. The performances are consequently evaluated. The ModifiedBoostARoota feature selection algorithm's performance on high dimensional datasets is assessed [7].

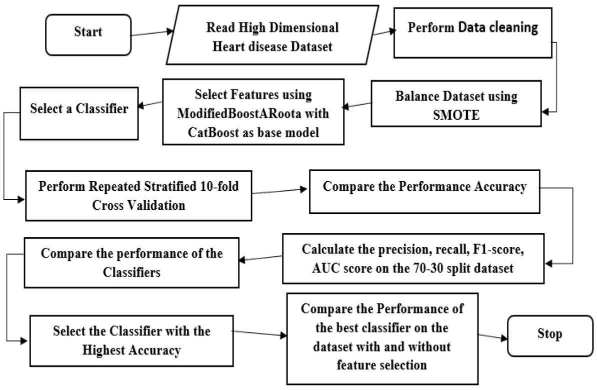
Start
Read High Dimensional Heart disease Dataset
Perform Data cleaning
Balance Dataset using SMOTE
Select Features using ModifiedBoostARoota with CatBoost as base model
Select a Classifier
Perform Repeated Stratified 10-fold Cross Validation
Compare the Performance Accuracy
Calculate the precision, recall, F1-score, AUC score on the 70-30 split dataset
Compare the performance of the Classifiers
Select the Classifier with the Highest Accuracy
Compare the Performance of the best classifier on the dataset with and without feature selection
Stop

Fig. 1. Methodology.

III. Data Sets

Two datasets were used in this work. The first high-dimensional dataset, the Arrhythmia heart dataset [26] in the UCI Machine Learning Repository, consists of ECG signals data with 279 attributes and 452 instances. Among the attributes, 206 contained linear values, and the rest are nominal. The instances of the dataset belonged to sixteen groups or classes [17], [26]. Class 1 referred to normal beats. Class 2 to Class 15 referred to different types of Arrhythmias. Unclassified beats were grouped as in Class 16 [17], [26]. There are 245 instances of normal types, and 207 instances of the abnormal types. In this work, these instances are grouped into two classes: i) normal and ii) arrhythmia.

The second dataset used is the Z-Alizadeh Sani dataset [27] in the UCI Machine Learning Repository that consists of 54 features related to coronary artery disease and 303 instances. The dataset contains ECG, demographic, laboratory, echo, symptom and examination data of the patients [27]. A patient is categorized as normal, if his/her diameter narrowing is less than 50%; otherwise she or he has CAD.

IV. Results and Discussion

ModifiedBoostARoota algorithm (MBAR) is a wrapper method for a feature selection devised by the authors (Anuradha and David) in their previous work [7], which is mentioned in Fig. 2. MBAR was developed by modifying BoostARoota (BAR) algorithm. BAR was published in Python package Index (PyPI)) and devised by Chasedehan [7]. Catboost is used as the base model in MBAR [7]. In this article, MBAR algorithm is used for feature selection. The experiment was carried out using python on a system with 4 GB RAM and ubuntu operating system.

Algorithm ModifiedBoostARoota [7]:

1. Compute shadow feature (by shuffling original features at random) for each feature in the dataset and merge the shadow features with the dataset to form an extended dataset of 'n' features.

2. Using any Tree-Based models, compute the Feature Importance (FI) of all features in the extended dataset.

3. Assign rank, r_i for all features $i =$ to1 $to\ n$.

4. If FI of original feature < FI of the corresponding shadow feature, then eliminate that original feature and its shadow feature.

5. If FI of any feature is insignificant then remove that feature.

6. Compute fscore for each feature in the extended dataset,
$fs_i = \frac{r_i}{FI_i}$, i=1 to n

7.Compute weighted harmonic mean,

$$m = \frac{\sum r_i}{\sum fs_i} , i = 1\ to\ n$$

8. For any feature ´ i ´ in the extended dataset, if $i < whm$, eliminate the feature ´ i ´ .

9. If fs of any original feature < fs of its corresponding shadow feature, then eliminate that original feature. Also, if fs of any feature is insignificant then remove that feature.

10. Repeat steps 1 to 9 until in each iteration at least 10% of the features are eliminated or if maximum iterations have not been completed. Else, return the remaining features and stop.

Fig. 2. ModifiedBoostARoota algorithm for feature selection.

Initially, in the Arrhythmia dataset, missing values was filled with mean values 36, 49, 37, -14, 75 in columns c10, c11, c12, c13, c14. The normal class was defined as 0 and all other classes in the target variable were grouped as 1. There are 245 instances of class 0 and 207 instances of class 1 [26].

Applying Catboost classifier on Arrhythmia dataset with all features, the stratified ten-fold cross validation accuracy score with three repeats was 83.93%. After balancing the dataset with SMOTE, we get 245 instances of both classes. Then, applying Catboost classifier on the dataset with all features, the stratified ten-fold cross validation accuracy score with three repeats was 85.44%.

Using ModifiedBoostARoota algorithm (MBAR) for feature selection on the unbalanced dataset, one gets 64 features being selected. Upon applying various classifiers namely XGBoost, Logistic Regression, Catboost, Decision Tree Classifier, Gaussian Naive Bayes, Extra Trees, K Nearest Neighbors, Random Forest and Support Vector Classifier, it was observed that Catboost yields highest accuracy of 85.77%.

TABLE I: Performance of the Classifiers by Repeated Stratified K-Fold CV on Arrhythmia Dataset

Classifiers	**Accuracy**
XGBoost	84.27%
Logistic Regression	73.20%
Decision Tree Classifier	72.86%
Gaussian Naïve Bayes	73.61%
K Nearest Neighbors	71.84%
Random Forest	86.06%
Extra Trees	85.51%
Support Vector Classifier	75.37%
CatBoost	86.33%

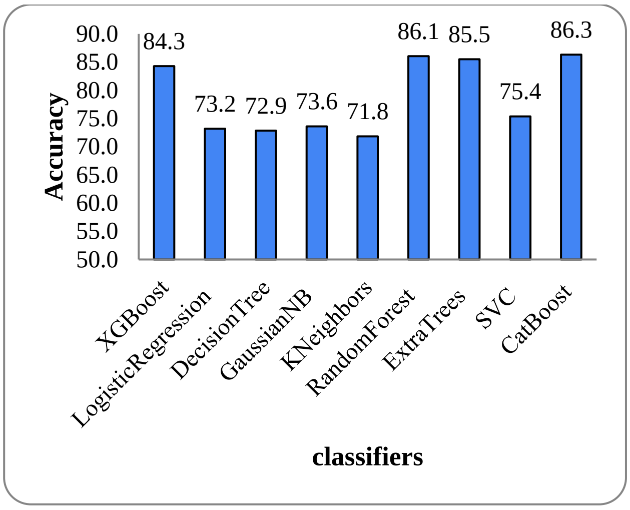

Fig. 3. Comparison of Classifiers by repeated stratified k-fold CV on Arrhythmia dataset after SMOTE and Feature Selection using MBAR.

After balancing the dataset with SMOTE and selecting features using MBAR, Table I displays the accuracy obtained by various classifiers after performing stratified ten-fold cross validation with three repeats. On comparing the performance of all classifiers, CatBoost gives the highest accuracy of 86.33%. Fig. 3 shows that the Tree-Based models performed better on Arrhythmia dataset.

TABLE II: A Comparison of the Models Applied on Arrhythmia Dataset by Various Authors

Author	Classifier on Arrhythmia dataset	Accuracy
Mustaqeem *et al*.	FI by RF+MLP	78.2
Mustaqeem *et al*.	Wrapper FS+SVM(OAO)	92.07
Khare *et al*.	Rank corr + PCA + SVM	85.98
Yang *et al*.	MKDF-WKNN	73.01
Yilmaz *et al*.	Fscore+LSSVM	82.09
Jadhav *et al*.	BPNN	82.22
Ayar *et al*.	GA+DT	86.96
Mitra *et al*.	CFS+IBPNN+LM	87.71
Shimpi *et al*.	PCA+ SVM	91.2
Anuradha and David	MBAR+Catboost	85.77
Anuradha and David	MBAR+Catboost (balanced with SMOTE)	86.33

Table II shows a comparison of the models applied on Arrhythmia dataset by various authors. On comparing the performance of other models proposed by various authors detailed in section II, Fig. 4 shows that the proposed model, MBAR+Catboost (balanced with SMOTE) performs generally on par with all models; however, more accurately than those of the Mustaqeem *et al*.'s first method, Khare *et al*., Yang *et al*., Jadhav *et al*. and Yilmaz *et al*. used on Arrhythmia dataset.

Performing 70-30 split of the balanced Arrhythmia dataset with features selected by MBAR, and analyzing the performance of various classifiers, one gets CatBoost classifier displaying higher performance compared to the other classifiers. Table III shows the precision, recall and f1-score of the classifiers considered in the comparison. It shows that Random Forest and CatBoost have very close values of precision, recall and f1-score.

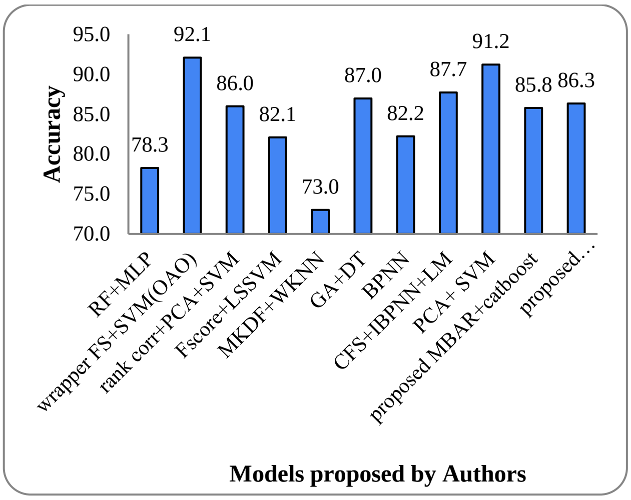

Fig. 4. Comparison of models by various authors on Arrhythmia dataset.

TABLE III: Performance of Classifiers after SMOTE, Feature Selection by MBAR and Applying Train-Test Split on Arrhythmia Dataset

Classifiers	Class	Precision	Recall	F1-score	Accuracy
XGBoost	0	0.79	0.83	0.81	80.95%
	1	0.83	0.79	0.81	
Logistic Regression	0	0.67	0.72	0.69	69.39%
	1	0.72	0.67	0.69	
Decision Tree	0	0.68	0.76	0.72	70.75%
	1	0.75	0.66	0.70	
Gaussian NB	0	0.68	0.93	0.79	75.51%
	1	0.90	0.59	0.71	
K Nearest Neighbors	0	0.67	0.89	0.76	73.47%
	1	0.85	0.59	0.70	
Random Forest	0	0.82	0.82	0.82	82.31%
	1	0.83	0.83	0.83	
Extra Trees	0	0.81	0.80	0.81	81.63%
	1	0.82	0.83	0.82	
CatBoost	0	0.82	0.85	0.83	83.67%
	0	0.85	0.83	0.84	

Fig. 5 shows the Receiver operating characteristic curve of classifiers, considered in this study, on Arrhythmia dataset.

The AUC score of CatBoost (CB) is higher than other classifiers. In the Z-Alizadeh Sani dataset [27], the target variable has 216 instances of class 1 and 87 instances of the class 0.

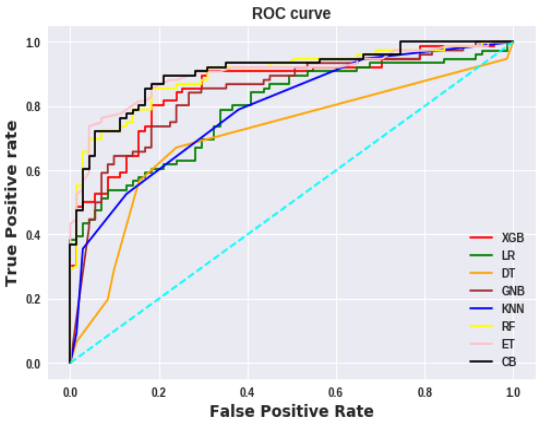

Fig. 5. Receiver operating characteristic curve of classifiers on Arrhythmia dataset.

Classification on the Z-Alizadeh Sani dataset with all features resulted in Catboost yielding a higher accuracy of 87.78%.

On performing feature selection with ModifiedBoostARoota (MBAR) on the imbalanced z-Alizadeh Sani dataset out of 55 features, 12 features were selected. On applying classifiers on these selected features, Catboost gave a better accuracy of 89.44%.

After balancing the dataset with SMOTE, the authors get 216 instances of both classes. The classification accuracy by Catboost applied on all features was 92.14%.

TABLE IV: COMPARISON OF CLASSIFIERS AFTER APPLYING SMOTE, FEATURE SELECTION BY MBAR AND REPEATED STRATIFIED K-FOLD CV APPLIED ON Z-ALIZADEH SANI DATASET

Classifier on Z-Alizadeh Sani dataset	Accuracy
XGBoost	91.21
Logistic Regression	91.98
Decision Tree	86.81
Support Vector Machine	78.32
Gaussian Naïve Bayes	89.28
K-Nearest Neighbors	74.93
Random Forest	92.52
Extra Trees	92.22
CatBoost	92.76

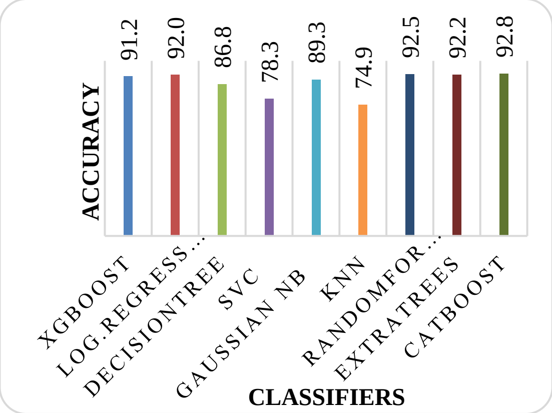

Fig. 6. Performance of classifiers on SMOTE-MBAR and repeated stratified k-fold CV applied Z-Alizadeh Sani dataset.

On performing feature selection by MBAR on the balanced dataset, 21 features were selected. Applying various classifiers using stratified ten-fold cross-validation with three repeats on the balanced dataset, the Catboost model outperformed others yielding an accuracy of 92.76%. Table VI displays the accuracy yielded by various classifiers by repeated stratified k-fold Cross Validation applied on the balanced Z-Alizadeh Sani dataset. Fig. 6 shows that most of the classifiers have performed equally well and Catboost leads by a small margin.

Table V displays the models proposed by various authors mentioned in Section II and exhibits the accuracy obtained by the models they used. Fig. 7 compares the performance of various models proposed by other authors. The proposed model MBAR and Catboost when applied on the balanced dataset outperform other authors' models by yielding an accuracy of 92.76%.

TABLE V: COMPARISON OF THE PROPOSED MODEL WITH EARLIER MODELS ON Z-ALIZADEH SANI DATASET

Authors	Models	Accuracy
dahal *et al.*,	SVM	89.47
koluisa *et al.*,	ensemble	88.38
koluisa *et al.*,	LDA-SVM	92.74
Cuvitoglu *et al.*,	ANN	85
R. Alizadehsani *et al.*,	SMO	92.09
Proposed-MBAR- Catboost	10-fold CV	89.44
Proposed-MBAR- Catboost (balanced with smote)	10-fold CV	92.76

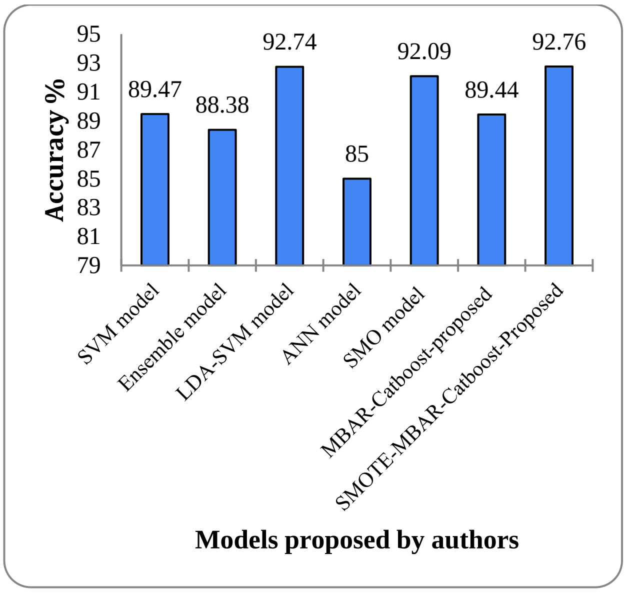

Fig. 7. Comparison of models by various authors on Z-Alizadeh Sani dataset.

Table VI shows the precision, recall and F1 score of the classifiers applied on 70-30 split of the balanced and selected features of Z-Alizadeh Sani dataset. CatBoost classifiers outperforms all other classifiers taken into comparison in this study. Fig. 8 showcases the ROC curve of the various classifiers considered in this study and finds CatBoost displaying better score that the others.

As CatBoost shows better performance compared to other classifiers, on modelling CatBoost on the 70-30 train-test split of the balanced Z-Alizadeh Sani dataset, we find as shown in Table VII that, MBAR-CatBoost combination demonstrates better performance compared to the performance of the classifier applied on the dataset with no feature selection.

TABLE VI: PERFORMANCE OF CLASSIFIERS AFTER SMOTE AND FEATURE SELECTION BY MBAR ON THE TRAIN-TEST SPLIT Z-ALIZADEH SANI DATASET

Classifiers	**class**	**Precision**	**Recall**	**F1-score**	**Accuracy**
XGBoost	0	0.94	0.97	0.95	95.38%
	1	0.97	0.94	0.95	
Logistic Regression	0	0.92	0.94	0.93	93.08%
	1	0.94	0.92	0.93	
Decision Tree	0	0.86	0.89	0.88	87.69%
	1	0.89	0.86	0.88	
Gaussian NB	0	0.90	0.89	0.90	90%
	1	0.90	0.91	0.90	
K Nearest Neighbors	0	0.73	0.84	0.78	76.92%
	1	0.82	0.70	0.75	
Random Forest	0	0.94	0.95	0.95	94.62%
	1	0.95	0.94	0.95	
Extra Trees	0	0.93	0.97	0.95	94.62%
	1	0.97	0.92	0.95	
CatBoost	0	0.94	0.98	0.96	96.15%
	1	0.98	0.94	0.96	

TABLE VII: COMPARISON OF THE CLASSIFICATION PERFORMANCE WITH AND WITHOUT FEATURE SELECTION AND APPLYING TRAIN-TEST SPLIT ON Z-ALIZADEH SANI DATASET

Z-Alizadeh Sani dataset	**class**	**Precision**	**Recall**	**F1-score**	**Accuracy**
No features Selection and Catboost Classifier	0	0.93	0.98	0.95	95.38%
	1	0.98	0.92	0.95	
With feature selection by **MBAR** and **CatBoost** Classifier	0	0.94	0.98	0.96	**96.15%**
	0	0.98	0.94	0.96	

Table VIII shows the Area-Under-the-Curve (AUC) scores of the various classifiers applied on the feature-selected balanced datasets considered in this study. It shows that the Tree-Based models have better AUC scores compared to other models, and CatBoost also outperforms all models.

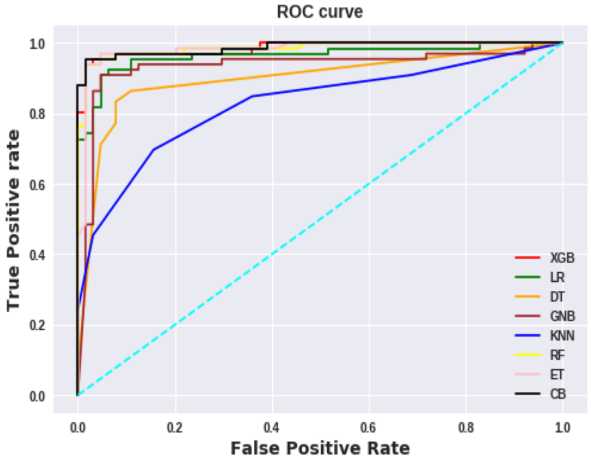

Fig. 8. Receiver operating characteristic curve of classifiers on Z-Alizadeh Sani dataset.

TABLE VIII: COMPARISON OF AUC SCORES OF VARIOUS CLASSIFIERS ON Z-ALIZADEH SANI DATASET AND ARRHYTHMIA DATASET

Classifier	**Z-Alizadeh Sani dataset- AUC score**	**Arrhythmia Dataset- AUC score**
XGBoost	0.985	0.859
Logistic Regression	0.965	0.795
Decision Tree	0.898	0.707
Gaussian NB	0.934	0.840
K Nearest Neighbors	0.826	0.788
Random Forest	0.985	0.896
Extra Trees	0.982	0.899
CatBoost	**0.987**	**0.904**

Table IX shows the comparison of the performance of the CatBoost Classifier on the Arrhythmia dataset with and without feature selection. The performance of MBAR with Catboost on the dataset is higher than that without feature selection.

TABLE IX: COMPARISON OF THE CLASSIFICATION PERFORMANCE WITH AND WITHOUT FEATURE SELECTION AND APPLYING TRAIN-TEST SPLIT ON ARRHYTHMIA DATASET

Arrhythmia DS	**class**	**Precision**	**Recall**	**F1-score**	**Accuracy**
No features Selection and Catboost Classifier	0	0.80	0.83	0.81	81.63%
	1	0.84	0.80	0.82	
With feature selection by **MBAR** and **CatBoost** Classifier	0	0.82	0.85	0.83	**83.67%**
	0	0.85	0.83	0.84	

Therefore, both Table VII and Table IX evidence that classification done on feature-selected datasets yields high performance compared to datasets with all features considered.

V. Limitations

ModifiedBoostARoota (MBAR) can be tried on more high-dimensional datasets. Due to time constraints and the non-availability of high-dimensional datasets on heart disease, only two high-dimensional heart datasets were used in this research article.

VI. Conclusion

In this work, feature selection by ModifiedBoostARoota (MBAR) was applied on high dimensional datasets namely, Arrhythmia dataset and Z-Alizadeh Sani dataset. Various classifiers namely XGBoost, Logistic Regression, CatBoost, Decision Tree Classifier, Gaussian Naive Bayes, K Nearest Neighbors, Random Forest, Extra Trees and Support Vector Classifier were used on both the datasets. Their performances by repeated stratified k-fold cross-validation and by 70-30 train-test split were observed on both datasets.

The accuracy yielded by classifiers when applied on features selected with MBAR was better than the accuracy obtained without feature selection. Moreover, on balancing both the datasets with SMOTE, the performance of the classifiers increased. Performing stratified 10-fold cross-validation with three repeats on the balanced Arrhythmia dataset with all the above-mentioned classifiers, the CatBoost model outperformed others by yielding an accuracy of 86.33%. Similarly, on the balanced Z-Alizadeh Sani dataset, the accuracy obtained by MBAR with Catboost was 92.76%. The precision, recall, and f1-score of the classifiers were compared and the highest performance was exhibited by CatBoost. The classification done by CatBoost on both the datasets with features selected by MBAR yielded a better performance as compared to datasets with no feature selection.

Thus, by selecting the prominent features and using a strong classifier, a correct prediction of heart diseases can be performed, thereby saving human lives and preventing death.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Anuradha.P had conducted the research, analyzed the performance of the models and wrote the paper.

Dr. Vasantha Kalyani David had guided towards the research work. Both the authors approve the final version.

References

[1] American Heart Association. [Online]. Available: https://www.heart.org/en /health-topics /arrhythmia/about-arrhythmia

[2] E. J. S. Luz, W. R. Schwartz, G. Cámara-Chávez, and D. Menotti, "ECG-based heartbeat classification for arrhythmia detection: A survey," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 144-164, 2016.

[3] National Heart, Lung, and Blood Institute (NHLBI). What Is Coronary Heart Disease? [Online]. Available: https://www.nhlbi.nih.gov/health-topics/coronary-heart-disease

[4] Coronary Artery Disease - Coronary Heart Disease. (2015). [Online]. Available: https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease/coronary-artery-disease

[5] J. Brownlee. (2016). How machine learning algorithms work (they learn a mapping of input to output). *Machine Learning Algorithms*. [Online]. Available: https://machinelearningmastery.com/how-machine-learning-algorithms-work

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

[7] Anuradha. P and V. K. David, "Feature selection using ModifiedBoostARoota and prediction of heart diseases using gradient boosting algorithms," in *Proc. 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 2021, pp. 19-23.

[8] A. Mustaqeem, S. M. Anwar, M. Majid, and A. R. Khan, "Wrapper method for feature selection to classify cardiac arrhythmia," in *Proc. 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC*), 2017, pp. 3656-3659.

[9] A. Mustaqeem, S. M. Anwar, and M. Majid, "Multiclass classification of cardiac arrhythmia using improved feature selection and SVM invariants," *Computational and Mathematical Methods in Medicine*, vol. 1, pp. 1-10, 2018.

[10] S. Khare, A. Bhandari, S. Singh, and A. Arora, "ECG arrhythmia classification using Spearman rank correlation and support vector machine," in *Proc. the International Conference on Soft Computing for Problem Solving (SocProS 2011)*, K. Deep, A. Nagar, M. Pant, J. Bansal, Eds. *Advances in Intelligent and Soft Computing*, Springer, India, vol. 131, 2012.

[11] F. Yang, J. Du, J. Lang, W. Lu, L. Liu, C. Jin, and Q. Kang, "Missing value estimation methods research for arrhythmia classification using the modified kernel difference-weighted KNN algorithms," *BioMed Research International*, vol. 2020, p. 9, 2020.

[12] E. Yılmaz, "An expert system based on fisher score and LS-SVM for cardiac arrhythmia diagnosis," *Computational and Mathematical Methods in Medicine*, vol. 2013, p. 6, 2013.

[13] S. M. Jadhav, S. L. Nalbalwar, and A. A. Ghatol, "Modular neural network-based arrhythmia classification system using ECG signal data," *International Journal of Information Technology and Knowledge Management*, vol. 4, no.1, pp. 205-209, 2011.

[14] M. A. Khan and Y. Kim, "Cardiac arrhythmia disease classification using LSTM deep learning approach," *Computers, Materials & Continua*, vol. 67, no.1, pp. 427–443, 2021.

[15] M. Mitra and R. K. Samanta, "Cardiac arrhythmia classification using neural networks with selected features," *Procedia Technology*, vol. 10, pp. 76-84, 2013.

[16] P. Shimpi, S. Shah, M. Shroff, and A. Godbole, "A machine learning approach for the classification of cardiac arrhythmia," in *Proc. 2017 International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 603-607,2017.

[17] M. Ayar and S. Sabamoniri, "An ECG-based feature selection and heartbeat classification model using a hybrid heuristic algorithm," *Informatics in Medicine Unlocked*, vol. 13, pp. 167-175, 2018.

[18] B. Kolukisa *et al*., "Evaluation of classification algorithms, linear discriminant analysis and a new hybrid feature selection methodology for the diagnosis of coronary artery disease," in *Proc. 2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 2232-2238.

[19] B. Kolukisa *et al*., "Coronary artery disease diagnosis using optimized adaptive ensemble machine learning algorithm," *International Journal of Bioscience, Biochemistry, and Bioinformatics*, vol. 10, no. 1, 2020.

[20] A. Gupta, R. Kumar, H. S. Arora *et al*., "C-CADZ: Computational intelligence system for coronary artery disease detection using Z-Alizadeh Sani dataset," *Appl. Intell*., 2021.

[21] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm," *Comput Methods Programs Biomed*, vol. 141, pp. 19-26, 2017.

[22] R. O. Bonow, D. L. Mann, D. P. Zipes, and P. Libby, *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine*, 9th ed. Elsevier Science, 2011.

[23] K. R. Dahal and Y. Gautam, "Argumentative comparative analysis of machine learning on coronary artery disease," *Open Journal of Statistics*, vol. 10, no. 4, pp. 694-705, 2020.

[24] A. Cüvitoğlu and Z. Işik, "Classification of CAD dataset by using principal component analysis and machine learning approaches," in *Proc. 2018 5th International Conference on Electrical and Electronic Engineering (ICEEE)*, 2018, pp. 340-343.

[25] R. Alizadehsani, M. J. Hosseini, Z. A. Sani, A. Ghandeharioun, and R. Boghrati, "Diagnosis of coronary artery disease using cost-sensitive algorithms," in *Proc. 2012 IEEE 12th International Conference on Data Mining Workshops*, 2012, pp. 9-16.

[26] Arrhythmia Dataset. (1998). [Online]. Available: https://archive.ics.uci.edu/ml/datasets/arrhythmia

[27] Z-Alizadeh Sani Dataset. (2017). [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani

Anuradha. P is a research scholar at the Department of Computer Science at Avinashilingam Institute for Home Science and Higher Education for Women, deemed to be University, Coimbatore, India. She has done the MCA and is KSET and UGC NET qualified. She is a professor and the head of the Department of Computer Science at Indian Academy Degree College (Autonomous), Bangalore, Karnataka, India. Her areas of interest include object-oriented programming, design and analysis of algorithms, database management system, operating system, systems programming, advanced java programming and data science.

Vasantha Kalyani David is a professor and the head of the Department of Computer Science at Avinashilingam Institute for Home Science and Higher Education for Women, deemed to be University, Coimbatore, India. Earlier she is a mathematician with a master of philosophy in mathematics and later did research in computer science.

Dr. Vasantha Kalyani David has published many papers in areas of soft computing. Her research interests, include neural networks, artificial intelligence, fuzzy logic, genetic algorithms, cellular automata, theoretical computer science, and automata theory. She has authored a book on “Pattern Recognition Using Neural and Functional Networks”.