CHAPTER 5

FEATURE SELECTION BY MODIFIEDBOOSTAROOTA FOR HEART DISEASE PREDICTION

This chapter discusses the proposed wrapper feature selection method, ModifiedBoostARoota (MBAR) with CatBoost as base model. The proposed MBAR algorithm's process of efficiently selecting the optimal features and the performance of CatBoost classifier on the selected features of both low and high dimensional heartdatasets are explained in detail. Also, it presents the effectiveness of the MBAR with CatBoost classifier compared to the existing feature selection algorithms on distinct Heart Datasets in terms of different evaluation metrics.

5.1 INTRODUCTION

The medical industry is encountering novel issues such as emerging diseases, expenses, new medicines, rapid decisions, and so on. Because clinical decision-making necessitates the most precise diagnosis, this is an arduous and time-consuming procedure for clinicians. The medical sector can benefit from an automated process that aids in an illness diagnosis. Physical exertion, tobacco use, diabetes, and excessive body fat all contribute to the epidemic of heart disease that plagues modern society. The most common types of heart disease are angina, myocardial infarction, cardiovascular disease, and CAD. Since heart disease is more common in developed nations, it is crucial toperfect early detection techniques.

The emergence of cutting-edge technologies has allowed for the early prediction of cardiac illnesses, which in turn has led to the mitigation of heart disease risk. Many researchers over the past few years have built automated systems to anticipate the onset of cardiac conditions from massive datasets using data mining and ML algorithms. Pre-processing data, extracting features, selecting features, and classifying data were all steps in such systems. Feature selection was critical since it defined which features will be used to predict cardiovascular disease with the highest accuracy and lowest misclassification rate. The high-dimensional dataset makes heart disease prediction more difficult, and the resulting predictions are less accurate. The efficiency of heart disease prediction can be enhanced by lowering the amount of features in the heart disease dataset.

So many algorithms for selecting aspects that are more significant to causing heart disease have been developed, and they can be broadly categorized as either filter, wrapper, or embedding techniques. Wrapper methods, in contrast to filters, select a more precise and computationally expensive subset of features. Wrapper algorithms make advantage of efficient search processes to find candidate subsets, as an exhaustive search of each conceivable feature subset is not practical for high-dimensional datasets. Wrapper-based feature selection techniques can make use of Boosting, a popular algorithm in ensemble classification, to fine-tune the feature space. Also, heuristic search and nature-inspired algorithms selected the best features for disease prediction with faster convergence speeds, but this faster convergence produces premature convergence. In nature-inspired algorithms, processing a large population takes more running time to select the best feature while meeting the necessary objectives. High precision in subset selection is achieved at the tradeoff of higher iteration with this approach.

As a result, this study presents a novel wrapper approach technique called 'ModifiedBoostARoota' (MBAR) for selecting important qualities for the prediction of heart disease with reduced computing cost. MBAR creates a shadow feature for all original features in the dataset by randomly shuffling the feature values. The Feature Importance (FI) scores of all features by the CatB base model are taken and all featuresare ranked based on their FI values. The Feature Score(FS) of each feature is computed as ratio its rank to its FI. The weighted harmonic mean of all features is computed as ratio of sum of all ranks to sum of all FS. Features with insignificant FS as well as with score less than the harmonic mean are eliminated. This process is repeated until the terminating condition is reached. Then, CatB and XGB are modeled on the selected set of features. Experimental outcomes of MBAR compared with contemporary methods of choosing features show that the suggested approach delivers improved prediction performance on a number of heart illness datasets.

5.2 OVERVIEW OF FEATURE SELECTION ALGORITHMS

Feature Selection (FS) seeks to streamline the data collection process by eliminating superfluous characteristics from the dataset. The ML task would be better ifthe dataset contains less number of features and unnecessary features can be eliminated. Selecting the appropriate feature is known as feature selection or variable selection. Some feature selection strategies include filtering, wrapping, and embedding.

Filter Method

Using statistical methods, this technique determines which traits are most closely related to one another before discarding the one with the lowest correlation. No special ML algorithms are required. These methods have less time complexity. When compared to the wrapper method, the filter method is noticeably quicker.

Wrapper Method

The wrapper method involves several combinations of feature subsets. It depends on a specific machine-learning algorithm. The model-specific subset with the highest performance will be chosen. When the model is processing a large number of characteristics, the computation time will increase.

Embedded Method

Both filter and wrapper quality are included. Feature selection occurs during iterative model training, or when the model is being constructed.

5.2.1 Tree-based Feature Importance

Tree-based ensembles' inherent FI scores are useful beginning points for feature selection. This is due in great part to the following factors: First, these models are flexible in that they work with a wide variety of data types and sizes, including numeric, categorical, and missing information, and they are also scale-invariant. Second, it is possible to determine intrinsic significance scores without adding any more time or effort to the training of the model.

In a DT, the importance of a feature is measured by the total value of a node- splitting criterion (such IG or the Gini Index) for which the feature is responsible. The splitting criterion produced by a feature is typically described in ensembles as the sum or average across all trees. The relevance scores of two or more redundant features in an ensemble of trees, like a RF, will be distributed uniformly due to feature subsampling and bootstrapping (Alsahaf et al. 2022).

If such ratings are used as the basis for feature selection in large datasets, it is possible that a feature with many duplicates will be selected by mistake. If significance scores from tree models are to be used effectively in feature selection, these issues must

be resolved. Sample enhancement and reweighting are two methods that can help withthis. Boosting methods improve the performance of underperforming classifiers by iteratively increasing their sample weights depending on the results of prior boosting rounds. Classifiers in subsequent iterations of the algorithm prioritize previously misclassified samples (Alsahaf et al. 2022). Each classifier then casts a vote that determines the ensemble's final verdict. The FI scores (and ranks) produced by the boosted classifier are modified by sample re-weighting. Because of their performance in correctly classifying samples that were misclassified in earlier boosting rounds, several characteristics that initially ranked poorly become top-ranked positions in later boosting rounds.

From this perspective, this research proposes MBAR for feature selection with a CatB classification in predicting heart diseases from high-dimensional datasets.

5.3 PROPOSED METHODOLOGY

This section briefly describes the proposed ModifiedBoostARoota (MBAR) feature selection algorithm and the overall methodology used for heart disease prediction. MBAR is applied to the low-dimensional dataset namely, South African (SA) heart dataset, Cleveland (Clev) heart dataset, Statlog (Stat) heart dataset, and the high- dimensional Heart Datasets namely, Arrhythmia Heart Dataset and Z-Alizadeh Sani (Z- Ali) heart dataset, for selecting the appropriate features for predicting the heart disease.

5.3.1 ModifiedBoostARoota (MBAR) Algorithm for Feature Selection

MBAR is a wrapper based method that encloses a boosting algorithm, as described in Algorithm 5.1. The MBAR algorithm is a modified version of the ‘BoostARoota (BAR)’ algorithm (Chasedehan, 2018). MBAR differs from BoostARoota by utilizing a different base model and using different feature elimination process in the algorithm. Algorithm 5.1 clearly describes the steps involved in the feature selection of Heart Datasets.

Algorithm 5.1 **ModifiedBoostARoota (MBAR)**

1. Create an extended dataset of 'n' features by computing the shadow feature (by randomly rearranging the original features) for each feature in the dataset and combining it with the original dataset.

2. Employing any Tree-Based models, figure out the Feature Importance (FI) of each and every feature in the augmented dataset.
3. All features should be given a rank, r_i is i=1 to n.
4. If the original feature's FI is less than the shadow feature's FI, then both the original and shadow features can be removed.
5. If a feature's FI is negligible, it should be removed.
6. For each feature in the augmented dataset, determine its fscore (fs) using,

$$fs_i = \frac{r_i}{FI_i}, i = 1, \dots, n \quad (5.1)$$

7. Compute weighted harmonic mean as:

$$whm = \frac{\sum r_i}{\sum fs_i}, i = 1, \dots, n \quad (5.2)$$

8. If $fs_i < whm$, then remove feature $'i\ '$ from the enlarged dataset.
9. A feature is superfluous if its fs score is lesser than or equal to the fs score of its shadow. Eliminate a feature if its fs score is negligible (zero).
10. The process loops back to step 1 if 10% or more features are deleted in each iteration, or if the maximum number of iterations has not been reached. Otherwise, return the remaining (chosen)features and end the process.

5.3.2 Overall framework

The methodology used in this stage of the research for effective feature selection and prediction of heart disease is depicted in Figure 5.1. The experiment employs two algorithms existing BoostARoota (BAR) and proposed ModifiedBoostARoota (MBAR): the former uses XGBoost as the base model, while the other uses CatB. Then, the relevant characteristics of each heart dataset are produced by applying both the feature selection techniques to the datasets. That is, MBAR with XGBoost as base model and MBAR with CatB as base model, as well as BAR with XGBoost as base model and BAR with CatB as base model were experimented. Time complexity of the original BAR method is compared to that of the new MBAR algorithm.

The classifiers CatB, XGBoost, and MVE, (described in section 4.1.3), are modelled on the selected subset of features by MBAR. Three rounds of ten-fold stratified cross-validation are used to compare the classifiers' accuracy scores. The top classifier is

chosen based on a comparison of its performance metrics with those of the others. The effectiveness of the classifiers is compared before and after feature selection using the BAR and MBAR algorithms.

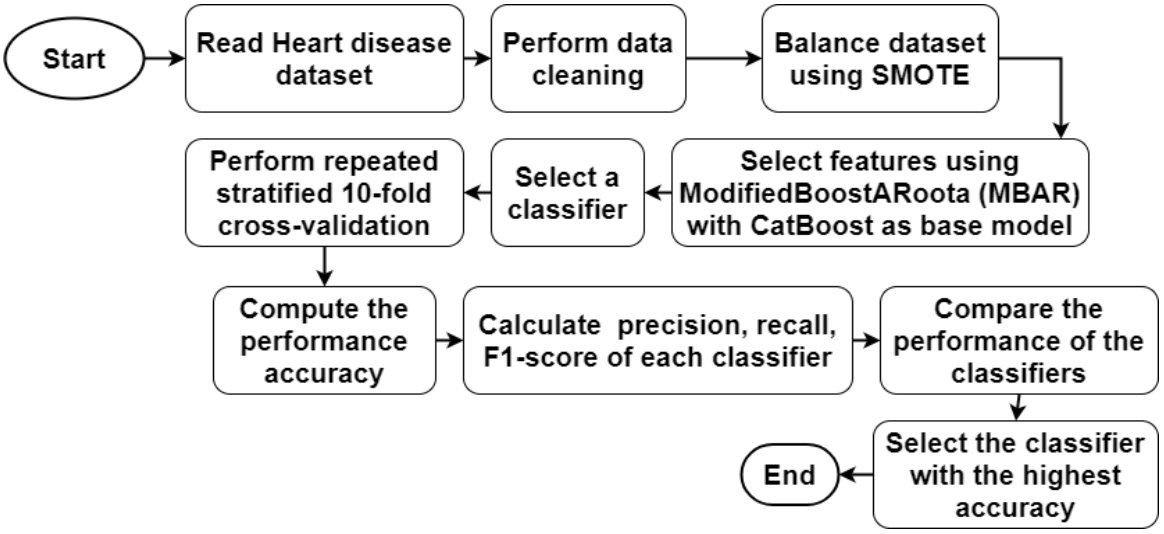

Figure 5.1 Methodology for Feature Selection by MBAR and subsequent Classification of Heart Datasets

5.4 RESULTS AND DISCUSSION

The experiment was executed using Python on Ubuntu platform on a system with 4 GB RAM and i5 processor. The experiments were performed on three different low-dimensional Heart Datasets, and two different high-dimensional Heart Datasets. The details of all the datasets are given in Chapter 3. In addition, Chapter 3 explains the evaluation criteria that were applied to evaluate the proposed algorithms.

5.4.1 Evaluation of MBAR on Low-Dimensional Heart Datasets

Initially, any missing values are inserted into all datasets. Then numeric values are normalized using the Min-Max scaling function mentioned in Eq.5.3.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{5.3}$$

Low dimensional heart datasets are used with the existing feature selection technique, BoostARoota (BAR), with XGBoost as the basis model and BAR with CatB as the base model to choose the pertinent features. Then the proposed feature selection algorithm, MBAR with XGBoost as base model and MBAR with CatB as base model is

used on the heart datasets and the best subset of features of each dataset are obtained. Table 5.1 displays a time complexity comparison between the original BAR algorithm and the new MBAR approach. In Table 5.1, the proposed model MBAR-CB with CatB Classifier (CBC) demonstrates the least prediction time across all the three heart datasets.

Table 5.1 Execution Time Comparison of Existing BAR and Proposed MBAR Algorithms

Datasets	Feature Selection Algorithm & Classifier	Training time (in seconds)	Prediction time (in seconds)
Cleveland Heart Dataset	**MBAR-CB, CBC**	1.736	**0.001**
	BAR-CB, CBC	1.954	0.002
	MBAR-XGB, XGBC	0.034	0.004
	BAR-XGB, XGBC	0.044	0.002
SA Heart Dataset	**MBAR-CB, CBC**	1.331	**0.001**
	BAR-CB, CBC	1.673	0.013
	MBAR-XGB, XGBC	0.051	0.003
	BAR-XGB, XGBC	0.084	0.002
Statlog Heart Dataset	**MBAR-CB, CBC**	2.494	**0.002**
	BAR-CB, CBC	2.626	0.003
	MBAR-XGB, XGBC	0.032	0.003
	BAR-XGB, XGBC	0.036	0.002

CatB Classifier (CBC) is modelled on features selected by MBAR-CatB and BAR-CatB on all three low-dimensional datasets, and XGB Classifier (XGBC) is modelled on features selected by MBAR-XGB and BAR-XGB. Table 5.1 demonstrates that MBAR-XGB combination with XGBoost classifier entails reduced Training time and the MBAR-CB combination with CatB requires shorter prediction time when compared to classifier combinations with BAR technique.

Table 5.2 displays performance metrics of the classifiers CatB and XGBoost when modelled on the features subsets given by BAR and MBAR in combination with base models CatB and XGB. It is evident from Figure 5.2 that the classifiers combination with MBAR outperforms the other combinations considered on all the three heart datasets.

Table 5.2 Performance Metrics of Classifiers after Feature Selection on Heart Datasets

Datasets	Feature Selection Algorithm & Classifier	Recall	Precision	Accuracy
Cleveland Heart Dataset	**MBAR-CB, CBC**	87%	96%	**91.8%**
	BAR-CB, CBC	87%	90%	88.52%
	MBAR-XGB, XGBC	87%	96%	**91.8%**
	BAR-XGB, XGBC	87%	84%	85.25%
SA Heart Dataset	**MBAR-CB, CBC**	48%	65%	**74.19%**
	BAR-CB, CBC	55%	59%	72.04%
	MBAR-XGB, XGBC	61%	63%	**75.27%**
	BAR-XGB, XGBC	58%	55%	69.89%
Statlog Heart Dataset	**MBAR-CB, CBC**	77%	95%	**87.04%**
	BAR-CB, CBC	81%	91%	87.04%
	MBAR-XGB, XGBC	73%	95%	**85.19%**
	BAR-XGB, XGBC	69%	95%	83.33%

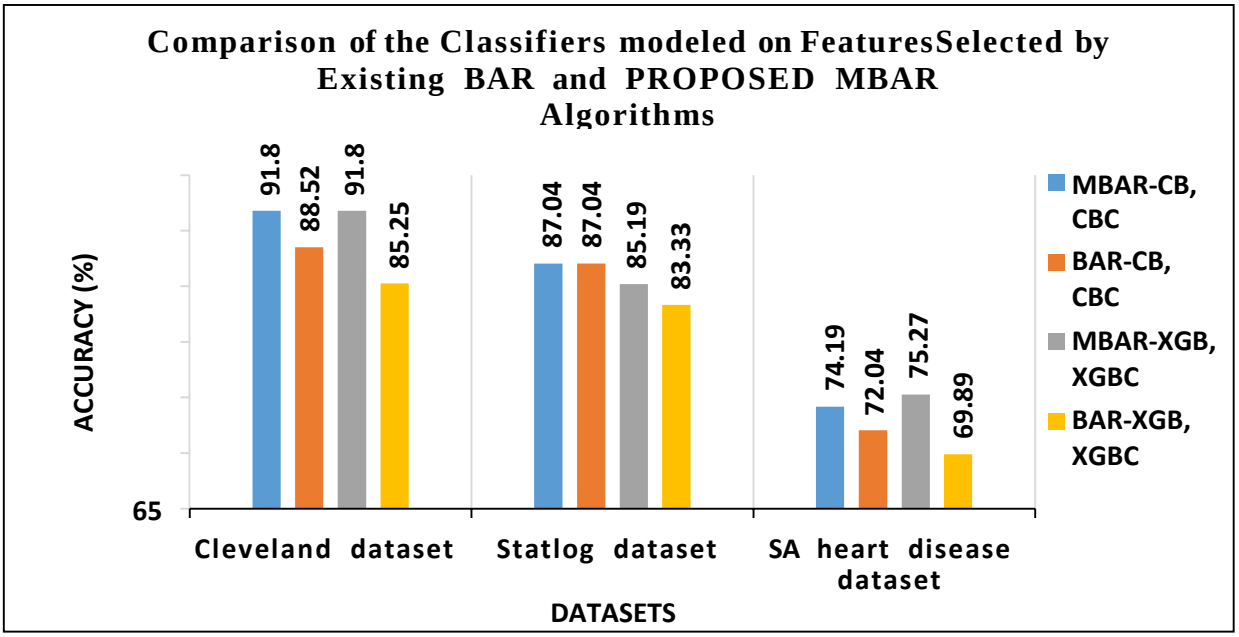

Figure 5.2 Performance of Classifiers modelled on features selected by BAR and MBAR

5.4.2 Evaluation of MBAR on High-Dimensional Heart Datasets

Two high-dimensional datasets, Arrhthmia dataset and Z-Alizadeh Sani heart dataset, are used in the experiments. The missing values in the Arrhythmia dataset were first populated with the mean values (36, 49, 37, -14, and 75) in features c10, c11, c12,and c14. The target variable has class 0 was designated as the normal category, and class 1 represented all disease categories. Class 0 has 245 occurrences and Class 1 has 207 occurrences.

Table 5.3 depicts the total number of features selected by MBAR from the high-dimensional Heart Datasets after balancing with SMOTE.

Table 5.3 Number of MBAR Selected Features of the High-dimensional Heart Datasets

High-dimensional heart datasets	Total number of features in the dataset	Number of selected features by MBAR
Arrhythmia heart dataset	279	64
Z-Alizadeh Sani heart dataset	55	12

Figure 5.3 shows that out of a number of different classifiers used, CatB had the greatest accuracy at 86.33%. In addition to XGBoost and DT Classifier, other classifiers tested included LR, CatB, Extra Trees, KNN, RF, and Support Vector Classifier.

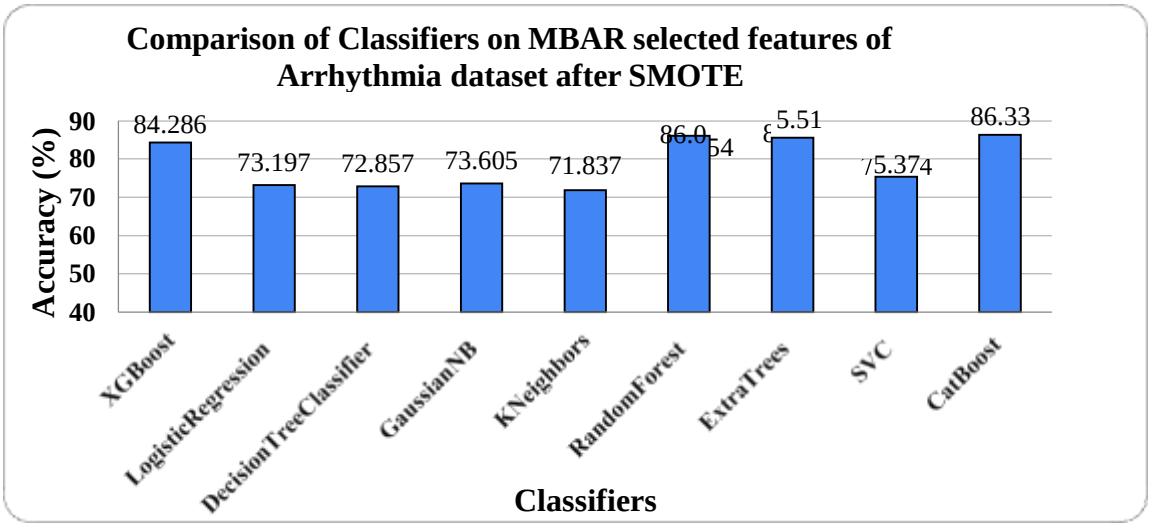

Figure 5.3 Comparison of Classifiers on MBAR Selected Features of Arrhythmia Dataset

When comparing the classifiers accuracy on MBAR selected features of the Arrhythmia dataset, as shown in Figure 5.3, CatB gets the best accuracy, at 86.33%. Results on the Arrhythmia dataset favor Tree-Based models.

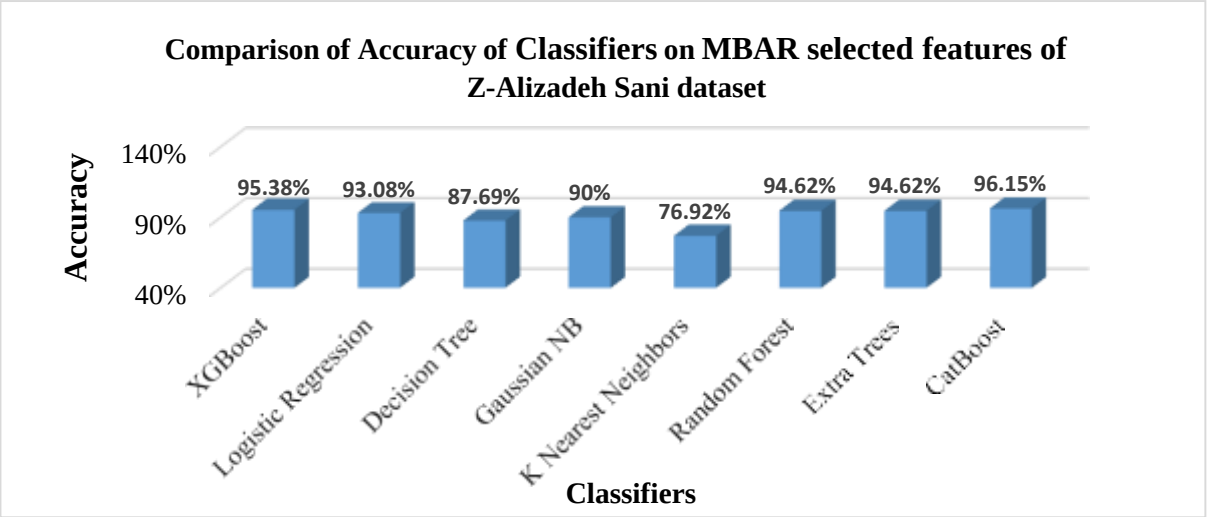

Figure 5.4 Comparison of Classifiers accuracy on MBAR selected features of Z-Alizadeh Sani dataset

In Figure 5.4, on comparing the performance of all classifiers on MBAR selected features of Z-Alizadeh Sani dataset, CatB gives the highest accuracy of 96.15%. This research shows that compared to other classifiers, CatB performs the best.

Table 5.4 Performance Metrics of MBAR with Cat Boost on High-Dimensional Heart Datasets

Methods	Arrhythmia Dataset				Z-Alizadeh Sani Dataset			
	precision	**recall**	**f1-score**	**accuracy**	**precision**	**recall**	**f1-score**	**accuracy**
Without feature selection and with CatBoost classifier	84%	80%	82%	81.63 %	98%	92%	95%	95.38%
With MBAR selected Features and CatBoost classifier	86%	81%	83%	**86.3%**	98%	94%	96%	**96.15%**

Table 5.4 displays the performance metrics for CatB classification using MBAR on the balanced Arrhythmia and Z-Alizadeh Sani dataset . It is clear that MBAR with CatB outperforms classification without feature selection on these high-dimensional heart datasets.

Table 5.5 Comparison of Proposed Model with Related Work on Low-Dimensional Heart Datasets

Heart Disease Dataset	Feature Selection method	Classifier	Accuracy
Cleveland HeartDisease Dataset	Information gain and backward elimination (Khemphila & Boonjing, 2011)	Artificial Neural Networks	92%
	Genetic Algorithm (Gokulnath et al., 2019)	Support Vector Machine	88%
	Average of the FI done by Relief-F, Gain Ratio, Information Gain, and Chi-Square (Kolukisa et al., 2018)	Naive Bayes	86%
	Proposed ModifiedBoostARoota	**CatBoost**	**92%**
Statlog Heart Disease Dataset	F-score (Divya & Agarwal, 2014)	Least Square Twin SVM	86%
	RF (Hera et al., 2022)	Multi-Tier Ensemble (MTE)	84%
	Infogain, Correlation and ReliefF (Zahangir et al., 2019)	RF	83%
	Proposed ModifiedBoostARoota	**CatBoost**	**87%**
SA heart Dataset	Infogain, Correlation and ReliefF (Zahangir et al., 2019)	RF	71%
	MannWhitney-Wilcoxon Test and LR (Babic et al., 2017)	DT	74%
	Average of the FI done by Relief-F, Gain Ratio, Information Gain, and Chi-Square (Kolukisa et al., 2018)	SVM	75%
	Proposed ModifiedBoostARoota	**CatBoost**	**75%**

Table 5.6 Comparison of Proposed Model with Related Work on High-Dimensional Heart Datasets

Heart Disease Dataset	Feature Selection method & Classifier	Accuracy
Arrhythmia Heart Dataset	Spearman Rank correlation, SVM (Khare et al., 2012)	86%
	Principal Component Analysis (PCA),XGBoost (Iyer et al., 2021)	83.2%
	Feature Importance by Random Forest, Multi-Layer Perceptron (MLP) (Mustaqeem et al., 2017)	78%
	Fisher score, Least-squares support-vector machines (LS-SVM) (Yilmaz, 2013)	82%
	Modified Kernel Difference-Weighted KNN (Yang et al., 2020)	73%
	Proposed ModifiedBoostARoota (MBAR) , CatBoost	**86%**
Z-Alizadeh Sani Heart Dataset	Linear Discriminant Analysis, SVM (Kolukisa et al., 2018)	93%
	Artificial Bee Colony, Sequential Minimal Optimization (SMO) (Kilic & Keles, 2018)	89.4%
	Principal Component Analysis (PCA) and t-test, Artificial Neural Networks (ANN) (Cüvitoğlu and Z. Işik, 2018)	85%
	Genetic Algorithm, Neural Networks (Arabasadi et al., 2017)	94%
	Chi square, SVM (Dahal & Gautam, 2020)	89%
	Proposed ModifiedBoostARoota (MBAR) , CatBoost	**96.15%**

The recommended technique is compared to the findings of other researchers' investigations in Table 5.5. The proposed feature selection by MBAR and classificationby CatB exhibits better accuracy compared to other models.

5.5 CHAPTER SUMMARY

The MBAR feature selection approach with the CatB classifier presented in this chapter outperforms the standard BoostARoota (BAR) strategy in terms of prediction time efficiency. The datasets are initially pre-processed by filling missing values and continuous variables are normalized. The datasets are then balanced using SMOTE. On the Cleveland heart dataset, the proposed model MBAR with CatB achieves an impressive 91.8% accuracy, while on the Statlog heart dataset it achieves 87% and on the SA Heart dataset it achieves 74%. On the Arrhythmia dataset, it improves accuracy to 86%, while on the Z-Alizadeh Sani dataset, it improves accuracy to 96%. In addition, the suggested model is compared to others in the literature and is shown to perform well.

Thus, the proposed MBAR with CatB model shows high performance when applied to both the low-dimensional and high-dimensional Heart Datasets. In the next phase, ensemble classifier is experimented along with the feature selection algorithm MBAR to achieve greater performance.