IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 17, 2021, accepted February 25, 2021, date of publication March 4, 2021, date of current version March 16, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3064084*

# Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques

**ABID ISHAQ[1], SAIMA SADIQ[1], MUHAMMAD UMER[1,4], SALEEM ULLAH[1], SEYEDALI MIRJALILI[2,3,5], (Senior Member, IEEE), VAIBHAV RUPAPARA[6], AND MICHELE NAPPI[7], (Senior Member, IEEE)**

[1]Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan
[2]Center for Artificial Intelligence Research and Optimization, Torrens University Australia, Brisbane, QLD 4006, Australia
[3]Yonsei Frontier Lab, Yonsei University, Seoul 03722, South Korea
[4]Department of Computer Science and Information Technology, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan
[5]King Abdulaziz University, Jeddah 21589, Saudi Arabia
[6]School of Computing and Information Sciences, Florida International University, Miami, FL 33199, USA
[7]Department of Computer Science, University of Salerno, 84084 Fisciano, Italy

Corresponding authors: Vaibhav Rupapara (vaibhav.rupapara.sept@gmail.com) and Michele Nappi (mnappi@unisa.it)

This work was supported by the Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan.

**ABSTRACT** Cardiovascular disease is a substantial cause of mortality and morbidity in the world. In clinical data analytics, it is a great challenge to predict heart disease survivor. Data mining transforms huge amounts of raw data generated by the health industry into useful information that can help in making informed decisions. Various studies proved that significant features play a key role in improving performance of machine learning models. This study analyzes the heart failure survivors from the dataset of 299 patients admitted in hospital. The aim is to find significant features and effective data mining techniques that can boost the accuracy of cardiovascular patient's survivor prediction. To predict patient's survival, this study employs nine classification models: Decision Tree (DT), Adaptive boosting classifier (AdaBoost), Logistic Regression (LR), Stochastic Gradient classifier (SGD), Random Forest (RF), Gradient Boosting classifier (GBM), Extra Tree Classifier (ETC), Gaussian Naive Bayes classifier (G-NB) and Support Vector Machine (SVM). The imbalance class problem is handled by Synthetic Minority Oversampling Technique (SMOTE). Furthermore, machine learning models are trained on the highest ranked features selected by RF. The results are compared with those provided by machine learning algorithms using full set of features. Experimental results demonstrate that ETC outperforms other models and achieves 0.9262 accuracy value with SMOTE in prediction of heart patient's survival.

**INDEX TERMS** Data mining, heart disease classification, machine learning, cardiovascular disease, feature selection, SMOTE.

## I. INTRODUCTION

According to WHO, Heart Diseases are a leading cause of death worldwide [1]. It is quite difficult to identify the cardiovascular disease (CVD) because of some contributory factors which contribute to CVD like high blood pressure, cholesterol level, diabetics, abnormal pulse rate, and many other factors [2]. Sometimes CVD symptoms may vary for different genders. For example, a male patient is more likely to have chest pain while a female patient has some other symptoms with chest pain like chest discomfort: such as nausea, extreme fatigue, and shortness of breath [3]. Researchers have been exploring a wide range of techniques to predict heart diseases but the disease prediction at an early stage is not very efficient due to many factors,including but not limited to complexity, execution time, and accuracy of the approach [4]. As such, proper treatment and diagnosis can save many lives [5].

One American dies every 36 seconds due to CVD [6]. More than .665 million people die due to heart disease which 1 in every 4 deaths [7]. Cardiovascular disease costs a lot

The associate editor coordinating the review of this manuscript and approving it for publication was Ramesh Babu N.

to the US healthcare system. In the years 2014 and 2015, it cost about $219 billion per year in terms of healthcare services, medicine, and lost productivity due to death [8]. Early diagnosis can also help to prevent heart failure which can lead to the death of a person. Angiography is considered as the most precise and accurate method for the prediction of cardiac artery disease (CAD) [9], but it is very costly which makes it less accessible to low-income families.

A number of factors such as blood pressure, cholesterol, creatine, etc., affect heart health, so it makes it difficult to diagnose. The authors in [10] analyzed different factors that cause heart disease and identified controllable factors such as alcohol usage, smoking, diabetics, high cholesterol, and limited physical activity. In the modern era, electronic health records (EHRs) are also helpful for clinical and research purposes [11]. The physical examination might have some errors and in the case of heart disease, these minor errors can cost a life in the future. Machine learning-based expert systems effectively diagnose CVD and as a result death ratio is reduced [12].

Data mining plays an immense role in extracting useful information from big data. It is widely used in almost every field of life like medicine, engineering, business, and education. Data mining is used to explore the data to extract the hidden crucial decision making information from the collection of the past repository for future. A variety of machine learning algorithms have been used to understand the complexity and non-linear interaction between different factors by decreasing the error in prediction and factual outcomes [13]. Due to ever increasing medical data, we need to leverage on machine learning algorithms to assist medical healthcare professionals in analyzing data and making accurate and precise diagnostic decisions. In medical data mining, different classification algorithms are used to predict the CVD in patients and death predictions due to the heart attack [14].

Ahmad *et al.* [15] released a dataset consisting of medical records of heart patients having heart failure previously collected at Institute of Cardiology and Allied hospital Faisalabad, Pakistan. Authors predicted mortality rate by applying Cox regression. They also highlighted the patterns of survival using Kaplan-Meier Plots. It is notable that the have made the dataset publicly available for the scientific community. Subsequently, Zahid *et al.* [16] explored the same dataset and proposed two different gender-based models to predict mortality. Afterwards, Chicco and Jurman [17] predicted performance of machine learning using only two features of the same dataset. Even though aforementioned researchers showed interesting results by applying standard statistical techniques, such methods are inefficient for large-scale datasets leaving room for other machine learning algorithms.

This motivated our attempts to help healthcare professionals by developing machine learning techniques in the diagnosis of CVD patients' survival. We employed nine machine learning models: Decision Tree (DT) [18], Adaptive Boosting model (AdaBoost) [19], Logistic Regression (LR) [20], Stochastic Gradient Descent (SGD) [21], Random Forest (RF) [22], Gradient Boosting classifier (GBM) [23], Extra Tree Classifier (ETC) [24], Gaussian Naive Bayes (G-NB) [25] and Support Vector Machine (SVM) [26]. Synthetic Minority Oversampling Technique (SMOTE) is applied to handle class-imbalance problem. This study contributes to the literature in the following areas:

- Designed an effective decision support system that can effectively diagnose the survival of heart failure patients.
- Performance of tree-based, regression-based, and statistical-based models is compared using SMOTE technique in predicting survival of heart patients.
- To investigate the major risk factors, significant features are identified from the dataset that also affect the performance of the machine learning algorithm.

The rest of the paper organised as follows: Section II describes the heart related work that gives a brief description of related literature. Section III describes the dataset, preprocessing and visualisation of data to find the hidden pattern that is present in the dataset. It also describes the different algorithms used in this research. Section IV describes the discussion and analysis of the result. Conclusion and future work is presented in section V.

## II. RELATED WORK

Data mining with the help of machine learning is very useful for solving different kinds of problems. In medical data mining, healthcare data is difficult to be manually handled as it has vast data sources. Advancement in artificial intelligence also inducted precise and accurate systems for the medical application while dealing with sensitive medical data [27]. Heart disease is a leading cause of death even in developed areas [28]. Machine learning models have been widely used in identifying risks at early stages of heart disease. Smoking, age, diabetes and hypertension are considered as risk factors for heart disease [29].

Muthukaruppan and Er [30] proposed a Particle Swarm Optimization (PSO)-based fuzzy expert system for the detection of CVD. Rules were extracted from the decision tree and then converted into fuzzy rules. They have achieved 93.27% accuracy by the fuzzy expert system. In their work, a small number of rules were extracted on the small-size dataset. Alizadehsani *et al.* [31] applied an ensemble-based learning approach. In their research, they used the dataset which was obtained from the Rajaie Cardiovascular Medical and Research Centre and comprises of 303 instances. Authors used the begging C45 ensemble learning approach for CVD prediction. They have achieved 68.96% accuracy for diagnosis of stenosis in the Right Coronary Artery (RCA), 61.46% accuracy in Left Circumflex (LCX), and 79.54% accuracy in Left Anterior Descending (LAD). Another group of researchers improved the results by applying the SVM model and achieved 80.50% accuracy for RCA, 86.14% accuracy for LAD and 83.17% accuracy for LCX [32].

Manogaran *et al.* [33] employed Multiple Kernel Learning (MKL) with Adaptive Neuro-Fuzzy Inference System (ANFIS) for the diagnosis of heart disease using the KEGG

metabolic reaction network dataset and achieved robust results. In [34], Manogaran *et al.* studied different kinds of heart diseases. They proposed an ensemble learning framework of different neural network models and a method of aggregating random under-sampling. To enhance the performance of the classification algorithms they used preprocessing steps with feature selection. They used different kinds of unidirectional and bidirectional neural networks models and the result proved that the ensemble classifiers with BiGRU or BiLSTM with a CNN model outperformed. Tama *et al.* [35] proposed the two-tier ensemble model in which some classifiers are exploited as base classifiers of another ensemble. The proposed stacked architecture is built by blending the class labels prediction of Gradient Boosting Machine (GBM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). Their proposed detection model is evaluated on four different kinds of datasets. They also used particle swarm optimization-based feature selection techniques. Their proposed model performed better with respect to the 10-fold cross-validation. Authors only considered the stacking of tree-based models. Other regression-based and statistical-based could be tested to improve model results.

Melillo *et al.* [36] proposed an automatic classifier for the patients with high risk which separates them from the low-risk patients. In their study classification and regression tree (CART) performed better with 93.3% sensitivity and 63.5% specificity. They analyzed only 12 low-risk patients and 34 high-risk patients. A bigger dataset needs to be explored to test the effectiveness of their proposed approach.

Guidi *et al.* [37] scrutinized the clinical support system (CDSS) for the analysis of heart failure. They used different machine learning classifiers in their research and compared their performance. With 87.6% accuracy, random forest, and CART performed best.

Parthiban and Srivatsa [38] research focused on the patient who had heart issues with diabetes. They used different kinds of predictive features such as blood pressure, blood sugar, and age. They achieved 94.60% accuracy by the SVM classifier. Dataset was imbalanced and the authors did not use any approach to handle this problem. Al Rahhal *et al.* [39] utilized a deep neural network (DNN) model for the classification of ECG signals to study the top set of features.They allowed expert interaction at each iteration during training which can cause biases.

Shah *et al.* [40] proposed a system to study different conditions that can affect the heart and primary factors for the deaths. Different supervised machine learning algorithms were used such as Decision Tree (DT), Naïve Bayes (NB), RF, and KNN. Out of 76 attributes, only 14 attributes were used because the accurate and efficient system with less number of attributes is their research goal. Out of four supervised machine learning classifiers KNN outperformed. Ensemble approaches could be applied to improve the classification results. Mohan *et al.* [41] proposed a hybrid model for heart disease prediction. Authors also proposed a novel feature selection method to improve training of Machine Learning models and achieved 88% accuracy. More feature engineering techniques and machine learning models could be analyzed to improve performance. Geweid and Abdallah [42] designed an optimized and improved SVM model using ECG-signals for heart disease identification. More advanced machine learning models with the combination of signal processing applications need to be explored.

A comprehensive literature survey showed that existing approaches performed well in the prediction of heart disease on different datasets. However, different optimization techniques have been used to improve several measures such as accuracy, precision, and recall. In this research, the main goal is to highlight a comparison of different machine learning techniques to select the most suitable method for heart disease survival prediction. To the best of our knowledge, it is the first attempt to analyze all features of the dataset [15] using machine learning models in predicting heart patient's survival.

**TABLE 1. Dataset specifications.**

| Sr No. | Attributes | Description | Range | Measured In |
|---|---|---|---|---|
| 1 | Time | Followup period | 4-285 | Days |
| 2 | Event (target) | If the patient died in the followup time | 0,1 | Boolean |
| 3 | Gender | Man or woman | 0,1 | Binary |
| 4 | Smoking | If the patient smokes | 0,1 | Boolean |
| 5 | Diabetics | If the patient has diabetics | 0,1 | Boolean |
| 6 | B.P | If the patient has blood pressure issue | 0,1 | Boolean |
| 7 | Anaemia | Decrease in red blood cell or haemoglobin | 0,1 | Boolean |
| 8 | Age | Age of the patient | 40-95 | Years |
| 9 | Ejection fraction | Percentage of blood leaving the heart at each concentration | 14-80 | Percentage |
| 10 | Sodium | Level of sodium in the blood | 114-148 | mEq/L |
| 11 | Creatinine | Level of creatinine in the blood | .50-9.40 | mg/dL |
| 12 | Platelets | Platelets in blood | 25.01-850.00 | kiloplatelets/mL |
| 13 | CPK (creatinine Phospho. . . .) | Level of CPK enzyme in the blood | 23-7861 | Mcg/L |

## III. MATERIALS AND METHODS

### A. DATASET DESCRIPTION

In this research, the Heart-failure-clinical-records-dataset [15] is derived from the UCI machine learning repository [43]. The dataset contains the medical records of 299 patients who had heart problems, collecting during the follow-up period where every patient profile has 13 clinical features. Out of 299 records, 194 are men, and 105 are women. The ages of all the patients are above 40 years. In target class, 1 is for deceased and 0 is for alive. All 299 patients who had left ventricular systolic dysfunction and had heart failure in the past were in the class III or IV of NYHA. The overview of the data set is given in the Table. 1.

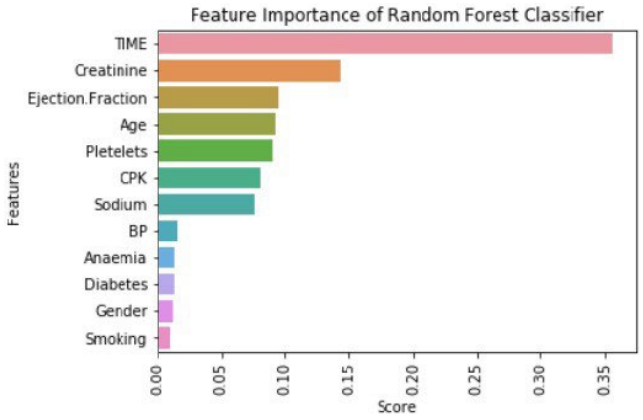


**FIGURE 1.** Important features by RF classifier.

### B. FEATURE IMPORTANCE

Data visualization assist with explaining the hidden patterns that present inside the dataset. It helps to qualitatively get more information about the dataset by visualizing the attributes characteristics. RF was used to employ feature ranking. Figure 1 shows the feature importance predicted by the RF. RF clearly identifies Time, Creatinine, Ejection fraction, Age, Platelets, CPK and Sodium as the most relevant features.

### C. CLASSIFIERS

Classification, a supervised machine learning model is utilized for predicting the result from the data. This work proposes a technique for the prediction of heart disease using classification methods, and to improve the classification accuracy using an ensemble of classifiers. The data has been divided into a training set and a test set, and individual classifiers are trained using a train set. The efficiency of the classifiers is tested with the test data. The details of several machine learning classifiers is discussed in Table 2.

### D. SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE)

SMOTE technique is an oversampling method and has been widely used to deal with class imbalanced data in medicine [44]. SMOTE increases the number of data instances by generating random synthetic data of minority class from its nearest neighbours using Euclidean distance. New instances become similar to the original data because they are generated on the basis of original features [45]. SMOTE is not the best option in dealing with high-dimensional data as it can create additional noise. In this study, new training dataset is generated using SMOTE technique. SMOTE increased data samples from 97 instances to 300 instances for each class.

### E. EVALUATION MATRICES

There are some performance evaluation methods for the machine learning models. The blend of different evaluation tools is expected to endorse the development of analytical research [46]. In this research, four basic

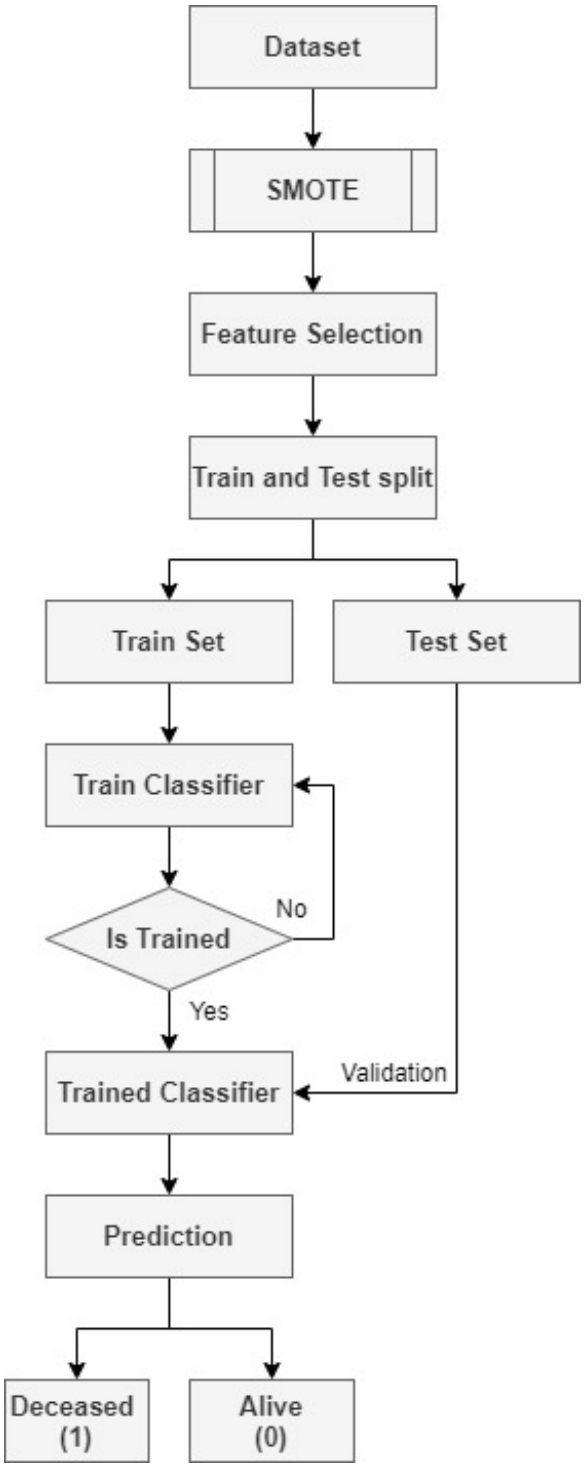


**FIGURE 2.** Flowchart of proposed heart failure patient's survival prediction framework.

metrics (accuracy, precision, recall, F-Score) will be examined for the difference in machine learning-based algorithms.

Confusion matrix [47] helps us to calculate all four metrics. The elements of the confusion matrix are true positive (TP), true negative (TN), false positive (FP) and false negative (FN). If data is related to the medical false negative is the

**TABLE 2.** Machine learning models.

| Reference | Model | Description |
|---|---|---|
| [18] | DT | Decision Tree is a classification algorithm that works well on both forms of data i.e., categorical and numerical. Generally speaking, a decision tree is utilized for the creation of tree-like structures. A decision tree is simple and easy to implement so it is widely used for medical data analysis. The decision tree has three nodes. (1) Root node (it is the main node, functions of other nodes based on it) (2) Interior node (it handles various types of attributes) (3) Leaf node (it is also called as end node, it is the final node which represent the results of each test) |
| [19] | AdaBoost | AdaBoost is the abbreviation of adaptive boosting. AdaBoost is normally used in conjunction with the other algorithms to enhance their performance. To train weak learners into strong learners it works on boosting. Every tree in the AdaBoost classifier is dependent on the outcome error rate of the last built tree. |
| [20] | LR | Logistic regression usually deals with the classification problems. It is a predictive analysis algorithm and statistical model and based on the concept of probability. It is usually used to analyze binary data in which one or more variables work to find output. It produces the connection between the categorical dependent variable and one or more than one independent variable by approximation probability by utilising a logistic regression sigmoid function. |
| [21] | SGD | Stochastic Gradient Descent combines multiple binary classifiers in the one-versus-all method. SGD has been widely used for large dataset because in each iteration it uses all the samples. The working principle is quite similar to the regression technique so it is quite easy to implement and understand. Hyper parameters of SGD need to be correctly valued to obtain accurate results. In terms of feature scaling sensitivity of SGD is high. |
| [22] | RF | Random forest is a tree based ensemble learning model, which produces accurate predictions by combining many weak learners. The bagging technique is used by this model to train a variety of decision trees using various bootstrap samples. In random forest, a bootstrap sample is derived by subsampling the training dataset with replacement, where the size of the sample is same as that of training data set |
| [23] | GBM | In a gradient boosting machine many weak classifiers work together to create a strong learning model. GBM usually works on the principle of decision tree. In this it creates every independent tree so; it is a costly and time-consuming choice. As it is clear from its name it improves the weak learning algorithms after a series of tweaks to it that increases the strength of the algorithm. This method of improving the strength of the learning algorithm is termed as PAC (probability approximately correct learning). Due to this quality it works well on the un-processed data. It handles the data missing values efficiently. |
| [24] | ETC | Extra trees classifier working is quite similar to the random forest and only different from it in method of construction of trees in the forest. Every decision tree in the extra tree classifier is made from the original training sample. Random samples of k best feature are utilised for decision and Gini index is used to select the top feature to split the data in the tree. These random samples of feature indication to the creation of multiple de-correlated decision trees. |
| [25] | G-NB | Gaussian Naive Bayes is a variant of Naive Bayes and worked on Gaussian distributions and utilised for the continuous data. G-NB involves prior and posterior probability of the classes in the data. It is used with the features having continuous values. It is also supposed that all the features are following a Gaussian or normal distribution. |
| [26] | SVM | Support Vector Machine is a supervised learning technique and based on mathematical models. It is applied for regression and classification problems. It performs classification by constructing high dimensional hyper-planes also called decision planes. Hyper-plane distinguish one type of data from another type. |

**TABLE 3.** Performance evaluation measures.

| | |
|---|---|
| **Accuracy =** | $\frac{\text{Number of correctly classified predictions}}{\text{Total predictions}}$ |
| **Precision =** | $\frac{TP}{TP+FP}$ |
| **Recall =** | $\frac{TP}{TP+FN}$ |
| **F-Score =** | $2* \frac{precision.recall}{precision+recall}$ |

most critical prediction. The measures of the performance are given in Table 3.

## IV. ANALYSIS AND DISCUSSION OF RESULTS

In this section, experimental design and results of all experiments for heart patients' survival prediction are discussed. Firstly, we present the results with full set of features followed by the results with the significant set of features. The dataset, containing 13 features about body features, clinical features and lifestyle features. Some of these features are binary such as anaemia, diabetes, blood pressure, smoking and Gender. Death event feature is taken as a target class in binary classification task which tells if a patient is survived or died before 130 days of follow up period. Specification of the dataset is presented in Table 1. SMOTE is applied to make dataset balanced. Machine learning models have been trained on the balanced dataset and evaluated on accuracy, precision, recall and F-Score. Flowchart of the proposed methodology is presented in Figure 2.

### A. EXPERIMENTAL DESIGN

Supervised machine learning models have been conducted in order to analyze the performance of the models. Data has been split into the train set and test set as 70:30 ratio. This ratio is practiced in several literature's for classification tasks and help to avoid overfitting [48]. Performances of the machine learning classifiers are tested using different performance evaluation metrics. All the experiments have been conducted in a python environment using different libraries on an 2 *GB* Dell PowerEdge $T430$ graphical processing unit on $2x$ Intel Xeon 8 Cores running at $2.4Ghz$ machine which is equipped with 32 GB DDR4 Random Access Memory (RAM).

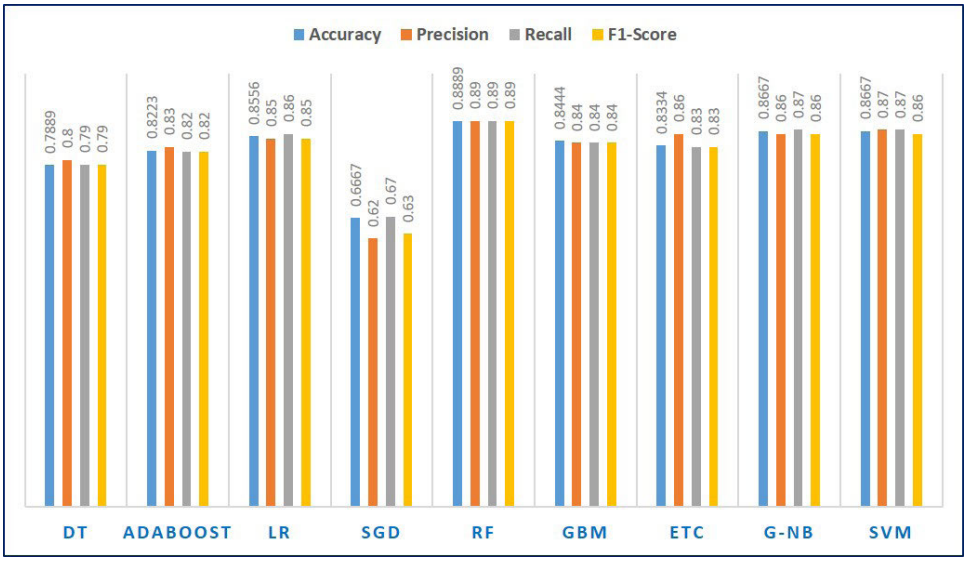


**FIGURE 3.** Performance of classifiers on full set of features.

### B. EXPERIMENTAL RESULTS ON FULL SET OF FEATURES

A comparative analysis of supervised machine learning classifiers has been performed on full set of features of heart-failure-clinical-record-dataset. Some classifiers showed good results on evaluation metrics while some showed poor performance. This work has applied tree-based, regression based, and statistical-based models for the prediction of heart failure survival. Tree-based ensemble models include DT, RF and ETC. Tree-based boosting models AdaBoost and GBM. Regression-based include LR and SGD. and Statistical-based include G-NB and SVM. Table 4 presents the performance evaluation of machine learning models on full set of features. As per the results in Table 4, the LR classifier has achieved good results with 0.8556 accuracy, 0.85 precision, 0.86 recall and 0.85 F-Score. SVM and G-NB were second good classifiers with 0.8667 accuracy and 0.86 F-Score. RF, a tree-based classifier, outperformed among all nine classifiers using all features and obtained 0.8889 accuracy 0.89 value for precision, recall and F-Score.

The worst classifier was SGD for heart failure survival prediction with 0.6667 accuracy, 0.62 precision, 0.67 recall and 0.63 F-Score. Performance comparison of all models is presented in Figure 3.

### C. EXPERIMENTAL RESULTS WITH SMOTE

SMOTE is a powerful solution to the class imbalance problem and have shown robust results in various domains. SMOTE algorithm adds synthetic data to the minority class to make a balanced dataset. Table 5 shows the result of machine learning classifiers using SMOTE technique on all 13 features of heart-failure-record dataset. From Table 5, it is clear that performance of tree-based classifiers significantly improve with the SMOTE in all evaluation matrices. Performance of DT improved from 0.79 accuracy to 0.8278

**TABLE 4.** Classification result of all machine learning models using all features without SMOTE.

| Models | Accuracy | Precision | Recall | F-Score |
|---|---|---|---|---|
| DT | 0.7889 | 0.80 | 0.79 | 0.79 |
| AdaBoost | 0.8223 | 0.83 | 0.82 | 0.82 |
| LR | 0.8556 | 0.85 | 0.86 | 0.85 |
| SGD | 0.6667 | 0.62 | 0.67 | 0.63 |
| RF | 0.8889 | 0.89 | 0.89 | 0.89 |
| GBM | 0.8444 | 0.84 | 0.84 | 0.84 |
| ETC | 0.8334 | 0.83 | 0.83 | 0.83 |
| GNB | 0.8667 | 0.86 | 0.87 | 0.86 |
| SVM | 0.8667 | 0.87 | 0.87 | 0.86 |

accuracy with SMOTE. AdaBoost showed good performance and obtained 0.8852 accuracy, 0.89 precision, 0.89 recall and 0.89 F-Score with balanced dataset. Similarly RF achieved 0.9180 with accuracy and 0.82 F-Score and improved results with SMOTE. ETC with full features and SMOTE showed 10% improvement in results as compared to those results achieved without applying SMOTE. ETC achieved highest results with 0.9262 accuracy, 0.93 precision, 0.93 recall and 0.93 F-Score. Boosting algorithm build trees by reducing errors from previously built weak learners. Up-sampling of the similar data does not show any impact in improving of results [49]. That is the reason that GBM did not show any improvement with SMOTE. Performance evaluation of machine learning models with SMOTE has been shown in Figure 4.

It can be clearly observed that performance of regression-based (LR and SGD) models and statistical-based (G-NB and SVM) models have been decreased with SMOTE. SMOTE performed well with tree-based classifiers for the prediction of heart patient's survival. Accuracy comparison of the classifiers with SMOTE and without SMOTE is presented in Figure 5.

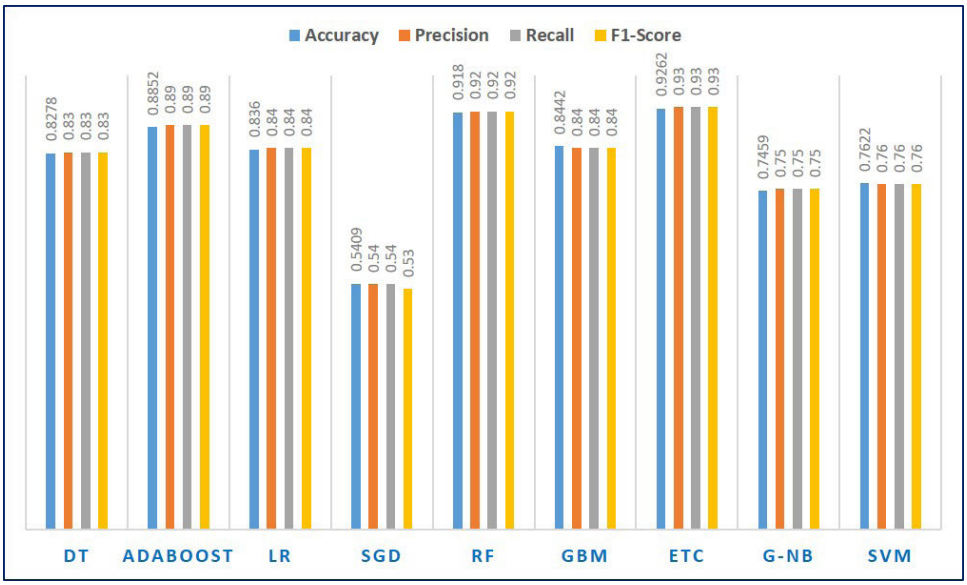


**FIGURE 4.** Performance of classifiers on full set of features with SMOTE.

**TABLE 5.** Classification results of all machine learning models using all features with SMOTE.

| Models | Accuracy | Precision | Recall | F-Score |
|---|---|---|---|---|
| DT | 0.8278 | 0.83 | 0.83 | 0.83 |
| AdaBoost | 0.8852 | 0.89 | 0.89 | 0.89 |
| LR | 0.8360 | 0.84 | 0.84 | 0.85 |
| SGD | 0.5409 | 0.54 | 0.54 | 0.53 |
| RF | 0.9180 | 0.92 | 0.92 | 0.92 |
| GBM | 0.8442 | 0.84 | 0.84 | 0.84 |
| ETC | 0.9262 | 0.93 | 0.93 | 0.93 |
| GNB | 0.7459 | 0.75 | 0.75 | 0.75 |
| SVM | 0.7622 | 0.76 | 0.76 | 0.76 |

**TABLE 6.** Accuracy of all machine learning models using nine significant features with SMOTE.

| Models | Accuracy |
|---|---|
| DT | 0.8778 |
| AdaBoost | 0.8852 |
| LR | 0.8442 |
| SGD | 0.5491 |
| RF | 0.9188 |
| GBM | 0.8852 |
| ETC | 0.9262 |
| GNB | 0.7540 |
| SVM | 0.7622 |

### *D. EXPERIMENTAL RESULTS ON IMPORTANT FEATURES SELECTED BY RF*

In this experiment important features selected by RF are investigated by machine learning classifiers with SMOTE technique. First classifiers were trained and tested by removing least important features identified by RF. Performance of the classifiers were pretty good by removing last four features that are: Anaemia, Diabetes, Gender, and Smoking. Further removal of features started a decrease in performance. Accuracy results by removing four least important features are presented in Table 6. The accuracy result of LR showed 1% improvement with 9 significant features and achieved 0.8442 accuracy. GBM showed significant improvement using 9 significant features and showed 4% improvement in accuracy result and achieved 0.8852 accuracy. Performance comparison of full-set of features with 9 significant features identified by RF is presented in Figure 6.

### *E. IMPACT OF SIGNIFICANT FEATURES*

Experimental results demonstrated that supervised machine learning models can efficiently predict heart failure patients survival. Tree-based algorithms showed good performance on imbalanced and balanced dataset with SMOTE technique. RF clearly identified Time, Creatinine, Ejection fraction, Age, Platelets, CPK and Sodium as significant features as shown in Figure 1. Results showed that tree-based algorithms outperformed using nine features identified by RF using SMOTE technique. This aspect is useful in patient care as doctors can predict patient's survival by just analyzing nine significant features.

ETC outperformed other models with 0.9262 accuracy using nine significant features with SMOTE technique for the prediction of heart patient's survival. ETC selects a random subset of features like RF for node splitting. It is different from RF in a way that it makes trees from complete data samples by selecting cut points for nodes randomly. While RF selects the cut point for the node using local samples. Trees of training set labels can be made independent by setting the value of k as 1 [50]. RF produces constant approximation whereas ETC produces multi-linear approximation. Such additional randomization in the ensemble smoothed the decision boundaries and also a reason for better performance

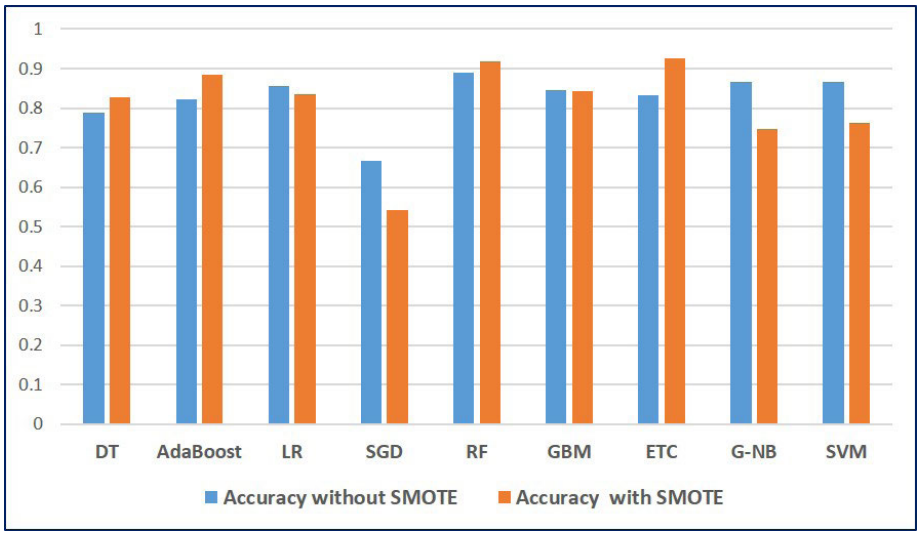


**FIGURE 5.** Accuracy comparison of classifiers with SMOTE and without SMOTE.

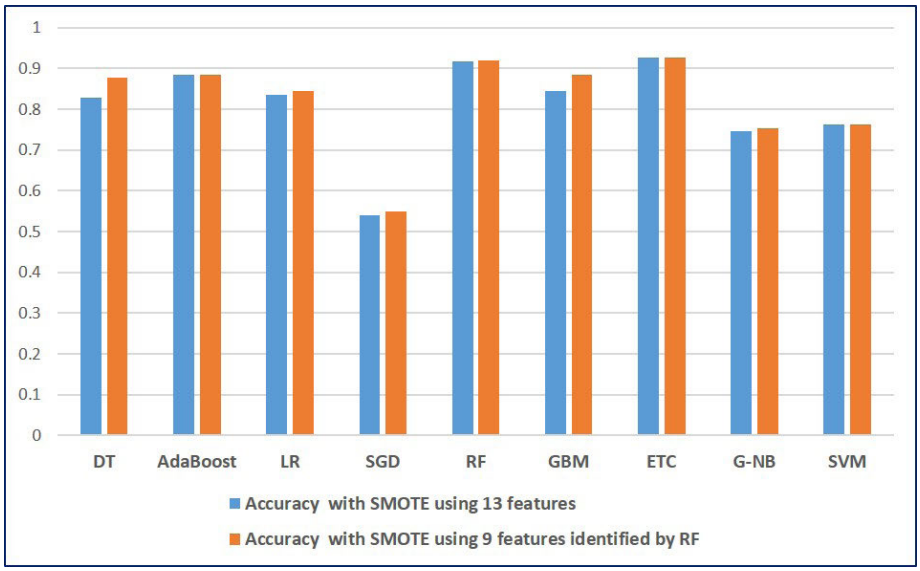


**FIGURE 6.** Accuracy comparison of classifiers with SMOTE using full-set of features and 9 features identified by RF.

than RF. [51] also showed that ET outperformed RF in terms of accuracy.

## V. CONCLUSION

Processing raw health data of heart information using machine learning algorithms will help in saving the lives of heart patients. By analyzing factors contributing to heart failure, mortality rate can be controlled by adopting preventive measures. In this study, an effective and efficient machine learning based technique is suggested for the prediction of heart patients' survival. Machine learning techniques include LR, AdaBoost,RF, GBM, G-NB and SVM. SMOTE is applied to deal class imbalance problem. Furthermore, RF employed feature ranking. According to RF, most significant features are: Time, Creatinine, Ejection fraction, Age, Platelets, CPK and Sodium. Performance of machine learning models are compared on a full set of features and selected features from Heart-failure-clinical-records-dataset. Thus experimental results proved that tree-based with feature selection are highly effective in achieving highest accuracy. SMOTE technique significantly improved performance of tree-based classifiers in predicting heart patient's survival. ETC with SMOTE showed highest result in all evaluation measures and achieved 0.9262 accuracy, 0.93 precision, 0.93 recall and 0.93 F-Score.

This work has the potential to improve the health care system, and become a useful tool for health care providers in predicting survival of heart failure. It will also help physicians in understanding that if a patient of heart failure will survive, they can focus on major risk factors. The future work of this research can be performed with multiple combinations of machine learning models to benefit from their advantages

combined. To improve the performance of machine learning models, better feature selection techniques can be devised. In this case, meta-heuristics can be used due to NP-hard nature of feature selection problems.

## REFERENCES

[1] WHO. *The Top 10 Causes of Death*. Accessed: Dec. 30, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] C. Fryar, T.-C. Chen, and X. Li, "Prevalence of uncontrolled risk factors for cardiovascular disease: United states, 1999-2010," in *NCHS Data Brief*, vol. 103. Aug. 2012, pp. 1–8.

[3] Medical Professionals. *Cardiovascular Diseases*. Accessed: Dec. 29, 2020. [Online]. Available: https://www.mayoclinic.org/medical-professionals/cardiovascular-diseases

[4] L. A. Allen, L. W. Stevenson, K. L. Grady, N. E. Goldstein, D. D. Matlock, R. M. Arnold, N. R. Cook, G. M. Felker, G. S. Francis, P. J. Hauptman, and E. P. Havranek, "Decision making in advanced heart failure: A scientific statement from the American heart association," *Circulation*, vol. 125, p. E587, Apr. 2012.

[5] Q. K. Al-Shayea, "Artificial neural networks in medical diagnosis," *Int. J. Comput. Sci.*, vol. 8, no. 2, pp. 150–154, 2011.

[6] Centers for Disease Control and Prevention. *Underlying Cause of Death 1999-2019*. Accessed: Dec. 28, 2020. [Online]. Available: https://wonder.cdc.gov/wonder/help/ucd.html

[7] S. S. Virani, A. Alonso, E. J. Benjamin, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, A. R. Chang, S. Cheng, F. N. Delling, and L. Djousse, "Heart disease and stroke statistics—2020 update: A report from the American heart association," *Circulation*, vol. 141, pp. E139–E596, Mar. 2020.

[8] E. J. Benjamin, P. Muntner, A. Alonso, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, A. R. Chang, S. Cheng, S. R. Das, and F. N. Delling, "Heart disease and stroke statistics—2019 update: A report from the American heart association," *Circulation*, vol. 139, pp. e56–e528, Mar. 2019.

[9] 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, Apr. 2017.

[10] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7675–7680, May 2009.

[11] B. Chapman, A. D. DeVore, R. J. Mentz, and M. Metra, "Clinical profiles in acute heart failure: An urgent need for a new approach," *ESC Heart Failure*, vol. 6, no. 3, pp. 464–474, Jun. 2019.

[12] S. I. Ansarullah and P. Kumar, "A systematic literature review on cardiovascular disorder identification using knowledge mining and machine learning method," *Int. J. Recent Technol. Eng.*, vol. 7, no. 6S, pp. 1009–1015, 2019.

[13] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" *PLoS ONE*, vol. 12, no. 4, 2017, Art. no. e0174944.

[14] M. M. A. Mary, "Heart disease prediction using machine learning techniques: A survey," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 8, no. 10, pp. 441–447, Oct. 2020.

[15] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, "Survival analysis of heart failure patients: A case study," *PLoS ONE*, vol. 12, no. 7, Jul. 2017, Art. no. e0181001.

[16] F. M. Zahid, S. Ramzan, S. Faisal, and I. Hussain, "Gender based survival prediction models for heart failure patients: A case study in Pakistan," *PLoS ONE*, vol. 14, no. 2, Feb. 2019, Art. no. e0210602.

[17] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, p. 16, Dec. 2020.

[18] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees* (Statistics/Probability Series). 1984.

[19] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *J.-Jpn. Soc. Artif. Intell.*, vol. 14, nos. 771–780, p. 1612, 1999.

[20] C. R. Boyd, M. A. Tolson, and W. S. Copes, "Evaluating trauma care: The TRISS method," *J. Trauma, Injury, Infection, Crit. Care*, vol. 27, no. 4, pp. 370–378, Apr. 1987.

[21] W. A. Gardner, "Learning characteristics of stochastic-gradient-descent algorithms: A general study, analysis, and critique," *Signal Process.*, vol. 6, no. 2, pp. 113–133, Apr. 1984.

[22] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[23] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, pp. 1189–1232, Oct. 2001.

[24] A. Sharaff and H. Gupta, "Extra-tree classifier with metaheuristics approach for email classification," in *Proc. Adv. Comput. Commun. Comput. Sci.* Singapore: Springer, 2019, pp. 189–197.

[25] A. Pérez, P. Larrañaga, and I. Inza, "Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes," *Int. J. Approx. Reasoning*, vol. 43, no. 1, pp. 1–25, Sep. 2006.

[26] B. Schölkopf, C. Burges, and V. Vapnik, "Incorporating invariances in support vector learning machines," in *Proc. Int. Conf. Artif. Neural Netw.* Berlin, Germany: Springer, 1996, , pp. 47–52.

[27] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine," *Database*, vol. 2020, Jan. 2020.

[28] R. Gupta, "Recent trends in coronary heart disease epidemiology in India," *Indian heart J.*, vol. 60, no. 2, pp. B4–B18, 2008.

[29] X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang, and Q. Wang, "A hybrid classification system for heart disease diagnosis based on the RFRS method," *Comput. Math. Methods Med.*, vol. 2017, pp. 1–11, Jan. 2017.

[30] S. Muthukaruppan and M. J. Er, "A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease," *Expert Syst. Appl.*, vol. 39, no. 14, pp. 11657–11665, Oct. 2012.

[31] Z. Sani, R. Alizadehsani, J. Habibi, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, F. Khozeimeh, and F. Alizadeh-Sani, "Diagnosing coronary artery disease via data mining algorithms by considering laboratory and echocardiography features," *Res. Cardiovascular Med.*, vol. 2, no. 3, p. 133, 2013.

[32] R. Alizadehsani, M. H. Zangooei, M. J. Hosseini, J. Habibi, A. Khosravi, M. Roshanzamir, F. Khozeimeh, N. Sarrafzadegan, and S. Nahavandi, "Coronary artery disease detection using computational intelligence methods," *Knowl.-Based Syst.*, vol. 109, pp. 187–197, Oct. 2016.

[33] G. Manogaran, R. Varatharajan, and M. K. Priyan, "Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system," *Multimedia Tools Appl.*, vol. 77, no. 4, pp. 4379–4399, Feb. 2018.

[34] A. Baccouche, B. Garcia-Zapirain, C. C. Olea, and A. Elmaghraby, "Ensemble deep learning models for heart disease classification: A case study from Mexico," *Information*, vol. 11, no. 4, p. 207, Apr. 2020.

[35] B. A. Tama, S. Im, and S. Lee, "Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble," *BioMed Res. Int.*, vol. 2020, Apr. 2020, Art. no. 9816142.

[36] P. Melillo, N. De Luca, M. Bracale, and L. Pecchia, "Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 3, pp. 727–733, May 2013.

[37] G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, "A machine learning system to improve heart failure patient assistance," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 6, pp. 1750–1756, Nov. 2014.

[38] G. Parthiban and S. K. Srivatsa, "Applying machine learning methods in diagnosing heart disease for diabetic patients," *Int. J. Appl. Inf. Syst.*, vol. 3, no. 7, pp. 25–30, Aug. 2012.

[39] M. M. A. Rahhal, Y. Bazi, H. AlHichri, N. Alajlan, F. Melgani, and R. R. Yager, "Deep learning approach for active classification of electrocardiogram signals," *Inf. Sci.*, vol. 345, pp. 340–354, Jun. 2016.

[40] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," *Social Netw. Comput. Sci.*, vol. 1, no. 6, pp. 1–6, Nov. 2020.

[41] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[42] G. G. N. Geweid and M. A. Abdallah, "A new automatic identification method of heart failure using improved support vector machine based on duality optimization technique," *IEEE Access*, vol. 7, pp. 149595–149611, 2019.

[43] A. Asuncion and D. Newman, "UCI machine learning repository," Tech. Rep., 2007.

[44] R. Blagus and L. Lusa, "Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models," *BMC Bioinf.*, vol. 16, no. 1, pp. 1–10, Dec. 2015.
[45] N. V. Chawla, "Data mining for imbalanced datasets: An overview," in *Data Mining and Knowledge Discovery Handbook*. Springer, 2009, pp. 875–886.
[46] T.-S. Lim, W.-Y. Loh, and Y.-S. Shih, "A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms," *Mach. Learn.*, vol. 40, no. 3, pp. 203–228, 2000.
[47] A. M. Hay, "The derivation of global estimates from a confusion matrix," *Int. J. Remote Sens.*, vol. 9, no. 8, pp. 1395–1398, Aug. 1988.
[48] A. Yousaf, M. Umer, S. Sadiq, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Emotion recognition by textual tweets classification using voting classifier (LR-SGD)," *IEEE Access*, vol. 9, pp. 6286–6295, 2021.
[49] A. T. Kalai and R. A. Servedio, "Boosting in the presence of noise," *J. Comput. Syst. Sci.*, vol. 71, no. 3, pp. 266–290, Oct. 2005.
[50] P. Geurts and G. Louppe, "Learning to rank with extremely randomized trees," *J. Mach. Learn. Res.*, vol. 14, pp. 49–61, Jan. 2011.
[51] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006.

**ABID ISHAQ** received the M.Sc. degree in telecommunication from the Government College University, Lahore, in 2009. He is currently pursuing the Master of Computer Science degree with KFUEIT. His research interests include data mining, mainly working on AI and deep learning-based text mining, and data science management technologies.

**SAIMA SADIQ** is currently pursuing the Ph.D. degree in computer science with the Khwaja Fareed University of Engineering and IT (KFUEIT). She is also working as an Assistant Professor with the Department of Computer Science, Government Degree College for Women. Her recent research interests include data mining, machine learning, and deep learning-based text mining.

**MUHAMMAD UMER** received the B.S. degree from the Department of Computer Science, Khwaja Fareed University of Engineering and IT (KFUEIT), Pakistan, in October 2018, where he is currently pursuing the Ph.D. degree in computer science. He is also working as a Research Assistant with the Fareed Computing and Research Center, KFUEIT. His research interests include data mining, mainly working machine learning and deep learning-based IoT, text mining, and computer vision tasks.

**SALEEM ULLAH** was born in Ahmedpur East, Pakistan, in 1983. He received the B.Sc. degree in computer science from The Islamia University Bahawalpur, Pakistan, in 2003, the M.I.T. degree in computer science from Bahauddin Zakariya University, Multan, in 2005, and the Ph.D. degree from Chongqing University, China, in 2012. From 2006 to 2009, he worked as a Network/IT Administrator in different companies. From August 2012 to February 2016, he worked as an Assistant Professor with The Islamia University Bahawalpur. Since February 2016, he has been working as an Associate Dean of the Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan. He has almost 14 years of industry experience in field of IT. He is an Active Researcher in the field of adhoc networks, IoT, congestion control, data science, and network security.

**SEYEDALI MIRJALILI** (Senior Member, IEEE) is currently an Associate Professor and the Director of the Centre for Artificial Intelligence Research and Optimization, Torrens University Australia. He is internationally recognized for his advances in swarm intelligence and optimization, including the first set of algorithms from a synthetic intelligence standpoint—a radical departure from how natural systems are typically understood—and a systematic design framework to reliably benchmark, evaluate, and propose computationally cheap robust optimization algorithms. He has published over 200 publications with over 25 000 citations and an H-index of over 55. As the most cited researcher in robust optimization, he is in the list of 1% highly-cited researchers and named as one of the most influential researchers in the world by Web of Science in Computer Science and Engineering. He is an Associate Editor of several journals, including *Neurocomputing*, *Applied Soft Computing*, *Advances in Engineering Software*, *Applied Intelligence*, and IEEE Access.

**VAIBHAV RUPAPARA** received the Master of Science degree in computer science from Florida International University, Miami, FL, USA. He has worked on different domain, including finance and healthcare. His expertise contributed towards achieving high quality, scalable deliverability with security. His research interests include machine learning, AI, and deep learning.

**MICHELE NAPPI** (Senior Member, IEEE) received the laurea degree *(cum laude)* in computer science from the University of Salerno, Italy, in 1991, the M.Sc. degree in information and communication technology from I.I.A.S.S. E.R. Caianiello, in 1997, and the Ph.D. degree in applied mathematics and computer science from the University of Padova, Italy, in 1997. He was one of the founders of the spin off BS3 (biometric system for security and safety), in 2014. He is currently a Full Professor of computer science with the University of Salerno. He is a Team Leader of the Biometric and Image Processing Laboratory (BIPLAB). He is the author of more than 180 papers in peer-reviewed international journals, international conferences, and book chapters. His research interests include pattern recognition, image processing, image compression and indexing, multimedia databases and biometrics, human–computer interaction, and VR/AR. He is also member of TPC of international conferences. He is a GIRPR/IAPR Member. He received several international awards for scientific and research activities. He is the co-editor of several international books. He serves as an associate editor and a managing guest editor for several international journals. He has been the President of the Italian Chapter of the IEEE Biometrics Council.