A REAL TIME HEART FAILURE PREDICTION TECHNIQUE USING DATA MINING AND SMOTE METHOD

Y. Sri Lalitha
Department of IT, Gokaraju Rangaraju Institute of Engineering and Technology
srilalitham.y@gmail.com

N. V. Ganapathi Raju
Department of IT, Gokaraju Rangaraju Institute of Engineering and Technology.

Abstract:
The significant global mortality & morbidity are attributable to the cardiovascular-CVD disease. Predicting who will or will not survive cardiac illness is an enormous issue in clinical data analytics. The health business generates vast volumes of raw data, and data mining could take this data and turn it into insights that help guide decision-making. Numerous investigations demonstrated that crucial elements help enhance ML model performance. Its research looks at the 299 hospitalized individuals who survived cardiac failure. The goal is to improve the accuracy of survivor prediction for cardiovascular patients by identifying important variables & effective data mining methods. This research uses nine classification models to predict patient survival: DT, Adaptive boosting classifier (AdaBoost), LR, SGD, RF, GBM, ETC, G-NB, & SVM. The synthetic Minority Oversampling Method addresses the issue of class imbalance (SMOTE).The characteristics with the highest rankings chosen by RF are then used to train ML models. Machine learning-ML methods that use a more comprehensive set of features are also used for comparison. According to the data obtained from the experiments, ETC is the most effective model for predicting the survival of heart patients, with an accuracy value of 0.9262 compared to SMOTE.

Keywords: Synthetic Minority Oversampling Technique,Heart failure prediction, ML.

Introduction:

The World Health Organization (WHO) reports that cardiovascular disease is the primary reason for death across the globe. Identifying cardiovascular disease-CVD is challenging due to the various risk factors that lead to CVD, including but not limited to high blood pressure, high cholesterol levels, diabetes, irregular pulse rate, and many more. Men and women may experience CVD symptoms differently [1]. When comparing chest pain in men and women, men are more likely to have chest pain, while women are more likely to experience other symptoms, such as nausea, acute exhaustion, and shortness of breath. Researchers for predicting cardiac illnesses have investigated numerous methods. However, earlier disease predictions are inefficient for several reasons. These reasons include the approach's complexity, execution time, & accuracy [2]. That's why it's essential to have a good diagnosis and treatment as soon as possible.

Approximately every 36 seconds, an American loses their life to cardiovascular disease. Heart disease is the leading cause of death in the United States, accounting for over 0.66 million fatalities yearly. The American healthcare system bears a hefty burden due to cardiovascular-CVD disease [3]. About $219 billion was spent annually on healthcare, medication, and lost productivity because of death in 2014 and 2015.Preventing heart failure, a leading cause of death, also requires early identification. Even though angiograms are the gold standard for predicting coronary artery disease, their high price makes them out of reach for many people with modest incomes [4].

Heart health is complicated to detect because of the many elements that influence it, including blood pressure, cholesterol, creatine, and so on. Alcohol consumption, smoking, diabetes, high cholesterol, and lack of physical activity were all recognized as modifiable risk factors for cardiovascular disease by the authors. EHRs, or electronic health records, are an advanced tool they may use in the clinic & in the lab. Even a tiny mistake during the physical examination could have fatal consequences in cardiac illness [5]. Reducing CVD mortality with the help of machine learning-ML-powered specialist networks has been demonstrated. They can mine large amounts of data for valuable insights via data mining. In addition to its obvious applications in the medical and scientific communities, industry & academia also use it extensively [6].

Data mining is a technique used to sift through large amounts of historical facts to unearth previously undiscovered but vital data that can inform future decisions [7]. Reduced error in prediction & actual results has been achieved by applying several machine learning-ML algorithms, which have been utilized better to grasp the complexity & non-linear interaction between various elements [8].They can use machine learning-ML algorithms to aid medical personnel in data analysis & diagnosis as the volume of medical data grows [9]. Predictions of cardiovascular disease (CVD) in individuals & mortality rates after a heart attack can be made using a variety of classification algorithms implemented in medical data mining [10]. Medical records of hf patients were obtained at the Institute of Cardiology & Allied hospital in Faisalabad, Pakistan, & were made public by Ahmad et al. The researchers predicted mortality rates using Cox regressions. There was also the usage of Kaplan-Meier plots to demonstrate the shifts in survival probabilities over time [11].

Making the datasets available to the public dramatically serves the scientific community. Next, Zahid et al. looked into the same data sets, but this time they offered two different strategies for predicting mortality depending on gender. Later, Chicco and Jurman utilized only two characteristics from the same datasets to make predictions about ML performance [7]. Although the researchers mentioned above demonstrated promising outcomes using traditional statistical methods, these approaches are not optimal for large-scale datasets, giving rise to the need for new ML algorithms [12]. Given this, we set out to aid doctors in the survival prediction of CVD patients by creating ML strategies. Researchers used nine different ML models, including DT (AdaBoost), LR, RFO, DE, KNN, ETC, G-NB, & SVM; to remedy this issue of class difference, we employ the SMOTE. In the mentioned ways, this research adds to the existing body of knowledge:

Developed a diagnostic decision-support system for heart problems that improve patients' chances of survival. The SMOTE method is used to examine the accuracy of several models for predicting the survival of cardiovascular patients, including those depending on trees, regression, & statistics [13]. They extracted vital characteristics from the datasets that impact the ML algorithm's performance and are then used to research the most critical risk factors [14].

After this point, this is how the article was laid out: Outlining the research in question, Section II summarises the associated study that focuses on the heart. The dataset, its pre-processing, & its visualization are all described in Section III, along with their respective roles in discovering the underlying pattern [15]. The many algorithms employed in this study are also detailed. The outcome & evaluation are discussed in Section IV. Section V presents a summary & suggestions for further research [16].

Literature survey:

Assisting data mining with ML is a powerful tool for addressing many issues. Due to the sheer volume of data, healthcare information can be challenging to process manually in medical data mining. In addition, AI development has introduced precise & accurate solutions for medical applications, which is especially important when dealing with private medical data. Even in industrialized countries, cardiovascular disease is a significant killer. The earlier detection of cardiovascular illness risks using ML models has gained considerable traction. Some things that can put you at risk for cardiovascular disease are smoking, being older, having diabetes, or having high blood pressure.

During Cardiovascular detections, Muthukaruppan & Er [17] suggested fuzzy expert systems dependent on Particle Swarm Optimization-PSO. The decision tree's rules were parsed out & then translated into fuzzy rules. Their use of the unclear experts' systems has resulted in an accuracy rate of 93.27 percent. They could extract a small number of rules from the small datasets used in their investigation.

Alizadehsani et al [18]. made utilized a technique called ensemble-based learning. Our researchers used a dataset containing 303 cases they received through the Rajaie Cardiovascular Medical & Research Centre. The researchers predicted CVD using a nascent ensemble learning called C45. Researchers diagnosed stenosis in the RCA with 68.96% accuracy, the LCX with 61.46% accuracy, and the LAD with 79.54% accuracy (LAD). Using the SVM classifier, another team of researchers enhanced the outcomes, achieving 80.50% accuracy for RCA, 86.14% accuracy for LAD and 83.17% accuracy for LCX.

When utilizing the KEGG metabolic reaction networks datasets for detecting cardiac illness, Manogaran et al. [20] combined MKL with Adaptive Neuro-Fuzzy Inference System-ANFIS & obtained reliable findings. Manogaran and colleagues examined various heart conditions. In addition to a technique for aggregated random under-sampling, they suggested an ensemble learning framework of several neural network structures. Preprocessing techniques, including feature extraction, were utilized to improve the efficiency of the classification techniques. They used many different types of unidirectional & bidirectional neural network models. Found that CNN-structured ensemble classifications using BiGRU or BiLSTM worked best.

The two-tier ensemble paradigm, first by Tama et al., uses some classifiers as the base classifications for other ensembles. EGB, GBM, & DE class labels are used to create the suggested stacking architectures (XGBoost). On four distinct types of datasets, they assess their suggested detection models. Additionally, they employed methods for selecting features depending on particle swarm optimization-PSO. In comparison to the 10-fold cross-validations, their suggested framework outperformed higher. The stacking of tree-based models was the authors' only consideration. To enhance the performance of the models, they might attempt further statistical & regression-based approaches.

Melillo et al. presented an artificial classification to distinguish high-risk illnesses from low-risk ones. With a 93.3% sensitivity and a 63.5% specificity, the classifications & regression trees (CART) method excelled in their experiment. Only 12 patients at low risk were studied, while they studied 34 at high risk. To verify the efficacy of their strategy, they should investigate a larger dataset.

The CDSS used by Guidi et al. for cardiac dysfunction analysis was subjected to scrutiny. In their study, they utilized numerous machine learning classifications & evaluated their results. Random forest & the Confusion Analysis and Reduction Technique (CART) achieved the highest accuracy (87.6%).

s.no	Attributes	Description	Range	Measured In
1.	Time analysis	Follow up period	40-300	Hours
2.	Gender	Male or female	10,20	Decimal
3.	smoking	If the user smokes	10,20	Decimal
4.	Diabetics	If the user has diabetics	10,20	Decimal
5.	Age	Patient age	20 to 75	Years

Table 1: Parameters for describing a datasets.

Proposed methodology:

This study utilizes the Heart-failure-clinical-records-datasets obtained from the UCI ML repository. There are 299 individuals with cardiac issues whose medical records were collected during the follow-up period, and each of their profiles includes 13 clinical characteristics. There were 194 males & 105 females total out of 299 entries. Everyone in this group of patients is over the age of 40. A value of 1 indicates death, and 0 shows life in the target class. All 299 individuals with a history of heart failure & left ventricular systolic dysfunction were classified as NYHA functional class III or IV. The Table1 provides a summary of the entire dataset.

Through data visualization, they can understand previously unseen relationships within a dataset. Visualizing the characteristics that define the datasets is a great way to learn more about them. We use RF to rate features.RF's anticipated feature relevance. As determined by RF, the most critical components are time, creatinine, ejection fraction, age, platelets, CPK, & sodium.

For this kind of prediction from data, known as "classification," a supervised machine learning model is used. This study suggests a strategy to increase classification accuracy using an ensemble of classification models, which might be utilized to diagnose heart disease. Each classifier is trained on its subset of data from a train set, and then the subsets are combined to form a test set. The effectiveness of the classifications is measured by their ability to successfully separate training data from test data. It provides descriptions of some popular categories used in ML.

The SMOTE methodology is an oversampling strategy that has seen extensive medical use to address unbalanced class data. SMOTE generates random synthetic data of the minority class from its nearest neighbors utilizing Euclidean distance, thus increasing the total number of data examples. When new instances are produced based on the properties of the original data, they resemble the actual data in many ways. Due to the noise it introduces, SMOTE is not preferable when working with high-dimensional data. In this research, the SMOTE method is used to produce a brand-new set of training data. SMOTE doubled the data samples for each class from 97 to 300.

There are a few conventional ways to evaluate the efficacy of machine learning models. The expansion of work analysis seems to be predicted by combining several assessment techniques. This study would compare different ML based algorithms depending on four primary metrics: accuracy, precision, recall, and F-Score. They can determine these four measures using the confusion matrix.

True positives (TP), false negatives (TN), false positives (FP), and false negatives (FN) make up the four quadrants of the confusion matrix (FN). The false negative is the most critical forecast if the data affects medicine.

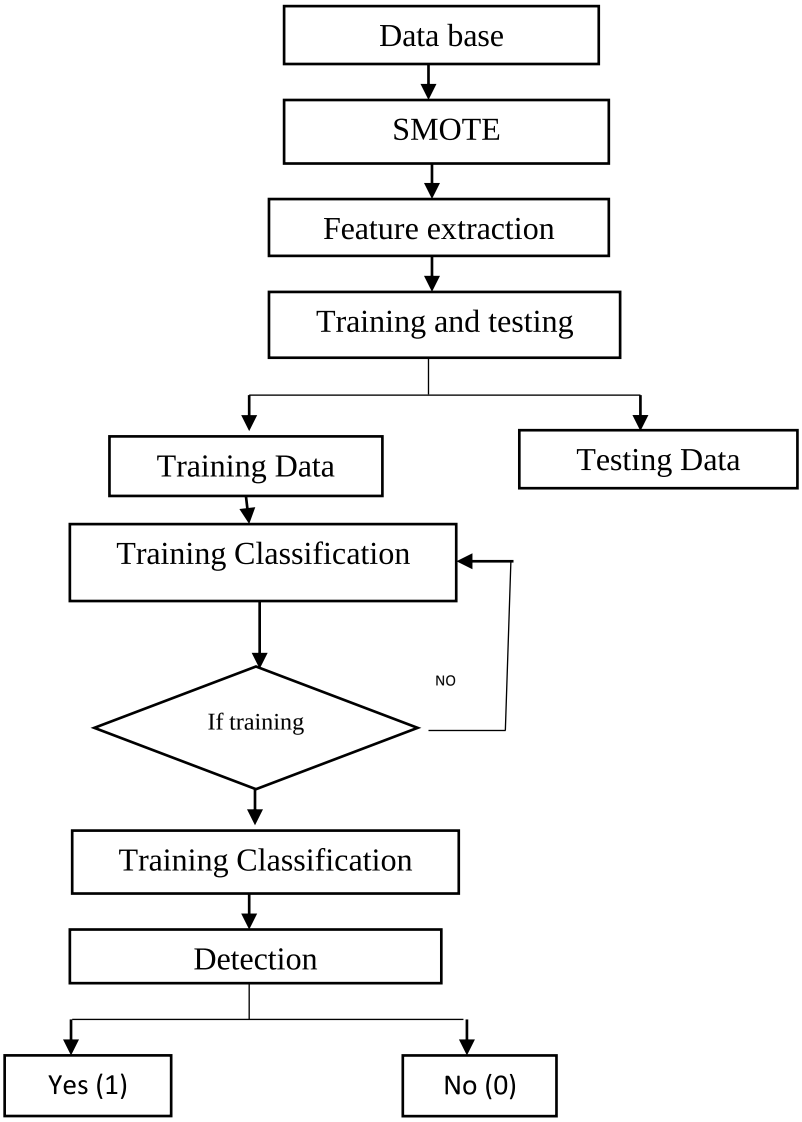

FIGURE 1. Proposed methodology block diagram.

Results and analysis:

The discussion centers on the methodology used to forecast the likelihood of survival for cardiac patients. At first, we show the outcomes with the complete collection of features, and then we offer the results with the critical set of characteristics. The dataset has 13 characteristics that describe a person's body, health, & habits. The anaemia, diabetes, blood pressure, smoking, & gender categories were all examples of these traits that can only take on one of two possible values. When performing a binary classifications activity, the characteristic denoting a patient's survival or death within the first 130 days of follow-up is used as the target class. Table 1 contains the dataset's specifications. If a dataset is unbalanced, it can be fixed by using SMOTE. All the ML models were trained on the balanced dataset & tested for accuracy, precision, recall, & F-score. You can see a flowchart of the suggested procedure in Figure 1.

Researchers have used supervised machine learning to evaluate the efficacy of these models. A 70:30 split between the two data sets will serve as the train set and the test set, respectively. They can avoid overfitting with this ratio, which has been used in several studies for classification tasks. Additional performance evaluation measures are used to assess the classifiers' abilities in machine learning. All the tests have been run in a Python environment using various libraries on a computer with a 2 GB Dell PowerEdge T 430 GPU, 2x Intel Xeon 8 Cores clocking in at 2.4 GHz, and 32 GB of DDR4 Random Access Memory (RAM).

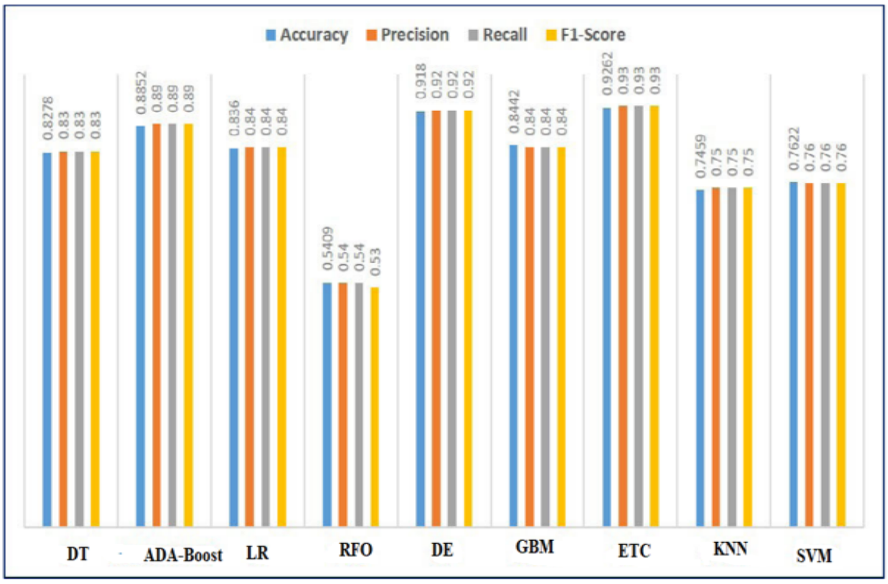

FIGURE 2: Classifier accuracy when given a complete dataset.

The complete collection of characteristics from the heart failure-clinical-record-dataset has been used to compare supervised ML classifiers. Depending on the measure used for assessment, the performance of various classifications ranged from excellent to poor. The authors of this study use tree-based, regression-based, & statistical models to forecast how long patients with heart failure will live. DT, DE, & ETC are all examples of ensemble models dependent on trees. AdaBoost & GBM are two examples of tree-based boosting models.

LR & RFO are examples of regression-based methods, while KNN & SVM are examples of statistical methods. The overall performance of ML models on the complete feature set is shown in Table 2. Table 2 shows that the LR classification performed well, with an accuracy of 0.7556, precision of 0.75, recall of 0.76, & F-score of 0.75. At 0.7667 accuracy and 0.76 F-score, SVM & G-NB were the second-best classifications. Among the nine categories tested with all characteristics, the tree-based classification DE had the highest accuracy (0.7779), recall (0.79), & F-Score (0.79).

Models	**Accuracy**	**Precision**	**Re-call**	**F-score**
DT	0.6889	0.70	0.69	0.69
AdaBoost	0.7223	0.73	0.72	0.72
LR	0.7556	0.75	0.76	0.75
RFO	0.5667	0.52	0.57	0.53
DE	0.7779	0.79	0.79	0.79
GBM	0.7444	0.74	0.74	0.74
ETC	0.7334	0.73	0.73	0.73
KNN	0.7667	0.76	0.77	0.76
SVM	0.7667	0.77	0.77	0.76

Table 2. Classifier results of all ML models without SMOTE.

With an accuracy of 0.5667, a precision of 0.52, a recall of 0.57, & F-score of 0.53, RFO was the lowest classification for predicting heart failure mortality. Figure 2 shows the results of a comparative of the different models' performances.

SMOTE is an effective method for addressing the issue of class differences, with proven success in a wide range of contexts. The SMOTE algorithm uses fabricated data to determine the minority group's contribution to the database. The classification trained with the SMOTE method for ML is presented in Table 3 for the 13 characteristics of the heart-failure-record database. It might see in Table 3. The SMOTE dramatically boosts the efficiency of tree-based classifications across the board. With SMOTE, DT's performance went from 0.69 to 0.7278 accuracy. AdaBoost performed well, achieving 0.7852 accuracy, 0.79 precision, 0.79 recall, & 0.89 F-score with a balanced dataset.

Similarly, RF enhanced their outcomes with SMOTE to the tune of a 0.8180 F-Score & a 0.72 accuracy rating. The results of ETC with all of its characteristics enabled plus SMOTE were 10% better than those obtained without SMOTE. Outcomes for ETC were the best, with an F-Score of 0.83. A precision score of 0.83, a recall score of 0.83, and an accuracy score of 0.8262. Trees are constructed via a boosting algorithm by minimizing the mistakes made by weaker learners. There is no evidence that up-sampling similar data improve findings. It explains why GBM did not enhance when treated with SMOTE. Figure 3 shows the results of SMOTE's performance evaluations of ML models.

In contrast to SMOTE, regression-based (LR and RFO) and statistical-based (KNN and SVM) models perform worse. In a study predicting survival rates for heart patients, SMOTE outperformed tree-based classifiers.

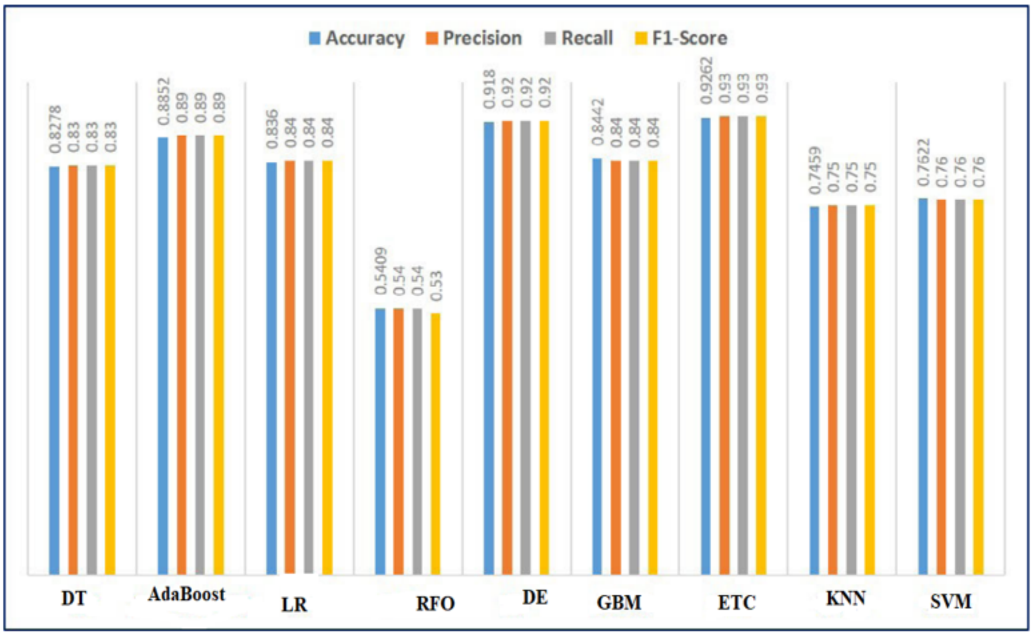

FIGURE 3. SMOTE classifier performance.

Models	Accuracy	Precision	Re-call	F-score
DT	0.7278	0.73	0.73	0.73
AdaBoost	0.7852	0.79	0.79	0.79
LR	0.7360	0.74	0.74	0.75
RFO	0.4409	0.44	0.44	0.43
DE	0.8180	0.82	0.82	0.82
GBM	0.7442	0.74	0.74	0.74
ETC	0.8262	0.83	0.83	0.83
KNN	0.6459	0.65	0.65	0.65
SVM	0.6622	0.66	0.66	0.66

Table 3. Classifier results of all ML models with SMOTE.

Conclusion:

Using machine learning algorithms to process raw health data about the heart will assist save the lives of heart patients. It can lower the death rate by looking at what causes heart failure and taking preventive measures. In this study, an effective and efficient method based on machine learning is suggested for determining whether a heart patient will live or die. Some strategies for machine learning are LR, AdaBoost, RF, GBM, G-NB, and SVM. SMOTE is used to solve the problem of a class imbalance. RF also used something called "feature ranking." RF says that the most important factors are time, creatinine, ejection fraction, age, platelets, CPK, and sodium. A complete set of features and some features from the Heart-failure-clinical-records-dataset are used to compare the performance of machine learning models. So, the results of the experiments showed that tree-based models with feature selection are a perfect way to get the highest accuracy. The SMOTE technique made tree-based classifiers much better at predicting the survival of heart patients. ETC with SMOTE did the best on all evaluation measures, scoring 0.9262 for accuracy, 0.93 for precision, 0.93 for recall, and 0.93 for F-Score. This work could enhance the healthcare framework & become a beneficial tool for doctors & nurses to utilize in figuring out who will survive heart failure. It would also help doctors determine if a heart failure patient lived or died so users can focus on the most significant risk factors. Future work on these studies can be done using different combinations of ML models to take advantage of their benefits. It can improve methods for selecting characteristics to make ML models work effectively. The fact that characteristics extraction problems are NP-hard means that meta-heuristics can be employed in this case.

References:

1. 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.
2. 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.
3. Q. K. Al-Shayea, ‘‘Artificial neural networks in medical diagnosis,’’ Int. J. Comput. Sci., vol. 8, no. 2, pp. 150–154, 2011.
4. 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.
5. 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.
6. 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.
7. L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees (Statistics/Probability Series). 1984.
8. 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.
9. 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.
10. 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.
11. L. Breiman, ‘‘Random forests,’’ Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
12. J. H. Friedman, ‘‘Greedy function approximation: A gradient boosting machine,’’ Ann. Statist., pp. 1189–1232, Oct. 2001.
13. . 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.

14. 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.
15. R. Gupta, ‘‘Recent trends in coronary heart disease epidemiology in India,’’ Indian heart J., vol. 60, no. 2, pp. B4–B18, 2008.
16. 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.
17. 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.
18. 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.
19. 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.
20. 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.
21. P. Melillo, N. De Luca, M. Bracale, and L. Pecchia, ‘‘Classification tree for risk assessment in patients suffering from congestive heart failure via longterm heart rate variability,’’ IEEE J. Biomed. Health Informat., vol. 17, no. 3, pp. 727–733, May 2013.
22. 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.
23. 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.