Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2020.Doi Number

Novel Feature Reduction (NFR) Model with Machine Learning and Data Mining Algorithms for Effective Disease Risk Prediction

Syed Javeed Pasha[1], E. Syed Mohamed[2]

[1]Department of Computer Applications
[2]Department of Computer Science and Engineering
[1, 2] School of Mathematical and Computer Sciences, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India

Corresponding author: Syed Javeed Pasha (syedjaveedrs@gmail.com)

ABSTRACT Presently, the application of machine learning (ML) and data mining (DM) techniques have a vital role in healthcare systems and wisely convert all obtainable data into beneficial knowledge. It is proven from the literature works that a chance of 12% error remains in the diagnosis of the diseases by the medical practitioners. Moreover, for effective disease risk prediction in medical analysis, more emphasis is accorded to the area under the curve (AUC) with accuracy as an evaluation metric. However, the role of the AUC has not been previously characterized notably. In this research article, a novel feature reduction (NFR) model that is aligned with the ML and DM algorithms is proposed to reduce the error rate and further improve the performance. The proposed NFR model comprises of two approaches and uses the AUC in addition to the accuracy to achieve a robust and effective disease risk prediction. The first approach is based on a heuristic process evaluating performance by reducing features with respect to the improvement in the AUC besides the accuracy as evaluation metrics, working to obtain the best subset of highly contributing features in the prediction. The second approach evaluates the accuracy and AUC of all individual features and forms the subsets with the highest accuracies, AUCs, and least difference between them, which are combined in various combinations to achieve the best-reduced set of highly relevant features. For this purpose, the benchmarked public heart datasets of the ML repository of the University of California, Irvine (UCI) are tested; the results are promising. The highest accuracy and AUC achieved with the proposed NFR model are 95.52% and 99.20% with 41.67% feature reduction, respectively. The accuracy is 4.22% higher than recent existing research with a significant improvement of 25% in the performance of the running time of the algorithm.

INDEX TERMS AUC, cardiovascular, data mining, disease risk prediction model, feature selection, health care, machine learning, NFR, UCI ML repository

I. INTRODUCTION

The steadily escalating and unanticipated mortality folio of individuals of various age groups, which are recorded daily, has become a major concern. Studies suggest that cardiovascular diseases (CVDs), such as heart illnesses and strokes, are a leading cause of these deaths. CVDs cause the deaths of 17.9 million individuals each year and account for 31% of all deaths worldwide [1]. CVD and stroke cause substantial health and financial drains in the United States and worldwide. The statistical system Apprise presents the newest data on a variety of clinical heart and vascular ailment situations (stroke, congenital heart disease, coronary heart disease [CHD], heart failure [HF], and cardiopathy) and the associated outcomes (importance of caution, measures and monetary expenses). Since 2007, the yearly publications of Apprise have been mentioned more than 20,000 times in various studies. In 2017, Apprise was accessed more than 106,500 times between January to July. In addition, chronic disease treatment comprises 80% of American medical care fees, and the percentage of Americans who have at least one or more chronic diseases is 50% according to a report by McKinsey. The occurrence of

persistent diseases is expanding with changes in expectations for everyday comforts. Every year, 2.7 trillion USD is spent on treatment of chronic ailments in the United States; this sum involves 17.6% of the yearly GDP [2]. In numerous nations, medical services for chronic diseases are essential. Hence, risk assessment of chronic illnesses has become indispensable.

From the disease history of patients compiled in electronic health records (EHR), the test outcomes and statistical data have helped to reduce the expenses of medical case studies by empowering us to obtain potential data-driven answers. Diagnosis of the diseases by the medical practitioners leaves behind a chance of 12% error [3]. To reduce this error rate and to further improve the performance, machine learning (ML) algorithms (e.g., logistic regression and random forest) are more frequently employed for prediction by conventional disease risk models. Particularly, supervised learning algorithms are used to prepare models that utilize training data with labels. In the clinical datasets, patients are categorized as susceptible and immune. In preventive medicinal fields, these categories are important and broadly examined.

In ML and data mining (DM), a feature is a distinct quantifiable characteristic of a phenomenon that is being witnessed [4]. Features are usually numeric. Selecting informative, selective, and independent features is a vital step to improving the performance of ML algorithms. Selecting a subgroup of effective features (attributes, predictor) to build a model is known as "feature selection". Furthermore, the datasets usually contain few features, which are repetitive or noneffective, and removing them will not cause a substantial loss of information; this premise is also referred to as "feature selection" [5]. Irrelevant and redundant features are treated as two different types of features, since one relative feature can be repetitive in the presence of another feature with an intense correlation [6].

The mechanism of a feature selection algorithm combines an assessment process that reduces the error rate and an exploration procedure for recommending novel feature subgroups. The selection of an assessment metric profoundly impacts the algorithm, which differentiates among various feature selection algorithms [7].

Wrapper, filter, and hybrid are the three existing models commonly utilized for feature selection. However, these models have several drawbacks. The wrapper model employs ML classifiers but relies only on accuracy as a measurement for evaluation eventuating in imprecise results. The disadvantages of the filter model is that it is a feature-based feature selection model and does not employ ML algorithms at all, instead provides results by incorporating statistical techniques and measurements based on the input variables or features and target variables, thus also, impeding the decline in the error rate of 12% in the prediction that can be decreased by means of ML algorithms. A combination of both the wrapper and filter models is the Hybrid model. The other shortcomings of the recent research are as follows: firstly, in risk assessment of medical disease datasets, the accuracy alone is not sufficient as a metric for evaluation. In the instance of imbalanced datasets that contain a huge divergence between the positive and negative cases, the resultant accuracy is highly misleading, thus providing the wrong diagnosis of the patients causing serious damage or even loss of life. Literature studies suggest the use of the area under the curve (AUC) as an essential evaluation metric for disease risk prediction, specifically in the medical field. In the medical analysis, the AUC is equally significant for determining the performance of the ML algorithms in datasets of healthcare. Secondly, a larger difference between the AUC and accuracy is a sign of a poor classification; a smaller difference leads to a more precise disease risk prediction result. In the existing studies, feature selection models related to particularly the heart disease datasets are not many and most of these approaches do not utilize the AUC as an evaluator but use only the accuracy. Since the current study addresses datasets from the medical database, the AUC has also been employed in it with the accuracy to achieve a robust prediction.

The objective of this research article is to present a Novel Feature Reduction (NFR) model that improves the accuracy and AUC by reducing the number of features without omitting relevant features. The proposed NFR model is aligned with five ML classification algorithms— logistic regression (LR), support vector machine (SVM), boosted regression tree (BRT), stochastic gradient boosting (SGB), and random forest (RF) —for datasets of healthcare to enhance the diagnosis capabilities by creating smaller feature subsets from a higher number of features. The proposed model comprises of two approaches, where i) the first approach is based on a heuristic process evaluating the prediction performance by reducing features vis-à-vis the improvement in the AUC simultaneously with the accuracy as evaluation metrics, in order to acquire the best subset of highly contributing features. ii) The second approach evaluates the accuracy and AUC of all individual features attained on aligning with ML classifiers and forms the subsets with the highest accuracies, AUCs and least difference between them. These subsets are then combined in various combinations to achieve the best reduced set of highly contributing features in the disease risk prediction.

The contributions of the proposed model include: first, diminishing the chance of the error rate that remains in the diagnosis of the diseases by the medical practitioners. Second, provides with a more accurate prediction since the AUC is employed with the accuracy, which is an essential evaluation metric specifically for datasets of healthcare but has not been previously characterized notably. Third, the proposed model identifies the reduced set of the most important and highly contributing features from the dataset that play a major role in the disease risk prediction. Fourth, the model produces improvement in the convergence speed,

i.e., the time taken by the algorithms to run is greatly reduced as the number of features is reduced. This outcome results in another key benefit of the NFR model. Finally, since susceptible patients frequently require costly healthcare, one of the applications of this study is to recognize irrelevant features to diminish their medical costs. The NFR model also benefits in reducing the economic burden on patients, since in general, fewer features mean fewer medical diagnostic tests required to predict the risk of disease. The NFR model employs heart datasets of Cleveland, Hungarian, Statlog, and Switzerland that are taken from the ML Repository of the University of California, Irvine (UCI) and the results are promising.

The other sections of this research article are presented as follows: Section II discusses the standard associated approaches with and without feature selection models for similar datasets. Section III presents the details of the datasets. The Methodology and Approach is explained in detail in Section IV. The experimental outcomes and analysis are detailed in Section V. Section VI ends with Conclusion and Future Work.

II. RELATED WORK

Due to the formulation and progress in numerous measurement practices, medical data contain relevant features that are useful, relevant features that are less useful, i.e., irrelevant features, and redundant features that are not useful. Identification of these features is important for denoting the domain properly. Since the narrative of the prediction column is not affected by irrelevant features, labeling it is a nontrivial assignment. A suitable assortment of the maximum pertinent features affects the work of classification models. The size of the datasets can be decreased by removing the redundant features. From the group of all features, feature selection is specified as a way to extract a subgroup of the most relevant features.

Here, recent research is presented in two subsections. The first subsection deals with the feature selection techniques used in heart disease risk prediction and heart disease prediction models on the same datasets that do not employ feature selection are laid out in the other subsection. Additionally, in the first subsection, works are listed together for the datasets of Cleveland, Hungarian, and Statlog that are available in the ML repository of the UCI. Works employing the Switzerland heart dataset from the same repository are added along with the aforementioned datasets in the second subsection.

A. FEATURE SELECTION TECHNIQUES IN HEART DISEASE RISK PREDICTION

In this article, notable research related to disease risk prediction has been examined. Utilizing the Cleveland heart dataset of the UCI [8][9], El-Bialy *et al.* [10] used fast decision trees (FDT) with four features and obtained 77.55% accuracy. They also used pruned C4.5 trees and attained 78.54% accuracy with five features. Paul *et al.* [11] utilized a fuzzy decision support system (FDSS) that includes rules based on weighted fuzzy derives from a genetic algorithm (GA). They retrieved eight effective features and obtained 80% accuracy. Established on the PI-Sigma model, Burse *et al.* [12] designed a multilayer Pi-Sigma neuron model (MLPSNM) for heart disease diagnosis. Their model, which asserts to make the engineering and computation elementary, uses a k-fold validation method for its affirmation. They elected four attributes that are adjacent to the hyperplane in the support vector machine - linear discriminant analysis (SVM-LDA) method and obtained 88.32% accuracy. Moreover, Amin *et al.* [13] observed that the highest accuracy of 87.4% was attained utilizing nine important features and the DM technique Vote, which is a hybrid technique that is obtained by merging naive Bayes (NB) and logistic regression (LR). They also applied the k-nearest neighbor (KNN), SVM, neural network (NN), and decision tree (DT) classifiers for the same heart dataset. According to Hager *et al.* [14], two variants of algorithms employed in feature selection, such as relief and univariate algorithms, were utilized to sort key features. They applied a DT, SVM, random forest (RF), LR, ML algorithms with and without feature selection. Cross-validation and hyperparameter tuning were applied to further enhance the performance of disease risk accuracy. They obtained an accuracy of 88.4% using seven features obtained via the univariate feature selection technique applied with LR. They also achieved 89.9% accuracy with the LR and the same number of features using the feature selection technique Relief. Beulah *et al.* [15] applied ensemble classification, which merges manifold classifiers that are low performing, such as NB, Bayes net (BN), RF, and multilayer perceptron (MP), and then carried out feature selection to further enhance the accuracy. Their results were reformatory using a majority vote with NB, BN, RF, and MP, and the highest achieved accuracy was 85.48%. Gupta *et al.*[16] presented a machine intelligence framework (*MIFH)*, which employs the factor analysis of mixed data (FAMD) to extract and derives features from the heart disease dataset to train the machine learning predictive models. Their framework attained an accuracy of 91.80% and an AUC of 91.61% with 10 features using FAMD+RF.

For the Hungarian dataset of UCI [8], El-Bialy *et al.* [10] adopted six features using C4.5 with pruned trees and achieved 78.54% accuracy. Using two features with fast decision trees (FDT), they were able to attain an accuracy of 78.23% in patients with cardiovascular ailments. Shah *et al.* [17] applied probabilistic principal component analysis (PPCA) to extract the most important features from the dataset. The eight features subset was applied with the radial basis function (RBF) kernel-based SVM, and they achieved an accuracy of 85.82%. Javeed *et al.* [18] designed an ensemble gain ratio feature selection (EGFS) model employing an RF as an ensemble algorithm and a gain ratio

algorithm in order to extract features that aid in performance improvement. Via KNN, LR, and NB, they applied their model to medical datasets of UCI to attain an accurate disease risk prediction. Saqlain *et al.* [19] used mean Fisher score grounded (MFSFSA), forward, and reverse feature selection algorithms to obtain the most vital feature subset from the dataset. They considered the dimension and value of the Matthews correlation coefficient and the Fisher score to form the feature subset. The subset of seven features was applied with the RBF kernel-based SVM, and an accuracy of 84.52% was achieved.

For the Statlog heart dataset of UCI [9], Archana *et al.* [20] introduced a principal component analysis (PCA) based optimal feature subset selection (FSS) for fuzzy extreme learning machine (FELM), the PF-FELM approach, to solve weighted classification problems. Their method works in four steps, the PCA, FSS, fuzzification, and classification. Filter-based ranking algorithms are employed by the FSS to filter an optimal feature order that is based on the highest number of occurrences. The PF-FELM attained the highest accuracy of 87.65% with a subset of 11 features for the Statlog heart dataset. Furthermore, Amin *et al.*[13] designed prediction models employing seven classification techniques: k-NN, DT, NB, LR, SVM, NN and Vote (a hybrid technique with NB and LR) with various combinations of features. They inferred that the best-performing subset of nine significant features via Vote achieved an accuracy of 87.4% in heart disease prediction. Al-Attar *et al.* [21] framed a novel Parasitism-Predation Algorithm (PPA) that is based on a multi-heuristic approach merging the cat swarm optimization (CSO), cuckoo search (CS), and crow search algorithm (CSA). It filtered the features and obtained a subset to boost up the classification accuracy via KNN. With a subset of four features, they attained an accuracy of 86.17% and a runtime of 49.13 sec. A technique utilizing a genetic algorithm and an RF known as the bio inspired ensemble feature selection (BEFS) model was developed by Javeed *et al.* [22] to identify the most relevant features from the disease databases. They further applied RF and LR on various combinations of those features until the subset of features contributing most in enhancing the risk prediction was attained by them. Zahangir *et al.* [23] presented three feature ranking models for disease datasets. Their models utilize a suitable feature ranking algorithm followed by applying the RF algorithm on the first positioned features. They further performed 10-fold cross-validation test for enhancing their predictor. They inferred that their Model III HtS that employed ReliefAttributeEval ranker performed better than the other models and obtained the highest accuracy of 83.3% and an AUC of 86.9% on reducing one feature. The drawback of their proposal is that it is not benchmarked for performance with any existing research, except for the baseline model.

An analytical study of the above listed models depicts a large scope for improvement in the prediction accuracy. Moreover, most of the models did not utilize the AUC, which is a more essential evaluation metric than the accuracy especially for the medical disease datasets for a robust and more accurate disease risk prediction. Furthermore, most of the previous prediction models lack the observation of the time taken to run the algorithms.

B. HEART DISEASE PREDICTION TECHNIQUES

Works without feature selection on the Cleveland heart dataset UCI [8] include research by Xu *et al.* [24] who recommended the Structural Least Square Twin SVM, with the Structural Twin SVM and Least Square Twin SVM methods, and utilized previous structural evidences of data to obtain 87.82% accuracy to classify arrhythmia data. Mokeddem *et al.* [25] proposed a fuzzy clinical decision support system (CDSS) using an RF and C5.0. With C5.0, rules are generated and 90.5% accuracy was attained with this system. Sabahi *et al.* [26] employed a bimodal fuzzy analytic hierarchy process (BFAHP) that legitimizes the risk elements in the dataset with statistical data to procure the fuzzy rules by applying the Bayesian formula. The BFAHP attained 87.31% accuracy. Shilaskar *et al.* [27] used a cluster-based hybrid sampling method, which required minimal sampling time. A confidence measure-based (CMB) technique was applied to assess the test set, and they attained 77.5% accuracy. Maji *et al.* [28] proposed a hybridization technique for a DT and an artificial neural network (ANN) and obtained 78.14% accuracy and 77.4% accuracy, respectively. 10-fold cross-validation was used to substantiate their results, and the experiments were carried out in WEKA. They also noted that hybrid-DT outperforms the ANN algorithm in predicting heart ailment risk. Moloud *et al.* [29] computed a new nested ensemble nu-Support Vector Classification (NE-nu-SVC) technique for diagnosis. Their model is based on traditional ML and ensemble learning methods. They employed the GA algorithm, and the ClassBlancer and Resample methods were utilized to balance the datasets. Four kernel functions, the linear, polynomial, radial basis (RBF), and sigmoid were aligned with their base algorithm. An accuracy of 84.51% was attained with the NE-nu-SVC+Sigmoid. A hybrid random forest with a linear model (HRFLM) was developed by Senthilkumar *et al.*[30] aimed to enhance the prediction accuracy with ML classification algorithms. An accuracy of 88.7% was obtained through their prediction model for the heart disease. Liaqat *et al.* [31] introduced a χ^2 statistical model addressing the refinement of features and eradicating the problems of prediction model i.e., the problems of underfitting and overfitting. A deep neural network (DNN) is employed for the purpose of an exhaustive search strategy. Their model attained a prediction accuracy of 91.57%. Thippa *et al.* [32] proposed an adaptive genetic algorithm with fuzzy logic (AGAFL) model. Their model comprises of a rough set theory and a fuzzy rule based classification module. An accuracy of 90% was obtained by the model. Kanti *et al.*[33]

introduced a hybrid predictive model that works to optimizes at two levels for diagnosing clinical datasets. In Level-1 optimization, identification of a parallelly optimal proportion (Popt) is carried out for training and test sets for each of the dataset on a parallel machine. The best training set (Tbest) for Popt is again searched parallelly. In level-2 optimization, the rule set (R) generated by the Perfect Rule Induction by Sequential Method (PRISM) learner on Tbest is refined via parallel genetic algorithm. Their model obtained an accuracy of 89.81% and an AUC of 91.10%. Furthermore, Zhu *et al.* [34] proposed an improved discrete artificial fish-swarm algorithm joined by margin-distance-minimization for ensemble pruning (IDAFMEP), in which initially low-performing classifiers are prepruned to improve their performance with margin distance minimization (MDM). The resultant ensemble, which is based on MDM and uses the proposed improved discrete artificial fish swarm algorithm (IDAFSA), achieved enhanced results for datasets of the UCI, including the Cleveland dataset.

For the Hungarian dataset of UCI [8], works without feature reduction include research by Fernandez-Delgado *et al.* [35], who attained an accuracy of 58% using an RF classifier. According to Mokeddem *et al.* [25], the rules were obtained with C5.0 to form the fuzzy CDSS. They obtained 85.71% accuracy. Sabahi *et al.* [26] designed the bimodal fuzzy analytic hierarchy process (BFAHP), which works with the potency of the risk elements in the dataset, to tally the fuzzy rules using a Bayesian formula. The BFAHP attained an accuracy of 86.57%. Repaka et al. [36] used NB, multilayer perception (MLP), and sequential minimal optimization (SMO) classifiers to obtain accuracies of 81.11%, 77.4%, and 84.07%, respectively, using all features. Tanveer *et al.* [37] analyzed eight different forms of the twin support vector machine (TWSVM) with other classifiers, and using the Friedman Rank (FRank), verified the statistical test for various datasets of UCI and the Hungarian dataset. The TWSVM_m attained 79.6% accuracy. Perales-González *et al.* [38] developed a hierarchical ensemble approach—the boosting ridge extreme learning machine (BRELM), which stimulates diversity in the constituents of an ensemble using the loss function in the one-hidden-layer feed-forward network version of extreme learning machine (ELM). Their layout managed to obtain 81.28% accuracy.

For the Statlog heart dataset of UCI [9], Shuo *et al.* [39] proposed an improved classification approach for the prediction of diseases based on the classical Iterative Dichotomiser 3 (Id3) algorithm. The improved Id3 algorithm adopts a heuristic approach to develop rules to inculcate the classifier models. They attained an accuracy of 77.78%. Two efficient cost-sensitive (CS) classification models were built by Shichao *et al.* [40] that were based on KNN. With the Direct-CS-KNN and Distance-CS-KNN, they obtained an AUC of 76.42 and 76.88, respectively. Soraya *et al.* [41] constructed a classifier that merges the Maximum Relevance Maximum Diversity (MRMD) method and diversity measures. Their MRMD-II model is based on a greedy search algorithm and assesses the diversity and accuracy determining an optimal classifier ensemble. They attained an accuracy of 85.58% with MRMD-II, an accuracy of 84.07% with MRMD-MLP, an accuracy of 82.59% with MRMD-SVM, and an accuracy of 84.81% with MRMD-J48.

A semi-supervised rough fuzzy Laplacian Eigenmaps (SSRFLE) approach was presented by Minghua *et al.* [42]. Their model works to construct a set of semi-supervised fuzzy similarity granules to assess the similarity between samples. Then via building a neighborhood rough fuzzy set model of these granules, the degrees of both the samples of the similar class are evaluated. The SSRFLE yielded an accuracy of 78.68%. Yijie *et al*. [43] developed a fuzzy SVM built on Linear Neighborhood Representation (FSVM-LNR). Features are supplied to the FSVM-LNR for prediction and they obtained an accuracy of 84.81%. Himansu *et al.* [44] proposed a Neuro-Fuzzy model, which allows the participation of all features in the fuzzification process. They employed the heart dataset for their purpose and attained an accuracy of 85.83%.

Utilizing the Switzerland heart dataset [8], P. K. Anooj [45] adopted a weighted fuzzy rule-based clinical decision support system (CDSS) for the diagnosis of heart disease database. Their CDSS is divided into two phases. In phase one, data mining techniques are used to form the weighted fuzzy rules. In phase two, a fuzzy rule-based DSS is built on the basis of the fuzzy rules. A performance of 51.22% for accuracy was obtained with their system. Fernandez-Delgado *et al.* [35] employed RF, J48, C5.0 Tree, direct kernel perceptron (DKP_C), and SVM_C for the heart dataset that yielded an accuracy of 39.8%, 33.3%, 32.5%, 33.9%, and 35.5%, respectively. Via the probabilistic principal component analysis (PPCA), Saqlain *et al.* [17] used parallel analysis (PA) to determine the selection of projection vectors. The features are then fed to RBF kernel based SVM for classification and prediction of the disease datasets. Their technique attained an accuracy of 91.30% for all the features. Animesh *et al.* [46] presented an automatic fuzzy diagnostic system built on a modified dynamic multi-swarm particle swarm optimization (MDMS-PSO) and utilized GA for risk prediction of the heart disease. Following data preprocessing by statistical methods such as correlation coefficient, R-Squared, and weighted least squared (WLS), GA was employed to form the weighted fuzzy rules. MDMS-PSO was then used for the optimization of membership functions (MFs) of the fuzzy system and an ensemble fuzzy system was generated based on the fuzzy rules by combining the different local fuzzy systems. They obtained an accuracy of 89.47%. Sabahi *et al.* [26] built the bimodal fuzzy analytic hierarchy process (BFAHP) that employs the bayesian formula to calculate the fuzzy probability of risk factors. In their model, a reciprocal comparison matrix is computed based on the fuzzy validities and fuzzy probabilities. These fuzzy validities and fuzzy probabilities are then grouped in

pairs for each risk factor. They attained an accuracy of 85.22%. A hybrid adaptive genetic algorithm with fuzzy logic (AGAFL) model was presented by Thippa *et al.* [32]. They used a rough set theory and a fuzzy-based classification module. The AGAFL classifier was employed and an accuracy of 89% was obtained by the model. For the diagnosis of clinical datasets, a hybrid technique was designed by Kanti *et al.* [33] to optimize the prediction. They suggested a two level framework in which Popt is labelled in the test and training sets during level-1 optimization. A parallel exploration is run on the best training set (Tbest) for Popt. The level-2 optimization employs a parallel genetic algorithm to filter the rule produced by the PRISM learner on Tbest. The model yielded an accuracy of 57.82% and an AUC of 88.30%.

Along with further scope for improvement in the prediction accuracy, examining these models demonstrates that apart from not reducing the number of features, most of them did not utilize the AUC that is an essential evaluation metric for a robust disease risk prediction and is equally significant for determining the performance of the ML algorithms in datasets of healthcare. Also, some of the studies do not balance the datasets that are imbalanced, such as the Switzerland heart dataset, which results in a misleading prediction. Moreover, most of the previous prediction models did not indicate the decrease in time taken to run the algorithms. Furthermore, ML algorithms are preferable in healthcare datasets over deep learning (DL) algorithms that are slow in terms of performance.

III. DATASET DESCRIPTION

In this research, the NFR model employs heart datasets of Cleveland, Hungarian, Statlog, and Switzerland taken from the ML Repository of UCI. While the databases have 76 raw attributes/features (including the prediction label), only 14 of them are used in all published experiments[8][9]. Of these 14 features, eight features are categorical and six features are numeric, as shown in Table I. For the experimental purpose, the datasets are divided into training and test sets in the ratio of 70:30.

The "prepossessed Cleveland heart dataset" consists of 303 records out of which six have missing values, while the

TABLE I
ATTRIBUTE DESCRIPTION OF HEART DATASETS FROM THE UCI ML REPOSITORY

Feature Number	Feature Name	Description	Type
1	pt_age	patient age	Numeric
2	pt_sex	gender (1--male, 0--female)	categorical
3	pt_cp	chest ache/pain types 1--normal 2--atypical 3 --non-anginal_ strain/pain 4 --asymptomatic	categorical
4	pt_trestbps	resting/normal blood-pressure	Numeric
5	pt_chol	serum-cholesterol (mg/dL)	Numeric
6	pt_fbs	fasting_ blood_sugar over 120 mg/dL (1-- yes, 0-- no)	categorical
7	pt_restecg	rest electrocardiographic output 0- normal/regular 1- abnormal in ST-T wave 2- positive hypertrophy in left-side ventricular	categorical
8	pt_thalach	highest heartbeat rate	Numeric
9	pt_exang	exercise/workout generated angina/pain (here, 1--yes, 0--no)	Numeric
10	pt_oldpeak	exercise generated ST stress/depression related to rest	Numeric
11	pt_slope	maximum workout ST segment slope 1- up, 2- flat, 3- down	categorical
12	pt_ca	count of major vessels (0-3) dyed with fluoroscopy	Numeric
13	pt_thal	defect type: 3- normal, 6- fixed, 7- reversible	categorical
14	pt_num	prediction attribute: shows heart ailment status 0- absent, 1,2,3,4- present	categorical

Hungarian dataset consists of 294 records out of which 34 of the records have missing values and 261 records are complete. The Statlog dataset has 270 records and no missing values. Except for the Switzerland dataset, the other three datasets are balanced. The Switzerland database has more number of missing values. It contains 123 data instances. The features in this particular dataset are decreased from 13 to 12 since the column of the feature 'thal' does not contain any information and is either left blank or is filled with a '?'. In this study, the Switzerland heart dataset is balanced via the SMOTE algorithm. On applying the SMOTE algorithm, the records increased from 123 to 226, as shown in Table II.

TABLE II
PROPERTIES OF UCI ML EXPERIMENTAL HEART DISEASE DATASETS

Dataset Name	Is the dataset balanced?	No. of records	No. of features	No. of classes
Cleveland	Yes	303	13	5
Hungarian	Yes	294	13	5
Statlog	Yes	270	13	2
Switzerland	No	123 & 226 (**After SMOTE**)	13	5

A. DISEASE RISK PREDICTION

The prediction column depicts the heart ailment status. Experiments with the Cleveland, Hungarian, and Switzerland datasets have focused on differentiating presence of the disease (numbers 1, 2, 3, and 4) from no presence of the disease (number 0). For convenience of the experiment, the prediction column numbers 1 to 4 are treated as 1, which denotes appearance of a heart ailment, and 0, which denotes non-appearance of a heart ailment. The Statlog has two classes, and the two classes are treated as 1, which denotes appearance of a heart ailment, and 0, which denotes non-appearance of a heart ailment.

B. EVALUATION METHOD

To assess the datasets, the AUC and accuracy of the classifiers, as expressed in equation (1), are used in the proposed model. Algorithms are evaluated using one of the standard metrics, which is known as the receiver operating characteristic (ROC) curve that is obtained by the values of the true positive rate (TPR, as defined in equation (2), also referred to as the "sensitivity") and the false positive rate (FPR, as defined in equation (3), which is 1 - specificity). The accuracy obtained by the algorithm is measured in this ROC curve, which is the AUC. Hence, a larger area indicates a more accurate prediction by the algorithm.

$$\text{Accuracy} = \frac{TP+TN}{(TP+FP+TN+FN)} \quad (1)$$

$$\text{TPR} = \frac{TP}{P} \quad (2)$$

$$\text{FPR} = 1 - \frac{TN}{N} \quad (3)$$

Where TP and TN are groups that are correctly categorized as true positives and true negatives by the algorithm. FP and FN are groups that are wrongly categorized, i.e., wrongly labeled as positives, and wrongly labeled as negatives, while P and N refer to samples that are positive and negative, respectively. In medical data, more attention is paid to the AUC and accuracy. If the AUC is near 1, the model is better. Hence, the higher is the AUC value, the higher is the correctness of the disease risk prediction. Therefore, the proposed NFR model uses the AUC and accuracy as an evaluator in the risk assessment of diseases, while few studies associated with this field have characterized its role.

IV. METHODOLOGY AND APPROACH

A. PROPOSED NOVEL FEATURE REDUCTION (NFR) MODEL

Traditional feature selecting system involves four vital steps, viz., subgroup creation, subgroup valuation, stopping condition, and outcome assertion [31]. Filter, wrapper, and hybrid models are 3 types of feature selection algorithms [47]. However, these models have several drawbacks. The wrapper model employs ML classifiers but relies only on accuracy as a measurement for evaluation, eventuating in imprecise results. The disadvantages of the filter model is that it is a feature-based feature selection model and does not employ the ML algorithms at all, instead provides results by incorporating statistical techniques and measurements based on the input variables or features and target variables, thus also, hindering the lessening of 12% error rate in the prediction that can be decreased with the ML algorithms [3]. A combination of both the wrapper and filter models is the Hybrid model. The other shortcomings of these models and other recent research works are as follows: firstly, in risk assessment of medical disease datasets, the accuracy alone is not sufficient as a metric for evaluation. In the instance of imbalanced datasets that contain a huge divergence between the positive and negative cases, the resultant accuracy is highly misleading, thus providing the wrong diagnosis of the patients causing serious damage or even loss of life. Literature studies suggest the use of the AUC as an essential evaluation metric in determining the performance of the ML algorithms for disease risk prediction, specifically in the medical field.

Secondly, in ROC-based ranking, the difference between accuracy and the AUC of the feature should not be considerably high; hence, if the accuracy is 90% but the AUC is 52%, the classification presentation is poor [48]. A smaller difference leads to a more precise disease risk

prediction result. The other drawbacks of these models have already been discussed in detail in Section I, II.A and II.B. The current study addresses these gaps and the accuracy and the AUC and the number of features has been employed by the proposed model to assess the predictability while relating the performance of feature selection methods and to achieve a robust prediction. Accuracy and AUC observation values are worthy of comparison with heart ailment datasets. Feature reduction attains better accuracy with the AUC measurements when compared with other methods in this research area of heart ailment datasets. The proposed NFR model successfully reduces features, enhances the accuracies and AUCs of the algorithms, and greatly reduces the running time of the algorithms.

1) DATA PREPROCESSING: MEAN IMPUTATION AND DATA PARTITIONING

The original dataset cannot be utilized because it is in the prediction procedure due to absent data in the heart ailment datasets. In the datasets, the records with absent values cannot be eliminated because few of these records exist. The following steps are performed for data preprocessing:

Obtaining and replacing the absent values: Absent values are obtained and replaced with the mean of the particular column with simple functions in the R programing language. The pseudo-code is presented in Algorithm 2. A few predictor values of the dataset are changed from integer to factors for the purpose of making dummies, as represented in Algorithm 3.

Depending on the prediction column values, the input training datasets are divided into two subgroups, training and test sets in the ratio of 70:30, as shown in Algorithm 4. For the experimental purpose, in the prediction column, the value 0 means no disease, and the remaining values 1 to 4 mean the presence of disease.

2) DM AND ML ALGORITHMS USED

The proposed NFR model is aligned with the five ML algorithms Logistic regression (LR), support vector machine (SVM), boosted regression trees (BRT), stochastic gradient boosting (SGB), and random forest (RF). The proposed model is applied on the UCI ML repository CHD datasets is aligned with the ML algorithms to form the reduced set of highly contributing features, which aids in improved performance of the predictability of the disease risk.

The disease dataset *D*, contains of features set $F = \{f_1, f_2, f_3, \ldots, f_a, \ldots, f_x\}$ with x features and records $R = \{ r_1, r_2, r_3, \ldots, r_b, \ldots, r_y \}$ relevant to y subjects, with prediction class *C*.

Definition_1: A dataset *D* consists of records $R = \{ r_b \mid 1 \leq b \leq y \}$ where y is the number of subjects and, features set $F = \{ f_a \mid 1 \leq a \leq x \}$ where x is the number of features.

Definition_2: A record r_b is represented by feature values f_a, such as $r_b = \{ f_a \mid 1 \leq a \leq x \}$ *x* is the number of features in *D*. The value f_a is either numeric or categorical.

The design of the NFR model that is aligned with the ML and DM algorithms (*M*) with a reduced set of highly contributing features for effective disease risk prediction in records (*R*) of the dataset D is represented in equation (4).

$$P_h \rightarrow \min [F_{hc} \{ \max [P_{ts} [M (\sum_{h=1}^{y} \sum_{a=1}^{x} D (f_a \ r_b)]] \}] \quad (4)$$

Where 'F_{hc} 'is the best subset that contains the minimum number of highly contributing features in the disease risk prediction. The maximum effective disease risk prediction (P_{ts}) attained from the ML algorithms (*M*) is evaluated on the basis of highest performance (P_h) that is measured using accuracy and AUC. Section III (B) explains the metrics in detail. The NFR model consists of two approaches. They are as follows.

3) THE NFR MODEL: FIRST APPROACH

Highly contributing features are determined by various statistical techniques, for example, the correlation matrix, and weighted least squared (WLS) method. In this approach, a novel method is used to detect irrelevant, very weak, weak, and strong features in the training data. The first approach is based on a heuristic process evaluating performance by reducing features with respect to the improvement in the "AUC" along with the "accuracy" as "additional evaluation metrics", via the five aforementioned ML algorithms until the highly contributing features in the prediction are identified. Irrelevant features are features that do not contribute to the prediction. If they are removed, the accuracy and AUC of the prediction are increased. Very weak features are features that do not facilitate boosting the prediction accuracy and AUC, and no difference is observed if they are deleted or added. Weak features are features that contribute to emending the prediction accuracy and AUC; however, their contribution is substantially less when compared with strong features. If the weak features are deleted, then the prediction accuracy and AUC marginally reduces. The strong features contribute to the prediction accuracy and AUC. If they are deleted, the prediction worsens. Therefore, reduction of the extraneous features: irrelevant and very weak features increases the prediction accuracy and AUC. The time consumed by the process is significantly minimized.

The two approaches of the proposed NFR model are aligned with the five ML algorithms—LR, RF, BRT, SGB, and SVM—and are applied on the four UCI ML repository heart datasets, the Cleveland, Hungarian, Switzerland and Statlog while making a note of the best results. The experiment is conducted on a personal computer (Processor: Quad core (1.73 Giga Hertz) and Boost up to (2.93 Giga Hertz), 8 GB RAM, with 1 GB VRAM). In this work, R (and R Studio), which is a programming language, and an open source, with extensive usage among statisticians and data miners and aimed at data analysis and studies of scholarly literature databases, is utilized. A comparison is made of the results obtained with existing research and it is noted that the proposed model yields better results.

With the first approach of the proposed NFR model, as shown in Figure 1 and Algorithm 1 and 4, the analysis was conducted after preprocessing, i.e., by replacing absent values with the mean of the particular column, as presented in Algorithm 2 and 3, and after balancing the imbalanced dataset i.e., the Switzerland heart dataset using Algorithm 6. By consecutively adopting classification algorithms—LR, RF, BRT, SGB, and SVM—the AUCs and accuracies were noted. The experiment commenced with LR using all features, and the AUC and accuracy were recorded. The last feature was eliminated, and the modifications in the AUC and accuracy were observed. In this case, an increase in the accuracy and AUC was observed. This particular feature was deleted since it did not contribute to improving the accuracy and AUC; it was an irrelevant or redundant feature. Alternatively, if the accuracy and AUC decreased, then that feature was not deleted, which implied that it contributed to the purpose and was regarded as an effective feature, either strongly or weakly relevant. Consequently, if the AUC and accuracy were retained, then this particular feature was deleted, since, it also did not increase the AUC and accuracy. The feature was correspondingly rounded off as an irrelevant or redundant feature. Progress was made to the subsequent feature, and changes in the AUCs and accuracies were recorded. This process was consecutively repeated for all features. With the implementation of this procedure to the LR, RF, BRT, SGB, and SVM algorithms, the AUCs and accuracies were recorded for each feature, and the highest values were recorded.

The same process was applied to the Hungarian, Statlog, and Switzerland datasets, with the exception of obtaining the mean for the absent values of the Hungarian dataset, as the dataset does not contain any of these values, and running the SMOTE algorithm on the Switzerland dataset in order to balance it. Findings from these experiments include a substantial enhancement in AUCs and accuracies when adopting the LR, RF, BRT, SGB, and SVM algorithms with the first approach of the NFR model.

Algorithm 1 NFR Model**: First Approach:** Effective Disease Risk Prediction with reduced set of features with convergence speed on the heart disease datasets

Begin
Input:
Disease dataset, *D*, divided into training & testing set(70:30 ratio) and their prediction column
Set of all features F
Output:
Improved accuracy & AUC of the disease risk prediction with the reduced set of highly contributing features in the prediction.
Feature Reduction Method 1:
Read disease(heart) dataset *D*
Dim(*D*), Summary(*D*) command for checking the dimension and summary of the *D* (for data observation)
D_{MI} ← Data_MeanImputation(*D*)
Dc ← Data_ChangingPredictorValues(D_{MI})
// Splitting the *Dc* into training(D_{tr}) and test(D_{ts}) sets
(D_{tr} , D_{ts}) ← Dataset_SpilittingHoldout (D_c, split_ratio)
Performance(highest) Evaluation Metric P_h = { }
accuracy ← list ( ) //list to record results of ML algorithms accuracy
auc ← list ( ) //list to record results of ML algorithms auc
for ML_DM_Approach :(LR, RF, . . . , SVM) **do**
alg ← NFR_ML_DM_Approach(*i*)
For best reduced feature set f_{best} and ML_DM_Approach
On test set D_{ts} with f_{best}
Evaluation of performance on D_{ts},
P_{ts} (accuracy, auc) ← max (LR, RF, . . . , SVM)
end
F_{hc} ← f_{best}
P_h ← P_{ts} (accuracy, auc)
return (P_h , F_{hc})
End

Algorithm 2 Data_MeanImputation

Begin
Input:
Disease dataset, *D* and the associated prediction column
Output:
Mean imputed dataset D_{MI} without missing values.
(f_a, r_b) ← *D*
$F = \{ f_1, f_2, \ldots, f_a, \ldots, f_x \}$
$R = \{ r_1, r_2, \ldots, r_b, \ldots, r_y \}$
for each $f_a \in F$ where $1 \le a \le x$ **do**
for each $r_b \in R$ where $1 \le b \le y$ **do**
Replace missing values of column f_a with the mean value of the respected column of the *D*
D_{MI} ← Imputed dataset *D*
end
end
return D_{MI}
End

Algorithm 3 Data_ChangingPredictorValues(D_{MI}):

Changing a few predictor values of *D* from integer to factors (making dummies)

Begin

Input:

Disease dataset *D* and the associated prediction column

Output:

Changed dataset D_C after conversion of predictor values

$(f_i, r_j) \leftarrow D$

$F = \{ f_1, f_2, \ldots, f_a, \ldots, f_x \}$

$R = \{ r_1, r_2, \ldots, r_b, \ldots, r_y \}$

In *D*, features 'cp', 'thal', 'restecg', 'slope' need to be dummified as the distances in the values is random

Function (FC_n) for converting classes of predictor values

$FC_n \leftarrow$ function(obj, type) {

FN ← switch(type[*i*], ch = as.character, nm=as.numeric, fc = as.factor){

. . .

Obj[, i] = FN(obj[, i])

}

Obj

}

Cn.names ← function (row){

records $r_b \leftarrow$ gsub("f_a", "value", r_b)

}

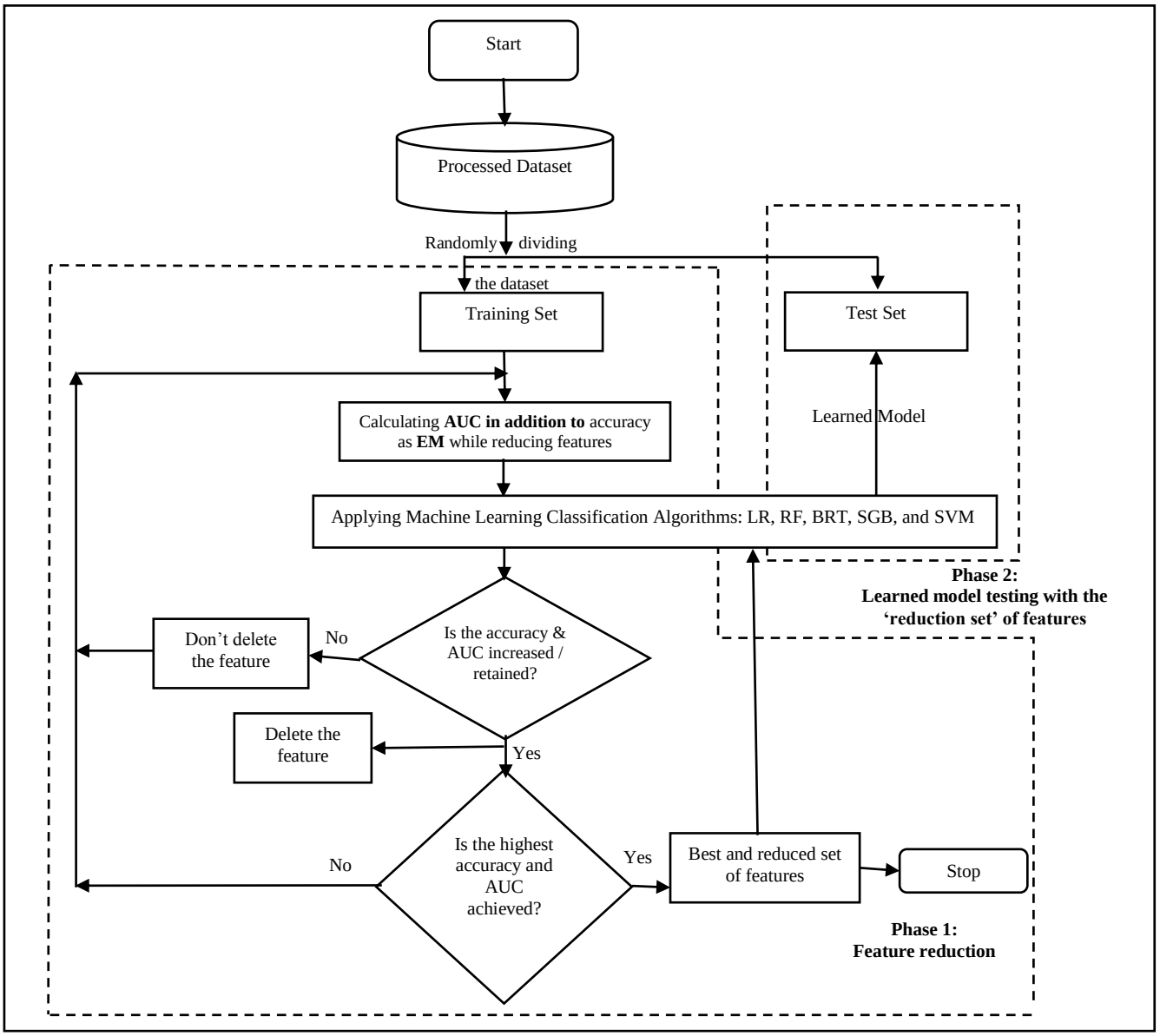

FIGURE 1. **Illustration of the proposed NFR model: First Approach (here 'EM' stands for Evaluation Metrics)**

18: //steps 9 to 17 can also be written as follows i.e., for changing a few predictor values from integer to factors(making dummies)

19: ch.class ← c("numeric", "factor", "numeric", "numeric", "factor", "factor", "numeric", "factor", "numeric", "factor", "factor", "factor", "factor")

20: $D_c \leftarrow C_n.f_n(D, \text{ch.class})$

21: *return* D_c

End

Algorithm 4 *NFR_ML_DM_Approach(i):* Effect of feature reduction on prediction performance P_{ts} on training set D_{tr}

Begin

1: **Input:**

2: Disease training dataset D_{tr} *obtained from* D and the associated prediction column

3: **Output:**

Trained ML_DM_model on D_{tr} with best reduced set features.

4: alg ← ML_DM_Approach (LR, RF, . . . , SVM) **do**

5: $alg_{acc,\ auc}$ ← ML_DM_Approach(LR, RF, . . . , SVM) **do** when $a = 13$ *//performance with all features*

6: **for**($a = 12$; $a <= 13$; a --) **do**

7: Calculating accuracy and AUC

8: $alg_{acc,\ auc}$ ← ML_DM_Approach(LR, RF, . . . , SVM)

9: **if** ($alg_{acc,\ auc}$ increased or no change then) {

10: *delete* (f_a) }

11: **else** {

12: *do not delete* (f_a) }

13: **if** ($alg_{acc,\ auc}$ highest) {

14: $P_{tr} \leftarrow \max(alg_{acc,\ auc})$ && $f_{best} \leftarrow \min(f_a)$ }

15: **end**

16: return P_{tr}

17: return f_{best}

End

4) THE NFR MODEL: SECOND APPROACH

With the advanced second approach of the proposed NFR model, as shown in Figure 2 and Algorithm 5 and 7. The second approach evaluates the accuracy and AUC of all individual features and forms their subsets with the highest accuracies, AUCs, and those with minimal difference attained between the two metrics. These subsets are then combined in various combinations to achieve the best reduced set of highly contributing features. The analysis was conducted on the four heart datasets, commencing with the Cleveland dataset after preprocessing, i.e., by replacing absent values with the mean of the particular column, as shown in Algorithm 2 and 3, and balancing the imbalanced datasets using Algorithm 6. By consecutively adopting classification algorithms—LR, RF, BRT, SGB, and SVM—the AUCs and accuracies were noted for the Cleveland heart dataset. In the second approach of the NFR model, the experiment initiated with LR using the first feature, and the AUC and accuracy were noted. Next, the second feature was applied with LR, and the AUC and accuracy were observed. Progress was made to the subsequent feature and the AUC and accuracy were recorded. This process was consecutively repeated for all features. The features were then grouped into various subsets. The first subset consisted of features with the highest AUC and highest accuracy and least difference between the two metrics, the second subset contained features resulting in the highest AUCs, the third subset included features with the highest accuracies, features that resulted in lower AUCs and accuracies formed the fourth subset, and so on. Subsequently, the AUC and accuracy were calculated with LR by employing all the features in the first subset, and the results were recorded. Similarly, the AUC and accuracy was calculated and noted for each of the second, third, fourth, until the last subset. In the next step, via LR the features of one subset were combined with those in each of the other subsets one at a time, and the AUC and accuracy were calculated and recorded for every such combination. After an extensive course of merging features in various combinations and recording their results, the best set was identified that contained the minimal number of features yet yielded the highest AUC and accuracy, as illustrated in Section V.

With the implementation of this procedure to the LR, RF, BRT, SGB, and SVM algorithms, the AUCs and accuracies were recorded, and the highest values were noted.

Algorithm 5 NFR Model**: Second Approach:** Effective Disease Risk Prediction with the reduced set of features with convergence speed on the heart disease datasets

Begin

1: **Input:**

2: Disease dataset, D, if D is imbalanced then balanced with SMOTE and then divided into training & testing sets (70:30 ratio) and their prediction column

3: Set of all features F

4: **Output:**

Improved accuracy & AUC of the disease risk prediction with reduced set of highly contributing features in the prediction with convergence speed.

5: **Feature Reduction Method 2:**

6: Read disease(heart) dataset D

7: Dim(D), Summary(D) command for checking the dimension and summary of the D (for data

observation)
8: D_{MI} ← Data_MeanImputation(D)
9: D_c ← Data_ChangingPredictorValues(D_{MI})
10: If (D is not balanced dataset){
11 SMOTE(D) }
12 D_{sm} ← D_c
13: // Splitting the D_c into training(D_{tr}) and test(D_{ts}) sets
14: (D_{tr}, D_{ts}) ← Dataset_SpilittingHoldout (D_{sm}, split_ratio)
15: Performance(highest) Evaluation Metric P_h = { }
16: accuracy ← list ( ) //list to record results of ML algorithms accuracy
17: auc ← list ( ) //list to record results of ML algorithms auc
18: **for** ML_DM_Approach_2 :(LR, RF, . . . , SVM) **do**
19: alg ← NFR_ML_DM_Approach_2(i)
20: For best highly contributing reduced feature set f_{best} and ML_DM_Approach_2
21: On test set D_{ts} with f_{best}
22: *Evaluation of performance on* D_{ts}, P_{ts} ← max(LR, RF, . . . , SVM)
23: **end**
24: F_{hc} ← f_{best}
25: P_h ← P_{ts} (accuracy, auc)
26: **return** (P_h , F_{hc})
End

Algorithm 6 SMOTE(D): Balancing the imbalanced dataset with SMOTE (Synthetic Minority Oversampling Technique)

Begin

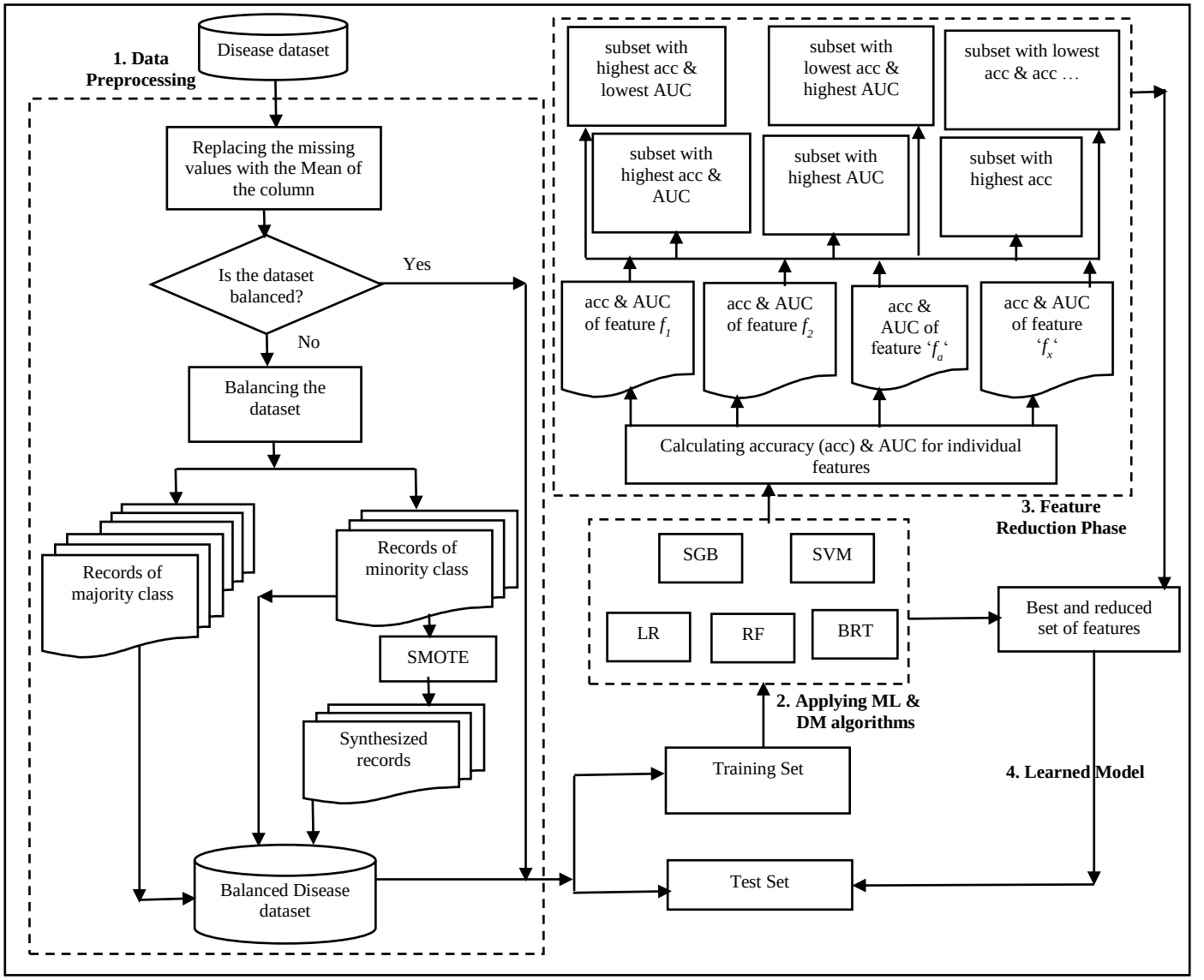

FIGURE 2. **Illustration the proposed NFR model: Second Approach (here 'acc' stands for Accuracy)**

1: **Input:**
2: Disease dataset, D and the associated prediction column
3: **Output:**
Balanced dataset D_{sm} after minority oversampling
4: $(f_i, r_j) \leftarrow D$
5: $F = \{ f_1, f_2, \ldots, f_a, \ldots, f_x \}$
6: $R = \{ r_1, r_2, \ldots, r_b, \ldots, r_y \}$
7: $D_c \leftarrow D$
8: num ← prediction *// prediction column*
9: library(DMwR) // R library used for SMOTE to balance the dataset
10: $D_c\$$ num ← as.factor($D_c\$$ num)
11: D_{sm} ← SMOTE(num ~ ., D_c , prec.over=1300, perc.under=110)
12: $D_{sm}\$$ num ← as.numeric($D_{sm}\$$ *num)*
13: print(prop.table(table(D_{sm} $ num))) // verifying that the dataset is balanced
14: return D_{sm}
End

Algorithm 7 *NFR_ML_DM_Approach_2(i):* Effect of feature reduction on prediction performance P_{ts} on training set D_{tr}

Begin
1: Input:
2: Disease training dataset D_{tr} *obtained from* D_{sm} and the associated prediction column
3: **Output:**
Trained/learned ML_DM_model on D_{tr} with best highly contributing reduced set features.
4: alg ← ML_DM_Approach_2(LR, RF, . . . , SVM) **do**
5: $alg_{acc,\ auc}$ ← ML_DM_Approach_2(LR, RF, . . . , SVM) **do** when $a = 13$ *//performance with all features*
6: **for***(a =13 ; a <=13 ; a --)* **do**
7: Calculating accuracy and AUC for every feature
8: $alg_{acc,\ auc}$ ← ML_DM_Approach(LR, RF, . . . , SVM)
9: subset1 ← max(acc and auc of f_a)
10: subse2 ← max(acc of f_a)
11: subse3 ← max(auc of f_a)
12: subse4 ← max(acc of f_a) && low (auc of f_a)
13: subse5 ← low(acc of f_a) && max (auc of f_a)
14: subse6 ← low(auc of f_a) && max (acc of f_a)
15: . . . *//various combination of features, and evaluating highly contributing features in the prediction*
16: **if** ($alg_{acc,\ auc}$ highest) {
18: P_{tr} ← max ($alg_{acc,\ auc}$) && f_{best} ← min *(f_a)*
19: *}*
20: **end**
21: return P_{tr}
22: return f_{best}
End

The same process was applied to the Hungarian, Statlog, and Switzerland datasets, with the exception of obtaining the mean for the absent values of the Hungarian dataset, as the dataset does not contain any of these values, and running the SMOTE algorithm on the Switzerland dataset in order to balance it. Findings from these experiments, as listed in detail in Section V, also include a substantial enhancement in AUCs and accuracies when adopting the LR, RF, BRT, SGB, and SVM algorithms with the first approach of the NFR model.

5) RUN TIME OF THE ALGORITHMS WITH AND WITHOUT FEATURE REDUCTION

The run times of the LR, RF, BRT, SGB, and SVM algorithms were recorded, as shown in Table III, and a comparison was made with and without the first and second approach of the NFR model.

V. RESULTS AND ANALYSIS

A. PERFORMANCE OF THE NFR MODEL

The performance of both the approaches of the NFR model with the ML algorithms is tested with the UCI datasets—Cleveland, Hungarian, Statlog, and Switzerland—and the highest AUCs and accuracies are recorded.

The feature reduction subsets obtained for the Cleveland heart dataset (*C*), Hungarian heart dataset (*H*), Statlog heart dataset (*S*), and Switzerland heart dataset (*W*) succeeding the alignment of the proposed model with the ML algorithms are drafted here. The subgroups formed with the NFR model range from the count of 4 to 11 features, with the feature pt_cp indicated as the strongest and most important feature. The features pt_ca, pt_thal, pt_slope, and pt_oldpeak are the most relevant features. The features pt_age, pt_sex, pt_exang, and pt_thalach are the next set of relevant features, while pt_trestbps, pt_restecg and pt_chol are the least repetitive features. The highest accuracies and AUCs are attained by using pt_ca, pt_slope, pt_age, pt_sex, pt_cp, pt_exang, pt_thalach, and pt_oldpeak, which can be concluded as a strong set of features.

The feature reduction subsets and results obtained with the proposed NFR model for the Cleveland heart dataset (*C*) are shown in Table III and Figures 3, 5, and 9.

C_1 = {pt_ca, pt_slope, pt_sex, pt_cp, pt_thalach, pt_thal, pt_exang, pt_oldpeak, pt_restecg}

C_2 = {pt_ca, pt_thal, pt_oldpeak, pt_cp, pt_age, pt_sex,

pt_slope, pt_chol}

C_3 = {pt_ca, pt_slope, pt_age, pt_sex, pt_cp, pt_chol, pt_thal, pt_exang, pt_fbs, pt_thalach, pt_oldpeak}

C_4 = {pt_ca, pt_slope, pt_age, pt_sex, pt_cp, pt_thal, pt_exang, pt_thalach, pt_oldpeak, pt_restecg}

C_5 = {pt_ca, pt_thal, pt_oldpeak, pt_cp, pt_age, pt_trestbps}

C_6 = {pt_ca, pt_thal, pt_oldpeak, pt_cp, pt_exang, pt_slope, pt_thalach}

C_7 = {pt_ca, pt_thal, pt_oldpeak, pt_cp, pt_thalach, pt_slope, pt_chol}

C_8 = {pt_ca, pt_thal, pt_oldpeak, pt_cp, pt_age, pt_sex, pt_trestbps, pt_exang, pt_slope, pt_restecg, pt_thalach}

All features with LR yielded an AUC of 91.62% and an accuracy of 86.52%, which increased to 92.68% and 92.53%, respectively, when the feature reduction subset C_1 that consists of the top nine features is applied via the NFR model. The NFR model with the BRT caused an increase in the AUC from 90.96% to 93.68% and, an increase in the accuracy from 84.27% to 88.89% for the ten features in C_4, and it increased the AUC to 92.68% and the accuracy to 85.56% via the top six features in C_5. On applying the NFR model with an RF, the best-performing eight features in C_2 produced an increase in the AUC from 89.53% to 92.63% and an increase in the accuracy from 80.89% to 87.78%. For the top eleven feature in C_3 an increase in the accuracy of 87.78% and an increase in the AUC of 94.18% was achieved. Furthermore, aligning the SGB with the proposed model increased the AUC from 90.14% to 91.04%, and the accuracy increased from 80.89% to 86.67% for the seven best features in C_6. After adopting SVM with the NFR model, an increase in the AUC from 88.26% to 90.43% and an increase in the accuracy from 79.78% to 83.33% was achieved with the seven best features in C_7. For the eleven features in C_8, 92.23% AUC and 87.78% accuracy were successfully attained.

Similarly, feature reduction subsets and their results for the Hungarian heart dataset (*H*), as shown in Table III and Figures 4, 5, and 9 are as follows:

H_1 = {pt_ca, pt_age, pt_slope, pt_cp, pt_fbs, pt_thal pt_exang, pt_sex}

H_2 = {pt_cp, pt_oldpeak, pt_sex, pt_thal, pt_chol, pt_age, pt_ca}

H_3 = {pt_exang, pt_cp, pt_age, pt_restbps}

H_4 = {pt_ca, pt_sex, pt_thal, pt_oldpeak, pt_cp, pt_age, pt_trestbps}

H_5 = {pt_ca, pt_sex, pt_thal, pt_oldpeak, pt_cp, pt_exang, pt_slope, pt_thalach}

$H_6 = C_4$

The top eight features in H_1 caused an increased AUC from 83.99% to 92.51%, and the accuracy increased from79.31% to 83.91% via LR with the NFR model. The application of an RF with the proposed model increased the AUC from 88.08% to 90.24% and increased the accuracy from 81.61% to 82.76% when the top seven features in H_2 were considered. The BTR with NFR model yielded an increase in AUC from 81.74% to 86.72% and an increase in accuracy from 79.31% to 85.06% for the four features in H_3 and an increase in AUC and accuracy of 89.14% and 83.91%, respectively, for the seven features in H_4. The application of SGB to the NFR model produced an increase in the AUC from 87.79% to 91.47% and an increase in accuracy from 81.61% to 85.06% for the eight best features in H_5. Moreover, using the proposed model with SVM, an increase in the AUC from 87.79% to 91.13% and an increase in the accuracy from 82.76% to 85.06% was steadily achieved for the ten best features in H_6.

The feature reduction subsets and results obtained with the proposed NFR model for the Statlog heart dataset (*S*), as shown in Table III and Figures 6, 8, and 9, are as follows:

S_1 = {pt_ca, pt_slope, pt_ age, pt_sex, pt_cp, pt_chol, pt_thal, pt_exang, pt_thalach, pt_trestbps, pt_oldpeak}

S_2 = {pt_ca, pt_slope, pt_age, pt_sex, pt_cp, pt_chol, pt_thal, pt_thalach, pt_trestbps}

S_3 = {pt_ca, pt_age, pt_cp, pt_thal, pt_sex, pt_oldpeak, pt_trestbps, pt_restecg, pt_slope}

S_4 = {pt_ca, pt_age, pt_slope, pt_cp, pt_fbs, pt_thal, pt_sex, pt_oldpeak, pt_trestbps}

S_5 = {pt_ca, pt_sex, pt_thal, pt_cp, pt_age, pt_trestbps, pt_restecg, pt_thalach, pt_slope, pt_fbs, pt_chol, pt_exang}

S_6 = {pt_ca, pt_exang, pt_thal, pt_cp, pt_age, pt_trestbps, pt_restecg, pt_thalach, pt_slope, pt_chol}

S_7 = {pt_ca, pt_thal, pt_oldpeak, pt_cp, pt_sex, pt_trestbps, pt_restecg, pt_thalach}

S_8 = {pt_ca, pt_thal, pt_slope, pt_oldpeak, pt_cp, pt_sex, pt_trestbps, pt_thalach}

S_9 = {pt_ca, pt_thal, pt_slope, pt_oldpeak, pt_cp, pt_age, pt_sex, pt_trestbps, pt_chol, pt_thalach}

All features with LR yielded an AUC of 89.81% and an accuracy of 83.95%, which increased to 90.12% and 85.19%, respectively, when the feature reduction subset S_1 that consists of the top eleven features is applied via the NFR model. The NFR model with the BRT caused an increase in the AUC from 81.74% to 91.79% and, an increase in the accuracy from 79.01% to 83.95% for the twelve features in S_5, and it increased the AUC to 90.83% and the accuracy to 83.95% via the top ten features in S_6. On applying the NFR model with an RF, the nine features in S_3 maintained the AUC as 89.38% and produced an increase in the accuracy from 85.19% to 87.65%.

Furthermore, aligning the SGB with the proposed model increased the AUC from 87.78% to 90.12%, and the accuracy increased from 79.01% to 86.42% for the eight best features in S_7. After adopting SVM with the NFR model, the AUC dropped slightly from 91.17% to 90.31% but an increase in the accuracy from 83.95% to 87.65% was achieved with the eight best features in S_8. For the ten

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED NFR MODEL ON THE **UCI ML DISEASE** DATASETS

Dataset	ML Algorithms	Performance metrics							
		AUC (%)		Accuracy (%)		Time(min)		No. of features	
		Without NFR	With Proposed NFR Model	Without NFR	With Proposed NFR Model	Without NFR	With Proposed NFR Model	Without NFR	With Proposed NFR Model
Cleveland	LR	91.62	**92.68**	86.52	**92.53**	0.05	0.05	13	**9**
	RF	89.53	**92.63**	80.89	**87.78**	0.06	**0.05**	13	**8**
	BRT	90.96	**93.68**	84.27	**88.89**	0.95	**0.77**	13	**10**
	SGB	90.14	**91.04**	80.89	**86.67**	9.20	**6.70**	13	**7**
	SVM	88.26	**92.23**	79.78	**87.78**	2.00	**1.90**	13	**11**
Hungarian	LR	83.99	**92.51**	79.31	**83.91**	0.08	**0.06**	13	**8**
	RF	88.08	**90.24, 89.23**	81.61	**82.76, 83.91**	0.06	**0.03, 0.04**	13	**7, 8**
	BRT	81.74	**86.72, 89.14**	79.31	**85.06, 83.91**	1.00	**0.47, 0.62**	13	**4, 7**
	SGB	87.79	**91.47**	81.61	**85.06**	9.60	**6.80**	13	**8**
	SVM	87.79	**91.13**	82.76	**85.06**	1.90	**1.70**	13	**10**
Statlog	LR	89.81	**90.12,** 88.64	83.95	**85.19, 85.19**	0.05	0.05	13	**11, 9**
	RF	89.38	**89.38,** 88.7	85.19	**87.65, 86.42**	0.05	**0.04**	13	9, 9
	BRT	81.74	**91.79, 90.83**	79.01	**83.95, 83.95**	1.01	**0.77**	13	**12, 10**
	SGB	87.78	**90.12**	79.01	**86.42**	7.66	**6.05**	13	**8**
	SVM	91.17	90.31, **91.11**	83.95	**87.65, 85.19**	1.82	**1.62**	13	**8, 10**
Switzerland	LR	88.55	**95.45**	88.06	**91.05**	0.54	**0.33**	12	**6**
	RF	79.95	**92.02, 93.76**	68.66	**86.57, 89.55**	0.005	**0.004, 0.004**	**9**	**2, 1**
	BRT	95.54	**99.2**	92.54	**95.52**	5.41	**4.06**	12	**7**
	SGB	95.81	**99.02**	92.54	**94.03**	52.87	**37.94**	12	**7**
	SVM	85.56	**91.76, 89.71**	65.67	**91.05, 92.54**	7.04	**2.55, 2.53**	12	**3, 3**

features in S_9, 91.11% AUC and 85.19% accuracy were successfully attained.

Similarly, feature reduction subsets and their results for the Switzerland heart dataset (W), as shown in Table III and Figures 7, 8, and 9, are as follows:

W_1 = {pt_exang, pt_ca, pt_slope, pt_thalach, pt_cp, pt_sex}
W_2 = {pt_age, pt_oldpeak}
W_3 = {pt_oldpeak}
W_4 = (pt_slope, pt_exang, pt_cp, pt_sex, pt_ca, pt_oldpeak, pt_thalach}
W_5 = {pt_oldpeak, pt_slope, pt_exang, pt_cp, pt_sex, pt_ca, pt_thalach}
W_6 = {pt_oldpeak, pt_exang, pt_ca}
W_7 = {pt_oldpeak, pt_exang, pt_cp}

The top six features in W_1 caused an increased AUC from 88.55% to 95.45%, and the accuracy increased from 88.06% to 91.05% via LR with the NFR model. The application of an RF with the proposed model increased the AUC from 79.95% to 92.02% and increased the accuracy from 68.66% to 86.57% when the top two features in W_2 were considered. With the best-performing one feature in W_3, the AUC increased to 93.76% and the accuracy increased to 89.55% successfully. The BTR with NFR model yielded the highest increase in AUC from 95.54% to 99.20% and an increase in accuracy from 92.54% to 95.52% for the seven features in W_4. The application of SGB to the NFR model produced an increase in AUC from 95.81% to 99.02% and an increase in accuracy from 92.54% to 94.03% for the seven best features in W_5. Moreover, using the proposed model with SVM, an increase in the AUC from 85.56% to 91.76% and an increase in the accuracy from 65.67% to 91.05% was steadily achieved for the three best features in W_6 and, an increase in AUC and accuracy of 89.71% and 92.54% was attained, respectively, for the three features in W_7. A comparison of the results obtained with the NFR model with the results obtained without the alignment of the proposed model indicated that the highest improvement recorded in the prediction accuracy, AUC, and feature reduction is 8%, 3.97%, and 46.15%, respectively, for the Cleveland heart dataset. Similarly, for the Hungarian heart dataset, the maximum improvement achieved in accuracy, AUC, and feature reduction was 5.75%, 8.52%, and 69.23%, respectively. Likewise, for the Statlog heart dataset, the maximum improvement achieved in accuracy, AUC, and feature reduction was 7.41%, 10.5%, and 38.46%, respectively. Lastly, for the Switzerland heart dataset, the maximum improvement achieved in accuracy, AUC, and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3028714, IEEE Access

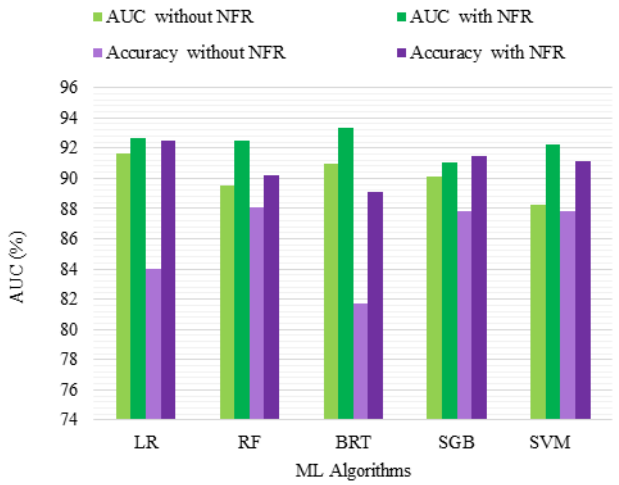

FIGURE 3. AUC and Accuracy Comparison of the proposed NFR model on the **Cleveland** dataset

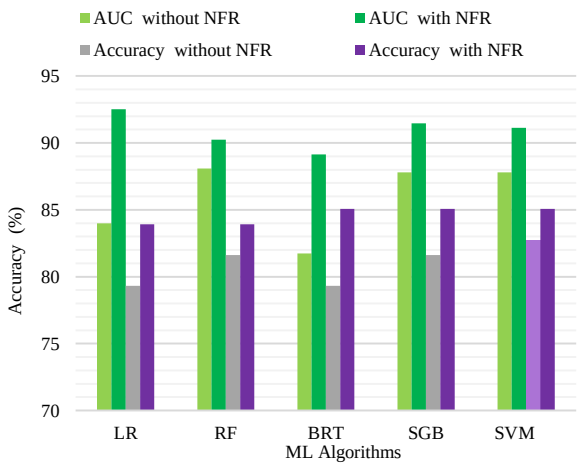

FIGURE 4. AUC and Accuracy Comparison of the proposed NFR model on the **Hungarian** dataset

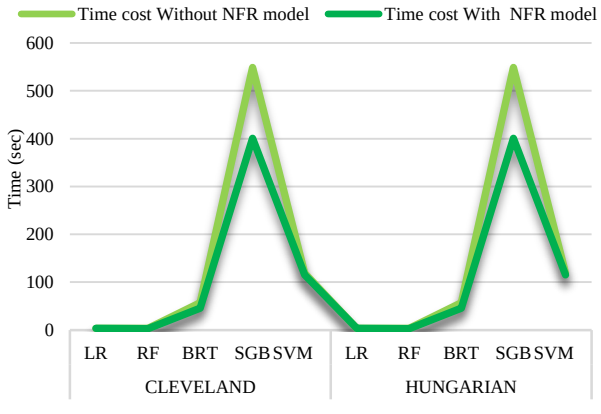

FIGURE 5. Run Time Comparison of the proposed NFR model on the **Cleveland** and **Hungarian** datasets

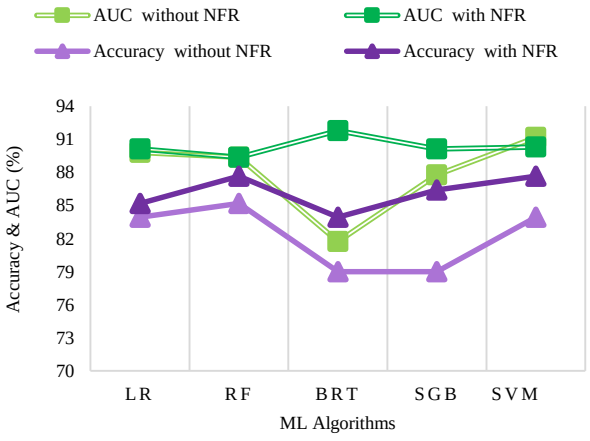

FIGURE 6. Accuracy and AUC Comparison of the **Statlog** dataset with the proposed NFR model

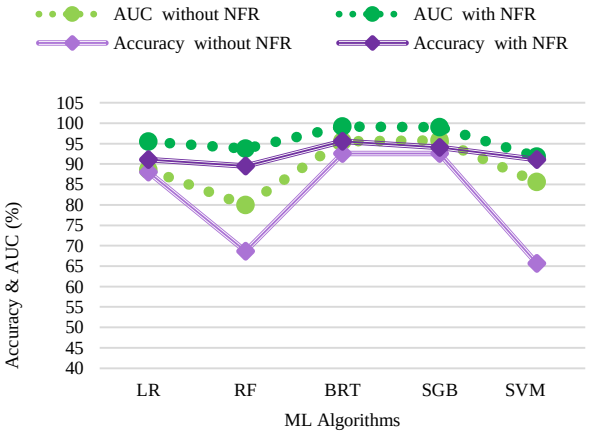

FIGURE 7. Accuracy and AUC Comparison of the **Switzerland** dataset with the proposed NFR model

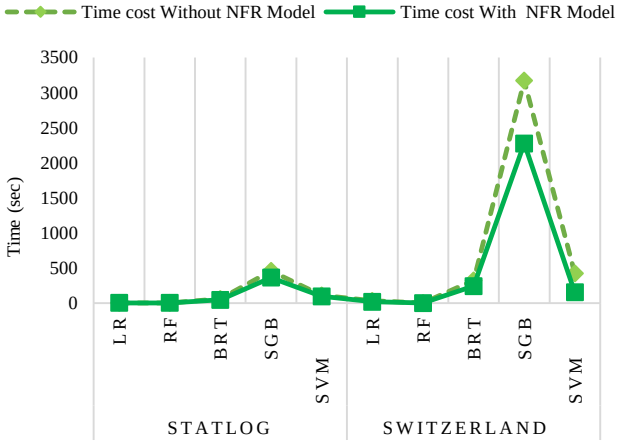

FIGURE 8. Run Time Comparison of the proposed NFR model on the **Statlog** and **Switzerland** datasets

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3028714, IEEE Access

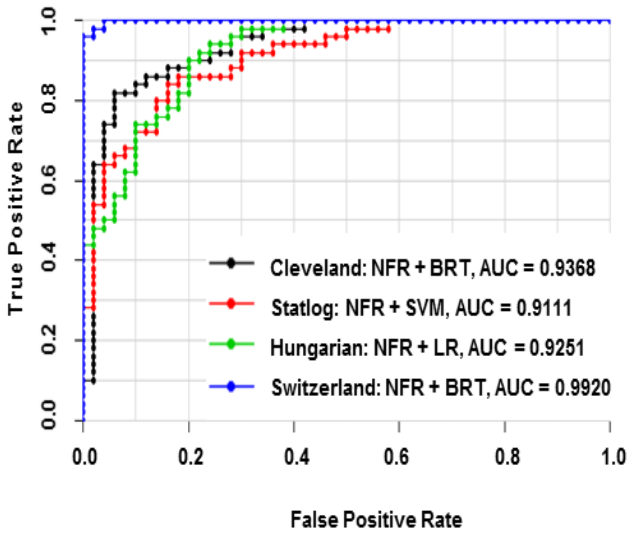

FIGURE 9. AUC Comparison of the UCI ML datasets with the proposed NFR model

feature reduction was 26.87%, 13.81%, and 91.66%, respectively, according to a comparison of the results achieved by the NFR model with the results attained without the proposed model.

It was also noted that the order of placing the features had an effect on the performance of the ML algorithms. When particular features were placed in the first positions, an increase in the percentage of the AUC and accuracy was observed, while positioning other features in the first brought a decrease in the results. This aspect of assigning and allocating the features in specific arrangements to maximize the prediction can be explored in the future.

For instance, after adopting SVM with the second approach of the NFR model, the AUC achieved with S_8 (ii) {pt_ca, pt_thal, pt_slope, pt_oldpeak, pt_cp, pt_age, pt_sex, pt_trestbps} was 90.31% and an accuracy of 83.95% was achieved with a running time of 1.636477 min. The same eight features with altered positions as in S_8 (iii) {pt_ca, pt_thal, pt_oldpeak, pt_cp, pt_age, pt_sex, pt_trestbps, pt_thalach}, resulted in an AUC of 90.74% and an accuracy of 85.19% with a running time of 1.608142 min.

B. RUN TIME COMPARISON

The running time, with and without the NFR model, is calculated and compared for LR, RF, BRT, SGB, and SVM algorithms. The time required for each of the algorithms adopted in the experiments was calculated with and without the NFR model. The outcomes were compared for the Cleveland, Hungarian, Statlog, and Switzerland heart datasets and are described in Table III and Figure 5 and 8. A considerable time reduction was procured when the NFR model was applied. With LR, the time listed for all features was 3.31 sec. The time with the NFR model for the Cleveland dataset declined to 3.09 sec for the nine features listed in C_1. For the Hungarian dataset, the time decreased from 5.01 sec to 3.28 sec for the eight features in H_1. Similarly, the NFR model with RF reduced the time from 3.38 sec to 2.14 sec for the top eight features in C_2 and decreased from 3.38 sec to 1.91 sec for the seven features in H_2. Subsequently, the BRT with the proposed model decreased the time from 56.70 sec to 45.89 sec for the top ten features in C_4 and decreased from 1.01 min to 28.33 sec for the four features in H_3. Aligning SGB with the NFR model, the time decreased from 9.15 min to 6.68 min for the top seven features in C_6 and decreased from 9.62 min to 6.81 min for the eight features in H_5. Further, in the prediction model with SVM, the run time recorded for all features was 1.98 min, which declined to 1.92 min for the top eleven features in C_8 and decreased from 1.99 min to 1.75 min for the top ten features in H_6.

With LR, the time listed for all features was 3.22 sec. The time with the NFR model for the Statlog dataset declined to 3.12 sec for the nine features listed in S_2. For the Switzerland dataset, the time decreased from 32.26 sec to 19.97 sec for the six features in W_1. Similarly, the NFR model with RF reduced the time from 3.21 sec to 2.19 sec for the top nine features in S_3 and decreased from 0.30 sec to 0.26 sec for one feature in W_3. Subsequently, the BRT with the proposed model decreased the time from 60.75 sec to 46.05 sec for the top ten features in S_6 and decreased from 324.30 sec to 243.31 sec for the seven features in W_4. Aligning SGB with the NFR model, the time decreased from 459.78 sec to 362.75 sec for the top eight features in S_7 and decreased from 3172.30 sec to 2276.68 sec for the seven features in W_5. Further, in the prediction model with SVM, the run time recorded for all features was 109.16 sec, which declined to 97.47 sec for the top eight features in S_8 and decreased from 422.16 sec to 153.25 sec for the top three features in W_6. The highest improvement was achieved for SGB, with a reduction of 14.93 min.

C. BENCHMARKING OF THE PROPOSED MODEL

The NFR model is benchmarked for performance with previous work on disease risk prediction using the Cleveland, Hungarian, Statlog, and Switzerland heart datasets. The achievements of the proposed and existing models, with accuracy as the criterion, are discussed here. However, the proposed model has assessed the AUC with accuracy, which has been prescribed as a standard measure in risk assessment by the medical field. Previously, many of the associated works have not characterized its role. Nonetheless, a higher AUC ascertains the meticulousness of the accuracy.

Tables IV, V, VI, and VII illustrate how the proposed model performs better than existing research. For the Cleveland heart dataset, the maximum improvement in the prediction accuracy achieved by the NFR model is 15.87% with 30.76% reduction in features, when compared with recent research studies between 2015 and 2020. The minimum improvement in accuracy obtained by the proposed model is 2.02% with 30.76% feature reduction. A comparison of recent work of the Hungarian heart dataset

indicates that the maximum improvement in the recorded prediction accuracy is 7.66% and the maximum feature reduction is 38.46%. Furthermore, the minimum improvement in the attained accuracy is 0.54% with 69.23% feature reduction. For the Statlog heart dataset, the maximum improvement in the prediction accuracy achieved by the NFR model is 7.41% with 38.46% reduction in features when compared with recent research studies between 2015 and 2020. The minimum improvement in accuracy obtained by the proposed model is 1.24% with 30.76% feature reduction. A comparison of recent work of the Switzerland heart dataset indicates that the maximum improvement in the recorded prediction accuracy is 26.87% and the maximum feature reduction is 91.66%. Furthermore, the minimum improvement in the attained accuracy is 1.49% with 41.67% feature reduction.

For the Cleveland heart dataset, as shown in Table IV, the NFR model aligned with LR yielded an accuracy of 92.53%, which is the highest among all twelve existing works listed here, the recorded AUC is 92.68%, and the attained feature reduction is 30.78%. The proposed model with an RF attained 87.78% accuracy, 92.63% AUC, and a feature reduction of 38.46%. These results exceed those of eight of previous studies with regard to risk prediction. Likewise, for the NFR model with BRT, 88.89% accuracy is achieved, the AUC is 93.68%, and the feature reduction is 23.08%. These results are better than those of nine existing works. The AUC acquired by the BRT is the highest recorded AUC among all

TABLE IV

COMPARISON OF THE PROPOSED MODEL WITH PREVIOUS WORK ON THE **CLEVELAND** DATASET

Source	Approach	Accuracy (%)	AUC (%)	No. of Features
Xu *et al. (*2015) [24]	LS-SVM classifier	83.70	-	13
El-Bialy *et al. (*2015) [10]	FDT and C4.5	78.54	-	4
Paul *et al.* (2016) [11]	FDSS using GA	80.00	-	8
Mokeddem *et al.* (2017) [25]	Fuzzy CDSS	90.51	-	13
Sabahi *et al.* (2018) [26]	BFAHP	87.31	-	13
Shilaskar *et al.* (2018) [27]	CMB classification	77.50	-	13
Burse *et al.* (2018) [12]	SVM-LDA	88.32	-	4
Amin *et al.(*2019) [13]	Vote	87.41	-	9
Maji *et al.* (2019) [28]	ANN	77.40	-	13
Maji *et al.* (2019) [28]	C4.5	76.66	-	13
Moloud *et al.(*2019) [29]	NE-Nu-SVC + Feature Selection + Sigmoid	84.51	89.00	13
Senthilkumar *et al.(*2019) [30]	HRFLM	88.40	-	13
Liaqat *et al.(*2019) [31]	χ^2 Statistical Model + DNN (k-fold)	91.57	-	13
Hager *et al.* (2019) [14]	Univariate + LR	88.40	-	7
Hager *et al.* (2019) [14]	Relief + LR	89.9	-	7
Beulah *et al.*(2019) [15]	Majority vote with NB, BN, RF, and MP	85.48	-	11
Thippa *et al.(*2019) [32]	AGAFL	90.00	-	13
Gupta *et al.(*2020) [16]	FAMD + RF	91.80	91.61	10
Kanti *et al.(*2020) [33]	Hybrid Model (PRISM + PGA)	89.81	91.10	13
Proposed NFR model	**NFR model + LR**	**92.53**	**92.68**	**9**

ML algorithms when applied with the NRF model. Furthermore, the proposed prediction model with SGB produced an accuracy of 86.67% and an AUC of 91.04% with the highest feature reduction of 46.15%. An accuracy of 87.78%, AUC of 92.23%, and feature reduction of 15.38% is procured with the NFR model for alignment with an SVM.

For the Hungarian heart dataset, as shown in Table V, the NFR model with SGB acquired the best results of 85.06% accuracy, 91.47% AUC, and a feature reduction of 38.46%. These results are higher than the nine previous studies provided in this paper. Aligning an SVM with the proposed model, an accuracy of 85.06%, AUC of 91.13%, feature reduction of 23.08% are achieved, which is also higher than nine existing works. Aligning LR with the NFR model produced 83.91% accuracy, 92.51% AUC, and a feature reduction of 38.46%. This result outperforms five of the previous associated works. The proposed prediction model with the BRT yielded 85.06% accuracy, while the AUC is 86.72% and the procured feature reduction is 69.23%, which is the highest reduction rate achieved among all models. The NFR model with RF attained 83.91% accuracy, 89.23% AUC, and 38.46% feature reduction.

For the Statlog heart dataset, as shown in Table VI, the NFR model aligned with SVM yielded an accuracy of 87.65%, which is the highest among all fourteen existing works listed here, the recorded AUC is 90.31%, and the attained feature reduction is 38.46%. The proposed model with an RF attained 87.65% accuracy, 89.38% AUC, and a feature reduction of 30.76%. These results also exceed all the previous studies with regard to risk prediction. The AUC acquired by the BRT is 91.79% that is the highest recorded AUC among all ML algorithms when applied with the NRF model. Furthermore, the proposed prediction model with SGB produced an accuracy of 86.42% and an AUC of 90.12% with the highest feature reduction of 38.46%. An accuracy of 85.19%, AUC of 90.12%, and feature reduction of 30.76% is procured with the NFR model for alignment with a LR.

TABLE V
COMPARISON OF THE PROPOSED MODEL WITH PREVIOUS WORK ON THE **HUNGARIAN** DATASET

Source	Approach	Accuracy (%)	AUC (%)	No. of Features
Fernandez-Delgado et al. (2014) [35]	Random Forest (53)	81.60	-	13
El-Bialy *et al. (*2015) [10]	C4.5	78.54	-	6
Mokeddem *et al*. (2018) [25]	Fuzzy CDSS	85.71	-	13
Shah *et al*. (2017) [17]	PPCA and RBF kernel-based SVM	85.82	-	8
Sabahi *et al*. (2018) [26]	BFAHP	86.57	-	13
Saqlain et al. (2019) [19]	MFSFSA, forward, and reverse feature selection algorithms	84.52	-	7
Repaka *et al*. (2019) [36]	MLP and AES	77.40	-	13
Repaka *et al*. (2019) [36]	SMO and AES	84.07		13
Repaka *et al*. (2019) [36]	BN and AES	81.11		13
Tanveer *et al*. (2019) [37]	TWSVM_m(TWSVM)	79.60	-	13
Perales-González *et al.* (2019) [38]	Hierarchical ensemble: BRELM	81.28	-	13
Proposed NFR model	**NFR model + SGB**	**85.06**	**91.47**	**8**

TABLE VI
COMPARISON OF THE PROPOSED MODEL WITH PREVIOUS WORK ON THE **STATLOG** DATASET

Source	Approach	Accuracy (%)	AUC (%)	No. of Features
Shuo *et al.*(2018) [39]	Improved ID3	77.78	-	13
Archana *et al.*(2018) [20]	PF-FELM	87.65	-	11

Amin *et al.*(2019) [13]	Vote	87.41	-	9
Shichao *et al.*(2019) [40]	Direct-CS-KNN	-	76.42	13
Shichao *et al.*(2019) [40]	Distance-CS-KNN	-	76.88	13
Soraya *et al.*(2019) [41]	MRMD-II-MLP	84..07	-	13
Soraya *et al.*(2019) [41]	MRMD-II-SVM	82.59	-	13
Soraya *et al.*(2019) [41]	MRMD-II-J48	84.81	-	13
Soraya *et al.*(2019) [41]	MRMD-II	85.58	-	13
Minghua *et al.*(2019) [42]	SSRFLE	78.68	-	13
Al-Attar *et al.(*2019) [21]	PPA	86.17	-	4
Yijie *et al.(*2019) [43]	FSVM-LNR	84.81	-	13
Zahangir *et al.(*2019) [23]	RF and ReliefAttributeEval	86.90	83.3.	12
Himansu *et al.(*2020) [44]	NF-LDA	85.83	-	13
Proposed NFR model	**NFR model + SVM**	**87.65**	**90.31**	**8**
Proposed NFR model	**NFR model + RF**	**87.65**	**89.38**	**9**

TABLE VII
COMPARISON OF THE PROPOSED MODEL WITH PREVIOUS WORK ON THE **SWITZERLAND** DATASET

Source	Approach	Accuracy (%)	AUC (%)	No. of Features
Anooj *et al.*(2012) [45]	Fuzzy rule-based CDSS	51.22	-	13
Fernandez-Delgado *et al.* (2014) [35]	Random Forest	39.80	-	13
Fernandez-Delgado *et al.* (2014) [35]	J48	33.30	-	13
Fernandez-Delgado *et al.* (2014) [35]	C5.0 Tree	32.50	-	13
Fernandez-Delgado *et al.* (2014) [35]	DKP_C	33.90	-	13
Fernandez-Delgado *et al.* (2014) [35]	SVM_C	35.50	-	13
Saqlain *et al.* (2017) [17]	PPCA	91.30	-	13
Animesh *et al.* (2018) [46]	Weighted Fuzzy system ensemble	89.47	-	13
Sabahi *et al.*(2018) [26]	BFAHP	85.22	-	13
Thippa *et al.(*2019) [32]	AGAFL	89.00	-	13
Kanti *et al.(*2020) [33]	Hybrid Model (PRISM + PGA)	57.82	88.30	13
Proposed NFR model	**NFR model + LR**	**91.05**	**95.45**	**6**
Proposed NFR model	**NFR model + BRT**	**95.52**	**99.20**	**7**

Proposed NFR model	**NFR model + SGB**	**94.03**	**99.02**	**7**
Proposed NFR model	**NFR model + SVM**	**91.05**	**91.76**	**3**

For the Switzerland heart dataset, as shown in Table VII, the NFR model with BRT acquired the best results of 95.52% accuracy, 99.2% AUC, and a feature reduction of 41.67%. These results are highest among all the eleven previous studies provided in this paper. Aligning an SGB with the proposed model, an accuracy of 94.03%, AUC of 99.02%, feature reduction of 41.67% are achieved, which is also higher than all the existing works. Aligning LR with the NFR model produced 91.05% accuracy, 95.45% AUC, and a feature reduction of 50%. This result also outperforms all of the previous associated works listed. The proposed prediction model with the SVM yielded 91.05% accuracy, while the AUC is 91.76% and the procured feature reduction is 75%. The NFR model with RF attained 89.55% accuracy, 93.76% AUC, and 91.66% feature reduction, which is the highest reduction rate achieved among all models.

VI. CONCLUSION AND FUTURE WORK

In this paper, the NFR model is proposed to improve the performance of the prediction of disease risk that is aligned with the ML algorithms for datasets of healthcare, thus facilitating to diminish the error rate of 12% that persists in the diagnosis even by medical experts. The NFR model uses AUC with the accuracy as evaluation metrics, which has not been used notably in existing studies. The proposed model comprises of two approaches. The first approach is based on a heuristic process, in which the performance is evaluated by carrying out feature reduction with respect to the improvement in the AUC and the accuracy together as evaluation metrics, producing the best subset of highly contributing features in the prediction. In the second approach, the AUC and accuracy are calculated for each individual feature of the disease datasets and the features are grouped into various subsets. These subsets are then superimposed on one another and the accuracy and AUC are calculated each time until the highest results are achieved. The five ML algorithms—LR, RF, SVM, BRT, and SGB—are aligned with the proposed model, and the most efficient reduced set of features is obtained with each of the algorithms. With the reduced number of features, it is inferred that the disease risk prediction considerably increases when measured with the accuracy and AUC for the UCI heart datasets. Compared with several existing studies, the proposed model achieves better performance. On applying the two approaches of the NFR model to the four datasets, the Switzerland heart dataset yields the best results with a maximum accuracy of 95.52% and a maximum AUC of 99.2% was achieved using the BRT algorithm with 41.67% feature reduction. A 25% of performance improvement was achieved in the run time of the algorithm.

The NFR model enables an effective analysis of medical data, which aids in speedy disease detection and patient care. By choosing the reduced and best features set, the suggested model lessens the number of diagnostic tests and becomes economical for patients, which expands the coverage of the community and increases the chances of saving lives. In the future, the NFR model can be conjointly applied with other classification algorithms and clustering algorithms for improved performance in effective disease risk prediction. Likewise, the NFR model can be implemented on datasets for an extensive range of diseases. Correspondingly, the model can be drafted into a web-enabled and smartphone application that is used for better assistance and service to healthcare communities.

ACKNOWLEDGMENTS

We would like to thank our Dean (Research), Dr. Raja Muhammad and the B.S. Abdur Rahman Crescent Institute of Science and Technology (Deemed to be University), Chennai, India for their support through PhD fellowship. This study employed the Data Science Research Center and Statistical Databases lab facility of the B.S. Abdur Rahman Crescent Institute of Science and Technology (Deemed to be University), Chennai, India. We extend our appreciation to Dr. Angelina Geetha for her valuable suggestions and review of this paper. We also extend our gratitude to the editor and anonymous reviewers of this journal for their suggestions.

REFERENCES

[1] "WHO | Cardiovascular diseases (CVDs)," *WHO*, 2018. [Online]. Available: https://www.who.int/cardiovascular_diseases/en/. [Accessed: 06-Dec-2018].

[2] P. Groves, B. Kayyali, D. Knott, and S. Van Kuiken, "Accelerating value and innovation," *"big data" Revolut. Healthc. Accel. value Innov.*, no. January, pp. 1–22, 2013.

[3] R. W. Brause, "Medical analysis and diagnosis by neural networks," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2199, pp. 1–13, 2001.

[4] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[5] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*. Kluwer Academic Publishers, 1998.

[6] S. Piramuthu and R. T. Sikora, "Iterative feature construction for improving inductive learning algorithms," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3401–3406, Mar. 2009.

[7] I. Guyon and A. Elisseeff, "AnIntroductionToVariableAndFeautureSelection.pdf," *Journal of Machine Learning Research*, 2003. [Online]. Available: http://www.jmlr.org/papers/v3/guyon03a.html. [Accessed: 07-Dec-2018].

[8] "UCI Machine Learning Repository: Heart Disease Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+Disease. [Accessed: 08-Dec-2018].

[9] "UCI Machine Learning Repository: Statlog (Heart) Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/statlog+(heart). [Accessed:

09-May-2020].

[10] R. El-Bialy, M. A. Salamay, O. H. Karam, and M. E. Khalifa, "Feature Analysis of Coronary Artery Heart Disease Data Sets," *Procedia Comput. Sci.*, vol. 65, no. Iccmit, pp. 459–468, 2015.

[11] A. K. Paul, P. C. Shill, M. R. I. Rabin, and M. A. H. Akhand, "Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease," *2016 5th Int. Conf. Informatics, Electron. Vision, ICIEV 2016*, pp. 145–150, 2016.

[12] R. B. Kavita Burse, Vishnu Pratap Singh, Kirar Abhishek Burse, *Various Preprocessing Methods for Neural Network Based Heart Disease Prediction*, vol. 851. Springer Singapore, 2019.

[13] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telemat. Informatics*, vol. 36, no. August 2018, pp. 82–93, 2019.

[14] H. Ahmed, E. M. G. Younis, A. Hendawi, and A. A. Ali, "Heart disease identification from patients' social posts, machine learning solution on Spark," *Futur. Gener. Comput. Syst.*, 2019.

[15] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics Med. Unlocked*, vol. 16, no. June, p. 100203, 2019.

[16] A. Gupta, R. Kumar, H. Singh Arora, and B. Raman, "MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis," *IEEE Access*, vol. 8, no. Ml, pp. 14659–14674, 2020.

[17] S. M. S. Shah, S. Batool, I. Khan, M. U. Ashraf, S. H. Abbas, and S. A. Hussain, "Feature extraction through parallel Probabilistic Principal Component Analysis for heart disease diagnosis," *Phys. A Stat. Mech. its Appl.*, vol. 482, pp. 796–807, 2017.

[18] S. J. Pasha and E. S. Mohamed, "Ensemble Gain Ratio Feature Selection (EGFS) Model with Machine Learning and Data Mining Algorithms for Disease Risk Prediction." *IEEE International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, pp. 590-596, *2020*

[19] S. M. Saqlain *et al.*, "Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines," *Knowl. Inf. Syst.*, vol. 58, no. 1, pp. 139–167, 2019.

[20] A. P. Kale and S. Sonavane, "PF-FELM: A Robust PCA Feature Selection for Fuzzy Extreme Learning Machine," *IEEE J. Sel. Top. Signal Process.*, vol. 12, no. 6, pp. 1303–1312, 2018.

[21] A. A. A. Mohamed, S. A. Hassan, A. M. Hemeida, S. Alkhalaf, M. M. M. Mahmoud, and A. M. Baha Eldin, "Parasitism – Predation algorithm (PPA): A novel approach for feature selection," *Ain Shams Eng. J.*, no. xxxx, 2019.

[22] S. J. Pasha and E. S. Mohamed, "Bio inspired Ensemble Feature Selection (BEFS) Model with Machine Learning and Data Mining Algorithms for Disease Risk Prediction," *IEEE 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, Pune, India, 2019, pp. 1-6pp., 2019.

[23] M. Z. Alam, M. S. Rahman, and M. S. Rahman, "A Random Forest based predictor for medical data classification using feature ranking," *Informatics Med. Unlocked*, vol. 15, no. January, p. 100180, 2019.

[24] Y. Xu, X. Pan, Z. Zhou, Z. Yang, and Y. Zhang, "Structural least square twin support vector machine for classification," *Appl. Intell.*, vol. 42, no. 3, pp. 527–536, 2015.

[25] S. A. Mokeddem, "A fuzzy classification model for myocardial infarction risk assessment," *Appl. Intell.*, vol. 48, no. 5, pp. 1233–1250, 2018.

[26] F. Sabahi, "Bimodal fuzzy analytic hierarchy process (BFAHP) for coronary heart disease risk assessment," *J. Biomed. Inform.*, vol. 83, no. July 2017, pp. 204–216, 2018.

[27] S. Shilaskar and A. Ghatol, "Diagnosis system for imbalanced multi-minority medical dataset," *Soft Comput.*, vol. 23, no. 13, pp. 1–11, 2018.

[28] S. Maji and S. Arora, "Decision Tree Algorithms for Prediction of Heart Disease," *Inf. Commun. Technol. Compet. Strateg.*, vol. 40, pp. 447–454, 2019.

[29] M. Abdar, U. R. Acharya, N. Sarrafzadegan, and V. Makarenkov, "NE-nu-SVC: A New Nested Ensemble Clinical Decision Support System for Effective Diagnosis of Coronary Artery Disease," *IEEE Access*, vol. 7, pp. 167605–167620, 2019.

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

[31] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, "An Automated Diagnostic System for Heart Disease Prediction Based on $\chi 2$ Statistical Model and Optimally Configured Deep Neural Network," *IEEE Access*, vol. 7, pp. 34938–34945, 2019.

[32] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, D. S. Rajput, R. Kaluri, and G. Srivastava, "Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis," *Evol. Intell.*, no. 123456789, 2019.

[33] B. K. Sarkar, "Hybrid model for prediction of heart disease," *Soft Comput.*, vol. 24, no. 3, pp. 1903–1925, 2020.

[34] X. Zhu, Z. Ni, L. Ni, F. Jin, M. Cheng, and J. Li, "Improved discrete artificial fish swarm algorithm combined with margin distance minimization for ensemble pruning," *Comput. Ind. Eng.*, vol. 128, no. December 2018, pp. 32–46, 2019.

[35] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, and D. Amorim Fernández-Delgado, "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?," *J. Mach. Learn. Res.*, vol. 15, pp. 3133–3181, 2014.

[36] A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, "Design and implementing heart disease prediction using naives Bayesian," *Proc. Int. Conf. Trends Electron. Informatics, ICOEI 2019*, vol. 2019–April, no. Icoei, pp. 292–297, 2019.

[37] M. Tanveer, C. Gautam, and P. N. Suganthan, "Comprehensive evaluation of twin SVM based classifiers on UCI datasets," *Appl. Soft Comput. J.*, vol. 83, p. 105617, 2019.

[38] C. Perales-González, M. Carbonero-Ruz, D. Becerra-Alonso, J. Pérez-Rodríguez, and F. Fernández-Navarro, "Regularized ensemble neural networks models in the Extreme Learning Machine framework," *Neurocomputing*, vol. 361, pp. 196–211, 2019.

[39] S. Yang, J. Z. Guo, and J. W. Jin, "An improved Id3 algorithm for medical data classification," *Comput. Electr. Eng.*, vol. 65, pp. 474–487, 2018.

[40] S. Zhang, "Cost-sensitive KNN classification," *Neurocomputing*, no. xxxx, 2019.

[41] S. Cheriguene, N. Azizi, N. Dey, A. S. Ashour, and A. Ziani, "A new hybrid classifier selection model based on mRMR method and diversity measures," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 5, pp. 1189–1204, 2019.

[42] M. Ma, T. Deng, N. Wang, and Y. Chen, "Semi-supervised rough fuzzy Laplacian Eigenmaps for dimensionality reduction," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 2, pp. 397–411, 2019.

[43] Y. Ding, J. Tang, and F. Guo, "Protein Crystallization Identification via Fuzzy Model on Linear Neighborhood Representation," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 5963, no. JUNE, pp. 1–1, 2019.

[44] H. Das, B. Naik, H. S. Behera, S. Jaiswal, P. Mahato, and M. Rout, "Biomedical data analysis using neuro-fuzzy model with post-feature reduction," *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, pp. 1–11, 2020.

[45] P. K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 24, no. 1, pp. 27–40, 2012.

[46] A. K. Paul, P. C. Shill, M. R. I. Rabin, and K. Murase, "Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease," *Appl. Intell.*, vol. 48, no. 7, pp. 1739–1756, 2018.

[47] A. Unler, A. Murat, and R. B. Chinnam, "Mr2PSO: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification," *Inf. Sci. (Ny).*, vol. 181, no. 20, pp. 4625–4641, 2011.

[48] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, 2005.

SYED JAVEED PASHA received a B.Sc. degree in Computer Science and a master's degree in Computer Applications in 2006 and 2009, respectively, from Osmania University, Hyderabad, Telangana State, India. Since 2017, he has been pursuing a Ph.D. as a full-time senior research scholar at the School of Mathematical and Computer Sciences, B. S. Abdur Rahman Crescent Institute of Science and Technology, Vandalur-48, Chennai, India.

He has two Scopus indexed research articles that were presented and published in international conferences, and achieved the best research article award in February 2020. He worked as teaching faculty at King Saud University, Riyadh, Kingdom of Saudi Arabia, from 2010-2016. His research areas include Data Mining, Data Analytics, Machine Learning and Big Data.

E.SYED MOHAMED received a Ph.D. (in 2015) from the School of Mathematical and Computer Sciences, B. S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India.

He is currently working as Head of the Department, Associate Professor at the School of Mathematical and Computer Sciences and also as an Assistant Dean (Research) at the B. S. Abdur Rahman Crescent Institute of Science and Technology, Vandalur-48, Chennai, India. He has ten research articles in various journals, including seven Scopus listed, and has presented ten papers in several international proceedings and conferences. He has guided several undergraduate and post-graduate students and is currently guiding six Ph.D. research scholars. Additionally, he has conducted and organized several workshops and conferences. His research areas include Soft Computing, Information Security, Web Technologies and Big Data & Data Visualization, Mathematical Modeling and Graphics and Multimedia.