Effective Feature Selection for Improved Prediction of Heart Disease

Ibomoiye Domor Mienye and Yanxia Sun(✉)

Department of Electrical and Electronic Engineering Science, University of Johannesburg, Johannesburg 2006, South Africa
ysun@uj.ac.za

Abstract. Heart disease is among the most prevalent medical conditions globally, and early diagnosis is vital to reducing the number of deaths. Machine learning (ML) has been used to predict people at risk of heart disease. Meanwhile, feature selection and data resampling are crucial in obtaining a reduced feature set and balanced data to improve the performance of the classifiers. Estimating the optimum feature subset is a fundamental issue in most ML applications. This study employs the hybrid Synthetic Minority Oversampling Technique-Edited Nearest Neighbor (SMOTE-ENN) to balance the heart disease dataset. Secondly, the study aims to select the most relevant features for the prediction of heart disease. The feature selection is achieved using multiple base algorithms at the core of the recursive feature elimination (RFE) technique. The relevant features predicted by the various RFE implementations are then combined using set theory to obtain the optimum feature subset. The reduced feature set is used to build six ML models using logistic regression, decision tree, random forest, linear discriminant analysis, naïve Bayes, and extreme gradient boosting algorithms. We conduct experiments using the complete and reduced feature sets. The results show that the data resampling and feature selection leads to improved classifier performance. The XGBoost classifier achieved the best performance with an accuracy of 95.6%. Compared to some recently developed heart disease prediction methods, our approach obtains superior performance.

Keywords: Feature selection · Heart disease · Machine learning · SMOTE-ENN · XGBoost

1 Introduction

Cardiovascular diseases such as heart diseases are the leading cause of death worldwide. According to the world health organization (WHO), heart diseases amount to one-third of worldwide deaths [1]. Early detection of heart diseases is usually challenging, but it is essential to patient survival. Therefore, several machine learning methods have been developed to predict heart disease risk [2, 3]. Usually, medical data contains several features, and some could be noisy, which can negatively impact the model's performance. An efficient feature selection approach could select the most informative feature set,

T. M. N. Ngatched and I. Woungang (Eds.): PAAISS 2021, LNICST 405, pp. 94–107, 2022.
https://doi.org/10.1007/978-3-030-93314-2_6

reduce the computation cost of making predictions, and enhance the prediction performance [4]. Therefore, feature selection is an essential step in most ML applications, especially in medical diagnosis.

Feature selection refers to obtaining the most suitable features while discarding the redundant ones [5]. Feature selection is usually achieved using a wrapper, filter, or embedded method. Wrapper methods perform feature selection via a classifier's prediction, while filter-based techniques score each feature and select the highest scores [6]. Meanwhile, embedded methods combine both wrapper and filter-based methods [7]. Also, having too many attributes in a model increases its complexity and could lead to overfitting. On the other hand, fewer features lead to ML models that are more effective in predicting the class variable. Therefore, this research aims to use the recursive feature elimination technique to obtain the most relevant features for detecting heart disease.

Recursive feature elimination is a type of wrapper-based feature selection method. It is a greedy algorithm used to obtain an optimal feature set [8]. The RFE employs a different ML algorithm to rank the attributes and recursively eliminates the least important attributes whose removal will improve the generalization performance of the classifier. The iterative process of eliminating the weakest attributes goes on until the specified number of attributes is obtained. The RFE's performance relies on the classifier used as the estimator in its implementation. Therefore, it would be beneficial to use different classifiers and compare the predicted features to obtain a more reliable feature set.

Our research aims to develop a feature selection approach to obtain the most informative features to enhance the classification performance of the classifiers. This research uses three base algorithms separately in the RFE implementation to predict the most relevant features. The algorithms include gradient boosting, logistic regression, and decision tree. A feature selection rule based on set theory is applied to obtain the optimal feature set. Then, the optimum feature set serves as input to the logistic regression (LR), decision tree (DT), random forest (RF), linear discriminant analysis (LDA), naïve Bayes (NB), and extreme gradient boosting (XGBoost).

Meanwhile, the class imbalance problem is usually considered when dealing with medical datasets because the healthy (majority class) usually outnumber the sick (minority class) [9]. Most conventional machine learning algorithms tend to underperform when trained with imbalanced data, especially in classifying samples in the minority class. Furthermore, in medical data, samples in the minority class are of particular interest, and the cost of misclassifying them is higher than that of the majority class [10]. Hence, this study employs the hybrid synthetic Minority Oversampling Technique-Edited Nearest Neighbor (SMOTE-ENN) to resample the data and create a dataset with a balanced class distribution.

The contributions of this research include the development of an efficient approach to detect heart disease, implement effective data resampling, select the most relevant heart disease features from the dataset, and compare the performance of different ML algorithms. The rest of this paper is structured as follows: Sect. 2 reviews some related works in recent literature. Section 3 briefly discusses the proposed approach and the various algorithms used in the study. Section 4 describes the dataset and performance assessment metrics used in this paper. Section 5 presents the results and discussion, while Sect. 6 concludes the article and discusses future research directions.

2 Related Works

Many research works have presented different ML-based methods to predict heart disease accurately. For example, in [11], a new diagnostic system was developed to predict heart disease using a random search algorithm to select the relevant features and a random forest classifier to predict heart disease. The random search algorithm selected seven features as the most informative features from the famous Cleveland heart disease dataset, which initially contained 14 features. The experimental results showed that the proposed approach achieved a 3.3% increase in accuracy compared to the traditional random forest algorithm. The proposed approach also obtained superior performance compared to five other ML algorithms and eleven methods from previous literature.

In [12], a deep belief network (DBN) was optimized to prevent overfitting and underfitting in heart disease prediction. The authors employed the Ruzzo-Tompa method to eliminate irrelevant features. The study also developed a stacked genetic algorithm (GA) to find the optimal settings for the DBN. The experimental results achieved better performance compared to other ML techniques. Similarly, Ishaq et al. [13] used the random forest algorithm to select the optimal features for heart disease prediction. They employed nine ML algorithms for the prediction task, including adaptive boosting classifier (AdaBoost), gradient boosting machine (GBM), support vector machines (SVM), extra tree classifier (ETC), and logistic regression etc. The study also utilized the synthetic minority oversampling technique (SMOTE) to balance the data. The experimental results showed that the ETC achieved the best performance with an accuracy of 92.6%.

Ghosh et al. [1] proposed a machine learning approach for effective heart disease prediction by incorporating several techniques. The research combined well-known heart disease datasets such as the Cleveland, Hungarian, Long Beach, and Statlog datasets. The feature selection was achieved using the least absolute shrinkage and selection operator (LASSO) algorithm. From the results obtained, the hybrid random forest bagging method achieved the best performance. Meanwhile, Lakshmananao et al. [14] developed an ML approach to predict heart disease using sampling techniques to balance the data and feature selection to obtain the most relevant features. The preprocessed data were then employed to train an ensemble classifier. The sampling techniques include SMOTE, random oversampling, adaptive synthetic (ADASYN) sampling approach. The results show that the feature selection and sampling techniques enhanced the performance of the ensemble classifier, which obtained a prediction accuracy of 91%.

Furthermore, Haq et al. [15] applied feature selection to the Cleveland heart disease dataset to obtain the most relevant features to improve the classification performance and reduce the computational cost of a decision support system. The feature selection was achieved using the sequential backward selection technique, and the classification was performed using a k-nearest neighbor (KNN) classifier. The experimental results showed that the feature selection step improved the performance of the KNN classifier, and an accuracy of 90% was obtained.

Mienye et al. [2] proposed an improved ensemble learning method to detect heart disease. The study employed decision trees as based learners in building a homogenous ensemble classifier which achieved an accuracy of 93%. In [3], the authors presented a heart disease prediction approach that combined sparse autoencoder and an artificial neural network. The autoencoder performed unsupervised feature learning to enhance

the classification performance of the neural network, and classification accuracy of 90% was obtained. Meanwhile, most of the heart disease prediction models in the literature achieved somewhat acceptable performance. Research has shown that datasets with balanced class distribution and optimal feature sets can significantly improve the prediction ability of machine learning classifiers [16, 17]. Therefore, this study aims to implement an efficient data resampling method and robust feature selection method to enhance heart disease prediction.

3 Methodology

This section briefly discusses the various methods utilized in the course of this research. Firstly, we discuss the hybrid SMOTE-ENN technique used to balance the heart disease data. Secondly, we provide an overview of the recursive feature elimination method and the proposed feature selection rule. Thirdly, the ML classifiers used in training the models are discussed.

3.1 Hybrid SMOTE-ENN

Resampling techniques are used to add or eliminate certain instances from the data, thereby creating balanced data for efficient machine learning. Conventional machine learning classifiers perform better with balanced training data. Oversampling techniques create new synthetic samples in the minority class, while undersampling techniques eliminate examples in the majority class [18]. Both techniques have achieved good performance in diverse tasks. However, previous research has shown that they perform excellent data resampling when both methods are combined [19].

This study aims to perform both oversampling and undersampling using the hybrid SMOTE-ENN method proposed by Batista et al. [20]. This hybrid method creates balanced data by applying both oversampling and undersampling. It combines the SMOTE ability to create synthetic samples in the minority class and the ENN ability to remove examples from both classes that have different class from its k-nearest neighbor majority class. The algorithm works by applying SMOTE to oversample the minority class until the data is balanced. The ENN is then used to remove the unwanted overlapping examples in both classes to maintain an even class distribution [21]. Several research works have shown that the SMOTE-ENN technique results in better performance than when the SMOTE or ENN is used alone [19, 22, 23].

3.2 Recursive Feature Elimination

Recursive feature elimination is a wrapper-based feature selection algorithm. Hence, a different ML algorithm is utilized at the core of the technique wrapped by the RFE. The algorithm iteratively constructs a model from the input features. The model coefficients are used to select the most relevant features until every feature in the dataset has been evaluated. During the iteration process, the least important features are removed. Firstly, the RFE uses the full feature set to calculate the performance of the estimator. Hence, every predictor is given a score. The features with the lowest scores are removed in every

iteration, and the estimator's performance is recalculated based on the remaining feature set. Finally, the subset which produces the best performance is returned as the optimum feature set [24].

An essential part of the RFE technique is the choice of estimator used to select the features. Therefore, it could be inefficient to base the final selected features using a single algorithm. Combining two or more algorithms that complement each other could efficiently produce a more reliable feature subset. Therefore, in this research, we aim to use gradient boosting, decision tree, and logistic regression as estimators in the RFE. We introduce a feature selection rule to obtain the most relevant features from the three predicted feature sets. The rule is that a feature is selected if it was chosen by at least two of the three base algorithms used in the RFE implementation. Assuming the final feature set is represented by A and the optimal feature set selected by gradient boosting, logistic regression and decision tree is represented by the set X, Y, and Z, respectively. Then, we can use set theory to define the rule as:

$$A = (X \cap Y \cap Z) \cup (X \cap Y) \cup (Y \cap Z) \cup (X \cap Z) \tag{1}$$

3.3 Logistic Regression

Logistic regression is a statistical method that applies a logistic function to model a binary target variable. It is similar to linear regression but with a binary target variable. The logistic regression models the probability of an event based on individual attributes. Since probability is a ratio, it is the logarithm of the probability that is modelled:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_m x_m \tag{2}$$

where π represents the probability of an outcome (i.e., heart disease or no heart disease), β_i denotes the regression coefficients, and x_i represents the independent variables [25].

3.4 Decision Tree

Decision trees are popular ML algorithms that can be used for both classification and regression tasks. They utilize a tree-like model of decisions to develop their predictive models. There are different types of decision tree algorithms, but in this study, we use the classification and regression tree (CART) [26] algorithm to develop our decision tree model. CART uses the Gini index to compute the probability of an instance being wrongly classified when randomly selected. Assuming a set of samples has J classes, and $i \in \{1, 2, \ldots, J\}$, then Gini index is defined as:

$$Gini = 1 - \sum_{i=1}^{J} p_i^2 \tag{3}$$

where p_i is the probability of a sample being classified to a particular class.

3.5 Random Forest

Random forest [27] is an ensemble learning algorithm that uses multiple decision tree models to classify data better. It is an extension of the bagging technique that generates random feature subsets to ensure a low correlation between the different trees. The algorithm builds several decision trees, and the bootstrap sample method is used to train the trees from the input data. In classification tasks, the input vector is applied to every tree in the random forest, and the trees vote for a class [28]. After that, the random forest classifier selects the class with the most votes. The difference between the random forest algorithm and decision tree is that it chooses a subset of the input feature, while decision trees consider all the possible feature splits. Different variants of the random forest algorithm [29–31] have been widely applied in diverse medical diagnosis applications with excellent performance.

3.6 Linear Discriminant Analysis

Linear discriminant analysis is a generalization of Fisher's linear discriminant, a statistical method used to compute a linear combination of features that separates two or more target variables. The calculated combination can then be utilized either as a linear classifier or for dimensionality reduction and then classification. LDA aims to find a linear function:

$$y = a_1 x_{i_1} + a_2 x_{i_2} + a_3 x_{i_3} + \cdots + a_m x_{i_m} \tag{4}$$

where $a^T = [\{a_1, a_2, \ldots, a_m\}]$ is a vector of coefficients to be calculated, whereas $x_i = \left[x_{i_1}, x_{i_2}, \ldots, x_{i_m}\right]$ are the input variables [32].

3.7 Naïve Bayes

Naïve Bayes classifiers are probabilistic classifiers based on Bayes' Theorem. They are called naïve because they assume the attributes utilized for training the model are independent of each other [33]. Assuming X is a sample with n attributes, represented by $X = (x_1, \ldots, x_n)$. To compute the class C_k that X belongs to, the algorithm employs a probability model using Bayes theorem:

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)} \tag{5}$$

The class that X belongs to is assigned using a decision rule:

$$y = argmax P(C_k) \Pi_{i=1}^{n} P(X_i|C_k) \tag{6}$$

where y represents the predicted class. The naïve Bayes algorithm is a simple method for building classifiers. There are numerous algorithms based on the naïve Bayes principle, and all of these algorithms assume that the value of a given attribute is independent of the value of the other attributes, given the class variable. In this study, we employ the Gaussian naïve Bayes algorithm, which assumes that the continuous values related to each class are distributed based on a Gaussian (i.e.normal) distribution.

3.8 Extreme Gradient Boosting

Extreme gradient boosting (XGBoost) is an implementation of the gradient boosting machine. It is based on decision trees and can be used for both regression and classification problems. The primary computation process of the XGBoost is the collection of repeated results:

$$\hat{y}_i^{(T)} = \hat{y}_i^{(0)} + \sum_{t=1}^{T} f_t(x_i) \tag{7}$$

where $f_0(x_i) = \hat{y}_i^{(0)} = 0$ and $f_t(x_i) = \omega_{q(x_i)}$. T represents the number of decision trees, $\hat{y}_i^{(T)}$ denotes the predicted value of the i_{th} instance, ω represents a weight vector associated with the leaf node, and $q(x_i)$ represents a function of the feature vector x_i that is mapped to the leaf node [34]. In the XGBoost implementation, the trees are added one after the other to make up the ensemble and trained to correct the misclassifications made by the previous models.

3.9 The Architecture of the Proposed Heart Disease Prediction Model

The flowchart of the proposed heart disease prediction method is shown in Fig. 1. Firstly, the heart disease dataset is resampled using the SMOTE-ENN method to create a dataset with a balanced class distribution. Secondly, the proposed feature selection method is used to select the optimal feature set, which is then split into training and testing sets. The training set is used to train the various classifiers, while the testing set is used to evaluate the classifiers' performance.

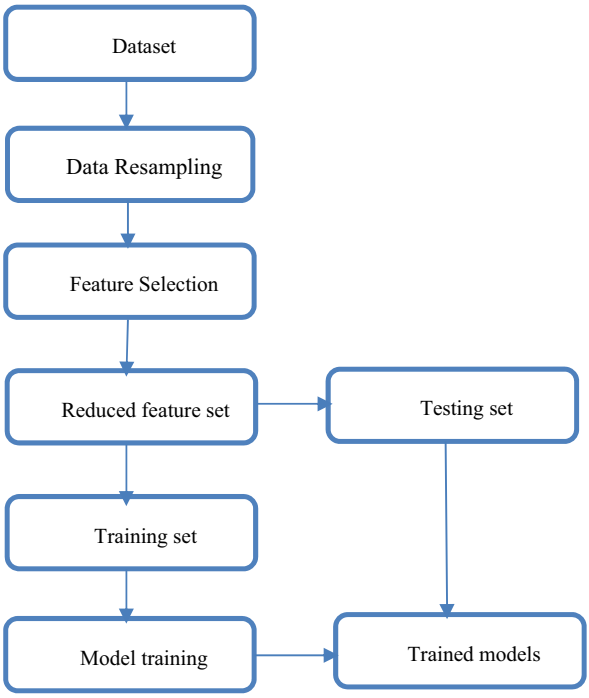

Fig. 1. Flowchart of the proposed heart prediction method

4 Dataset and Performance Metrics

The heart disease dataset used in this study contains 303 samples obtained from medical records of patients above 40 years old. The dataset was compiled at the Faisalabad Institute of Cardiology in Punjab, Pakistan [35]. It comprises 12 attributes and a target variable, including binary attributes such as anaemia, gender, diabetes, smoking, high blood pressure (HBP). Furthermore, the attributes include creatinine phosphokinase (CPK), which is the level of the CPK enzyme in the blood. Other features include ejection fraction, the amount of blood leaving the heart at every contraction, platelets, serum creatinine, etc. The full features are shown in Table 1.

Table 1. Features of the heart disease dataset.

S/N	Features	Code
1	Age	F1
2	Anaemia	F2
3	HBP	F3
4	Creatinine phosphokinase	F4
5	Diabetes	F5
6	Ejection fraction	F6
7	Gender	F7
8	Platelets	F8
9	Serum creatinine	F9
10	Serum sodium	F10
11	Smoking	F11
12	Time	F12
13	Death event (target variable)	F13

Meanwhile, the dataset is not balanced, as there are more samples in the majority class than the minority class. Hence, the need to efficiently balance the data to enhance the classification performance. Furthermore, this research utilizes performance metrics such as accuracy, precision, sensitivity, and F-measure. Their mathematical representations are shown below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{10}$$

$$Fmeasure = \frac{2 \times precision \times recall}{precision + recall} \tag{11}$$

where TN, TP, FN, and FP represent true negative, true positive, false negative, and false positive, respectively. Also, we utilize the receiver operating characteristic (ROC) curve and area under the ROC curve (AUC) to evaluate the performance of the various ML models.

5 Results and Discussion

This section presents the results obtained from the experiments. Firstly, the heart disease data is resampled using the SMOTE-ENN to create a dataset with a balanced class distribution. Secondly, the feature selection is performed using the proposed RFE technique. Though all the features are associated with heart disease, research has shown that reduced feature sets usually improve classification performance [36, 37]. The optimal feature set obtained by the RFE with gradient boosting estimator comprises the following: F1, F3, F5, F7, F8, F9, F10, F11, and F12.

The logistic regression estimator selected the following features: F1, F2, F3, F5, F7, F8, F9, F10, and F12, whereas the decision tree estimator selected F1, F2, F3, F5, F6, F7, F8, F9, F12. Therefore, applying the proposed feature selection rule gives the following features: F1, F2, F3, F5, F7, F8, F9, F10, F12, which is the final feature set. The selected features are used to build ML models. Tables 2 shows the performance of the classifiers when trained with the complete feature set. In contrast, Table 3 shows the performance when the algorithms are trained after the data has been resampled and the feature selection applied.

Table 2. Performance of the algorithms without feature selection and data resampling.

Algorithm	Accuracy	Sensitivity	Precision	F-measure	AUC
LR	0.822	0.746	0.765	0.755	0.857
DT	0.775	0.655	0.686	0.670	0.741
RF	0.830	0.753	0.770	0.761	0.880
LDA	0.810	0.686	0.703	0.694	0.867
NB	0.791	0.694	0.717	0.705	0.854
XGBoost	0.824	0.762	0.776	0.769	0.885

Table 3 shows that the reduced feature set enhanced the performance of the classifiers, and the XGBoost obtained the best performance with an accuracy, sensitivity, precision, F-measure, and AUC of 0.956, 0.981, 0.932, 0.955, and 0.970, respectively. Furthermore, the ROC curves of the various classifiers trained with the reduced feature set are shown in Fig. 2. The ROC curve further validates the superior performance of the XGBoost model trained with the reduced feature set.

Table 3. Performance of the algorithms after feature selection and data resampling.

Algorithm	Accuracy	Sensitivity	Precision	F-measure	AUC
LR	0.929	0.967	0.898	0.931	0.940
DT	0.877	0.913	0.887	0.900	0.870
RF	0.930	0.942	0.942	0.942	0.950
LDA	0.925	0.968	0.896	0.931	0.940
NB	0.904	0.928	0.914	0.921	0.900
XGBoost	0.956	0.981	0.932	0.955	0.970

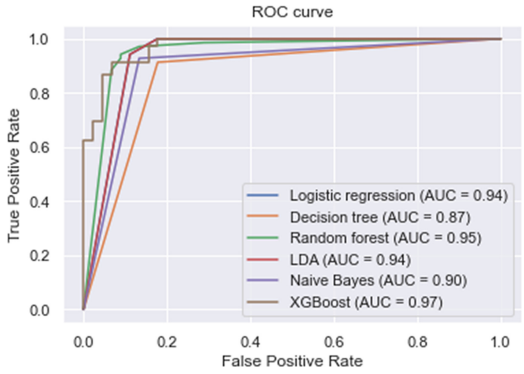

Fig. 2. ROC curves of the classifiers trained with the reduced feature set

Furthermore, we used the XGBoost model to conduct a comparative study with other recently developed research works, shown in Table 4. The comparative analysis is conducted to further validate the performance of our approach. We compare the XGBoost performance with recently developed methods, including an SVM and LASSO based feature selection method [38], XGBoost model [39], a hybrid random forest [40], a deep neural network (DNN) [41], a sparse autoencoder based neural network [3], an enhanced ensemble learning method [2], an improved KNN model [42], and an extra tree classifier with SMOTE based data resampling [13].

Table 4 further shows the robustness of our approach, as the XGBoost model trained with the reduced feature set outperformed the methods developed in the other literature. Furthermore, this research has also shown the importance of data resampling and efficient feature selection in machine learning.

Table 4. Performance comparison with other studies.

Reference	Algorithm	Accuracy	Sensitivity	Precision	F-measure
Li et al. [38]	SVM + LASSO	0.923	0.980	–	–
Tasnim and Habiba [39]	XGBoost	0.835	0.830	0.820	–
Pahwa and Kumar [40]	Hybrid RF	0.8415	–	–	–
Le et al. [41]	DNN	0.8382	0.9166	0.8627	0.8888
Mienye et al. [3]	Sparse autoencoder	0.900	0.910	0.890	0.900
Mienye et al. [2]	Ensemble classifier	0.930	0.910	0.960	0.930
Shah et al. [42]	KNN (k = 7)	0.907	–	–	–
Ishaq et al. [13]	ETC + SMOTE	0.926	0.930	0.930	0.930
This paper	XGBoost + RFE	0.956	0.981	0.932	0.955

6 Conclusion

In machine learning applied to medical diagnosis, data resampling and the selection of relevant features from the dataset is vital in improving the performance of the prediction model. In this study, we developed an efficient feature selection approach based on recursive feature elimination. The method uses a set theory-based feature selection rule to combine the features selected by three recursive feature elimination estimators. The reduced feature set then served as input to six machine learning algorithms, where the XGBoost classifier obtained the best performance. Our approach also showed superior performance compared to eight other methods in recent literature.

Meanwhile, the limitation of this work is that the proposed approach was tested on a single disease dataset. Future research would apply the proposed approach for the prediction of other diseases. Furthermore, future research could utilize evolutionary optimization methods such as a genetic algorithm to select the optimal feature set for training the machine learning algorithms, which could be compared with the method proposed in this work and tested on other disease datasets.

Acknowledgment. This work was supported in part by the South African National Research Foundation under Grant 120106 and Grant 132797 and in part by the South African National Research Foundation Incentive under Grant 132159.

References

1. Ghosh, P., et al.: Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access **9**, 19304–19326 (2021). https://doi.org/10.1109/ACCESS.2021.3053759

2. Mienye, I.D., Sun, Y., Wang, Z.: An improved ensemble learning approach for the prediction of heart disease risk. Inf. Med. Unlock. **20**, 100402 (2020). https://doi.org/10.1016/j.imu.2020.100402
3. Mienye, I.D., Sun, Y., Wang, Z.: Improved sparse autoencoder based artificial neural network approach for prediction of heart disease. Inf. Med. Unlock. **18**, 100307 (2020). https://doi.org/10.1016/j.imu.2020.100307
4. Saha, P., Patikar, S., Neogy, S.: A correlation - sequential forward selection based feature selection method for healthcare data analysis. In: 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), pp. 69–72 (2020). https://doi.org/10.1109/GUCON48875.2020.9231205
5. Kumar, S.S., Shaikh, T.: Empirical evaluation of the performance of feature selection approaches on random forest. In: 2017 International Conference on Computer and Applications (ICCA), pp. 227–231 (2017). https://doi.org/10.1109/COMAPP.2017.8079769
6. Hussain, S.F., Babar, H.Z.-U.-D., Khalil, A., Jillani, R.M., Hanif, M., Khurshid, K.: A fast non-redundant feature selection technique for text data. IEEE Access **8**, 181763–181781 (2020). https://doi.org/10.1109/ACCESS.2020.3028469
7. Pasha, S.J., Mohamed, E.S.: Novel Feature Reduction (NFR) model with machine learning and data mining algorithms for effective disease risk prediction. IEEE Access **8**, 184087–184108 (2020). https://doi.org/10.1109/ACCESS.2020.3028714
8. Zhang, W., Yin, Z.: EEG feature selection for emotion recognition based on cross-subject recursive feature elimination. In: 2020 39th Chinese Control Conference (CCC), pp. 6256–6261 (2020). https://doi.org/10.23919/CCC50068.2020.9188573
9. Mienye, I.D., Sun, Y.: Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. Inf. Med. Unlock. **25**, 100690 (2021). https://doi.org/10.1016/j.imu.2021.100690
10. Guan, H., Zhang, Y., Xian, M., Cheng, H.D., Tang, X.: SMOTE-WENN: solving class imbalance and small sample problems by oversampling and distance scaling. Appl. Intell. **51**(3), 1394–1409 (2020). https://doi.org/10.1007/s10489-020-01852-8
11. Javeed, A., Zhou, S., Yongjian, L., Qasim, I., Noor, A., Nour, R.: An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection. IEEE Access **7**, 180235–180243 (2019). https://doi.org/10.1109/ACCESS.2019.2952107
12. Ali, S.A., et al.: An optimally configured and improved deep belief network (OCI-DBN) approach for heart disease prediction based on ruzzo-tompa and stacked genetic algorithm. IEEE Access **8**, 65947–65958 (2020). https://doi.org/10.1109/ACCESS.2020.2985646
13. Ishaq, A., et al.: Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. IEEE Access **9**, 39707–39716 (2021). https://doi.org/10.1109/ACCESS.2021.3064084
14. Lakshmanarao, A., Srisaila, A., Kiran., T.S.R.: Heart disease prediction using feature selection and ensemble learning techniques. In: 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 994–998 (2021). https://doi.org/10.1109/ICICV50876.2021.9388482
15. Haq, A.U., Li, J., Memon, M.H., Hunain Memon, M., Khan, J., Marium, S.M.: Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection. In: 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), pp. 1–4 (2019). https://doi.org/10.1109/I2CT45611.2019.9033683
16. Kasongo, S.M., Sun, Y.: Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset. J. Big Data **7**(1), 1–20 (2020). https://doi.org/10.1186/s40537-020-00379-6

17. Kasongo, S.M., Sun, Y.: A deep learning method with filter based feature engineering for wireless intrusion detection system. IEEE Access **7**, 38597–38607 (2019). https://doi.org/10.1109/ACCESS.2019.2905633
18. Hasanin, T., Khoshgoftaar, T.M., Leevy, J.L., Bauder, R.A.: Severely imbalanced Big Data challenges: investigating data sampling approaches. J. Big Data **6**(1), 1–25 (2019). https://doi.org/10.1186/s40537-019-0274-4
19. Xu, Z., Shen, D., Nie, T., Kou, Y.: A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. J. Biomed. Inform. **107**, 103465 (2020). https://doi.org/10.1016/j.jbi.2020.103465
20. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. Newsl. **6**(1), 20–29 (2004). https://doi.org/10.1145/1007730.1007735
21. Fitriyani, N.L., Syafrudin, M., Alfian, G., Rhee, J.: HDPM: an effective heart disease prediction model for a clinical decision support system. IEEE Access **8**, 133034–133050 (2020). https://doi.org/10.1109/ACCESS.2020.3010511
22. Le, T., Vo, M.T., Vo, B., Lee, M.Y., Baik, S.W.: A Hybrid approach using oversampling technique and cost-sensitive learning for bankruptcy prediction. Complexity **2019**, e8460934 (2019). https://doi.org/10.1155/2019/8460934
23. Dogo, E.M., Nwulu, N.I., Twala, B., Aigbavboa, C.: Accessing imbalance learning using dynamic selection approach in water quality anomaly detection. Symmetry **13**(5), Art. no. 5 (2021). https://doi.org/10.3390/sym13050818
24. Koul, N., Manvi, S.S.: Ensemble feature selection from cancer gene expression data using mutual information and recursive feature elimination. In: 2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC), pp. 1–6 (2020). https://doi.org/10.1109/ICAECC50550.2020.9339518
25. Sperandei, S.: Understanding logistic regression analysis. Biochem. Med. (Zagreb) **24**(1), 12–18 (2014). https://doi.org/10.11613/BM.2014.003
26. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and regression trees. Wadsworth & Brooks, Monterey (1983). /paper/Classification-and-Regression-Trees-Breiman-Friedman/8017699564136f93af21575810d557dba1ee6fc6. Accessed on 05 Aug 2020
27. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
28. Mushtaq, M.-S., Mellouk, A.: 2 - Methodologies for subjective video streaming QoE assessment. In: Mushtaq, M.-S., Mellouk, A. (eds.) Quality of Experience Paradigm in Multimedia Services, pp. 27–57 Elsevier (2017). https://doi.org/10.1016/B978-1-78548-109-3.50002-3
29. Ke, F., Liu, H., Zhou, M., Yang, R., Cao, H.-M.: Diagnostic biomarker exploration of autistic patients with different ages and different verbal intelligence quotients based on random forest model. IEEE Access **9**, 1 (2021). https://doi.org/10.1109/ACCESS.2021.3071118
30. Cui, H., Wang, Y., Li, G., Huang, Y., Hu, Y.: Exploration of cervical myelopathy location from somatosensory evoked potentials using random forests classification. IEEE Trans. Neural Syst. Rehabil. Eng. **27**(11), 2254–2262 (2019). https://doi.org/10.1109/TNSRE.2019.2945634
31. Guo, C., Zhang, J., Liu, Y., Xie, Y., Han, Z., Yu, J.: Recursion enhanced random forest with an improved linear model (RERF-ILM) for heart disease detection on the internet of medical things platform. IEEE Access **8**, 59247–59256 (2020). https://doi.org/10.1109/ACCESS.2020.2981159
32. Ricciardi, C., et al.: Linear discriminant analysis and principal component analysis to predict coronary artery disease. Health Inf. J. **26**(3), 2181–2192 (2020). https://doi.org/10.1177/1460458219899210

33. Chen, S., Webb, G.I., Liu, L., Ma, X.: A novel selective naïve Bayes algorithm. Knowl.-Based Syst. **192**, 105361 (2020). https://doi.org/10.1016/j.knosys.2019.105361
34. Cui, L., Chen, P., Wang, L., Li, J., Ling, H.: Application of extreme gradient boosting based on grey relation analysis for prediction of compressive strength of concrete. Adv. Civil Eng. **2021**, e8878396 (2021). https://doi.org/10.1155/2021/8878396
35. Ahmad, T., Munir, A., Bhatti, S.H., Aftab, M., Raza, M.A.: Survival analysis of heart failure patients: a case study. PLoS ONE **12**(7), e0181001 (2017). https://doi.org/10.1371/journal.pone.0181001
36. Miao, J., Niu, L.: A survey on feature selection. Proc. Comput. Sci. **91**, 919–926 (2016). https://doi.org/10.1016/j.procs.2016.07.111
37. Mienye, I.D., Kenneth Ainah, P., Emmanuel, I.D., Esenogho, E.: Sparse noise minimization in image classification using Genetic Algorithm and DenseNet. In: 2021 Conference on Information Communications Technology and Society (ICTAS), pp. 103–108 (2021). https://doi.org/10.1109/ICTAS50802.2021.9395014
38. Li, J.P., Haq, A.U., Din, S.U., Khan, J., Khan, A., Saboor, A.: Heart disease identification method using machine learning classification in e-healthcare. IEEE Access **8**, 107562–107582 (2020). https://doi.org/10.1109/ACCESS.2020.3001149
39. Tasnim, F., Habiba, S.U.: A comparative study on heart disease prediction using data mining techniques and feature selection. In: 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), pp. 338–341 (2021). https://doi.org/10.1109/ICREST51555.2021.9331158
40. Pahwa, K., Kumar, R.: Prediction of heart disease using hybrid technique for selecting features. In: 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), pp. 500–504 (2017). https://doi.org/10.1109/UPCON.2017.8251100
41. Le, M.T., Thanh Vo, M., Mai, L., Dao, S.V.T.: Predicting heart failure using deep neural network. In: 2020 International Conference on Advanced Technologies for Communications (ATC), pp. 221–225 (2020). https://doi.org/10.1109/ATC50776.2020.9255445
42. Shah, D., Patel, S., Bharti, S.K.: Heart disease prediction using machine learning techniques. SN Comput. Sci. **1**(6), 1–6 (2020). https://doi.org/10.1007/s42979-020-00365-y