IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.17, No.1, January 2023, pp. 55~66
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.80214

Improving Cardiovascular Disease Prediction by Integrating Imputation, Imbalance Resampling, and Feature Selection Techniques into Machine Learning Model

Fadlan Hamid Alfebi[1], Mila Desi Anasanti*[2]
[1,2]Program Studi S2 Ilmu Komputer, Universitas Nusa Mandiri, Jakarta, Indonesia
[2]Department of Information Studies, University College London, London, United Kingdom
[2]Bart and London Genome Center, Queen Mary University of London, London, United Kingdom
e-mail: [1]14210242@nusamandiri.ac.id, ***[2]mila.mld@nusamandiri.ac.id**

Abstrak

Penyakit kardiovaskular (CVD) adalah penyebab utama kematian di seluruh dunia. Pencegahan primer adalah dengan prediksi awal timbulnya penyakit. Menggunakan data laboratorium dari National Health and Nutrition Examination Survey (NHANES) pada jangka waktu 2017-2020 (N= 7.974), kami menguji kemampuan algoritma machine learning (ML) untuk mengklasifikasikan individu yang berisiko. Model ML dievaluasi berdasarkan kinerja klasifikasinya setelah membandingkan empat teknik imputasi, tiga resampling ketidakseimbangan, dan tiga teknik pemilihan fitur.

Karena popularitasnya, kami menggunakan decision tree (DT) sebagai model dasar. Integrasi multiple imputation by chained equation (MICE) dan synthetic minority oversampling with Tomek link downsampling (SMOTETomek) pada model dasar meningkatkan area under the curve-receiver operating characteristics (AUC-ROC) dari 57% menjadi 83%. Menerapkan simultaneous perturbation feature selection and ranking (spFSR) mengurangi prediktor fitur dari 144 menjadi 30 fitur dan waktu komputasi sebesar 22%. Teknik terbaik diterapkan pada enam model ML, menghasilkan Xtreme gradient boosting (XGBoost) mencapai akurasi tertinggi sebesar 93% dan AUC-ROC sebesar 89%.

Keakuratan ML model kami dalam memprediksi CVD mengungguli studi sebelumnya. Kami juga menyorot penyebab penting CVD, yang dapat dieksplorasi lebih lanjut untuk efek potensialnya pada catatan kesehatan elektronik.

Kata kunci—*machine learning, cardiovascular disease, imputation, resampling, feature selection*

Abstract

Cardiovascular disease (CVD) is the leading cause of death worldwide. Primary prevention is by early prediction of the disease onset. Using laboratory data from the National Health and Nutrition Examination Survey (NHANES) in 2017-2020 timeframe (N= 7.974), we tested the ability of machine learning (ML) algorithms to classify individuals at risk. The ML models were evaluated based on their classification performances after comparing four imputation, three imbalance resampling, and three feature selection techniques.

Due to its popularity, we utilized decision tree (DT) as the baseline. Integration of multiple imputation by chained equation (MICE) and synthetic minority oversampling with Tomek link down-sampling (SMOTETomek) into the model improved the area under the curve-receiver operating characteristics (AUC-ROC) from 57% to 83%. Applying simultaneous

Received December 14th,2022; Revised January 30th, 2023; Accepted January 31th, 2023

perturbation feature selection and ranking (spFSR) reduced the feature predictors from 144 to 30 features and the computational time by 22%. The best techniques were applied to six ML models, resulting in Xtreme gradient boosting (XGBoost) achieving the highest accuracy of 93% and AUC-ROC of 89%.

The accuracy of our ML model in predicting CVD outperforms those from previous studies. We also highlight the important causes of CVD, which might be investigated further for potential effects on electronic health records.

Keywords—*machine learning, cardiovascular disease, imputation, resampling, feature selection*

1. INTRODUCTION

With an estimated 17.9 million deaths each year, cardiovascular disease (CVD) tops the list of causes of death worldwide [1]. The leading causes of this lethal disease, which is widely prevalent among people of all ages, include high blood pressure, obesity, high cholesterol, family history, smoking, and drinking [2]. The clinical presentation of CVD can range from asymptomatic to classic presentations.

The typical symptoms of CVD include typical anginal chest pain consistent with myocardial infarction (MI) and or acute cerebrovascular accident (CVA) presenting with focal neurological deficits of sudden onset [3]. This symptom may also be accompanied by nausea, vomiting, palpitations, diaphoresis, syncope, or even sudden death [4]. However, it is suggested to maintain a high level of suspicion in patients experiencing an acute MI [5]. It is essential to be aware of the possibility of variations in how symptoms present, particularly in patients with a known history of coronary artery disease (CAD) or MI and those with CVD risk factors [6].

Given that CVD is accompanied by various symptoms, rapid and precise identification is reasonably tricky for medical specialists. As a result, healthcare businesses are gathering enormous amounts of data to help medical experts comprehend the condition and ensure effective prevention. However, collecting data requires extensive processing to extract information efficiently. As a result, machine learning (ML) has become the most effective technique for processing data to advance the healthcare industry [7, 8].

Numerous studies have conducted various ML methods to predict CVD. Jaymin et al. used the random forest (RF) algorithm, logistic model tree, and J48 tree technique, each focused on a distinct technique [9]. The J48 tree method had an accuracy rate of 56.76% among them. Archana et al. used four algorithms and got the best in K-nearest neighbors (KNN) with an accuracy of 87% [10]. Alim et al. used logistic regression (LR), naïve bayes classifier (NBC), support vector machine (SVM), RF, and Gradient Boosting (GB), concluded with RF and stratified K-fold model being proposed as the better option of all with final accuracy of 86.12% [11]. In order to predict and diagnose cardiac disease, Kannan et al. used four ML algorithms: LR, RF, SVM, and Stochastic Gradient Boosting (SGB). Despite using 10-fold cross-validation in SVM and SGB models, the LR model performed the best with an 87% accuracy rate [12]. Utilizing the University of California Irvine (UCI) dataset with 14 attributes and various ML approaches, Atallah et al. achieved the best accuracy of 90% using the hard voting ensemble method [13]. Kohli et al. used Backward Elimination, an ML feature, to experiment with a dataset on cardiac illness. LR has the highest accuracy among the five algorithms used at 87.1% [14]. Abderrahmane et al. created a data processing and monitoring application where Spark MLlib and Spark streaming were used for data processing and achieved accuracy rates of 87.5%, sensitivity rates of 86.66%, and specificity rates of 88.37% [15]. Taking into account the works mentioned earlier, our study sought to increase the performance and produce more satisfying results in terms of addressing missing data problems, imbalanced data, and important features to predict CVD.

2. METHODS

2.1 Dataset

National Center for Health Statistics (NCHS) developed the National Health and Nutrition Examination Survey (NHANES), a program used to evaluate the population's health and nutritional status in the United States [16]. Medical professionals perform physical, physiological, dental, and other medically related assessments as part of the laboratory testing.

The NHANES dataset consists of five domains: demographic, dietary, examination, laboratory, and questionnaire. This study focuses on information obtained from laboratory test results, precisely information on the examination and laboratory domains, using a 2017–2020 timeframe survey [17]. We obtained 8.544 pre-processed samples to predict CVD with non-pregnant criteria and age over 20 years.

2.2 Data Preprocessing

2.2.1 Manual Filtering

To ensure the data used are relevant to the research being conducted, manual filtering was performed to eliminate irrelevant variables to the variable outcome of CVD diagnosis. Only 683 out of a total of 1.861 variables were deemed relevant to our study. We assessed the missing values in the dataset to determine the quality and consistency of the data obtained by the ML model. Eliminating variables and samples with above 60% missing values were done as extensive missing data tends to reduce a study's statistical power and produce biased estimates that lead to incorrect findings [18]. The final manual filtering procedure yielded 144 variables and 7.974 samples utilized in the model development process.

2.2.2 Class Labeling

Participants were categorized as having CVD (label = 1) if they reported getting any of the CVD characteristics defined by the enquiry: “Have you ever been told by a doctor that you had congestive heart failure, coronary heart disease, a heart attack, or a stroke?” Participants who responded "No" to all four questions were categorized as being exempt from the disease (label = 0). Note that these symptoms frequently represent CVD [19]. From a total of 7.974 samples, 750 individuals were classified as class 1, while 7.224 were classified as class 0.

2.2.3 Normalization

Min-max scaling normalization is one of the most widely recognized normalization approaches that scale data into specific ranges [20]. For each component, the base estimation of that element was transformed into 0, the most extreme value was transformed into 1, and all other values were transformed into decimals in the range of 0 and 1. We employed the MinMaxScaler function from the ‘scikit-learn’ Python module to transform all the numeric values.

2.3 Model Development

2.3.1 Missing Value Imputation

Numerous missing values exist in the 2017-2020 NHANES laboratory dataset, which can create biased estimates in many ML methods; hence, an imputation method must be utilized. Supplementary Table 1 describes all variables used in this research with their missing value percentages. To discover which method of imputation was the most effective, we evaluated several techniques: simple imputation using statistical mean values [21], multiple imputation by chained equation (MICE) [22], k-nearest neighbor (kNN) imputer [23], and missForest imputer [24]. We set mean imputation as a baseline, compared the results from these imputation methods in CVD prediction using the DT predictive model, and then brought the best imputation method onto the following model development stages.

2.3.2 Imbalanced Data Resampling

Class imbalance occurs when the number of instances in one class is disproportionately more significant than that of another class [25]. In such situations, classifiers tend to favour the majority class while ignoring the minority class. Resampling is one of the most critical strategies for resolving imbalanced data classification problems [26].

The 2017-2020 NHANES laboratory dataset exhibits highly imbalanced data, with an almost 90:10 ratio between classes 0 and 1 (controls vs cases). To address this issue, we compared the performance of the synthetic minority oversampling technique (SMOTE) [27] for oversampling method, instance hardness threshold (IHT) [28] for the undersampling method, and SMOTE with Tomek link (SMOTETomek) [29] for the combination of oversampling and undersampling method.

2.3.3 Feature Selection

Feature selection methods evaluate the importance of a feature or group of features based on a predetermined metric. The most important benefits of these strategies are: (a) to prevent overfitting and improve model performance, (b) to acquire a more precise and more comprehensive understanding of the problem, and (c) to create faster and more cost-effective prediction models [30]. We included all available features in the dataset for the initial run to establish a performance baseline and measure the processing time. In order to optimize the time required in ML training, we tried three methods that represent each feature selection widely used [31]: mutual information (MI) [32], random forest importance (RFI) [33], and simultaneous perturbation feature selection and ranking (spFSR) [34].

2.4 Machine Learning Predictive Models

2.4.1 Logistic Regression (LR)

LR uses a logistic function to presume that the target “y” is a member of the set {0,1} member and converts linear probabilities into logit; hence, it is more suited to classification problems than regression ones [35]. Consider the binary problem, assuming the positive class is marked as 1 and the negative class as 0. Since it returns a probability. LR is widely used in numerous practical applications, from data processing in industries, medical sciences, and many more.

2.4.2 Decision Tree (DT)

DT is a popular ML approach for classification since it resembles human reasoning and is simple to comprehend [36]. The purpose of the DT is to summarise a series of decision tree rules from a dataset containing features and labels by separating nodes based on certain features and displaying the algorithm for these rules using a tree diagram. Considering the latest data, the computational complexity of this method is modest [36].

2.4.3 Random Forest (RF)

RF is a form of DT that randomly restricts the features utilized for each split and functions by building numerous DTs during training. There is no correlation between each DT in an RF; after producing a number of trees, the final decision class is determined by a vote when a new sample is received, and each DT in the RF determines which category the sample belongs [33]. RF is regarded as one of the most effective algorithms that are insensitive to multicollinearity and tolerant of missing values and unbalanced data [33].

2.4.4 K-Nearest Neighbor (KNN)

KNN identifies a new data point in accordance with the majority of votes of its “k” neighbours, calculates the distance. The KNN is resource intensive, particularly when the size of the feature space is enormous [37].

2.4.5 Multilayer Perceptron (MLP)

MLP is a feed-forward neural network with numerous interconnected neurons. A neuron of one layer comes into contact via weighted connections with the neurons of its surrounding layers; however there is no link between the neurons of the same layer [38].

2.4.6 Xtreme Gradient Boosting (XGBoost)

Chen and Guestrin first suggested XGBoost in 2016 [39]. It has been regarded as a sophisticated estimator with exceptional classification and regression performance. It offers numerous enhancements over conventional gradient-boosting algorithms. Unlike the gradient boosting decision tree (GBDT), XGBoost's loss function employs regularization to prevent overfitting. As the objective function, XGBoost also employs a second-order Taylor series.

2.5 Performance Metrics

We assessed the performance evaluation in terms of the standard performance metrics widely used in ML studies, including the area under the receiver operating characteristic (AU-ROC) curve. Where TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative, the performance metrics are defined below.

2.5.1 Accuracy

Accuracy is a performance metric that measures the correctness of test data predictions [40]. It provides the proportion of accurate predictions while testing data, as shown in equation (1).

$$Accuracy = (TP + TN) \,/\, (TP + TN + FP + FN) \qquad (1)$$

2.5.2 Precision

As shown in equation (2), it is the proportion of correctly classified positive events relative to the total number of positive classifications.

$$Precision = TP \,/\, (TP + FP) \qquad (2)$$

A model with a low precision score likely produces many false positives. Concerns about class distribution are better addressed by this metric, along with recall, F1-score, and specificity [41].

2.5.3 Recall

It is the number of true positive cases that a model correctly identifies out of the overall number of true positive cases that are being examined, as shown in equation (3) [42].

$$Recall = TP \,/\, (TP + FN) \qquad (3)$$

2.5.4 F1 Score

It results from the harmonic mean of precision and recall as shown in equation (4):

$$F1\,Score = (2\ x\ Precision\ x\ Recall) \,/\, (Precision + Recall) \qquad (4)$$

A high F1 score is achieved by striking a balance between precision and recall. If the F1 score is low, it may indicate that one of these measures has been enhanced at the expense of the other [43].

2.5.5 Area under the Receiver Operating Characteristic Curve (AU-ROC Curve)

A ROC curve is a graph that displays the accuracy of a classification model across a range of cutoff values. Plotting the true positive rate (recall) versus the false positive rate (1−specificity) yields the resulting curve [44]. More items are labelled positive when the threshold is lowered, which increases both the false positive and true positive rates. Utilizing the AU-ROC curve offers an overall performance metric across all conceivable classification methods.

3. RESULTS AND DISCUSSION

We conducted step-by-step imputation, imbalance resampling, feature selection and classification methods into the ML baseline model using DT method. The best result from each step also served as a baseline for the next step. The results from each step are explained as follows:

3.1 Missing Value Imputation Result

In this imputation stage, simple mean imputation was set as a baseline for the DT model. All the imputation methods used in this study were compared with the baseline, as shown in Table 1. Any resampling and feature selection techniques have not been performed.

Table 1 Predictive Model Performance Comparison from Four Imputation Methods

Measurement	Simple	KNN	MICE	MissForest
Accuracy	0.84	0.84	0.87	0.85
Precision	0.21	0.21	0.32	0.22
Recall	0.24	0.23	0.36	0.25
F1-Score	0.23	0.22	0.34	0.23
AU-ROC	0.57	0.57	0.64	0.58
Average	0.42	0.41	0.51	0.43
Improvement	Base	- 2%	21%	2%

The MICE imputation method was superior against the others in all measurements, yielding 21% relative improvement against baseline performance; therefore, it will be utilized in further model development.

3.2 Imbalanced Data Resampling Result

An imbalanced ratio between classes is highly possible to bias the ML predictive model. In this particular attempt, we brought evidence that a resampling effort could give a significant leap to observed performances. The original dataset has a 10:90 imbalanced ratio between cases and controls; thereafter, we employed a more modest 25:75 resampled ratio. We did not apply a higher minority class increase than this to avoid excessively biased synthetic data. All available features and DT predictive models remain utilized in this experiment. Table 2 exhibits detailed results for every resampling method, where the baseline performance was set without applying any dataset resample.

Table 2 Predictive Model Performance Comparison of Three Resampling Methods Against Those Without Resampling

Measurement	None	SMOTE	IHT	SMOTETomek
Accuracy	0.87	0.87	0.88	0.88
Precision	0.32	0.73	0.74	0.74
Recall	0.36	0.77	0.75	0.78
F1-Score	0.34	0.75	0.75	0.76
AU-ROC	0.64	0.84	0.83	0.84
Average	0.51	0.79	0.79	0.80
Improvement	Base	56%	56%	58%

From the results, we observed that resampling remarkably elevates model performance in highly imbalanced datasets. Oversampling with SMOTE, downsampling with IHT, and the combination of oversampling and downsampling with SMOTETomek indicated similar good results in our experiment. However, the latest has a slight improvement over the others.

3.3 Feature Selection Result

With the goal of developing an accurate model relying on a limited set of available features, i.e. only involving features that have high correlation and impact with desired output class, we assessed the feature importance of the ML model for predicting CVD. Out of 144 features in the dataset, the top 30 most notable features were extracted using several feature selection methods, precisely mutual information (MI), random forest importance (RFI), and simultaneous perturbation feature selection and ranking (spFSR). We selected the 30 features based on cross-validation of the models; since models with fewer features showed significantly worse performance (>2% drop in AU-ROC score). Table 3 shows comparison results for the DT predictive model using full features against mentioned feature selection methods. Note that this development stage used the dataset imputed with MICE and resampled with SMOTETomek.

Table 3 Predictive Model Performance Comparison for Four Feature Selection Methods

Measurement	Full	MI	RFI	spFSR
Accuracy	0.88	0.83	0.87	0.87
Precision	0.74	0.65	0.73	0.73
Recall	0.78	0.68	0.76	0.77
F1 Score	0.76	0.66	0.74	0.75
AU-ROC	0.84	0.78	0.83	0.84
Average	0.80	0.72	0.78	0.79
Improvement	Base	- 10%	- 2%	- 1%

We also measured the time required for end-to-end predictive model training using mentioned feature selection methods, as shown in Table 4.

Table 4 Time Required in Model Training for Four Feature Selection Methods

Measurement	Full	MI	RFI	spSFR
Time in seconds	55	35	32	43
Improvement	Base	27%	36%	22%

Although RFI delivered the most efficient method, we decided to nominate spFSR as the preferable feature selection method. We considered the time difference in seconds less important than the highest accuracy achieved by spFSR.

3.4 Machine Learning Predictive Model Result

After we found the best method in imputation, resampling, and feature selection, the last stage of our model development was experimenting with the integration of MICE imputation, SMOTETomek resampling, and spFSR feature selection into six different ML methods. We used a stratified 10-fold cross-validation instead of a train-test split to avoid overfitting, as the dataset is considerably small. Table 5 lists the test result with various evaluation metrics.

Table 5 Machine Learning Performance Comparison using 10-fold Cross Validation for Cardiovascular Disease Prediction

Measurement	LR	DT	RF	KNN	MLP	XGB
Accuracy	0.78	0.87	**0.93**	0.83	0.82	**0.93**
Precision	0.62	0.73	**0.94**	0.63	0.69	0.92
Recall	0.26	0.76	0.76	0.77	0.52	**0.80**
F1 Score	0.36	0.74	0.84	0.69	0.59	**0.86**
AU-ROC	0.60	0.83	0.87	0.81	0.72	**0.89**
Average	0.48	0.79	0.87	0.75	0.67	**0.88**

Simplistic ML models like LR as expected gained very low performance in recall and F1 score, likely because the ratio between classes was still imbalanced and the small shift in resampling ratio did not take a significant effect. MLP and KNN showed relatively low scores as we only used default parameters and did not further optimize other hyperparameters. Basic tree algorithms like DT already have decent overall results. As a collection of DTs, the RF algorithm polished it up by achieving the best accuracy and precision among them all. On the top end, the ensemble ML model XGBoost successfully attained the highest scores in five out of six evaluation metrics and was elected the best predictive model in our experiment.

3.5 Feature Importance

Feature selection with spFSR shrunk down feature space into user desired number, in this case 30 best features out of 144 total features in the dataset, almost 20%. Table 6 summarises the top 5 features with the highest importance value for predicting CVD. The complete list of the 30 best features can be observed in Supplementary Table 1.

Table 6 Five Features with Highest Importance Factor

Feature	Description	Importance Factor
LBXVME	Blood Methyl t-Butyl Ether (pg/ml)	0,1023
LBXVMIK	Blood Methyl Isobutyl Ketone (ng/mL)	0,0735
LBXSF5SI	5,10-Methenyl-tetrahydrofolate (nmol/L)	0,0732
LBDSCHSI	Cholesterol, total (mmol/L)	0,0731
LBXMCVSI	Mean cell volume (fL)	0,0730

In this study, blood methyl t-butyl ether (MTBE) came out as the most influential factor from NHANES 2017-2020 laboratory data for predicting CVD by a more considerable margin than other factors. This was rather surprising since it was almost never mentioned in previous CVD studies. However, a recent study by Ren et al. showed for the first time how exposure to MTBE and the development of CVD are related [45]. Based on their experiment, MTBE was proven to affect the metabolism of glucose and lipids, causing obesity to develop, which is one of the widely known risk factors of CVD. Cholesterol unexpectedly only took the fourth position, with the importance factor slightly below blood methyl isobutyl ketone (MIBK) and 5,10-methenyl-tetrahydrofolate (MTH).

4. CONCLUSIONS

Our study conducted exhaustive experiments on improving the model development process to build a better predictive ML model in predicting CVD. Imputation using MICE gave 21% improvement over mean imputation, while resampling imbalanced data with

SMOTETomek achieved 58% relative improvement compared to data without resampling applied. The effort to further optimize model development employing spFSR feature selection from 144 features to 30 best features managed to reduce training time by 22% with a performance hit of only 1%. In ML model comparison using the combination from previous stages, XGBoost outperformed other ML models with 89% AUC-ROC and 93% accuracy, a slight win over RF with 88% AUC-ROC while the other four tested ML models cannot achieve AUC-ROC more than 79%. The utilization of advanced techniques in our model proved to be very effective and showed improved results in classification compared to the previous studies [9, 10, 11, 12, 13, 14] that only achieved the highest accuracy of 90% [14].

Identification of features which contributed most to the CVD prediction using NHANES 2017-2020 laboratory data resulting in blood MTBE as the main factor to the disease, while cholesterol in the fourth position below blood MIBK and blood MTH.

As shown in our analysis, our ML model showed promising results for predicting CVD in patients at-risk using only the 2017-2020 laboratory data timeframe. Expanding the timeframe for a better understanding of other factors would be suggested in the future. Also more timeframe surely provides more samples, giving better conditions in treating the dataset, i.e. we can raise the missing value threshold without sacrificing model performance.

REFERENCES

[1] Y. Ruan *et al.*, “Cardiovascular disease (CVD) and associated risk factors among older adults in six low-and middle-income countries: Results from SAGE Wave 1,” *BMC Public Health*, vol. 18, no. 1, p. 778, Jun. 2018, doi: 10.1186/s12889-018-5653-9.

[2] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, “Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes,” *BMC Med. Inform. Decis. Mak.*, vol. 10, no. 1, p. 16, 2010, doi: 10.1186/1472-6947-10-16.

[3] K. S. Yew and E. Cheng, “Acute stroke diagnosis,” *Am. Fam. Physician*, vol. 80, no. 1, p. 33, 2009.

[4] C. Kreatsoulas, H. S. Shannon, M. Giacomini, J. L. Velianou, and S. S. Anand, “Reconstructing angina: cardiac symptoms are the same in women and men,” *JAMA Intern. Med.*, vol. 173, no. 9, pp. 829–833, 2013.

[5] I. Kirchberger *et al.*, “Patient-reported symptoms in acute myocardial infarction: differences related to ST-segment elevation: the MONICA/KORA Myocardial Infarction Registry,” *J. Intern. Med.*, vol. 270, no. 1, pp. 58–64, 2011.

[6] J. Robson, L. Ayerbe, R. Mathur, J. Addo, and A. Wragg, “Clinical value of chest pain presentation and prodromes on the assessment of cardiovascular disease: a cohort study,” *BMJ Open*, vol. 5, no. 4, p. e007251, 2015.

[7] K. Chayakrit, Z. HongJu, W. Zhen, A. Mehmet, and K. Takeshi, “Artificial Intelligence in Precision Cardiovascular Medicine,” *J. Am. Coll. Cardiol.*, vol. 69, no. 21, pp. 2657–2664, May 2017, doi: 10.1016/j.jacc.2017.03.571.

[8] A. Rajkomar, J. Dean, and I. Kohane, “Machine learning in medicine,” *N. Engl. J. Med.*, vol. 380, no. 14, pp. 1347–1358, 2019.

[9] J. Patel, D. TejalUpadhyay, and S. Patel, “Heart disease prediction using machine learning and data mining technique,” *Hear. Dis.*, vol. 7, no. 1, pp. 129–137, 2015.

[10] A. Singh and R. Kumar, “Heart disease prediction using machine learning algorithms,” in *2020 international conference on electrical and electronics engineering (ICE3)*, 2020, pp. 452–457.

[11] M. A. Alim, S. Habib, Y. Farooq, and A. Rafay, "Robust heart disease prediction: a novel approach based on significant feature and ensemble learning model," in *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, 2020, pp. 1–5.

[12] R. Kannan and V. Vasanthi, "Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease," in *Soft computing and medical bioinformatics*, Springer, 2019, pp. 63–72.

[13] R. Atallah and A. Al-Mousa, "Heart disease detection using machine learning majority voting ensemble method," in *2019 2nd international conference on new trends in computing sciences (ictcs)*, 2019, pp. 1–6.

[14] P. S. Kohli and S. Arora, "Application of machine learning in disease prediction," in *2018 4th International conference on computing communication and automation (ICCCA)*, 2018, pp. 1–4.

[15] A. Ed-Daoudy and K. Maalmi, "Real-time machine learning for early detection of heart disease using big data approach," in *2019 international conference on wireless technologies, embedded and intelligent systems (WITS)*, 2019, pp. 1–5.

[16] C. for D. C. and P. (CDC), "Center for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS)," *National Health and Nutrition Examination Survey (NHANES)*, 2020. http://www.cdc.gov/nchs/nhanes/ about_nhanes.htm (accessed Oct. 01, 2022).

[17] C. for D. C. and P. (CDC), "Center for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS)," *National Health and Nutrition Examination Survey (NHANES)*, 2020. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/ default.aspx?Cycle=2017-2020 (accessed Oct. 01, 2020).

[18] H. Kang, "The prevention and handling of the missing data.," *Korean J. Anesthesiol.*, vol. 64, no. 5, pp. 402–406, May 2013, doi: 10.4097/kjae.2013.64.5.402.

[19] Center for Disease Control and Prevention (CDC)., "Center for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS)," *Indicator Definitions Cardiovascular Disease*, 2020. https://www.cdc.gov/cdi/definitions/ cardiovascular-disease.html (accessed Oct. 01, 2022).

[20] S. Jain, S. Shukla, and R. Wadhvani, "Dynamic selection of normalization techniques using data complexity measures," *Expert Syst. Appl.*, vol. 106, pp. 252–262, 2018.

[21] P. D. Allison, *Missing data*. Sage publications, 2001.

[22] D. B. Rubin and N. Schenker, "Multiple imputation for interval estimation from simple random samples with ignorable nonresponse," *J. Am. Stat. Assoc.*, vol. 81, no. 394, pp. 366–374, 1986.

[23] O. Troyanskaya *et al.*, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[24] D. J. Stekhoven and P. Bühlmann, "MissForest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[25] N. V Chawla, N. Japkowicz, and A. Kotcz, "Special issue on learning from imbalanced data sets," *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 1–6, 2004.

[26] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, 2017.

[27] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic

minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[28] M. R. Smith, T. Martinez, and C. Giraud-Carrier, "An instance level analysis of data complexity," *Mach. Learn.*, vol. 95, no. 2, pp. 225–256, 2014.

[29] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, 2004.

[30] G. George and V. C. Raj, "Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile," *arXiv Prepr. arXiv1109.1062*, 2011.

[31] M. D. Anasanti, K. Hilyati, and A. Novtariany, "The Exploring feature selection techniques on Classification Algorithms for Predicting Type 2 Diabetes at Early Stage," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 5, pp. 832–839, 2022.

[32] N. Barraza, S. Moro, M. Ferreyra, and A. de la Peña, "Mutual information and sensitivity analysis for feature selection in customer targeting: A comparative study," *J. Inf. Sci.*, vol. 45, no. 1, pp. 53–67, 2019.

[33] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[34] Z. D. Yenice, N. Adhikari, Y. K. Wong, V. Aksakalli, A. T. Gumus, and B. Abbasi, "SPSA-FSR: Simultaneous perturbation stochastic approximation for feature selection and ranking," *arXiv Prepr. arXiv1804.05589*, 2018.

[35] K. Siddique, Z. Akhtar, H. Lee, W. Kim, and Y. Kim, "Toward bulk synchronous parallel-based machine learning techniques for anomaly detection in high-speed big data networks," *Symmetry (Basel).*, vol. 9, no. 9, p. 197, 2017.

[36] P. Geurts, A. Irrthum, and L. Wehenkel, "Supervised learning with decision tree-based methods in computational and systems biology," *Mol. Biosyst.*, vol. 5, no. 12, pp. 1593–1605, 2009.

[37] L. Jiang, Z. Cai, D. Wang, and S. Jiang, "Survey of improving k-nearest-neighbor for classification," in *Fourth international conference on fuzzy systems and knowledge discovery (FSKD 2007)*, 2007, vol. 1, pp. 679–683.

[38] S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, classifiaction," 1992.

[39] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[40] R. R. Sanni and H. S. Guruprasad, "Analysis of performance metrics of heart failured patients using Python and machine learning algorithms," *Glob. Transitions Proc.*, vol. 2, no. 2, pp. 233–237, 2021, doi: https://doi.org/10.1016/j.gltp.2021.08.028.

[41] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.

[43] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[44] MathWork, "Detector Performance Analysis Using ROC Curves - MATLAB," *Www.Mathworks.Com*. http://www.mathworks.com/help/phased/examples/detector-

performance-analysis-using-roc-curves.html (accessed Nov. 19, 2022).

[45] Q. Ren, X. Xie, Y. Tang, Q. Hu, and Y. Du, “Methyl tertiary-butyl ether inhibits THP-1 macrophage cholesterol efflux in vitro and accelerates atherosclerosis in ApoE-deficient mice in vivo,” *J. Environ. Sci.*, vol. 101, pp. 236–247, 2021, doi: https://doi.org/10.1016/j.jes.2020.08.011.