Hindawi
Journal of Sensors
Volume 2022, Article ID 8400622, 13 pages
https://doi.org/10.1155/2022/8400622

*Research Article*

# A Machine Learning in Binary and Multiclassification Results on Imbalanced Heart Disease Data Stream

**Danish Hamid,[1] Syed Sajid Ullah,[2,3] Jawaid Iqbal,[1] Saddam Hussain,[4] Ch. Anwar ul Hassan,[1] and Fazlullah Umar[5]**

[1]*Capital University of Science & Technology, Islamabad 44000, Pakistan*
[2]*Department of Information and Communication Technology, University of Agder (UiA), N-4898 Grimstad, Norway*
[3]*Department of Electrical and Computer Engineering, Villanova University, Villanova, PA 19085, USA*
[4]*School of Digital Science, Universiti Brunei Darussalam, Jalan Tungku Link, Gadong, Bandar Seri Begawan BE1410, Brunei Darussalam*
[5]*Khana-e-Noor University, Pol-e-Mahmood Khan, Shashdarak, 1001 Kabul, Afghanistan*

Correspondence should be addressed to Saddam Hussain; saddam_1993@hotmail.com

Received 5 July 2022; Revised 5 August 2022; Accepted 10 August 2022; Published 20 September 2022

Academic Editor: Praveen Kumar Donta



In medical filed, predicting the occurrence of heart diseases is a significant piece of work. Millions of healthcare-related complexities that have remained unsolved up until now can be greatly simplified with the help of machine learning. The proposed study is concerned with the cardiac disease diagnosis decision support system. An OpenML repository data stream with 1 million instances of heart disease and 14 features is used for this study. After applying to preprocess and feature engineering techniques, machine learning approaches like random forest, decision trees, gradient boosted trees, linear support vector classifier, logistic regression, one-vs-rest, and multilayer perceptron are used to perform binary and multiclassification on the data stream. When combined with the Max Abs Scaler technique, the multilayer perceptron performed satisfactorily in both binary (Accuracy 94.8%) and multiclassification (accuracy 88.2%). Compared to the other binary classification algorithms, the GBT delivered the right outcome (accuracy of 95.8%). Multilayer perceptrons, however, did well in multiple classifications. Techniques such as oversampling and undersampling have a negative impact on disease prediction. Machine learning methods like multilayer perceptrons and ensembles can be helpful for diagnosing cardiac conditions. For this kind of unbalanced data stream, sampling techniques like oversampling and undersampling are not practical.

## 1. Introduction

The healthcare industry generates a huge amount of data about patients, illnesses, and diagnosis, but because it has not been properly analyzed, it does not convey the significance that it should. The leading cause of death has been heart disease. In accordance with the World Health Organization, cardiovascular diseases (CVDs), which claim an approximate 17.9 million lives each year [1], are the leading cause of death worldwide.

Coronary heart disease, cerebrovascular disease, rheumatic heart disease, and some other conditions are among the group of heart and blood vessel disorders known as CVDs. Heart attacks and strokes account for four out of every five CVD deaths, and one-third of these deaths happen before the age of 70 [2]. Sex, age, smoking, cholesterol, family history, high blood pressure, poor diet, obesity, inactivity, and alcohol consumption are the main risk factors for heart disease [3]. Hereditary risk factors like high blood pressure and diabetes also contribute to the disease. Obesity, poor eating habits, and physical inactivity are a few additional lifestyle factors that increase the risk.

The main signs and symptoms include palpitations, sweating, fatigue, shortness of breath, arm and shoulder pain, back pain, and chest pain. The most typical sign of poor heart blood flow or a heart attack is still chest pain. Angina is the name for this kind of chest pain [4]. There are various tests, including X-rays, MRI scans, and angiography, to diagnose the illness.

However, there are instances where there is a lack of resources in an emergency because medical equipment is available at crucial moments. Every second counts in the diagnosis and treatment of diseases like cardiovascular disease. The potential for big data analytics to enhance cardiovascular quality of care and patient outcomes is enormous [5] because the cardiac centers and OPDs generate enormous amounts of data related to the diagnosis of heart disease. However, because of noise, incomplete information, and inconsistency, it is difficult to make precise, accurate, and consistent decisions using that data. Artificial intelligence (AI) is now playing a significant role in cardiology thanks to enormous advancements in technology, storage, acquisition, and knowledge recovery [6–10]. Researchers have preprocessed data using a variety of data mining techniques in order to make decisions using various models of machine learning [11, 12].

The contents of this paper focus on the R&D of a decision support system to predict heart disease using 14 feature clinical data. Literature Review presents the related research up till now. The proposed research explains the loopholes in the previous research and discusses one of the right approaches to diagnose the disease accurately. Methodology and Results provides preprocessing techniques through data mining. It presents the analysis, precision, and accuracy of machine learning algorithms that can be effective to diagnose heart problems through clinical data. In the end, Conclusion describes the performance, analysis, and comparisons between different types of algorithms on the model.

*1.1. Motivation and Contributions.* Heart disease has historically been the main cause of death. The World Health Organization lists cardiovascular diseases (CVDs) as the number one killer in the world, claiming 17.9 million lives annually [1]. The group of heart and blood vessel disorders known as CVDs includes conditions like coronary heart disease, cerebrovascular disease, rheumatic heart disease, and others. Four out of every five CVD deaths result from heart attacks or strokes, and one-third of these deaths occur before the age of 70 [2].

Here are the following substantial contributions of the proposed work:

(1) We start by addressing the problem of datasets, which we later refined and standardized. This is one of the proposed work's major contributions. Following that, the datasets are used to for training and testing the classifiers to see which ones offer the highest accuracy

(2) We then use the correlation matrix to determine the best values or features

(3) In the third step, we used the preprocessed dataset with the machine learning approaches to achieve the highest accuracy possible by fine-tuning the parameters

(4) The accuracy, recall, precision, and *F*-measure of the proposed classifiers are assessed

(5) In evaluation to the state-of-the-art accuracy listed in "Figures 1 and 2," the proposed classifiers provide better accuracy

The remainder of the document is structured as follows: Section 2 described the literature review. Section 3 provides the methodology while Section 4 presents the algorithms. Sections 5 discusses the results and discussion. Finally, Section 6 concludes the research.

## 2. Literature Review

The classification of heart disease using data mining and machine learning has been the subject of numerous studies and methodologies [13]. Al-Janabi provided a thorough analysis of the research on the use of machine learning in the field of heart disease. The author opined that a dataset with adequate samples and accurate data must be used to create an effective model for predicting heart disease. The dataset should be preprocessed appropriately, as this is the step that will have the biggest impact on how well the machine learning algorithm uses the dataset.

In the study, the author advocated for the use of a suitable algorithm, such as a decision tree (DT) or artificial neural network (ANN), when creating a prediction model. Decision tree and artificial neural network (ANN) both performed well in the majority of method for estimating heart disease (DT). Using data analytics tools and algorithms for machine learning like artificial neural networks (ANN), decision trees, fuzzy logic, *K*-nearest neighbors (KNN), Naive Bayes, and support vector machines, Marimuthu et al. [14] proposed a prediction of the heart disease model (SVM). The performance of the algorithm and an overview of previous work are also discussed in the paper. Yadav et al. [15] suggested an architecture that involves preprocessing of the input data before training and testing on various algorithms. Author emphasis using AdaBoost to increase the presentation of every ML algorithm. The author also supported the idea of parameter tuning to get good accuracies.

Sharma et al. [16] recommended a deep learning approach to diagnose heart disease using the UCI dataset of heart diseases. They suggested that heart disease diagnosing is one of the key zones where deep neural networks can be applied to improve the quality of classification. They presented that Talos hyperparameter optimization is more efficient than the other model optimization techniques. The prognosis of heart disease using machine learning models with high certitude, precision, and recall was discussed by Ramalingam et al. [17]. These models included KNN, SVM, DT, and RF algorithms. The support vector machine (SVM) classification in their prediction model had the highest accuracy of 86% for heart diseases in the UCI machine learning repository.

Ravindhar et al. [18] used four machine learning algorithms and one neural network to compare performance measurements to cardiac disease identification. To be able to predict cardiac attacks, the authors evaluated the algorithms' accuracy, precision, recall, and F1 settings. The deep

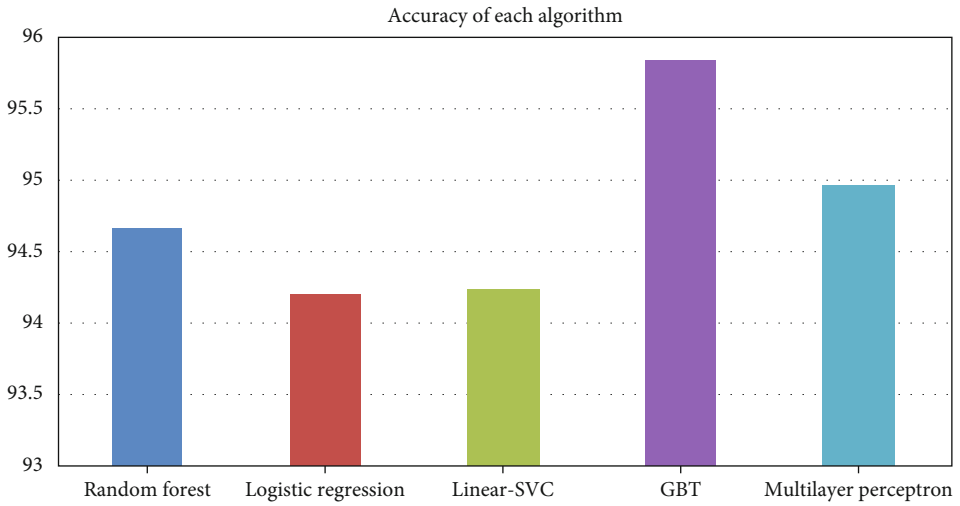


Figure 1: Binary classification algorithm accuracy.

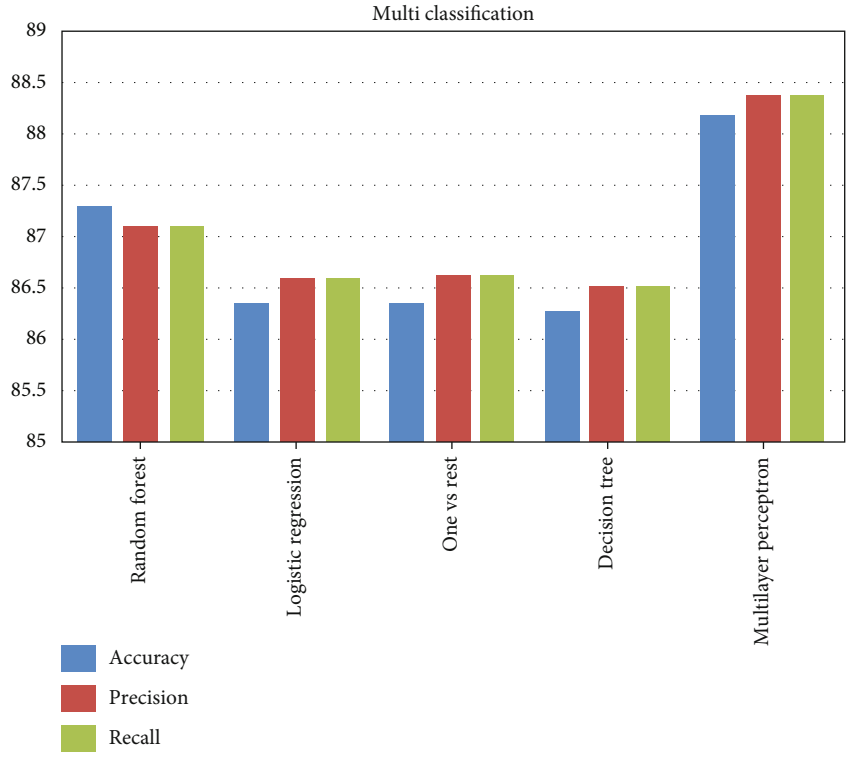


Figure 2: Multiclassification algorithm performance measures.

neural network algorithm achieved 98% accuracy in heart disease identification. In [19], Latha and Jeeva improves the prediction accuracy of heart disease using the ensemble classification models. In order to demonstrate the algorithm's value in early disease prediction, Latha and Jeeva [19] focuses on its application to a medical dataset. The study's findings show that ensemble techniques, like bagging and boosting, are useful for increasing the predictability of weak classifiers and perform admirably in calculating the risk of developing heart disease. Implementing feature selection improved the process' performance even further, and the results revealed a notable rise in prediction accuracy. Ensemble classification helped weak classifiers achieve an accuracy improvement of up to 7%.

The author of [20] compared ML classifiers on various datasets such as heart and diabetes datasets. The authors of [21] examined ML classifiers on medical insurance cost datasets. [22] used six popular data mining tools to categorize heart disease: using LR, KNN, SVM, RF, and KNIME, these tools were compared to six commonly used machine learning techniques. The most frequently encountered effective learning issue researched in the literature is single-label

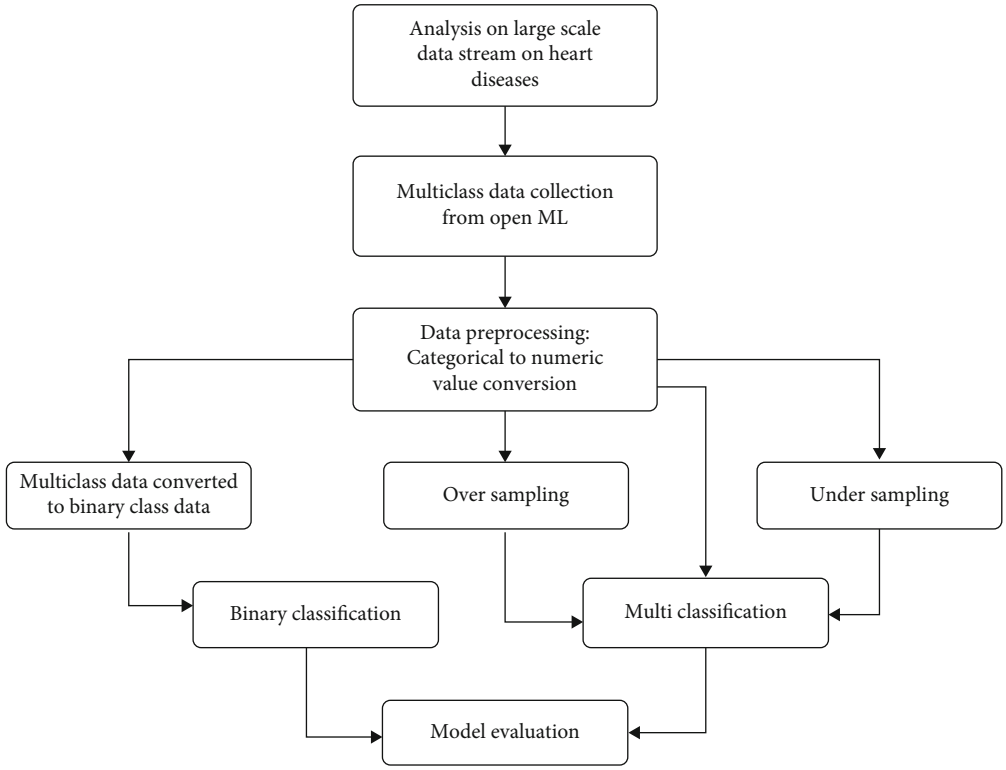


Figure 3: Working methodology.

classification. A training data classification algorithm is labeled for the instance that is the most ambiguous using the active learning strategy known as uncertainty sampling. Methods for sampling uncertainty are effective in terms of computation.

Despite the fact that they do not assess the candidate instance's future predictive informativeness on large amounts of unlabeled data, they have demonstrated good empirical performance [23]. A method for crystalline material prediction was proposed in [24] using the evolutionary optimization technique USPEX and machine-learning interatomic potentials consciously learning on the fly. [25] focused on the most effective methods to automatically represent configuration settings for the training set when developing moment tensor potentials, which were implemented in the MLIP package. [26] showed how to select automated hyperparameters solely through active learning. They enhanced the classification model hyperparameters that make up a super learner model using a Bayesian approach. To modify the solution close to the predicted optimum, they used simulations, deep learning training, and surrogate optimization.

They combined mixed data factor structure and RF-based MLA to create an autonomous systems framework in a paper [27]. RF was utilized to forecast disease by using the FAMD to find the relevant features. The proposed method had accuracy rates of 93.44 percent, sensitivity rates of 89.28 percent, and specificity rates of 96.96 percent. The same methodology was applied with a boosting hybrid model in [28], which resulted in accuracy of 75.9%. The boosting ensemble method was evaluated using the UCI laboratory dataset, with an ANN model attaining an accuracy of 82.5 percent and a hybrid model attaining a performance of 78.88 percent [29]. The prediction of heart disease is a research area that involves numerous researchers. Numerous aspects of cardiac illness are covered in their study. [30] finds that SVM performs better, averaging 96 percent accuracy. The DT model, according to the author in [31], consistently outperforms the NB and SVM models. According to its findings, SVM achieves an accuracy of 87%, DT achieves an accuracy of 90%, and LR achieves the highest accuracy in heart disease prediction when compared to DT, NB, SVM, and KNN, as shown in [32]. For the assessment of congenital heart disease, the RF-based framework's prediction accuracy is 97 percent [33], with a specificity of 88 percent and a sensitivity of 85 percent. With a specificity of 95% and a sensitivity of 93.5 percent, we used LR, EVF, MARS, and CART ML models in [34] to detect the co-occurrence of CVD and 94 percent.

Researchers put forth a number of ensemble and hybrid models for predicting cardiovascular disease in an effort to reach a better conclusion. On CVD datasets taken from the Mendeley Data Center, IEEE Data Port, and Cleveland datasets, respectively, the proposed models in [35] achieve 96, 93, and 88.24 percent accuracy. The author of [36] successfully combined the RF and LR models to predict heart disease with 88.7% accuracy. These studies' objective is to examine correlations between carotid plaque and coronary artery calcium in asymptomatic individuals, as well as their relationships to predict CVD occurrence risk [37]. The Internet of Things (IoT) and ML and deep learning are now widely used for disease detection and prediction. In [38], the author used

Table 1: Dataset description.

| No. | Feature name | Dataset description<br>Description | Value |
|---|---|---|---|
| 1. | Age | One of the most significant risk factors for heart disease is age. The likelihood of developing heart disease increases with age. | Integer value. |
| 2. | Sex | Displays the gender of the individual having heart disease or not. | Male = 1<br>Female = 0 |
| 3. | Chest pain type | It shows how often the person experiences chest pain. | Standard angina = 1<br>Angina atypical = 2<br>3 for nonanginal pain<br>Asymptotic = 4 |
| 4. | Resting blood pressure | Blood pressure is also one of the causes. Higher blood pressure occurs with other factors leads to the high risk. | It has whether integer or float value. |
| 5. | Cholesterol | The serum cholesterol is shown in mg/dl (unit). Your danger of a heart attack is also increased by having high blood levels of triglycerides, a form of fat connected to your diet. | It has whether integer or float value. |
| 6. | Fasting blood sugar | Compares a person's fasting blood sugar level to 120 mg/dl. A greater risk also results from an increase in blood sugar. | Value = 1 if fasting blood sugar > 120 (true)<br>If not, value is 0 (false) |
| 7. | Resting ECG | It displays resting electrocardiographic results. | Left ventricular hypertrophy = 2, ST − T wave abnormality = 1, and normal = 0 |
| 8. | Max heart rate achieved | This value is the highest heart rate achieved. It is linked with other factors such as blood pressure. | It has whether integer or float value. |
| 9. | Exercise-induced angina | It is the chest pain induced during the exercise. Although it usually starts in the center of your chest, it can also affect one or both shoulders, as well as your back, chest, jaw, or arm. | Yes = 1<br>No = 0 |
| 10. | Exercise-induced ST depression compared to rest | Integer values or float value | It has whether integer or float value. |
| 11. | Peak ST segment of exercise | When you see a horizontal or downward-sloping ST-segment depression of less than 1 mm 60–80 ms after the J point on a treadmill ECG stress test, it is deemed abnormal. | Upsloping = 1<br>Flat = 2<br>Downsloping = 3 |
| 12. | Major vessels colored by fluorescence in number | Integer values or float value | It has whether integer or float value. |
| 13. | Thalassemia | It displays thalassemia. | Normal = 3<br>Fixed defect = 6<br>Reversible defect = 7 |
| 14. | Heart disease diagnosis | The fact that the person is either healthy or struggling from heart disease is indicated. | Absence = 0<br>Present = 1, 2, 3, 4 |

mobile technology and the deep learning approach to predict heart disease with an accuracy of 94%. The author combines IoT with ML classifiers for early heart infection prediction in [39]. The goal is to show how ML can be used to resolve the issue. By examining hundreds of healthcare data sets, we use machine learning to analyze cases that are related to diseases and other health issues [40].

## 3. Methodology

On data streams with multiple classifications as well as binary data, we have applied machine learning techniques. The steps of our process following are displayed in "Figure 3."

*3.1. Dataset.* The large imbalanced heart disease data stream is gained from the OpenML repository. In the OpenML repository, various domain data streams are available. The imbalanced data stream consists of 14 attributes, 1000000 instances, and 5 target classes. The data stream is uploaded by Jan Van Rijn in 2014 in the OpenML repository. For binary classification, the multiclass data is converted into binary class by replacing target variable values 2, 3, and 4 with 1. The dataset description is in "Table 1."

*3.2. Data Descriptive Statistics.* Data descriptive statistics is described in Table 2, such as Min., Max., mean, standard deviation, and variance.

*3.3. Instance per Class.* "Figure 4" is a graphical representation of data distribution, including the number of instances in each class of heart disease dataset.

*3.4. Preprocessing.* The data stream consists of nominal and numerical values feature sets. Many ML algorithms do not

Table 2: Dataset descriptive statistics.

| | N | Minimum | Maximum | Mean | Std. deviation | Variance | Skewness | |
|---|---|---|---|---|---|---|---|---|
| Age | 1000000 | 27 | 82 | 54.38 | 9.082 | 82.475 | -0.195 | 0.002 |
| Sex | 1000000 | 0 | 1 | 0.68 | 0.486 | 0.217 | -0.775 | 0.002 |
| cp | 1000000 | 1 | 4 | 3.14 | 0.964 | 0.929 | -0.919 | 0.002 |
| trestps | 1000000 | 94.03 | 210.08 | 131.6239 | 17.53299 | 307.406 | 0.626 | 0.002 |
| chol | 1000000 | 100.76 | 529.66 | 246.3325 | 51.80040 | 2683.282 | 0.766 | 0.002 |
| rbs | 1000000 | 0 | 1 | 0.15 | 0.360 | 0.129 | 1.932 | 0.002 |
| Restecg | 1000000 | 0 | 2 | 0.99 | 0.991 | 0.982 | 0.028 | 0.002 |
| Thalach | 1000000 | 62.09 | 211.39 | 149.6228 | 22.88460 | 523.705 | -0.514 | 0.002 |
| Exang | 1000000 | 0 | 1 | 0.33 | 0.472 | 0.222 | 0.705 | 0.002 |
| Oldpeak | 1000000 | .64 | 6.96 | 1.0431 | 1.16101 | 1.348 | 1.173 | 0.002 |
| Slope | 1000000 | 1 | 3 | 1.61 | 0.632 | 0.399 | 0.535 | 0.002 |
| Ca | 1000000 | 0 | 3 | 0.72 | 0.965 | 0.932 | 1.108 | 0.002 |
| Thal | 1000000 | 3 | 7 | 4.74 | 1.932 | 3.734 | 0.236 | 0.002 |
| Valid N (listwise) | 1000000 | | | | | | | |

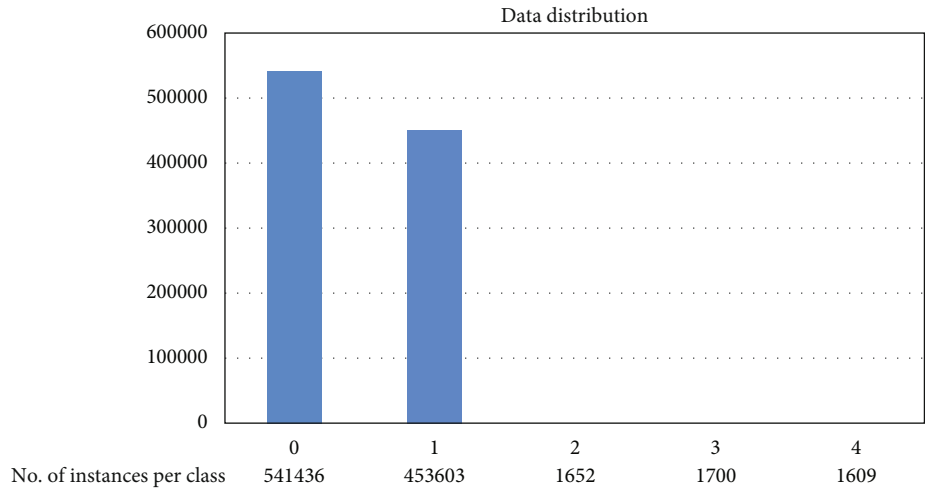


Figure 4: Dataset description.

process the nominal values; hence, these values need to be converted in numerical. In this approach, nominal values are replaced by the following table [1]. Also, some of the features in the dataset have relatively large values than the others which results in biased learning. In this approach, we have applied the Max Abs Scaler technique to the dataset [41].

*3.5. Feature Engineering.* Data mining techniques are used to create features from raw data using the process of feature engineering, which enhances the performance of ML algorithms. Feature importance provides the score for each feature of the dataset. The higher is the score, the more important feature is toward the target variable as shown in "Figure 5."

*3.6. Correlation Matrix with Heat Map.* It is simple to see which functionalities are most related to other characteristics or the target variable using a heat map [42]. Results are displayed in "Figure 6."

*3.7. Splitting.* For the purpose of gathering training and test data for the analysis process, splitting is used. The entire data stream is split into train and test sets, with training data accounting for 70% of the data and testing data for the remaining 30%.

*3.8. Classification.* The training data is trained by using seven different ML algorithms for binary and multiclassification. The detail of models is shown in Table 3 [2].

## 4. Algorithms

In this paper, some ML algorithms are applied to the large imbalanced data stream of heart diseases to see them.

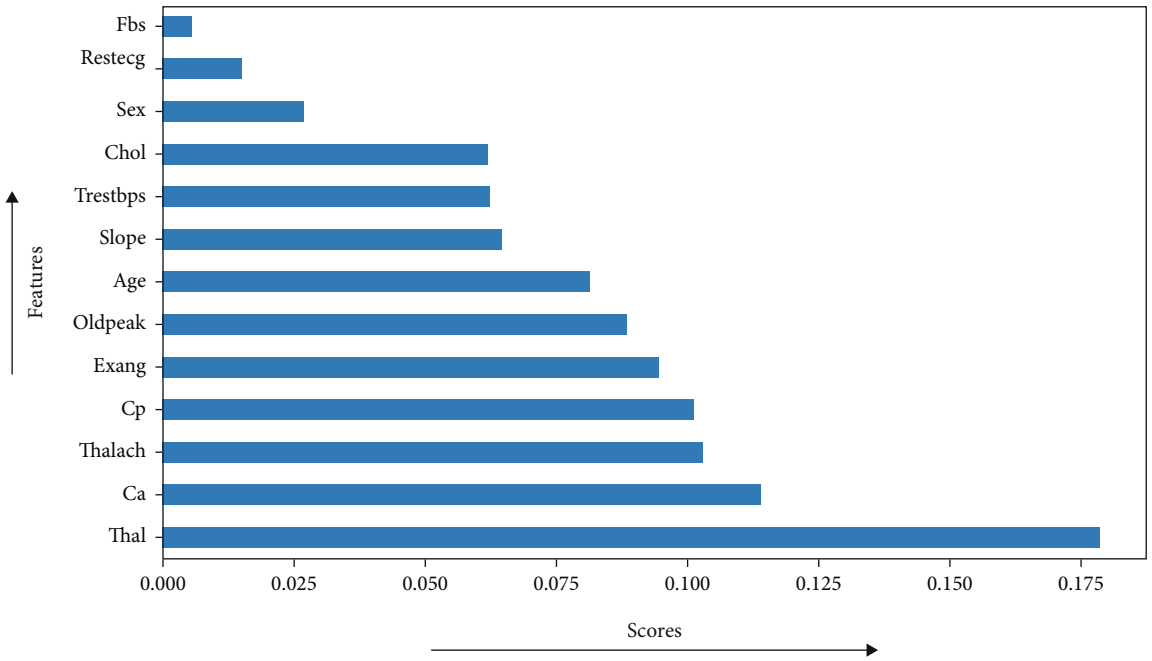


Figure 5: Feature importance.

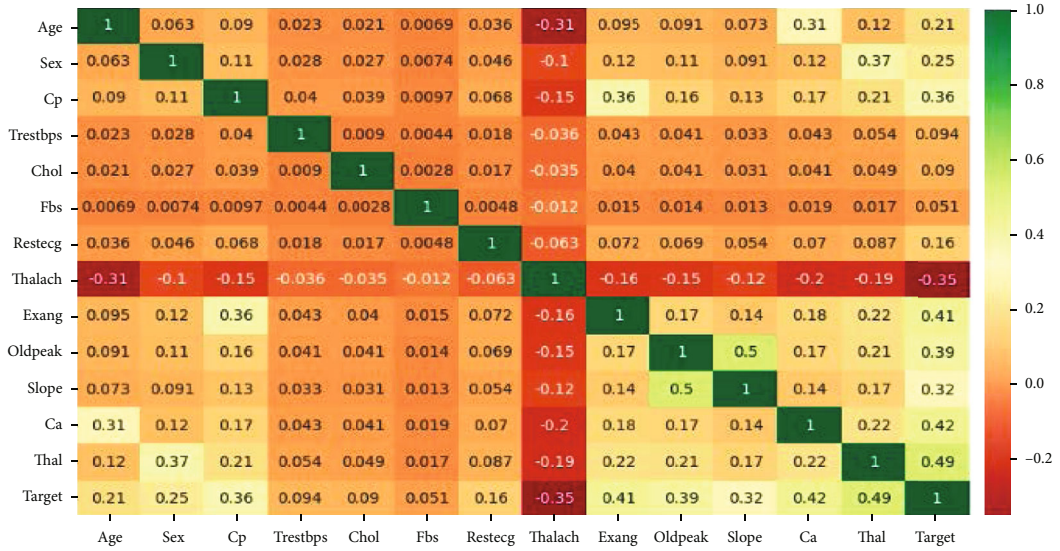


Figure 6: Heat map correlation matrix.

Table 3: Classification algorithms.

| Classification techniques | DT | RF | GBT | LSVC | LR | One-vs-rest | MLP |
|---|---|---|---|---|---|---|---|
| Binary classification | ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ |
| Multiclassification | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ |

*4.1. Decision Tree.* The most effective and well-liked tool for prediction and classification is the decision tree. By learning straightforward decision rules implied from data features, the decision tree forecasts the value of the target variable. In most cases, the decision rules are made up of if-then-else statements. The complexity of the rules as well as filter model increases with the depth of the tree [43].

*4.2. Random Forest.* One of the most well-liked and potent supervised machine learning algorithms, random forest or

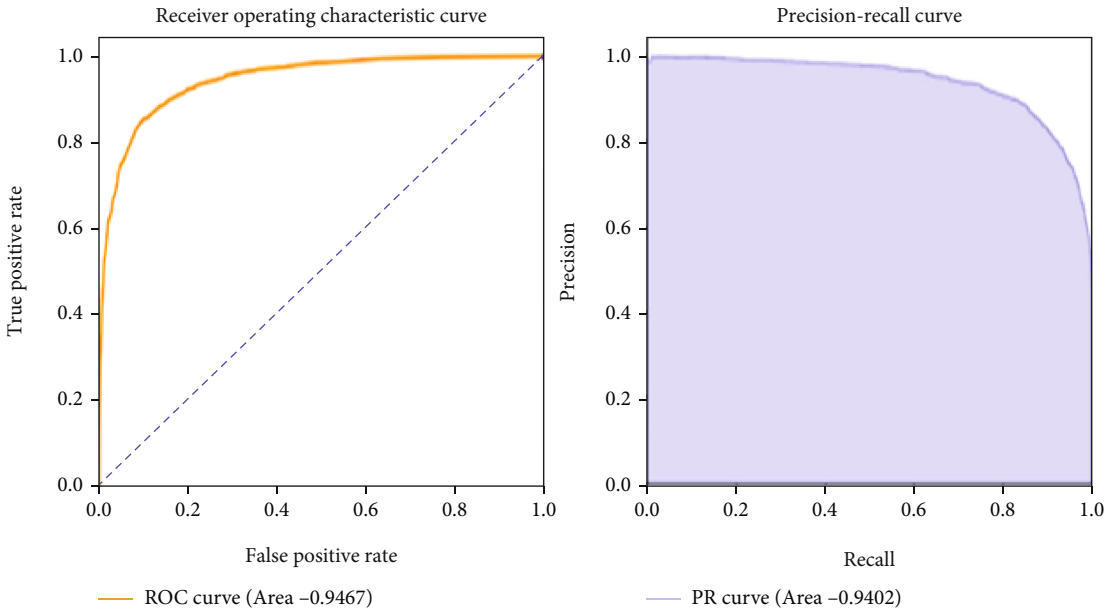


Figure 7: PR curve and ROC for random forest.

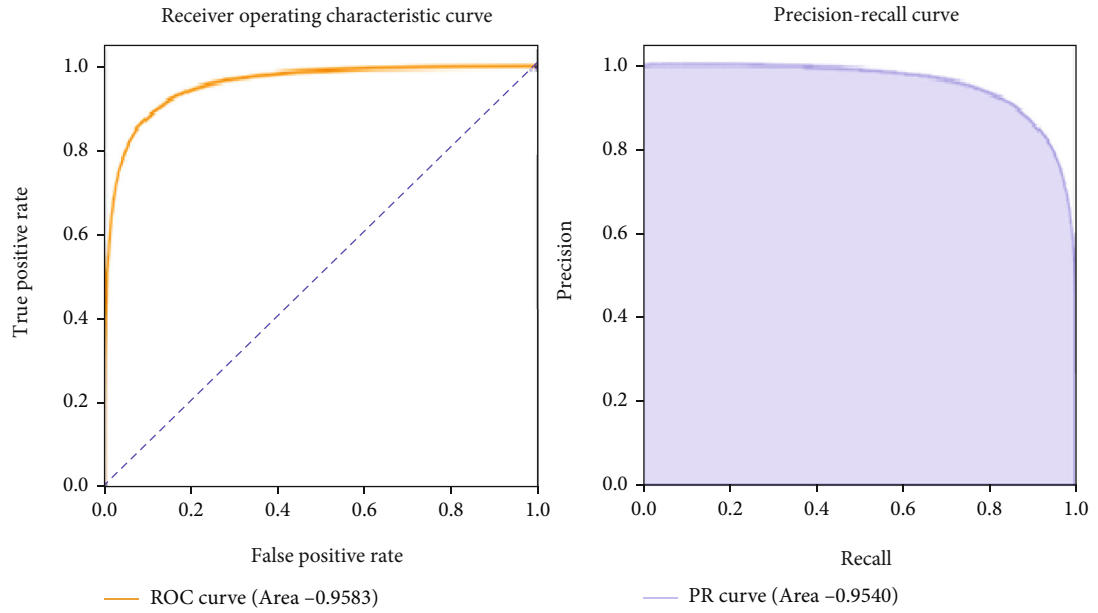


Figure 8: PR curve and ROC for GBTs.

extremely randomized forest can carry out both regression and classification tasks. It develops a decision-tree-filled forest. In general, a prediction is more accurate and robust the more trees there are in the forest. While regression takes the estimate of the outputs from various trees, classification uses a voting system to determine which class received the most votes from all the other trees within the forest. Additionally, it successfully manages higher dimensional large datasets. [44].

*4.3. Gradient Boosting Tree.* Gradient boosted tree learners are combined to form a strong learner, known as boosting. The GBT uses the same technique as AdaBoost in which equal weights are assigned to each of the observations. It decreases the masses of those observations which are easy to classify as well as increases of those which are difficult to classify. The second tree is grown using new weights. New predictions are made, and the process repeats itself until several iterations [45]. The gradient is different in such a way that it uses gradient in the loss function as in "Eq. (1)."

$$y = ax + b + e. \tag{1}$$

Here "$e$" indicates the error that means how much algorithms are good at predicting than the actual class.

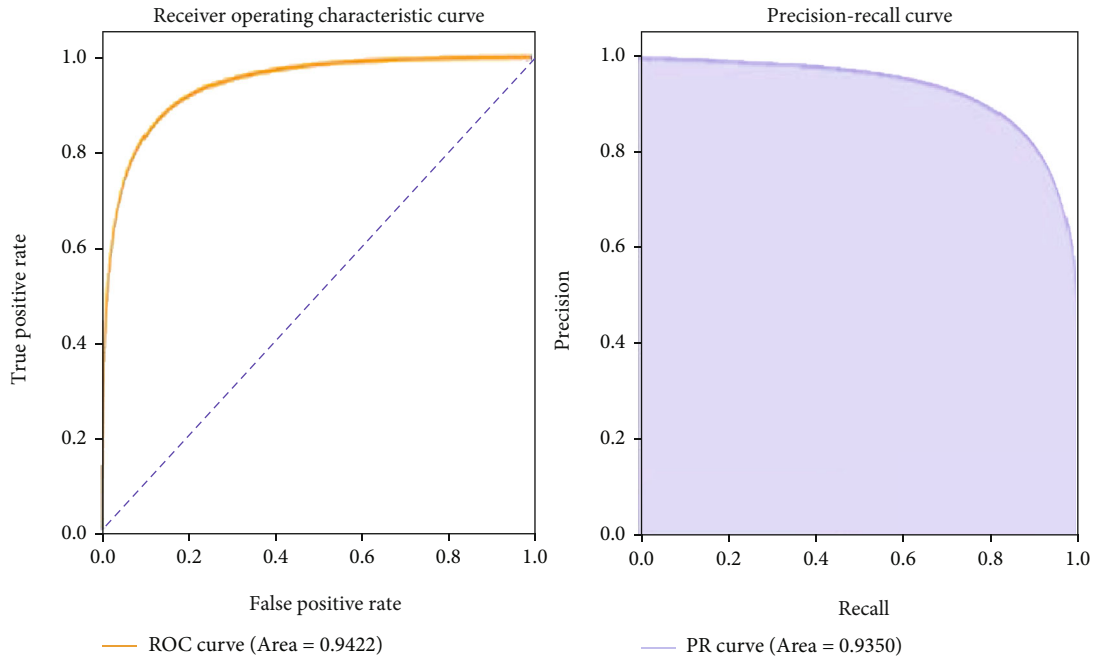


Figure 9: PR curve and ROC for linear support vector classifier.

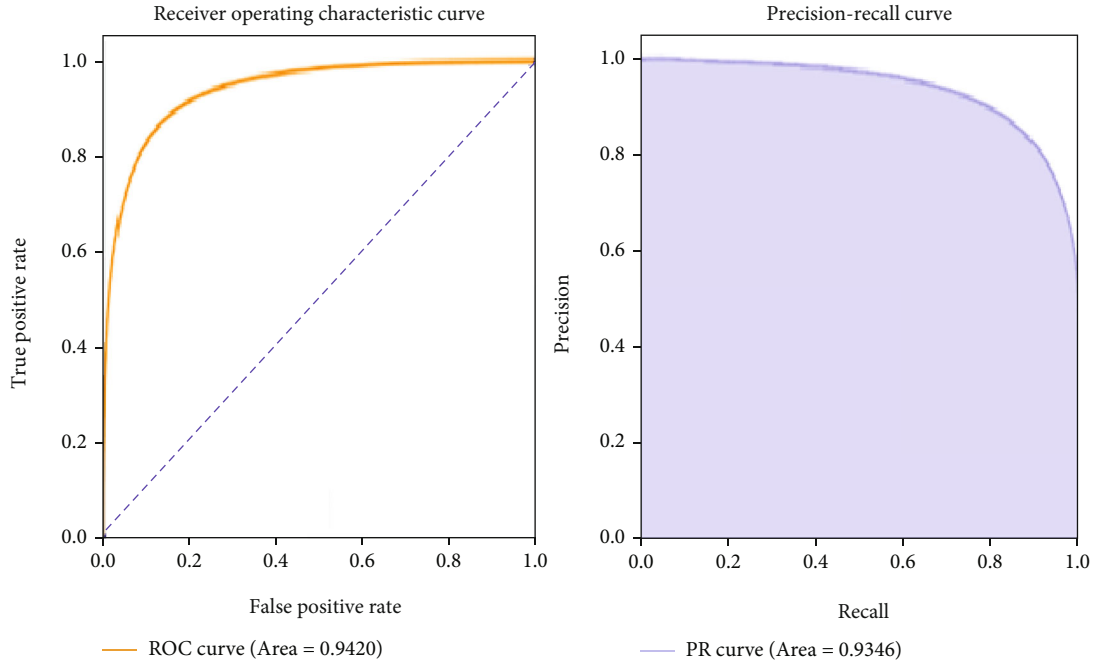


Figure 10: PR curve and ROC for logistic regression.

*4.4. Linear Support Vector Classifier.* The most effective technique which is used for binary classification is linear support vector classifier. Its objective is to learn from the data that is provided and returns the "best fit" hyperplane. The hyperplane is the decision boundary that helps classify the data points. The hyperplane dimension depends upon the features number. In this case, as the number of features is two, so hyperplane is just a straight separating line between the two classes. While building the SVC model, support vectors help in maximizing the margin so that a perfect boundary should be created [46].

*4.5. One-vs-Rest.* The one-vs-rest algorithm uses the problem transformation technique, in which a multiclass problem is divided into multiple binary problems [47]. It makes use of binary different classifiers for multiclass classification using a heuristic approach. The multiclass dataset is divided into various binary classification issues. Since there are an equal

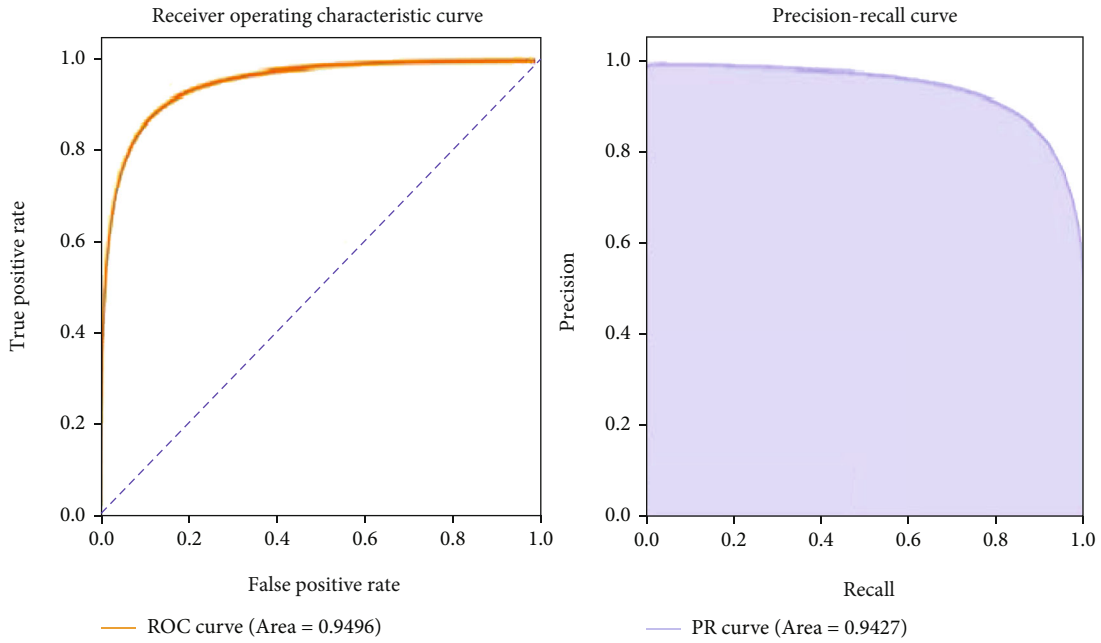


Figure 11: PR Curve and ROC for Multilayer Perceptron.

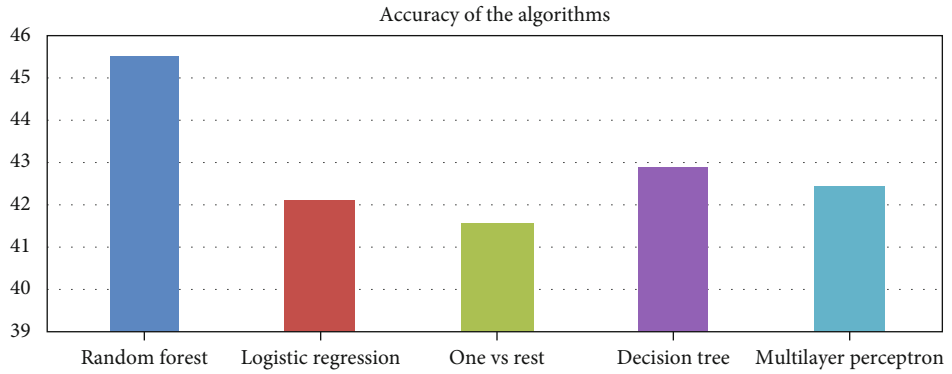


Figure 12: Oversampling algorithm accuracy.

variety of classes in the dataset, an equal number of models are created. The most certain model is then used to make predictions. Each model predicts response and membership probability. Class is chosen for which the respective model gave a positive response and the highest probability score [48].

*4.6. Logistic Regression.* Popular categorical response prediction techniques include logistic regression, which is a special case of generalized linear models that forecast the possibility of a target variable. It is the approach of choice for classification issues. A linear model or sigmoid function, which is a nonlinear function, is used to transform the output prediction. Logistic regression can be used for complex datasets where it can build more complex decision boundaries [49].

*4.7. Multilayer Perceptron.* A subclass of feedforward neural networks is the multilayer perceptron (ANN). It has several layers and produces a set of outputs from a set of inputs. It typically has an input data, a hidden layer, and an output layer or at least three layers of nodes. The hidden layer uses a linear combination of data with the weights and bias of each node and appears to apply the activation function to map the inputs to outputs. The input layer represents the input data. Backpropagation is used to train the network [50].

## 5. Result and Analysis

Results of binary and multiclassification are discussed in this section.

*5.1. Binary Classification Result.* Several classification approaches are used for the classification, and their performance is measured. Accuracy alone is not enough for the evaluation. Binary classification algorithm accuracy is shown in Figure 1, and for that, the values of precision and recall are calculated, and ROC and PR curves are generated. Below

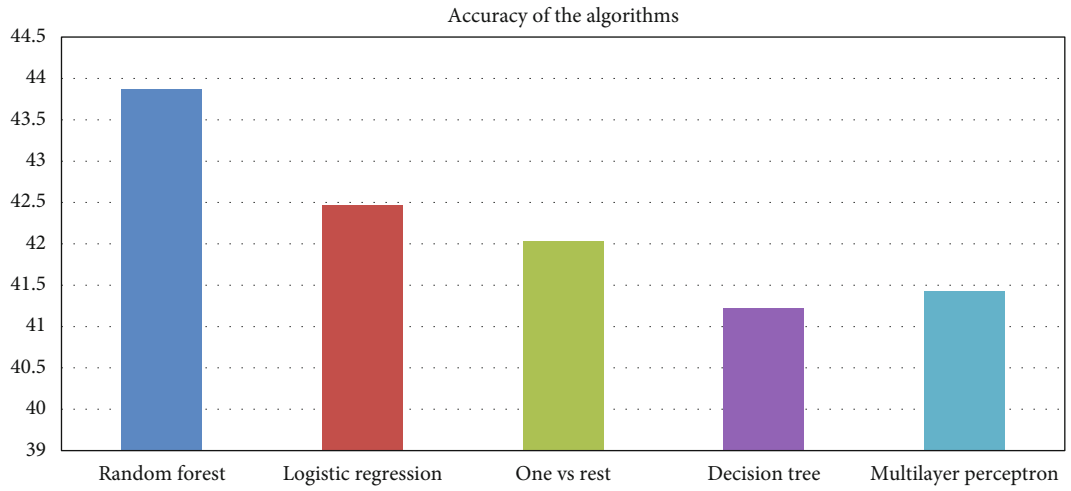


Figure 13: Undersampling algorithm accuracy.

diagrams illustrate the performance of our generated models (Figure 7–11) RF, GBT, LSVC, LR, and MLP, respectively.

Following are the PR and ROC curves for applied models:

*5.2. Multiclassification Result.* Numerous machine learning algorithms are used in multiclassification to analyze the heart data stream, and their evaluated accuracies are shown in Figure 2.

## 6. Sampling

As the multiclassification provides biased or wrong results due to imbalanced data in the data stream, sometimes, all the data of classes 2, 3, and 4 splits either into training or testing data. To handle this, oversampling and undersampling balancing techniques are applied to the data stream using the abovementioned classification algorithm table [2] and measured the accuracies.

*6.1. Oversampling.* Oversampling involves the random selection of examples from the minority class, with replacement and adding them to the training data set [51]. Results are shown in "Figure 12."

*6.2. Undersampling.* Undersampling involves the random selection of examples from the majority class and deleting them from the training dataset as in "Figure 13."

## 7. Conclusion

In this paper, some ML algorithms are practical to the large imbalanced data stream of heart diseases to see their behavior. In this approach, the heart disease dataset from the OpenML repository utilized for training in addition testing purposes. Classification of heart diseases trailed the steps of preprocessing and features engineering and data splitting, then classification, and evaluation. In the case of both binary classification and multiclassification, only the accuracy of multilayer perceptron improved by 3% after applying Max Abs Scaler, whereas the rest of the algorithms does not have such effect of Max Abs Scaler on their accuracies. Also, in both binary and multiclassification, the multilayer perceptron classifier performed adequately. For binary classification, the classification algorithms, random forest, logistic regression, GBT, linear SVC, and multilayer perceptron, provide high accuracy scores where the imbalance rate in the data stream is low, whereas in a multiclassification where imbalance rate in the data stream is high, the classification algorithms, random forest, logistic regression, decision tree, one vs rest and multilayer perceptron, provide fewer accuracy scores. Also, on this type of large imbalanced data stream, balancing techniques like oversampling and undersampling have an adverse effect on the accuracy of the data.

## Data Availability

The data used in this research can be obtained from the corresponding authors upon request.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References

[1] *World Health Organization. Cardiovascular Diseases (CVDs)*, 2022, https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1.

[2] *World Health Organization. Cardiovascular Diseases (CVDs)*, 2022, http://origin.who.int/cardiovascular_diseases/en/.

[3] 2022, https://www.heart.org/en/health-topics/high-blood-pressure/why-high-blood-pressure-is-a-silent-killer/know-your-risk-factors-for-high-blood-pressure.

[4] C. Balla, R. Pavasini, and R. Ferrari, "Treatment of angina: where are we?," *Cardiology*, vol. 140, no. 1, pp. 52–67, 2018.

[5] J. S. Rumsfeld, K. E. Joynt, and T. M. Maddox, "Big data analytics to improve cardiovascular care: promise and challenges," *Nature Reviews Cardiology*, vol. 13, no. 6, pp. 350–359, 2016.

[6] S. M. Arif, B. A. Bacha, S. S. Ullah, S. Hussain, and M. Haneef, "Tunable control of internet of things information hacking by

application of the induced chiral atomic medium," *Soft Computing.*, vol. 4, pp. 1–8, 2022.

[7] S. Hussain, I. Ullah, H. Khattak, M. A. Khan, C. M. Chen, and S. Kumari, "A lightweight and provable secure identity-based generalized proxy signcryption (IBGPS) scheme for Industrial Internet of Things (IIoT)," *Journal of Information Security and Applications.*, vol. 58, no. 58, article 102625, 2021.

[8] S. S. Ullah, S. Hussain, A. Gumaei, and H. AlSalman, "A secure NDN framework for Internet of Things enabled healthcare," *Continua*, vol. 67, no. 1, pp. 223–240, 2021.

[9] S. Hussain, S. Sajid Ullah, M. Shorfuzzaman, M. Uddin, and M. Kaosar, "Cryptanalysis of an online/offline certificateless signature scheme for Internet of Health Things," *Intelligent Automation & Soft Computing.*, vol. 30, no. 3, pp. 983–993, 2021.

[10] T. Hussain, D. Hussain, I. Hussain et al., "Internet of Things with deep learning-based face recognition approach for authentication in control medical systems," *Computational and Mathematical Methods in Medicine.*, vol. 2022, article 5137513, pp. 1–17, 2022.

[11] K. W. Johnson, J. T. Soto, B. S. Glicksberg et al., "Artificial intelligence in cardiology," *Journal of the American College of Cardiology*, vol. 71, no. 23, pp. 2668–2679, 2018.

[12] K. Gomathi and D. D. S. Priyaa, "Multi disease prediction using data mining techniques," *International Journal of System and Software Engineering*, vol. 4, no. 2, pp. 12–14, 2016.

[13] M. Aljanabi, H. M. Qutqut, and M. Hijjawi, "Machine learning classification techniques for heart disease prediction: a review," *International Journal of Engineering & Technology*, vol. 7, no. 4, pp. 5373–5379, 2018.

[14] M. Marimuthu, M. Abinaya, K. S. Hariesh, K. Madhankumar, and V. Pavithra, "A review on heart disease prediction using machine learning and data analytics approach," *International Journal of Computer Applications*, vol. 181, no. 18, pp. 20–25, 2018.

[15] K. K. Yadav, A. Sharma, and A. Badholia, "Heart disease prediction using machine learning techniques," *Information Technology In Industry*, vol. 9, no. 1, pp. 207–214, 2021.

[16] S. Sharma and M. Parmar, "Heart diseases prediction using deep learning neural network model," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 9, no. 3, pp. 2244–2248, 2020.

[17] V. V. Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques : a survey," *Engineering & Technology*, vol. 7, no. 2.8, pp. 684–687, 2018.

[18] N. V. Ravindhar, H. S. Anand, and G. W. Ragavendran, "Intelligent diagnosis of cardiac disease prediction using machine learning," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 11, pp. 1417–1421, 2019.

[19] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, article 100203, 2019.

[20] C. A. U. Hassan, M. S. Khan, and M. A. Shah, "Comparison of machine learning algorithms in data classification," in *2018 24th International Conference on Automation and Computing (ICAC)*, Newcastle Upon Tyne, UK, 2018.

[21] C. A. U. Hassan, J. Iqbal, S. Hussain, H. AlSalman, M. A. Mosleh, and S. Sajid Ullah, "A computational intelligence approach for predicting medical insurance cost," *Mathematical Problems in Engineering*, vol. 2021, Article ID 1162553, 13 pages, 2021.

[22] I. Tougui, A. Jilbab, and J. El Mhamdi, "Heart disease classification using data mining tools and machine learning techniques," *Health Technology*, vol. 10, no. 5, pp. 1137–1144, 2020.

[23] F. K. Nakano, R. Cerri, and C. Vens, "Active learning for hierarchical multi-label classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 5, pp. 1496–1530, 2020.

[24] E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, and A. R. Oganov, "Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning," *Physical Review B*, vol. 99, no. 6, p. 64114, 2019.

[25] I. S. Novikov, K. Gubaev, E. V. Podryabinkin, and A. V. Shapeev, "The MLIP package: moment tensor potentials with MPI and active learning," *Machine Learning: Science and Technology*, vol. 2, no. 2, 2021.

[26] O. Owoyele, P. Pal, and A. V. Torreira, "An automated machine learning-genetic algorithm framework with active learning for design optimization," *Journal of Energy Resources Technology*, vol. 143, no. 8, 2021.

[27] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, "MIFH: a machine intelligence framework for heart disease diagnosis," *IEEE Access*, vol. 8, pp. 14659–14674, 2020.

[28] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informatics in Medicine Unlocked*, vol. 20, no. 1, p. 100402, 2020.

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

[30] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, "A method for improving prediction of human heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2022, Article ID 1410169, 9 pages, 2022.

[31] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, "Multiple disease prediction using machine learning algorithms," *Materials Today: Proceedings.*, vol. 67, no. 6, 2021.

[32] C. Gupta, A. Saha, N. S. Reddy, and U. D. Acharya, "Cardiac disease prediction using supervised machine learning techniques," *Journal of Physics: Conference Series*, vol. 2161, no. 1, p. 012013, 2022.

[33] V. T. Truong, B. P. Nguyen, T. H. Nguyen-Vo et al., "Application of machine learning in screening for congenital heart diseases using fetal echocardiography," *The International Journal of Cardiovascular Imaging*, vol. 38, no. 5, pp. 1007–1015, 2022.

[34] A. S. Abdalrada, J. Abawajy, T. Al-Quraishi, and S. M. S. Islam, "Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study," *Journal of Diabetes & Metabolic Disorders*, vol. 21, no. 1, pp. 251–261, 2022.

[35] B. P. Doppala, D. Bhattacharyya, M. Janarthanan, and N. Baik, "A reliable machine intelligence model for accurate identification of cardiovascular diseases using ensemble techniques," *Journal of Healthcare Engineering*, vol. 2022, Article ID 2585235, 13 pages, 2022.

[36] A. Kondababu, V. Siddhartha, B. B. Kumar, and B. Penumutchi, "A comparative study on machine learning based heart disease prediction," *Materials Today: Proceedings.*, vol. 67, no. 6, 2021.

[37] E. F. Gudmundsson, G. Björnsdottir, S. Sigurdsson et al., "Carotid plaque is strongly associated with coronary artery calcium and predicts incident coronary heart disease in a population-based cohort," *Atherosclerosis*, vol. 346, pp. 117–123, 2022.

[38] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," *Computational intelligence and neuroscience*, vol. 2021, Article ID 8387680, 11 pages, 2021.

[39] A. Kishor and W. Jeberson, "Diagnosis of heart disease using Internet of Things and machine learning algorithms," in *In Proceedings of Second International Conference on Computing, Communications, and Cyber-Security*, pp. 691–702, Springer, Singapore, 2021.

[40] G. Gupta, U. Adarsh, N. S. Reddy, and B. A. Rao, "Comparison of various machine learning approaches uses in heart ailments prediction," *Journal of Physics: Conference Series*, vol. 2161, no. 1, p. 012010, 2022.

[41] 2022, https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.MaxAbsScaler.html.

[42] 2022, https://medium.com/mindorks/what-is-feature-engineering-for-machine-learning-d8ba3158d97a.

[43] 2022, https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e.

[44] 2022, https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/.

[45] 2022, https://en.wikipedia.org/wiki/Random_forest#:~:text=Random%20forests%20or%20random%20decision,prediction%20(regression)%20of%20the%20individual.

[46] 2022, https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab.

[47] 2022, https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

[48] 2022, https://en.wikipedia.org/wiki/Multiclass_classification#:~:text=%2Drest,all%20other%20samples%20as%20negatives.

[49] 2022, https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/.

[50] 2022, https://spark.apache.org/docs/2.1.0/ml-classification-regression.html#logistic-regression.

[51] 2022, https://spark.apache.org/docs/2.1.0/ml-classification-regression.html#multilayer-perceptron-classifier.