Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 7 (2025) 2385

Journal of the Nigerian Society of Physical Sciences

The effect of imbalance data mitigation techniques on cardiovascular disease prediction

Raphael Ozighor Enihe[a,*], Rajesh Prasad[b,c], Francisca Nonyelum Ogwueleka[c], Fatimah Binta Abdullahi[c]

[a]*Department of Computer Science, Baze University, Abuja, Nigeria*
[b]*Department of Computer Science & Engineering, Ajay Kumar Garg Engineering College, Ghaziabad, India*
[c]*Department of Computer Science, University of Abuja, Abuja, Nigeria*

Abstract

The prevalence of class imbalance is a common challenge in medical datasets, which can adversely affect the performance of machine learning models. This paper explores how several data imbalance mitigation techniques affect the performance of cardiovascular disease prediction. This study applied various data balancing techniques on a real-life cardiovascular disease (CVD) dataset of 1000 patient records with 14 features obtained from the University of Abuja Teaching Hospital Nigeria to address this problem. The data balancing techniques used include random under-sampling, Synthetic Minority Over-sampling Technique (SMOTE), Synthetic Minority Oversampling-Edited Nearest Neighbour (SMOTE-ENN), and the combination of SMOTE and Tomek Links undersampling (SMOTE-TOMEK). After applying these techniques, their performance was evaluated on seven machine learning models, including Random Forest, XGBoost, LightGBM, Gradient Boosting, K-Nearest Neighbours, Decision Tree, and Support Vector Machine. The evaluation metrics used are precision, recall, F1-score, accuracy, and receiver operating characteristic-area under the curve (ROC-AUC). Learning curve plots were also used to showcase the impact of the different data balancing techniques on the challenges of overfitting and underfitting. The results showed that the application of data balancing techniques significantly enhances the performance of machine learning models in heart disease prediction and effectively addresses the challenges of overfitting and underfitting with SMOTE-TOMEK, yielding the best-balanced fit as well as the highest precision, recall, F1-score, accuracy of 92%, and ROC-AUC of 96% on the Lightweight Gradient Boosting Machine (LightGBM) model. These results underscore the critical role of data balancing in predictive modelling for heart disease and highlight the effectiveness of specific techniques and models in achieving accurate, more reliable, and generalised predictions.

DOI:10.46481/jnsps.2025.2385

Keywords: Imbalance dataset, Cardiovascular disease prediction, SMOTE-TOMEK, Marchine learning, Overfitting and Underfitting

Article History :
Received: 24 September 2024
Received in revised form: 10 December 2024
Accepted for publication: 14 December 2024
Available online: 19 January 2025

Communicated by: Oluwatobi Akande

1. Introduction

Cardiovascular disease (CVD) is a major worldwide health issue; for every year, it is estimated to cause 31% of deaths globally (WHO, 2024). In low and middle-income nations especially, CVD causes a very heavy burden, with about 80%

*Corresponding author Tel. No.: +234-803-420-9386
Email address: raphael.enihe@bazeuniversity.edu.ng (Raphael Ozighor Enihe)

of all CVD fatalities reported occurring in these countries [1]. Similar to many developing countries, CVD remains a growing threat in Nigeria primarily due to factors such as urbanisation, efficient changes in life patterns, and longer years to live [2]. It is imperative to diagnose CVD risk early and provide precise predictions to ensure early interventions and assessments. Over the last few years, the next approaches of machine learning (ML) have demonstrated potential in improving CVD risk estimation, which appears to provide superiority to the conventional statistical techniques [3]. However, the general development of reliable and accurate ML models for CVD prediction encounters numerous challenges, among which is the class imbalance problem prominent in medical datasets.

The case of class distribution where one class is significantly smaller than the other class, for example, patients with CVD are fewer than healthy people, is typical for medical data [4, 5]. This can lead to the development of models that, although useful in the real world, offer very high true positive rates for the majority class and a very poor performance or identification of the minority class cases, even those that are likely to have more clinical relevance [6]. However, this has been seen as a challenge, and to resolve this, a number of data balancing approaches have been advocated for and have been used in different areas of healthcare predictive modelling. Accordingly, these methods are intended to provide a better distribution of data and seemingly lead to better performance and better generalisation of ML algorithms. Some of them are random sampling of the majority class, synthetic minority oversampling technique, and a few others are SMOTE-ENN and SMOTE-TOMEK [7, 8].

As these data balancing techniques have been demonstrated useful in numerous medical applications, their potential for CVD risk prediction as well as in the more diverse populations remains relatively uninvestigated. There is a dearth of such knowledge, particularly as it concerns the African population groups, which form the larger part of the global population, yet their representation in the informatics literature in the field of global health is limited [9].

To fill this knowledge gap, our study proposes to investigate the effect of several approaches to data imbalance handling on the performance of developing CVD prediction models. We use a data set from the University of Abuja Teaching Hospital in Nigeria and therefore add to the developing literature on CVD risk in Africans. Using four approaches that include random under-sampling, SMOTE over-sampling, SMOTE-ENN, and SMOTE-TOMEK, it is proposed to increase the predictive precision as well as the consistency of the ML models in the evaluation of CVD risk. Furthermore, we evaluate the performance of seven widely used ML algorithms: random forest, XGBoosting, light GBM, gradient boosting, KNN, decision tree, and SVM. These algorithms have shown significant success in other medical prediction activities, although the response of these algorithms to class imbalance and particularly when classifying CVD prediction remains an area of interest that needs further exploration [10].

For the extensive analysis of the effectiveness of the data balancing techniques on model performance, we have used evaluation metrics such as precision, recall, F1-score, accuracy, and ROC-AUC. Also, using Learning Curve plots, we show the impact of various data balancing approaches on the problems of overfitting and underfitting, which arise in the process of developing an ML model [11]. By following this extensive approach, our effort intends to provide significant findings in the testing efficiencies of the data balancing methods for CVD prediction. These findings should be of valuable use in establishing a further understanding of the variables involved as well as more precise and justifiable expected patterns that will one day enhance the cardiovascular health of Nigeria and perhaps other comparable nations.

This paper is segmented into five sections, with section 1 detailing the introduction, section 2 presenting the literature review, and section 3 talking about the materials and methods. Here the collected data was described, how the data was processed explained, and the different algorithms used for the model's development for the different data balancing techniques explained, in addition to the explanation of the performance evaluation metrics. In section 4, the experiment performed was explained in detail together with the results obtained from them. Moreover, some comparison among the results is made in order to evaluate their applicability concerning several criteria. Lastly, section 5 provides a conclusion and research implications as well as introduces directions for future research.

2. Review of related works

Ref. [12] focused on novel techniques to predict heart disease using machine learning dependent on artificial neural networks (ANNs) to examine different issues like data intricacy, elements of selection, as well as over-learning. The researchers aimed to enhance the precision of early diagnosis, thereby aiding in medical decision-making and customizing treatment strategies. To balance the dataset, the study used the SMOTE algorithm and the Edited Nearest Neighbours (ENN) algorithm. They used different machine learning techniques such as logistic regression, decision trees, random forests, gradient-boosting decision trees, XGBoost, SVM, and ANN. Their results revealed that ANN obtained satisfactory performance, with an accuracy of 0.808 and a recall of 0.81. This research has therefore shown that ANN has the possibility of enhancing the diagnosis and management of heart diseases. The application of the SMOTE data balancing technique has contributed to the performance accuracy of the model. This research did not showcase the effect of overfitting or underfitting on the proposed models; also, the proposed models performance can be enhanced to achieve higher accuracy using a more effective data balancing technique.

Machine learning was used by Ref. [13] as a vital tool to detect a heart disease because of its severe effects on the health of individuals. This work employed oversampling techniques, attribute reduction, the Classification And Regression Tree (CART) decision tree classifier, and rule reduction by optimising hyperparameters for improved prediction and to determine relative attributes influencing heart malfunctions. The researcher did show that by using the SMOTE over-sampling,

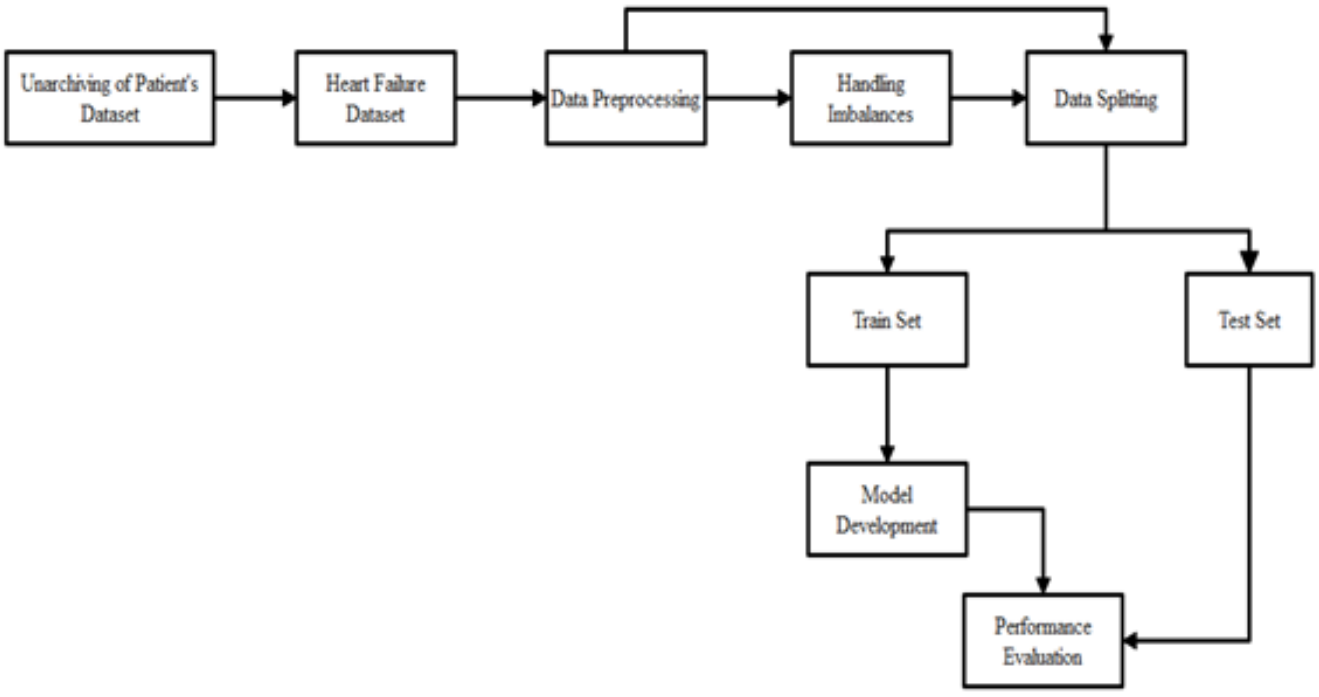

Figure 1. Methodology.

the study was able to achieve over 11% increase in overall accuracy, 75% increase in precision, 62% increase in recall, and 71% increase in F1 score when compared to using the unsampled datasets. It is noteworthy that the study achieved the highest performance of 0.801 accuracy, 0.79 precision, 0.832 recall, 0.807 f1 score, and 0.68 AUC ROC. The work showed how this algorithm is the perfect match for dealing with highly imbalanced datasets, along with its ability to detect key features linked to heart disease.

Ref. [14] proposed a framework to predict heart disease, a major killer, in the world using machine learning approaches. The proposed Decision Tree (DT) algorithm was implemented using the Cleveland Heart Disease dataset, with 14 significant quantitative features. Aware of the fact that one class may contain much more samples than the other, the synthetic minority oversampling technique (SMOTE) was used to balance the data. The researcher used the Weka tool to show that tuning the distribution to its optimal state successfully improves the classifications, with the best score of 73.3% for the DT algorithm when using unbalanced data and the best score of 68.6% for the sap algorithm when using balanced data. The improvement resulting from the equal distribution of the dataset improved the classification performance from 73.3% to 91.4%. This work also emphasises the need for mitigating bias in data to enhance the accuracy of heart disease diagnosis prediction.

Ref. [15] examined a machine learning algorithm to predict some gentle coronary susceptibility, a major killer in global society. Their study used classification algorithms like random forest, logistic regression (LR), K-nearest neighbour (KNN), decision tree, Xtreme Gradient Boosting (XGboost), convolutional neural network (CNN), and Kaggle dataset based on the parameters like age, gender, and cholesterol level. For handling the issue related to data imbalance, they used methods such as random undersampling, oversampling, SMOTE, and density-based SMOTE (DBSMOTE) and found better model performance. After the data balancing step, the CNN algorithm achieved an accuracy of 90%, which is higher compared to the results of other algorithms: XGBoost with 89.83% and LR with 89.82%. This work affirms the possibility of using machine learning in the identification of risk factors for heart diseases and calls for the adoption of the method in the healthcare sector for early prognosis.

Ref. [16] proposed heart disease prediction via online consultation mechanically aided by the Support Vector Machine (SVM) machine learning model. The study employed data to train models on features such as patients age, gender, blood pressure, cholesterol levels, and medical history. Patients offered similar information during consultations in the form of symptoms, lifestyles, and some kind of medical tests where they were given estimated heart disease risks. On their part, they pointed out that Support Vector Machine (SVM) was better than most other models with a prediction accuracy of 89% on the heart disease factor. This approach showcases the potential of machine learning in increasing the precision of heart disease diagnoses and the utilisation of remote consultations to lessen the load on health care facilities. The drawback of this study is its deficiency in level of prediction accuracy, which can be attributed to an inbalanced dataset and can be enhanced by

applying data balancing.

Ref. [17] carried out a study on cardiovascular disease prediction and how the application of machine learning can enhance its diagnosis, eventually diminishing mortality rates. To increase classification accuracy, they developed the k-modes with Huang's initialisation strategy. Models used in the study include Random Forest, Decision Tree, Multilayer Perceptron (MP), and XGBoost, in all of which hyperparameter optimisation through GridSearchCV was applied. Their models performance lies between 86.37 and 87.28% accuracy when tested on a Kaggle dataset of 70000 samples; the Multilayer Perceptron came out on top with an accuracy of 87.28% and an AUC of 0.95. MP usage shows that this technique can be useful in enhancing cardiovascular disease prediction. The prediction accuracy of this study requires improvement; also, the effects of overfitting and underfitting were not considered in this study; these can be handled through the application of data balancing.

In their review of heart disease prediction and selection of features for better ML model performance [18]. They evaluated the effect of feature selection on the performance of their models on two datasets: CVD and Framingham heart disease. The studies started with the feature transformation, cleansing, and balancing with random down sampling and went through feature selection based on the ANOVA-F test. It was noted that revealing the age, hypertension, glucose, prior heart disease, and blood pressure statistics, the number of associated factors can be considered to be targeted in relation to heart diseases. The study compared full and reduced features with and without preprocessing the data to examine the consequent accuracy, where use of reduced features gave slightly better estimates of accuracy, going from 0.73 to 0.75 on the CVD dataset and from 0.66 to 0.71 on the Framingham dataset. The results suggested that feature selection can be beneficial in increasing the level of accuracy, on the other hand, while at the same time minimising the computational costs. Balancing the dataset makes the model more reliable and also enhances the prediction accuracy. The authors proposed further studies by combining machine learning and deep learning to enhance the prediction of the model and use other feature selection methodologies for the more relevant datasets to diagnose heart disease. This study has a deficiency in the prediction accuracy, which can be improved on with a more effective data balancing technique.

Ref. [19] also paid attention to the problem of imbalanced datasets in heart disease prediction and presented a new divide-and-conquer data balance program based on the K-Means clustering algorithm. This method divides the data into different sections to increase the performance of the classifier, which is less prone to the risk of getting overfit and underfit, mainly focusing on accuracy, precision, and recall values, which are vital in giving correct medical predictions. In contrast to previous research, which frequently assessed models on a single data set, this study used two data sets to provide consistent, stable model performance with accuracy increases ranging from 81% to 90%. The result of the statistical analysis further supported the reliability of the model with a confidence interval of 0.95% for AUC at a range of 0.8187 to 0.8411. The balancing of the dataset also contributed to the high performance accuracy of this study. Furthermore, the combination of Explainable AI (XAI) was used to examine feature contributions in the Random Forest model, which was deemed the best on the comprehension of the subject matter specialist, which in turn enhances the model's interpretability and possible clinic application. The prediction accuracy of this model requires further improvement using a more effective data balancing technique.

In their study, Ref. [20] look at the large health threat made by CVDs, which are now the leading global killers. Since a list of complications linked to CVDs comprises hypertension, coronary heart disease, heart failure, angina, myocardial infarction, as well as stroke, the authors stress the significance of prevention and effective early diagnosis. The study therefore recommended the use of a supervised machine learning method to create accurate prediction models for CVDs, especially given that using SMOTE is well known to handle imbalanced data sets. The following key risk factors were chosen as inputs for the binary classification and for training multiple ML models with and without SMOTE: The study confirmed that the higher level of accuracy, recall, and specificity were observed for the model that applied SMOTE with 10-fold cross-validation, having accuracy of 87.8%, recall of 88.3%, precision of 88%, and AUC of 98.2%. The authors suggested that the stacking ensemble model provides great potential as a valid tool for the prognosis of CVDs. The prediction accuracy of this model can be enhanced.

Table 1 showcases the summary of the related works reviewed, considering their achievements and challenges.

3. Materials and methods

Cardiovascular disease datasets having 15 features, collected from records of 1000 patients of the University of Abuja Teaching Hospital, which contained features related to demographic, clinical, and lifestyle factors, along with the binary target variable indicating the presence or absence of CVD, were utilised in this study. The dataset was preprocessed to handle missing values, normalise features, and encode categorical variables. To evaluate the effectiveness of data imbalance correction techniques, we employed three approaches: oversampling, undersampling, and SMOTE. Comparisons were drawn with models developed using the baseline models. Figure 1 shows an overview of the adopted methodology.

3.1. Data collection and description

The dataset utilised was extracted from the patient's record of the University of Abuja Teaching Hospital (UATH), which contains medical information for one thousand (1000) patients. It consists of information on 348 females and 652 males who had come for medical help in the hospital. The dataset consists of a total of 1000 data points with 15 features. Table 2 summarizes the features of the dataset. However, four of the fifteen features are numeric, while the remaining eleven are boolean. Therefore, the statistical information of the numerical attributes is tabulated in Table 3. Following that, the dataset was imported into Jupyter Notebook and was subjected to exploratory

Table 1. Summary of related works reviewed.

S/N	Paper & Year	Research Outcome	Research Limitations
1	[19]	Applied a new divide-and-conquer data balance technique based on the K-Means clustering algorithm to achieve highest accuracy of 90%	The study did not showcase the effect of data balancing on overfitting/underfitting. Also a more effective data balancing technique can be applied to enhance the accuracy.
2	[12]	Applied SMOTE data balancing and attained prediction accuracy of 0.808.	Prediction accuracy requires improvement.
3	[15]	Applied Random undersampling, oversampleing, SMOTE, and DBSMOTE with CNN to attain highest accuracy of 90%	The accuracy can be enhanced. Also the effect of overfitting/underfitting were not showcased.
4	[20]	Utilized SMOTE with 10-fold cross-validation to attain highest accuracy of 87.8%.	The accuracy can be enhanced and also the effect of overfitting/underfitting showcased.
5	[17]	Multilayer Perceptron was applied to attain highest accuracy of 87.28%	There is need for data balancing and improvement of the prediction accuracy.
6	[16]	This study utilised SVM to train their data with selected features and attained a highest accuracy of 89%	Data balancing was not considered, neither was the effect of overfitting/underfitting showcased.
7	[18]	Applied SMOTE data balancing technique and attain highest accuracy of 91.4%	Did not showcase the effect of overfitting/underfitting.
8	[13]	Utilised SMOTE oversampling to achieve prediction accuracy of 0.801 which is 11% increase compared to the imbalanced data.	Prediction accuracy requires improvement.
9	[18]	Used random down-sampling for balancing their dataset and attain highest accuracy of 75%.	Prediction accuracy requires improvement using more

data analysis to ascertain its general characteristics and validity. Also, a correlation heatmap was developed, as depicted in Figure 2, to determine the degree of correlation among the attributes. Furthermore, an evaluation of the count of the target variable (CVD) showed the dataset is highly imbalanced in the ratio of 864 to 186. This will affect the performance of the models when developed. To validate these claims, seven different machine learning models were developed, and the performance of the models was evaluated.

3.2. Data processing

To process the dataset and make it fit for use data processing techniques such as label encoding, missing values, and duplicate data checks were performed. We took necessary steps to ensure the dataset was free of missing or null values, and also checked for duplicates to prevent duplicate data points and maintain accuracy and consistency in the data. Furthermore, categorical data were converted into numerical data using label encoding techniques.

3.3. Handling data imbalance

Data sampling techniques play a critical role in addressing class imbalance in machine learning datasets. In this study, we utilised four common sampling techniques, namely the Synthetic Minority Over-sampling Technique (SMOTE), Under Sampling, Over Sampling, and Minority Oversampling Technique, and Edited Nearest Neighbour (SMOTE-ENN), to compare with the performance of models developed without sampling. Figure 3 shows the data sampling techniques adopted.

3.3.1. Under sampling

Under-sampling require the number of instances in the majority class to be decreased so that the number of instances in both majority and minority classes can be equal. This technique creates a subset of the instances from the majority class equal to the size of the minority class [21]. Under-sampling is simple to implement and requires little computational power; however, reliance on only a sample of the data means that there could be informative instances that had to be removed from the majority class [22]. Therefore, undersampling is used in combination with other methods or compared to oversampling and SMOTE to judge the influence on the model [23, 24].

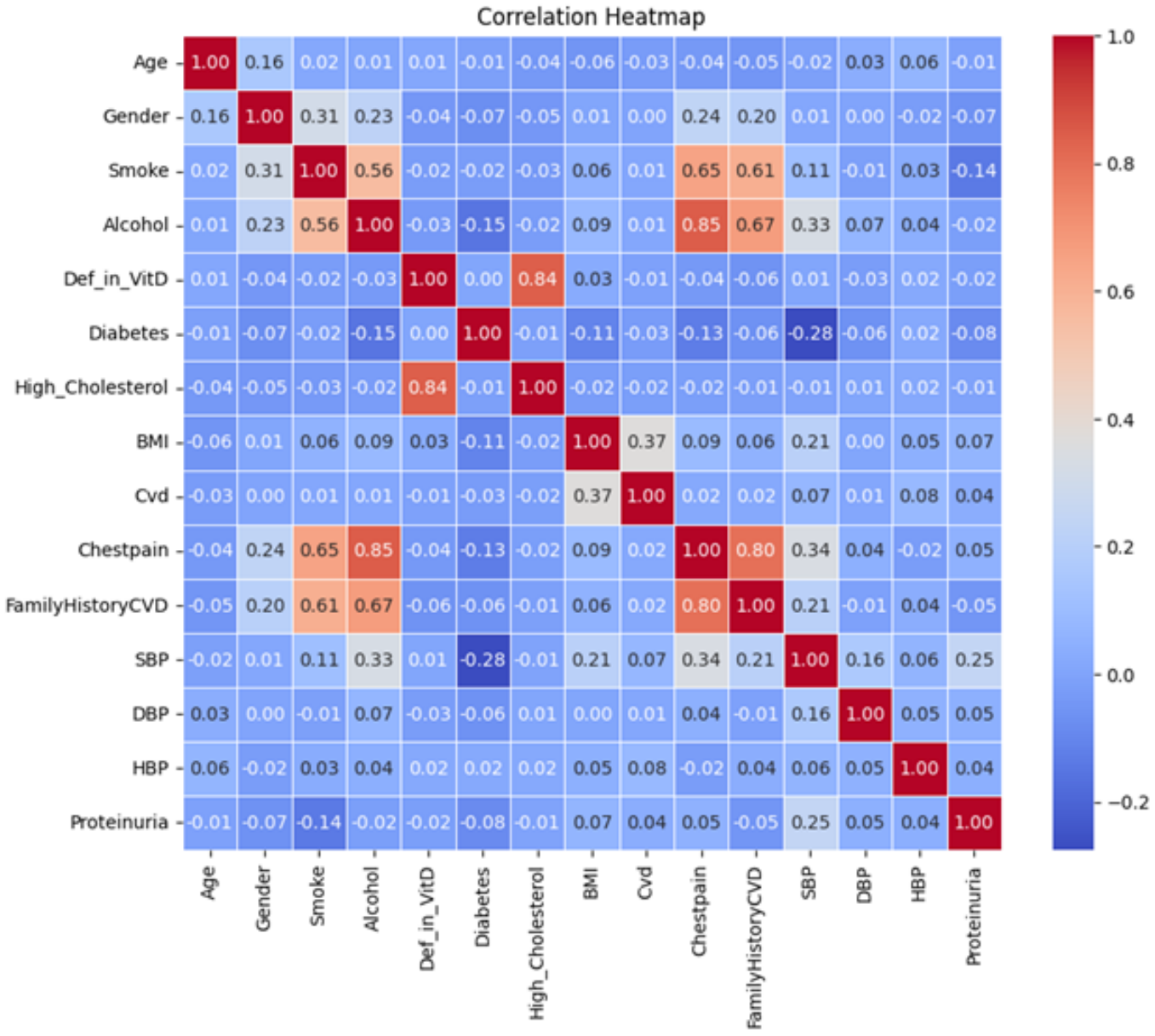

Figure 2. Correlation plot of the features.

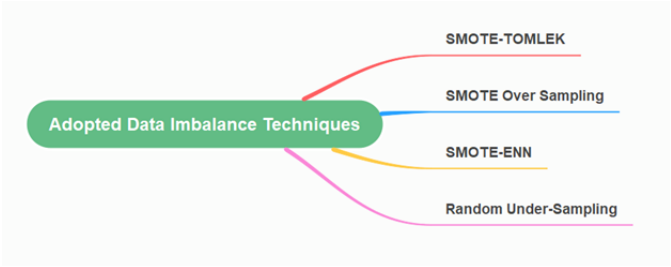

Figure 3. Data balancing techniques adopted.

3.3.2. *SMOTE over-sampling*

SMOTE is one of the most common approaches for handling the class imbalance problem through synthetic sample creation of the minority class. how the algorithm works is that it first picks up a sample in the minority class and then locates the k nearest neighbour in the feature space [25]. This then synthesises new instances along the line segments joining the minority class instance to its neighbours. The process of generating data is repeated until the majority-mixed minority class and majority classes are created in an appropriate proportion. SMOTE reduced overfitting and brought synthetic different factors to the minority class that improved the generalisation ability of the machine learning models. SMOTE synthesises new instances from the minority class by calculating a percentage distance between different instances of the minority class.

Table 2. Summary of the dataset.

SN	Feature	Data Type	Description
1	Age	Numeric	Age of the patient
2	Gender	Boolean	The gender of the patient (Male/Female)
3	Smoke	Boolean	If the patient smokes (Yes/No)
4	Alcohol	Boolean	If the patient consumes alcohol (Yes/No)
5	Def_in_VitD	Boolean	Deficiency in Vitamin D (Yes/No)
6	Diabetes	Boolean	If the patient has diabetes (Yes/No)
7	High_Cholesterol	Boolean	High cholesterol levels (Yes/No)
8	BMI	Numeric	Body Mass Index
9	CVD	Boolean	Cardiovascular Disease (Yes/No)
10	Chest pain	Boolean	Presence of chest pain (Yes/No)
11	FamilyHistoryCVD	Boolean	Family history of CVD (Yes/No)
12	SBP	Numeric	Systolic Blood Pressure
13	DBP	Numeric	Diastolic Blood Pressure
14	HBP	Boolean	High Blood Pressure (Yes/No)
15	Proteinuria	Boolean	Presence of protein in urine (Yes/No)

Table 3. Statistical description of numeric data.

	Age	BMI	SBP	DBP
Count	1000	1000	1000	1000
Mean	48.994	94.578	164.801	98.695
Std	14.96068	16.23433	33.7923	8.645734
Min	12	59	100	76
25%	36	83	122	92

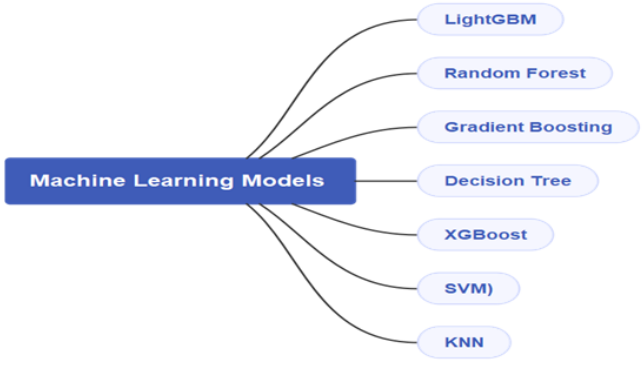

Figure 4. Machine learning models.

3.3.3. SMOTE-ENN

The SMOTE-ENN algorithm is an acronym for the names of two methods integrated into one: Synthetic Minority Oversampling Technique (SMOTE) and Edited Nearest Neighbour (ENN). It aimed to overcome the disadvantages of imbalanced datasets by balancing the class distributions and samples as well as removing noise samples. SMOTE (Synthetic Minority Oversampling Technique): By constructing synthetic instances from existing examples, SMOTE mainly targets the creation of synthetic examples for the minority class [26].

For every minority-class observation, it generates new synthetic data vectors based on the nearest neighbours of the observation. These synthetic examples are useful in balancing up the class distribution in the class. ENN (Edited Nearest Neighbor): It is a data filtering technique that determines noise or misclassification samples as the endnote. In the case of the observation, K-nearest neighbours are searched for and found by the ENN. For the observation, the majority of K nearest neighbours belong to a class different from the observation class; the observation is considered noisy and is disqualified from the set [26].

3.3.4. SMOTE-TOMEK

The SMOTE+TOMEK links are the integrating of SMOTE with TOMEK links. TOMEK links are the nearest neighbours; while one belongs to a particular class, the other belongs to an entirely different class. Using the SMOTE+TOMEK links approach, we will be able to eliminate TOMEK links, which can assist in decreasing overlapping in classes and augmenting the separability of the classes [27]. The steps involved in the working of the SMOTE-TOMEK Links technique are as follows: This involves the calculation of the nearest neighbour from within a cluster from the same cluster and the nearest neighbour from outside of the cluster of the instance in question in a given data set. Usually, a distance measure such as Euclidean distance can be used to identify such closest neighbours [28–30]. After that, each of the datasets is examined once again to check whether each of the two forms a Tomek link according to the given criteria. If a Tomek link is identified, then two occurrences are flagged for possible deletion from the data

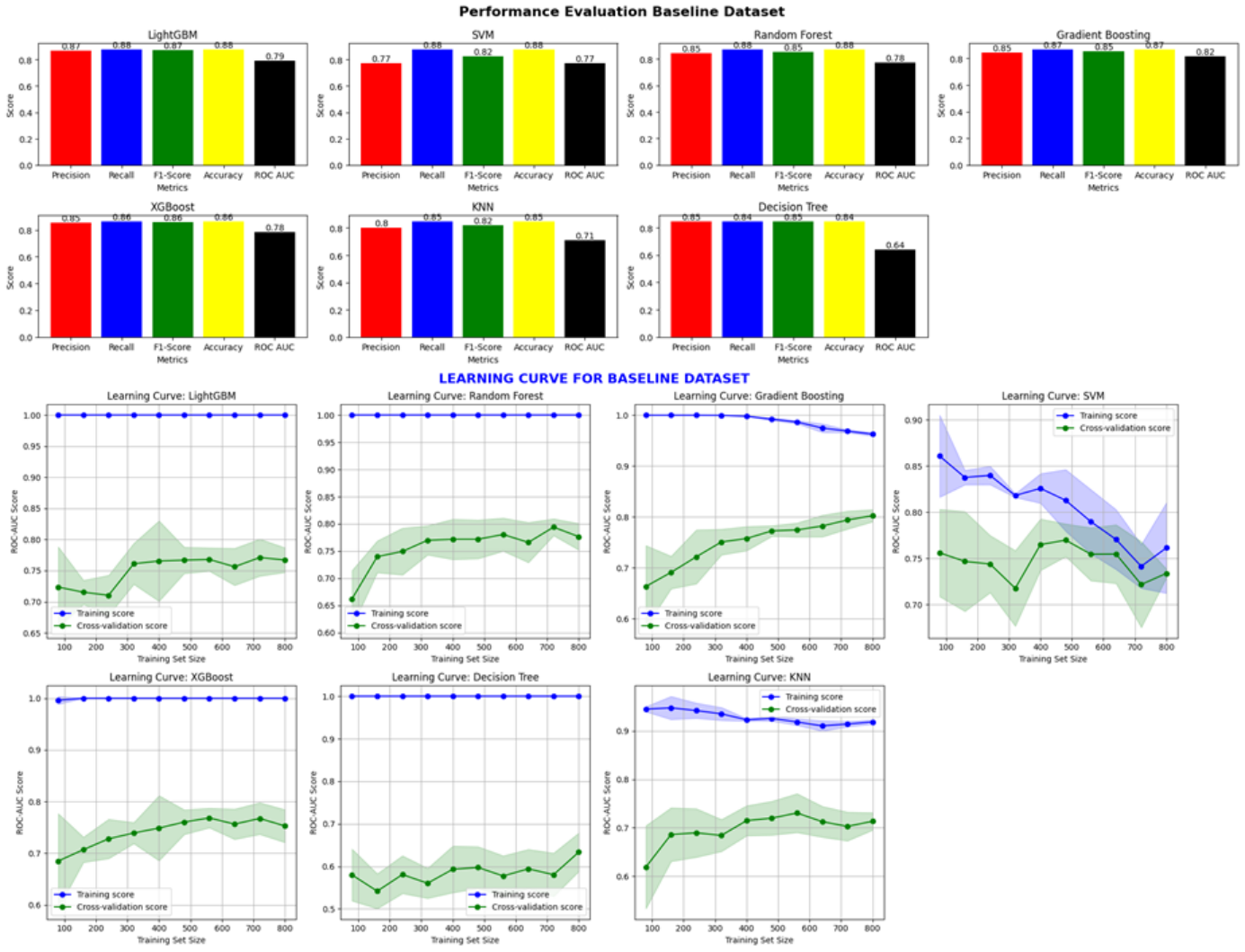

Figure 5. (a) Plot of performance metrics for baseline models. (b) Learning curve plot for baseline dataset

set [31]. Once all Tomek linkages within the dataset have been identified, the cases that make up these linkages are then regarded as ambiguous or noisy. These are then dropped from the dataset. This process of preparing the real data removes overlaps between the classes and makes the classifier more accurate [32].

3.4. Data splitting

After pre-processing and handling the data imbalances, the dataset is split into training and testing datasets in the ratio of 80 to 20. The adoption of this train-test split ratio is based on the fact that it gave us better results after experimenting with others such as 90:10, 70:30, 95:5, and 50:50. The training set is used to train the machine learning models, while the test set is used to evaluate the performance of the machine learning models. This research did not apply any feature selection approach.

3.5. Model development

Seven machine learning models were developed to compare the effectiveness of the different data imbalance techniques on the cardiovascular disease dataset. We utilised various classification algorithms implemented in popular machine learning libraries to train models on balanced datasets generated using random under-sampling, SMOTE over-sampling, SMOTE-ENN, and SMOTE-TOMEK techniques. The performance of each model was evaluated using precision, recall, F1-score, accuracy, and ROC-AUC. Figure 4 shows the carefully chosen classification algorithms used in developing our models.

3.5.1. LightGBM (Light Gradient Boosting Machine)

LightGBM is an optimised boosting algorithm for high speed, and it is mainly used for large scale data. It builds decision trees in a greedy approach, that is, to grow each leaf absolutely different and split nodes based on gradients of loss functions over a limited group of features. LightGBM also shines where large datasets are involved because it employs

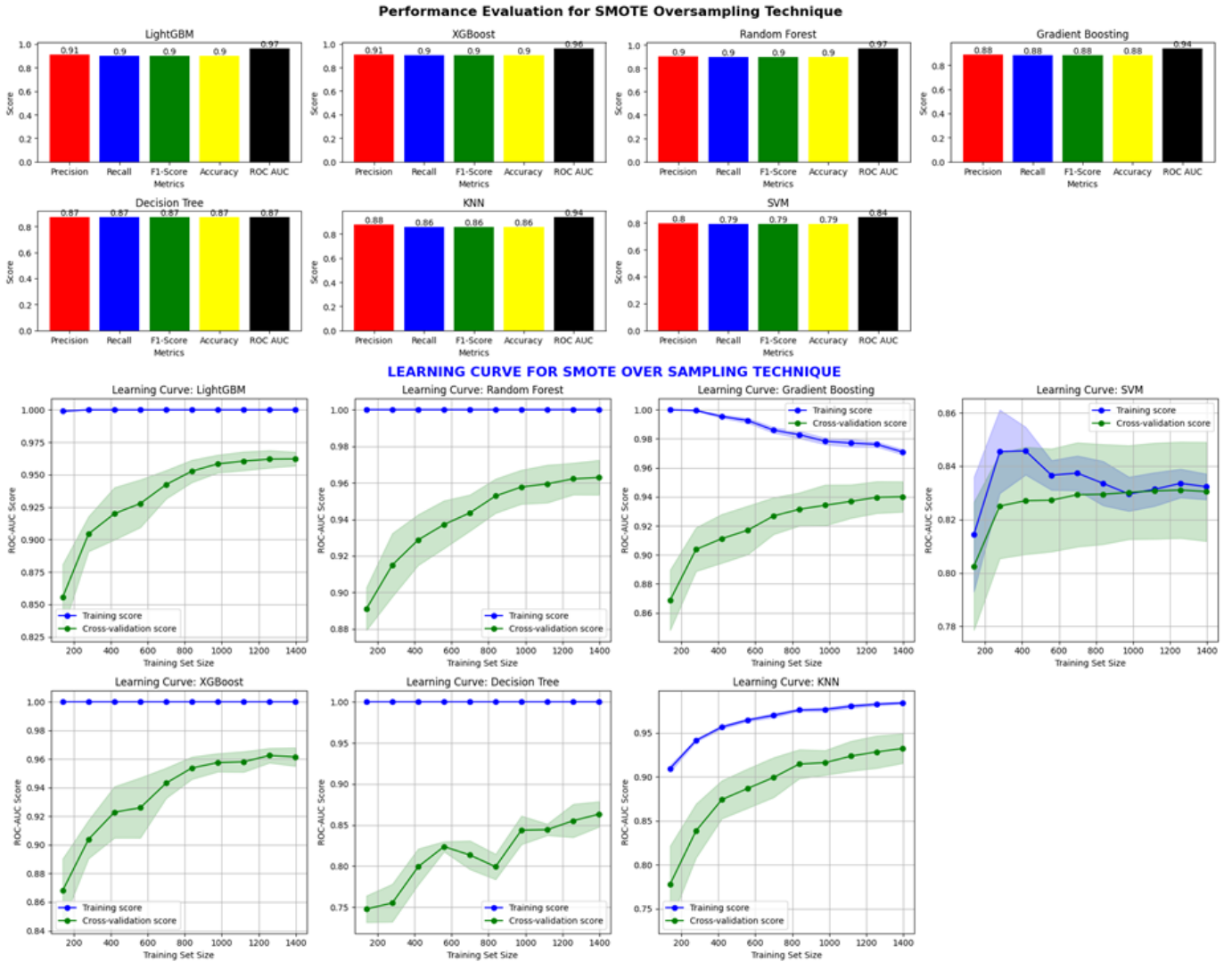

Figure 6. (a) Plot of performance metrics for SMOTE over sampling (b) Learning curve plot for SMOTE over sampling.

the histogram-based method besides effectively bundling features to lower computational needs. This makes it perfect for low-latency functions in tasks including the classification and ranking [33].

3.5.2. *Random forest*

Random forest is a method of ensemble learning model bagging that constructs a multitude of trees wherein each tree is grown on a different random boot strap sample of the data. It utilizes both random sampling of data points and random feature selection, subsequently combining multiple weak learning models to form a powerful single learning model. For classification the final prediction is the mode across trees and for regression, an average is taken. This method minimises overfitting and improves performance through the use of diverseness in the tree assembly [34].

3.5.3. *Gradient boosting*

This algorithm constructs the decision trees sequentially; every one of them tries to minimise the amount of errors made by the previous trees, targeting misclassified instances. With the help of the gradient descent technique, it applies to minimise the loss function, making it for both regression and classification types. But, when not well tuned, it can be easily overfitted and can consume even more computing power due to its sequential processing [35].

3.5.4. *Support Vector Machine (SVM)*

Support Vector Machine is a supervised learning model that seeks to generate a hyperplane or a set of hyperplanes in order to categorise the data. Alternatively, one tries to maximise the distance between different classes in order to make the decision line less sensitive to noise. In cases of nonlinear problems, SVM employs the method of kernels, by which data is transformed to a higher dimension where it becomes possible to draw a hyperplane. Thus, it is suitable where there are many

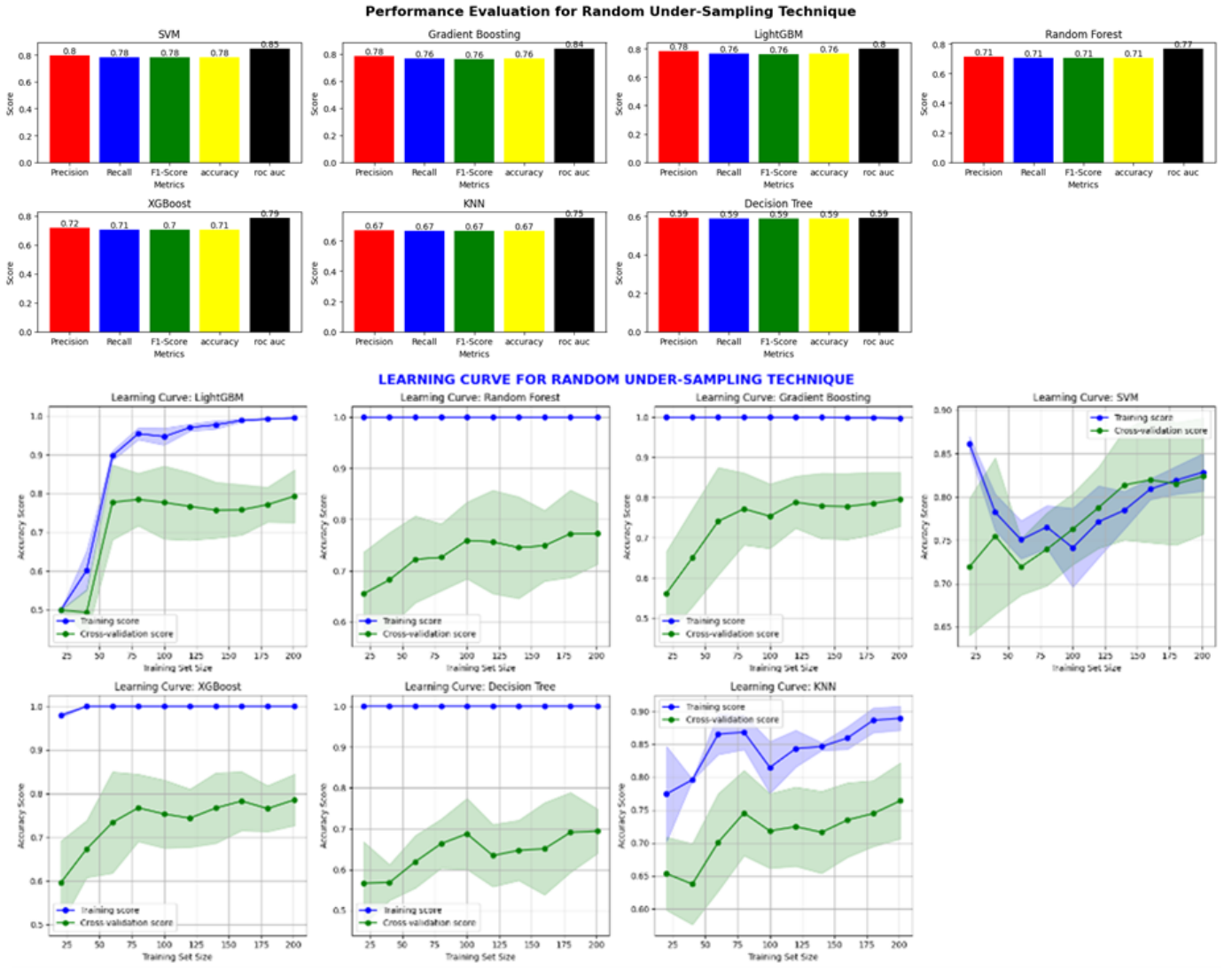

Figure 7. (a) Plot of performance metrics for random under-sampling (b) Learning curve plot for random under sampling.

central clusters of data, which enables decision-making [36].

3.5.5. *XGBoost (Extreme Gradient Boosting)*

The XGBoost, considering its name, is a more efficient version of the gradient boosting algorithm. While constructing the trees, like in the case of the usual gradient boosting, it constructs them step by step, but before building the new tree, it applies some kind of regularisation to prevent overfitting. To perform data operations in large amounts, XGBoost also uses parallelism and sparsity-aware algorithms and gets used broadly in Kaggle competition and real-world problems [37].

3.5.6. *Decision tree*

A decision tree is a type of machine learning model that is non-parametric, where the modeller breaks down the whole dataset into set levels with a space to split on the attribute that imparts the greatest information gain in the case of decision trees for classification and the lowest variance in the case of decision trees for regression. despite the model being easy to understand, it has the weakness of being very likely to overfit hence are applied in randomised forest and gradient boosting [38].

3.5.7. *K-Nearest Neighbours (KNN)*

KNN is selected as a basic, inductive instance learning method. It partitions data elements based on the ‘k’ nearest neighbours in the feature space and proceeds with the class that is most frequently occurring among these data points. Although KNN is easy to implement and understand, there are drawbacks of the algorithm: it is slow when used on large datasets, as it requires calculations of distance for each prediction; however, it is accurate where the data has a simple structure, such as in recommendation systems and pattern recognition [39].

The machine learning models were further developed using the specified algorithms and evaluated their performance on balanced datasets created using random under-sampling, SMOTE over-sampling, SMOTE-ENN, and SMOTE-TOMEK techniques.

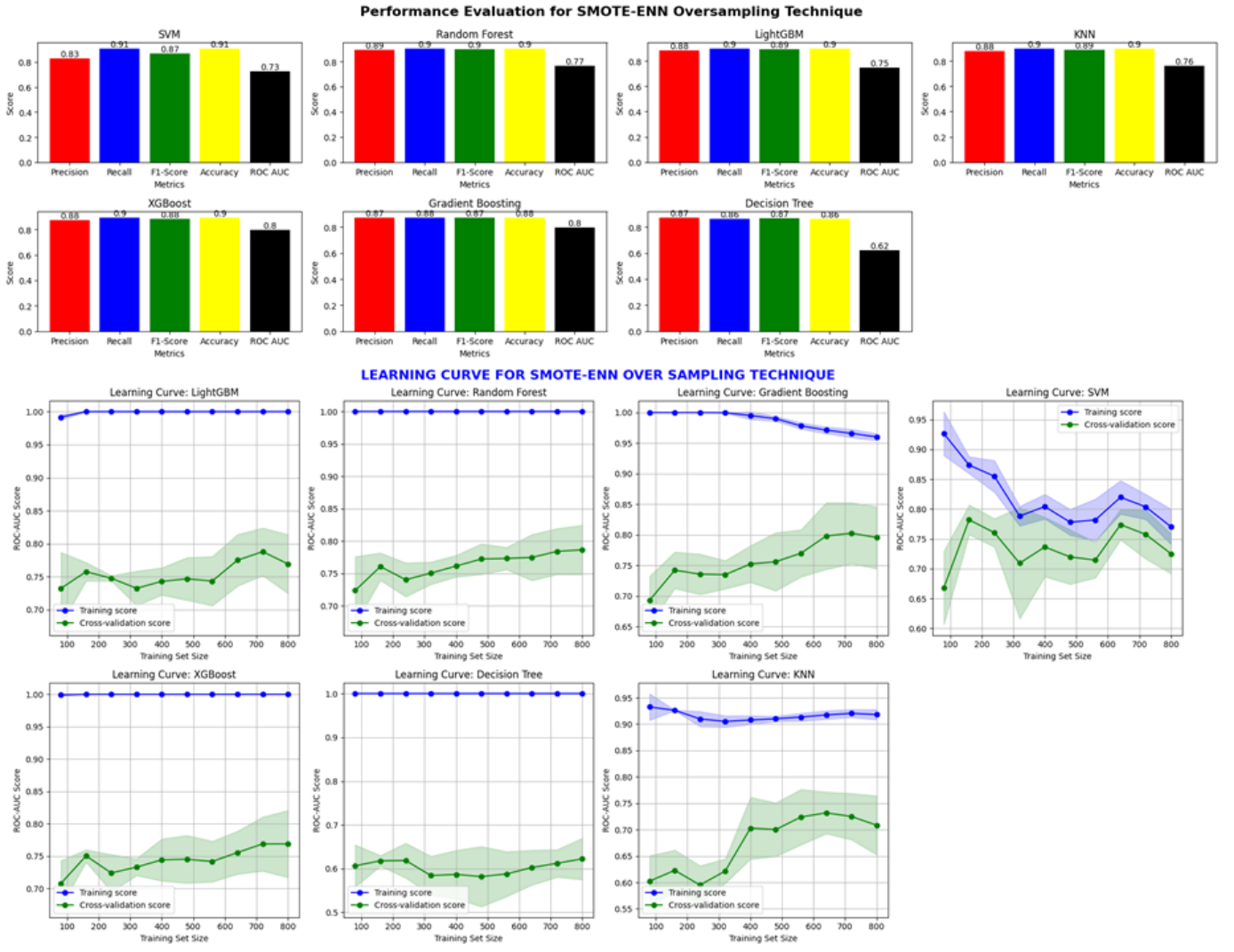

Figure 8. (a) Plot of performance metrics for SMOTE-ENN technique (b) Learning curve plot for SMOTE-ENN..

3.6. *Performance evaluation*

The performance evaluation of this study centres on articulating how the various forms of applied data imbalance mitigation techniques supplemented machine learning performance on CVD prediction. Accuracy, precision, F1-score, recall, and ROC-AUC were used in evaluating the performance of models developed in this study. The plots showing the different performance evaluation metrics for the seven chosen models were created for the different data balancing techniques as well as that of the baseline dataset (imbalanced dataset). The performance of the different models for the different data balancing techniques was compared to determine how the different data balancing techniques affected the prediction of CVD. Also, learning curves of the chosen models were plotted for the baseline dataset and the different data imbalance mitigation techniques and compared against each other to determine how the different datalancing techniques handle the issues of overfitting and underfitting. These visualisations help understand the model's performance, identify any improvement areas, and make informed decisions.

3.6.1. *Precision*

Precision is the ratio of actually positive cases, that is, the number of correct positive predictions to the total number of positive predictions. It is of great value when the cost associated with a false positive result is high, as used in the diagnosis of diseases. For example, when the model predicts a healthy person as sick, then it will lead to unnecessary treatments [40]. Precision is calculated using equation (1).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}. \tag{1}$$

3.6.2. *Recall (Sensitivity or True Positive Rate)*

Recall measures the ratio of true positives correctly flagged out of all actual positives in the population and is useful when false negatives are more costly, for example, when diagnosing a disease [41]. The process of calculating recall is represented

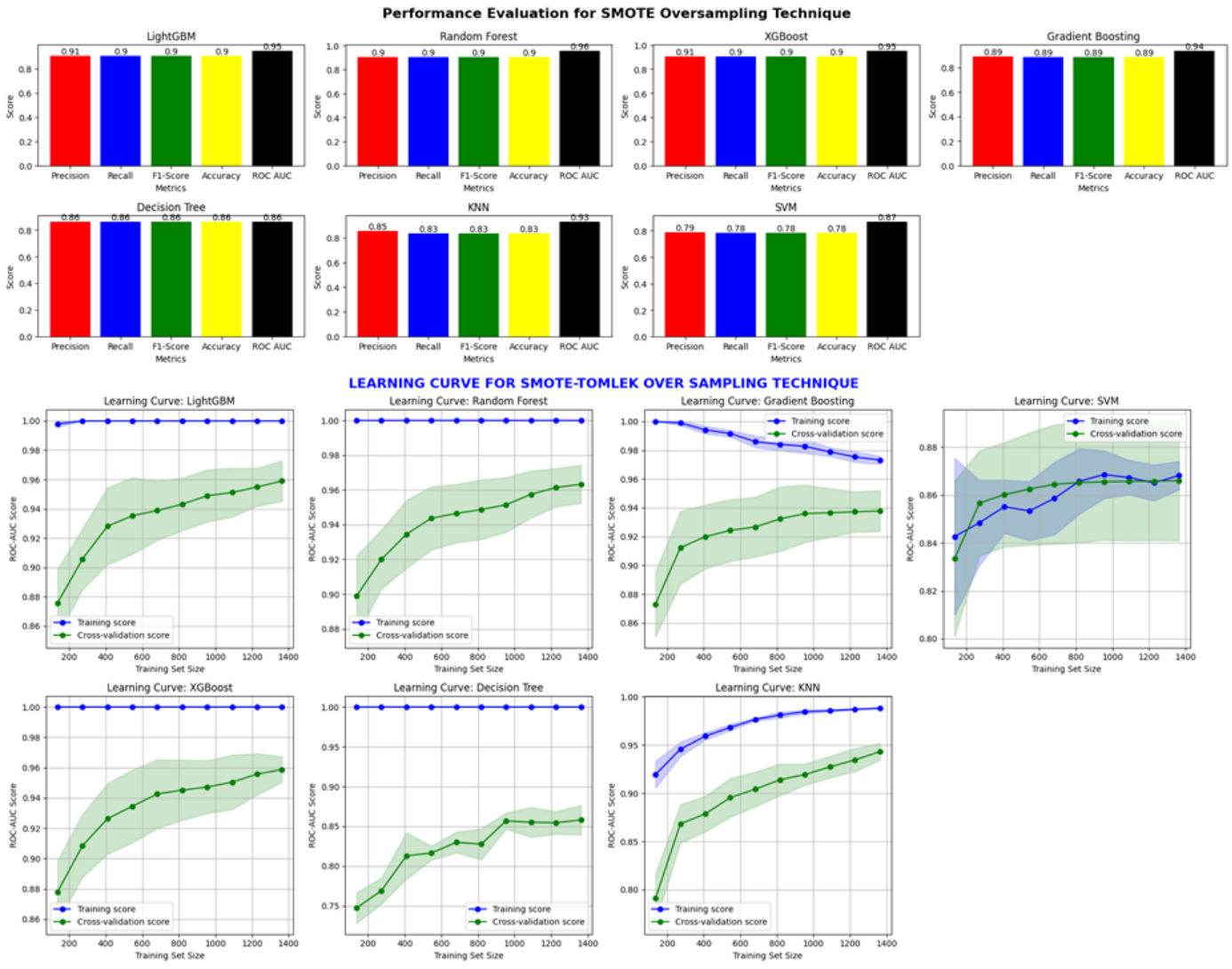

Figure 9. (a) Plot of performance metrics for SMOTE-TOMEK (b) learning curve plot for SMOTE-TOMEK.

in equation (1).

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}. \quad (2)$$

3.6.3. F1-Score

F1-scores are calculated as the harmony mean between precision and recall. It is useful when there is a significant skew in instances between the classes or both false positives and false negatives are costly. It is recommendable when you require equal accuracy as well as recall [42]. Equation (3) represents the F1-Score calculation.

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}. \quad (3)$$

3.6.4. Accuracy

Accuracy is the ratio of how many decisions were made correctly either positively (true positives) or negatively (true negatives) times one hundred, divided by the total number of cases of the decisions that were made. But it is not effective while there is an imbalance of datasets because accuracy can be deceptive [43]. Equation 4 is used to showcase the accuracy calculation process.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}. \quad (4)$$

3.6.5. ROC-AUC Value (Receiver Operating Characteristic—Area Under Curve)

The ROC-AUC quantifies how well a model separates the space of positive class from a space of negative class. AUC of 1 is a perfect classifier, and AUC of 0.5 is a random model. This is particularly helpful with an imbalanced dataset [44].

3.6.6. Learning curve

A learning curve is a diagram that represents the dependence of model effectiveness on the amount of training data used. It helps determine overfitting (high training performance

Table 4. Performance evaluation of models across data balancing techniques.

	SMOTE TOMLEK							SMOTE OVER SAMPLING						
Evaluation Matrices	LightGBM	Random Forest	Gradient Bossting	Decision Tree	XGBoost	SVM	KNN	LightGBM	Random Forest	Gradient Boosting	Decision Tree	XGBoost	SVM	KNN
Precision	0.92	0.92	0.88	0.84	0.91	0.81	0.89	0.91	0.9	0.88	0.88	0.91	0.8	0.88
Recall	0.92	0.92	0.88	0.84	0.91	0.81	0.88	0.9	0.9	0.88	0.88	0.9	0.79	0.86
F1-Score	0.92	0.92	0.88	0.84	0.91	0.81	0.88	0.9	0.9	0.88	0.88	0.9	0.79	0.86
Accuracy	0.92	0.92	0.88	0.84	0.91	0.81	0.88	0.9	0.9	0.88	0.88	0.9	0.79	0.86
ROC-AUC	0.96	0.96	0.94	0.84	0.96	0.86	0.95	0.97	0.97	0.94	0.88	0.96	0.84	0.94
	SMOTE-ENN							RANDOM UNDER SAMPLING						
Evaluation Matrices	LightGBM	Random Forest	Gradient Bossting	Decision Tree	XGBoost	SVM	KNN	LightGBM	Random Forest	Gradient Boosting	Decision Tree	XGBoost	SVM	KNN
Precision	0.91	0.9	0.91	0.89	0.91	0.9	0.9	0.82	0.81	0.73	0.76	0.76	0.77	0.73
Recall	0.87	0.84	0.84	0.8	0.85	0.74	0.74	0.82	0.8	0.73	0.75	0.76	0.75	0.73
F1-Score	0.89	0.86	0.87	0.84	0.87	0.79	0.79	0.82	0.8	0.73	0.74	0.76	0.74	0.73
Accuracy	0.87	0.84	0.84	0.8	0.85	0.74	0.74	0.82	0.8	0.73	0.75	0.76	0.75	0.73
ROC-AUC	0.79	0.82	0.82	0.69	0.78	0.82	0.77	0.85	0.82	0.79	0.75	0.84	0.84	0.8
	WITHOUT BALANCING													
Evaluation Matrices	LightGBM	Random Forest	Gradient Bossting	Decision Tree	XGBoost	SVM	KNN							
Precision	0.87	0.84	0.85	0.85	0.85	0.77	0.8							
Recall	0.88	0.88	0.87	0.84	0.86	0.88	0.85							
F1-Score	0.87	0.85	0.85	0.85	0.86	0.82	0.82							
Accuracy	0.88	0.88	0.87	0.84	0.86	0.88	0.85							
ROC-AUC	0.79	0.78	0.82	0.64	0.78	0.77	0.71							

but low validation) or underfitting (low training as well as validation performance) [45].

4. Results and discussion

In this section, the results obtained from carrying out the four data balancing techniques are presented. For each of the data balancing techniques used, the seven machine learning models were developed to evaluate the effect of the data balancing on the models' performance. In addition, learning curves were plotted for the seven models to evaluate how the issues of overfitting and underfitting were addressed across the several data balancing techniques. In a learning curve, a high training score and a much lower validation score indicate overfitting, while low training and validation scores are signs of underfitting.

4.1. Original imbalanced baseline data sets (Baseline)

During the first evaluation, no data balancing algorithm was used. The dataset had a class imbalance, with the majority class having a ratio of 874 to the minority class, which had a ratio of 126. Before model building, the dataset was preprocessed, and data splitting was carried out. After that, the model's performance was evaluated using various metrics, including precision, recall, F1-score, accuracy, and ROC-AUC. The results of this evaluation are visualised in Figure 5a, which shows the performance of the various models when trained and tested without data balancing. Figure 5b is a learning curve of the baseline dataset that showcased how the issue of overfitting/underfitting affected imbalanced data sets across the different models. The baseline dataset shows severe overfitting across most models, with LightGBM, Random Forest, Gradient Boosting, and XGBoost all exhibiting high training scores but significantly lower cross-validation scores, indicating poor generalization. SVM initially overfits but improves slightly with additional data, but Decision Tree and KNN suffer from significant overfitting and underfitting, respectively. Generally, the models struggle to generalise well, emphasising the importance of appropriate data balancing strategies to overcome these difficulties.

4.1.1. Oversampling using SMOTE

By implementing SMOTE oversampling on the minority class, we were able to achieve balance with the majority class. The results of our proposed models, developed using the oversampled dataset, are displayed in Figure 6a. The Figure provides insight into the performance of each individual model, including the precision, recall, F1-score, accuracy, and ROC-AUC of the model. The learning curve showing how the issue of overfitting/underfitting was addressed by the SMOTE Oversampling data balancing technique is displayed in Figure 6b. In comparison to the baseline dataset, SMOTE oversampling improves model generalisation, with models typically displaying high training scores and better cross-validation scores. Particularly, LightGBM and XGBoost exhibit good results with steady learning curves and high cross-validation scores, suggesting that data imbalance is handled well. The difference between the training and cross-validation scores is less noticeable, despite a modest overfitting tendency, indicating that SMOTE offers a more balanced dataset. When compared to predictive models trained on the baseline dataset, this leads to more robust and dependable models.

4.1.2. Random under sampling

This section demonstrates the results obtained after undersampling the majority dataset to match the minority data. To achieve this, the majority class was divided into four random sets, with one set used for model training. The distribution of the dataset was made equal for both majority and minority classes and was set at 126. Figure 7a shows the plot of the performance evaluation. While Figure 7b shows the learning curve of how random undersampling addresses the overfitting/underfitting issue across the seven models. Models that overfit the training data are typically the result of random undersampling, as seen by high training scores and significantly lower and more variable cross-validation scores. Particularly vulnerable to poor generalisation are models such as SVM, Decision Tree, and KNN, whose great sensitivity to the small amount of training data leads to inconsistent performance. Even while XGBoost and LightGBM perform somewhat better, they still have a noticeable overfitting problem. Random undersampling makes training easier when compared to the baseline dataset, but it usually results in less generalisation ability and more unpredictability in model performance.

4.1.3. SMOTE-ENN

The SMOTE-ENN technique was applied to the dataset, which resulted in the minority class being resampled to have 601 instances, while the majority class had 465 instances. The performance of the model after addressing the data imbalance is illustrated in Figure 8a. Figure 8b is a learning curve plot showcasing how the issue of overfitting/underfitting was addressed by the SMOTE-ENN data balancing technique. Though overfitting persists, models with the SMOTE-ENN oversampling strategy demonstrate some improvement in generalisation when compared to the baseline dataset. Slightly higher cross-validation scores indicate less overfitting for LightGBM, Random Forest, Gradient Boosting, and XGBoost. SVM exhibits baseline-like behaviour, with some overfitting at first and slight improvements. Even though there is a minor improvement in performance for most models, Decision Tree and KNN continue to be overfit and underfit, respectively, indicating that SMOTE-ENN is helpful but not a complete solution for all models.

4.1.4. SMOTE-TOMEK

After resampling the dataset using the SMOTE-TOMEK technique, the balanced state of 859 by 859 was achieved. The results of model performance after resampling is depicted in Figure 9a while Figure 9b is a learning curve plot used to showcase SMOTE-TOMEK effect on overfitting/underfitting. All models' generalization is greatly enhanced by the SMOTE-Tomek oversampling strategy, which also results in considerably closer alignment between training and cross-validation scores, which suggests less overfitting. Significant gains in cross-validation performance are demonstrated by LightGBM, Random Forest, Gradient Boosting, XGBoost, and SVM, indicating successful management of overfitting. KNN demonstrates a considerable decrease in underfitting, whereas Decision Tree also benefits, however some overfitting persists. SMOTE-Tomek offers the best data balancing overall, which improves generalization and performance in all areas.

4.2. Impact of data balancing techniques on model performance

Table 4 is a result summary table to aid easy comparisons of the performance of the respective models across the different data balancing techniques in terms of precision, recall, F1-score, accuracy, and ROC-AUC.

The experimental results demonstrate that all four data imbalance correction techniques: SMOTE over-sampling, random under-sampling, SMOTE-ENN, and SMOTE-TOMEK

improved the predictive performance of machine learning models on imbalanced CVD datasets compared to the baseline model trained on the original imbalanced dataset. However, the extent of improvement varied across techniques and models. SMOTE-TOMEK consistently outperformed other data balancing techniques, achieving higher accuracy, precision, recall, and F1 scores across different machine learning models.

4.3. Impact of data balancing techniques on overfitting/underfitting

With the baseline dataset, the learning curves showed increased distance between the training scores and the cross-validation scores across all models, which indicates significant overfitting. The SMOTE ENN technique was able to reduce the overfitting but was not significant enough across the models, as indicated by the increased cross-validation scores. The result from SMOTE Over Sampling shows that there are improvements in model generalisation indicated by the high training scores and better cross-validation scores alongside reduced overfitting. The result from the random undersampling technique showed overfitting and poor generalisation, as indicated by the inconsistent cross-validation scores across some of the models, though still better than that of the baseline dataset. While SMOTE-Tomek was the best in reducing overfitting and enhanced generalisation, as indicated by the cross-validation score's significant improvement.

5. Conclusion

In conclusion, addressing data imbalance is crucial for developing accurate predictive models for cardiovascular disease datasets. This study highlights the effectiveness of SMOTE over-sampling, random under-sampling, SMOTE-ENN, and SMOTE-TOMEK in mitigating the impact of class imbalance on predictive performance and addressing the issue of overfitting/underfitting. Among these techniques, SMOTE-TOMEK emerges as the most effective approach for improving the performance of machine learning models on imbalanced CVD datasets, achieving best fits, and effectively handling issues of underfitting and overfitting. It achieved the highest performance with accuracy, precision, recall, and F1 scores of 92% and an ROC-AUC of 96% when applied to Random Forest and LightGBM models. It also demonstrated the most balanced fit of the dataset, the best handling of overfitting and underfitting issues, and improved generalisation when combined with models like Random Forest, Gradient Boosting, and XGBoost. Future research could explore advanced data imbalance correction techniques, feature selection techniques, deep learning approaches, and ensemble methods to further enhance predictive performance in this domain.

Data availability

The link to the dataset used in this study is provided below: https://github.com/ambraph/RESEARCH2/blob/main/cvddataset.csv.

References

[1] M. Di Cesare, P. Perel, S. Taylor, C. Kabudula, H. Bixby, T. A. Gaziano, D. V. McGhie, J. Mwangi, B. Pervan, J. Narula, D. Pineiro & F. J. Pinto, "The heart of the world", Global Heart **19** (2024) 11. https://doi.org/10.5334/gh.1288.

[2] O. Olamide, O. Adebayo, A. Emmanuel, L. Eyitayo, O. Beatrice & M. Tomisin, "Prevalence and risk factors of cardiovascular diseases among the Nigerian population: A new trend among adolescents and youths", IntechOpen **2023** (2023) 1. https://doi.org/10.5772/intechopen.108180.

[3] S. Hossain, M. K. Hasan, M. O. Faruk, N. Aktar, R. Hossain & K. Hossain, "Machine learning approach for predicting cardiovascular disease in Bangladesh: evidence from a cross-sectional study in 2023", BMC Cardiovascular Disorders **24** (2024) 214. https://doi.org/10.1186/s12872-024-03883-2.

[4] W. W. Fan & C. H. Lee, "Classification of imbalanced data using deep learning with adding noise", Journal of Sensors **2021** (2021) 1. https://doi.org/10.1155/2021/1735386.

[5] I. Araf, A. Idri & I. Chairi, "Cost-sensitive learning for imbalanced medical data: a review", Artificial Intelligence Review **57** (2024) 80. https://doi.org/10.1007/s10462-023-10652-8.

[6] I. M. Alkhawaldeh, I. Albalkhi & A. J. Naswhan, "Challenges and limitations of synthetic minority oversampling techniques in machine learning" World Journal of Methodology **13** (2023) 373. https://doi.org/10.5662/wjm.v13.i5.373.

[7] A. Hassan, S. G. Ahmad, E. U. Munir, I. A. Khan & N. Ramzan, "Predictive modelling and identification of key risk factors for stroke using machine learning", Scientific Reports **14** (2024) 11498. https://doi.org/10.1038/s41598-024-61665-4.

[8] Q. Y. Yin, J. S. Zhang, C. X. Zhang & N. N. Ji, "A novel selective ensemble algorithm for imbalanced data classification based on exploratory undersampling", Mathematical Problems in Engineering **2014** (2014) 1. https://doi.org/10.1155/2014/358942.

[9] N. W. Minja, D. Nakagaayi, T. Aliku, W. Zhang, I. Ssinabulya, J. Nabaale, W. Amutuhaire, S. R. de Loizaga, E. Ndagire, J. Rwebembera, E. Okello & J. Kayima, "Cardiovascular diseases in Africa in the twenty-first century: Gaps and priorities going forward", Frontiers in Cardiovascular Medicine **9** (2022) 1008335. https://doi.org/10.3389/fcvm.2022.1008335.

[10] M. A. Sufian, W. Hamzi, S. Zaman, L. Alsadder, B. Hamzi, J. Varadarajan & M. A. K. Azad, "Enhancing clinical validation for early cardiovascular disease prediction through simulation, ai, and web technology", Diagnostics (Basel, Switzerland) 14 (2024) 1308. https://doi.org/10.3390/diagnostics14121308.

[11] C. Aliferis & G. Simon, "Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI", in *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences*, G. J. Simon & C. Aliferis, Eds., Health Informatics, Springer, Cham, 2024, pp. 477-524. http://dx.doi.org/10.1007/978-3-031-39355-6_10.

[12] Q. Chen & N. Ma, "Heart disease prediction method based on ANN", Highlights in Science, Engineering and Technology **85** (2024) 411. https://doi.org/10.54097/fgt46k23.

[13] A. J. Albert, R. Murugan & T. Sripriya, "Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology", Research on Biomedical Engineering **39** (2023) 99. https://doi.org/10.1007/s42600-022-00253-9.

[14] A. S. Jaddoa, "Heart disease prediction system using (SMOTE technique) balanced dataset and decision tree classifier", AIP Conference Proceedings **2834** (2023) 050006. https://doi.org/10.1063/5.0161558.

[15] R. Masram, S. K. Sharma & N. Kumar, "Heart disease identification methods using machine learning and efficient data balancing techniques", International Research Journal of Engineering and Technology **11** (2024) 377. https://www.irjet.net/archives/V11/i7/IRJET-V11I753.pdf.

[16] B. Duraisamy, R. Sunku, K. Selvaraj, V. V. R. Pilla & M. Sanikala, "Heart disease prediction using support vector machine", Multidisciplinary Science Journal **6** (2023) 2024ss0104. https://doi.org/10.31893/multiscience.2024ss0104.

[17] C. M. Bhatt, P. Patel, T. Ghetia & P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques", Algorithms **16** (2023) 88. https://doi.org/10.3390/a16020088.

[18] M. S. Pathan, A. Nag, M. M. Pathan & S. Dev, "Analyzing the impact of

feature selection on the accuracy of heart disease prediction", Healthcare Analytics **2** (2022) 100060. https://doi.org/10.1016/j.health.2022.100060.

[19] F. Yang, Y. Qiao, P. Hajek & M. Z. Abedin, "Enhancing cardiovascular risk assessment with advanced data balancing and domain knowledge-driven explainability", Expert Systems with Applications **255** (2024) 124886. https://doi.org/10.1016/j.eswa.2024.124886.

[20] E. Dritsas & M. Trigka, "Efficient data-driven machine learning models for cardiovascular disease risk prediction", Sensors **23** (2023) 1161. https://doi.org/10.3390/s23031161.

[21] J. Hoyos-Osorio, A. Alvarez-Meza, G. Daza-Santacoloma, A. Orozco-Gutierrez, G. Castellanos-Dominguez, "Relevant information undersampling to support imbalanced data classification", Neurocomputing **436** (2021) 136. https://doi.org/10.1016/j.neucom.2021.01.033.

[22] A. X. Wang, S. S. Chukova & B. P. Nguyen, "Synthetic minority oversampling using edited displacement-based k-nearest neighbors", Applied Soft Computing **148** (2023) 110895. https://doi.org/10.1016/j.asoc.2023.110895.

[23] D. H. Jeong, S. E. Kim, W. H. Choi & S. H. Ahn, "A comparative study on the influence of undersampling and oversampling techniques for the classification of physical activities using an imbalanced accelerometer dataset", Healthcare (Basel) **10** (2023) 1255. https://doi.org/10.3390/healthcare10071255.

[24] T. Wongvorachan, S. He & O. Bulut, "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining", Information **14** (2023) 54. https://doi.org/10.3390/info14010054.

[25] N. A. Azhar, M. S. Mohd Pozi, A. Mohamed Din & A. Jatowt, "An investigation of SMOTE-based methods for imbalanced datasets with data complexity analysis", IEEE Transactions on Knowledge and Data Engineering **35** (2022) 6651. https://doi.org/10.1109/TKDE.2022.3179381.

[26] N. Rout, D. Mishra & M. K. Mallick, "An advance extended binomial GLMBoost ensemble method with synthetic minority over-sampling technique for handling imbalanced datasets", International Journal of Electrical and Computer Engineering (IJECE) **13** (2023) 4357. https://doi.org/10.11591/ijece.v13i4.pp4357-4368.

[27] T. Sasada, Z. Liu, T. Baba, K. Hatano & Y. Kimura, "A resampling method for imbalanced datasets considering noise and overlap", Procedia Computer Science **176** (2020) 420. https://doi.org/10.1016/j.procs.2020.08.043.

[28] X. Yi, Y. Xu, Q. Hu & others, "ASN-SMOTE: a synthetic minority oversampling method with adaptive qualified synthesizer selection", Complex Intell. Syst. **8** (2022) 2247. https://doi.org/10.1007/s40747-021-00638-w.

[29] E. F. Swana, W. Doorsam & P. Bokoro, "Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset", Sensors **22** (2022) 3246. https://doi.org/10.3390/s22093246.

[30] Z. Xu, D. Shen, T. Nie & Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on random forest for medical imbalanced data", Journal of Biomedical Informatics **107** (2020) 103465. https://doi.org/10.1016/j.jbi.2020.103465.

[31] K. M. Hasib, S. Azam, A. Karim, A. A. Marouf, F. M. Javed Mehedi Shamrat & S. Montaha, "MCNN-LSTM: Combining CNN and LSTM to classify multi-class text in imbalanced news data", IEEE Access **11** (2023) 93048. https://doi.org/10.1109/ACCESS.2023.3309697.

[32] T. Ma, S. Lu & C. Jiang, "A membership-based resampling and cleaning algorithm for multi-class imbalanced overlapping data", Expert Systems with Applications **240** (2024) 122565. https://doi.org/10.1016/j.eswa.2023.122565.

[33] S. Luo & T. Chen, "Two derivative algorithms of gradient boosting decision tree for silicon content in blast furnace system prediction", IEEE Access **8** (2020) 196112. https://doi.org/10.1109/ACCESS.2020.3034566.

[34] A. F. Bulagang, G. W. Ng, J. Mountstephens & J. Teo, "A review of recent approaches for emotion classification using electrocardiography and electrodermography signals", Informatics in Medicine Unlocked **20** (2020) 100363. https://doi.org/10.1016/j.imu.2020.100363.

[35] A. Miller, J. Panneerselvam & L. Liu, "A review of regression and classification techniques for analysis of common and rare variants and gene-environmental factors", Neurocomputing **489** (2022) 466. https://doi.org/10.1016/j.neucom.2021.08.150.

[36] M. Mallik, A. K. Panja, C. Chowdhury, "Paving the way with machine learning for seamless indoor–outdoor positioning: A survey", Information Fusion **94** (2023) 126. https://doi.org/10.1016/j.inffus.2023.01.023.

[37] S. K. Kiangala & Z. Wang, "An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment", Machine Learning with Applications **4** (2021) 100024. https://doi.org/10.1016/j.mlwa.2021.100024.

[38] D. Packwood, L. T. H. Nguyen, P. Cesana, G. Zhang, A. Staykov, Y. Fukumoto & D. H. Nguyen, "Machine learning in materials chemistry: An invitation", Machine Learning with Applications **8** (2022) 100265. https://doi.org/10.1016/j.mlwa.2022.100265.

[39] S. Huang, M. Huang & Y. Lyu, "A novel approach for sand liquefaction prediction via local mean-based pseudo nearest neighbor algorithm and its engineering application", Advanced Engineering Informatics **41** (2019) 100918. https://doi.org/10.1016/j.aei.2019.04.008.

[40] T. F. Monaghan, S. N. Rahman, C. W. Agudelo, A. J. Wein, J. M. Lazar, K. Everaert & R. R. Dmochowski, "Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value", Medicina (Kaunas, Lithuania) **57** (2021) 503. https://doi.org/10.3390/medicina57050503.

[41] S. Orozco-Arias, J. S. Piña, R. Tabares-Soto, L. F. Castillo-Ossa, R. Guyot & G. Isaza, "Measuring performance metrics of machine learning algorithms for detecting and classifying transposable elements", Processes **8** (2020) 638. https://doi.org/10.3390/pr8060638.

[42] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, Pål Halvorsen & S. Parasa, "On evaluation metrics for medical applications of artificial intelligence", Scientific Reports **12** (2022) 5979. https://doi.org/10.1038/s41598-022-09954-8.

[43] H. Belyadi & A. Haghighat, "Supervised learning", in *Machine Learning Guide for Oil and Gas Using Python*, H. Belyadi & A. Haghighat, Eds., Gulf Professional Publishing, 2021, pp. 169–295. https://doi.org/10.1016/B978-0-12-821929-4.00004-4.

[44] T. Saito & M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets", PloS One **10** (2015) e0118432. https://doi.org/10.1371/journal.pone.0118432.

[45] O. A. Montesinos López, A. Montesinos López & J. Crossa, "Overfitting, model tuning, and evaluation of prediction performance", in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Springer, 2022, pp. 109–139. http://dx.doi.org/10.1007/978-3-030-89010-0_4.