Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 4, August 2025, Page. 1717-1728
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4753

Improving Extreme Gradient Boosting Model for Heart Disease Prediction Using SMOTE for Class Imbalance

Dini Rohmayani*[1], Castaka Agus Sugianto[2], Rangga Satria Perdana[3], Mohammed Mansoor Nafea[4]

[1,2]Informatics Engineering, Politeknik TEDC, Indoneisa
[3]Information Systems, Universitas Sangga Buana, Indonesia
[4]Computer Engineering Techniques, College of Technical Engineering University of Al Maarif, Iraq

Email: [1]dinirohmayani@poltektedc.ac.id

Received : May 19, 2025; Revised : Jun 30, 2025; Accepted : Jun 30, 2025; Published : Aug 18, 2025

Abstract

The goal of this study is to come up with an intelligent predictive model that can classify the severity of heart disease. The model will employ both XGBoost and oversampling to resolve the problem of data imbalance. In addition, the model will be implemented for real-world application using an interactive interface. The study uses the UCI Heart Disease dataset, which includes many clinical features. Preprocessing involves handling missing values, removal of features with a substantial fraction of missing values, and the use of SMOTE resampling for learning from class-balanced instances. The main classifier that was used for the research purposes was the XGBoost classifier, while the dataset was split 80:20 for training and testing purposes. For ease of individual-level real-time testing of the predictions, the model is implemented through Streamlit. The XGBoost model worked extraordinarily well, with the accuracy standing at 92%, as did precision along with recall, as well as the F1-score, being 92%. These findings clearly outperform other current studies of the same sort that have made use of alternative classifiers. In addition, its deployment using Streamlit makes it even more clinically applicable. Innovation The novelty of the research lies in the combined application of SMOTE with XGBoost, enabling effective classification under imbalanced conditions, along with the real-time implementation using Streamlit for user-level predictions. The model is of high value for early identification and stratification of the severity of heart disease in clinical decision support settings.

Keywords : *Heart Disease, Machine Learning, SMOTE, Streamlit, XGBoost*

1. INTRODUCTION

Cardiovascular disease remains the leading cause of death globally [1], [2]. According to a recent World Health Organization report, cardiovascular disease causes more than 18 million deaths per year, with the greatest effect seen in the Southeast Asia region, most notably in Indonesia. In the region, cardiovascular disease is accounting for about 3.9 million deaths per year, or 30% of the total death toll [3]. Ischemic heart disease and stroke are the top causes of death as identified in Indonesia by the Ministry of Health, taking up 19.42% and 14.38% of total mortality, respectively. Their chief causative factors comprise hypertension, hyperglycemia, tobacco use, and obesity [4]. Another important concern in the treatment of heart disease is the lack of early diagnosis, and notably so in low and middle-income nations [5], [6]. Their limited availability of proper healthcare and diagnostic equipment is one of the reasons for their low rates of early diagnosis. Additionally, the clinical data may always contain an unbalanced distribution of classes, with a very low number of positive cases of heart disease compared to negative ones [7]. Such a skew can negatively affect the performance of predictor models, since they become biased toward the dominant class [8].

Based on issue, machine learning has shown the potential to aid in the early diagnosis of heart disease by recognizing patterns in readily available clinical data that conventional methods struggle to detect [9], [10], [11]. These models can be seen to be capable of helping physicians make better decisions by supplying anticipatory analyses based on previous patient histories [12]. That being said, the effectiveness of such models depends heavily on the accuracy of the data used in the training process. A major issue is the uneven distribution of classes, in which cases of heart disease are greatly overshadowed by cases of patients who are disease free. The uneven distribution can lead to biased predictions, as the model leans toward the dominant class and doesn't give proper regard to the important minority cases [13]. To counter such a problem, work is in progress to improve the representation of the data before training, such that the uneven distribution of the classes is corrected in a way that the model can be made to learn from both classes equally, thereby allowing for more balanced and accurate predictions [14].

Research by [15] explored the prediction of heart disease by the use of several oversampling methods, such as SMOTE and ADASYN, combined with a Decision Tree classifier. The main objective of the study was to improve predictive performance by accomplishing a balance in the dataset before the training process of the model. Based on the study, both accuracy and sensitivity were greatly enhanced by the use of oversampling, with SMOTE performance outperforming other resampling methods. The main value of the work lies in the study of different oversampling methods to combat the problem of imbalance in heart disease prediction. However, a weakness of the work is the sole use of a single, comparatively simpler model (Decision Tree), leading to predictions that may be non-generalizable to other, more complicated datasets. The study also did not tackle ensemble or combinatorial models, nor did the study extensively evaluate the effect of resampling in the model's interpretability—a critical aspect in medical decision-making.

Research by [16] examined the effect of several resampling methods—namely, random oversampling, random undersampling, and SMOTE—on the performance of various cardiovascular disease prediction models. The experimental results suggested that the integration of SMOTE with Random Forest yielded the best classification performance in accuracy, sensitivity, and specificity. One of the strengths of the study lies in the careful evaluation of several resampling techniques as compared to the performance of a variety of classifiers. A major drawback, however, is the lack of higher-level feature selection and optimization, an operation that may be adopted to maximize the accuracy of the model. The dataset used in the study also had not been rigorously inspected in search of redundancy or noise, elements that tend to jeopardize the model's generalizability to actual clinical datasets.

Research by [17] suggested a hybrid approach that combines K-Means SMOTE with ensemble learning for the classification of Acute Coronary Syndrome (ACS). The authors utilized K-Means clustering to generate more representative synthetic samples, which were then used to train ensemble models. The method yielded improvements in the F1-score and showed improved resistance to overfitting. The contribution is significant in the area of combining clustering methods with oversampling, thus solving the problem of intra-class diversity. Nevertheless, the study is limited by its comparatively narrow scope of application, which is specifically tailored for ACS, making it less generalizable to wider classes of heart disease. Additionally, the study did not include a complete assessment of class-wise precision and recall, which are critical considerations in the context of imbalanced medical datasets.

In contrast to in the above-stated studies, the approach used in this study has a number of significant advantages. Firstly, whereas [15] focuses on using a single Decision Tree model, the proposed approach here merges an enhanced and more robust model structure that makes use of ensemble techniques, ensuring better generalization and greater prediction accuracy across different datasets. In addition, the performance of the proposed model is enhanced by the application of advanced feature selection and optimization processes, as compared to [16], where these techniques were not applied. This allows for improved interpretability of the model and reduced interference from noise, addressing some of the flaws realized in the dataset analysis of [16]. Moreover, unlike [17], which is specifically designed for Acute Coronary Syndrome (ACS), the approach established here is designed so that it can be more general and therefore be capable of classifying a broader spectrum of cardiovascular diseases. This flexibility allows the model to be applied in different clinical environments, making it more versatile and useful across a variety of heart disease types. Lastly, the combination of SMOTE with ensemble learning and careful evaluation of class-wise precision and recall ensures a more balanced performance, which was not fully explored in [17]. This makes the proposed approach not only more accurate but also more reliable in real-world applications, where both sensitivity to minority instances and interpretability are crucial.

2. METHOD

2.1. Proposed Algorithm

Based on Figure 1, the heart disease prediction framework proposed is divided into three main phases: Pre-processing, Over Sampling and Modeling, and Evaluation. During the pre-processing stage, the process starts with data gathering and the determination of object data, including the establishment of relevant medical attribute variables like age, cholesterol, blood pressure, and other cardiovascular risk factors. Then, during the attribute data stage, the above-stated attribute variables are chosen so that only relevant features are included in the modeling process. Next, data cleaning is carried out to deal with missing values, outliers, and inconsistencies to improve the quality of the data before model training starts.

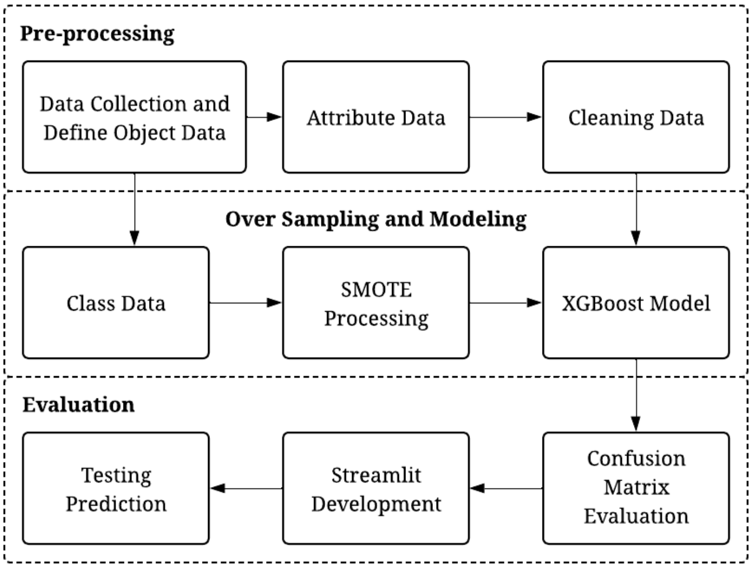


Figure 1. Flow of proposed method

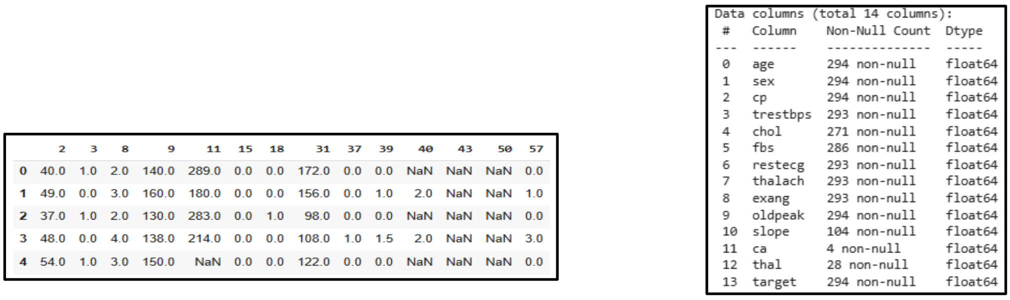
The next step includes Over Sampling and Modeling. In this step, the Class Data, which is often imbalanced as a result of a lack of heart disease cases, is subjected to the process of SMOTE (Synthetic Minority Over-sampling Technique). This method creates artificial samples of the minority class to correct the imbalance of the dataset, reducing the likelihood of bias of the prediction model for the majority class. Then, the balanced dataset is used for training the XGBoost model (Extreme Gradient Boosting) that is chosen for its precision, stability, as well as its ability to handle linear as well as non-linear trends.

In the Evaluation phase, the model is tested through a set of tests to evaluate its effectiveness. In the first step, the model is validated by comparing its predictions on unseen data for determining its fit for real-world deployment. In the second step, a Streamlit app is created for model deployment, hence providing an interactive interface for user participation. Finally, the model is evaluated quantitatively through the application of a confusion matrix, giving rise to key metrics of accuracy, precision, recall, and F1-score—key indicators for understanding the model's performance where class imbalances exist.

2.2. Data Collection and Define Object Data

The dataset used in this study was sourced from the UCI Machine Learning Repository, specifically the heart disease dataset, which can be accessed via https://archive.ics.uci.edu/dataset/45/heart+disease. The dataset comprises a total of 294 samples with 75 attributes (indexed from 0 to 74), though many of these features have a high percentage of missing values or are irrelevant for the classification task. For this research, a selection of 14 relevant attributes was made based on prior domain knowledge and the completeness of data entries. These selected features include: age, sex, cp (chest pain type), trestbps (resting blood pressure), chol (serum cholesterol), fbs (fasting blood sugar), restecg (resting ECG results), thalach (maximum heart rate achieved), exang (exercise-induced angina), oldpeak (ST depression), slope (slope of the peak exercise ST segment), ca (number of major vessels), thal (thalassemia), and target (presence of heart disease).

Initially, the target column in the dataset contains five distinct classes ranging from 0 to 4, where 0 indicates the absence of heart disease and values from 1 to 4 represent increasing levels of disease severity. To simplify the classification task and align with common clinical decision-making, the target was transformed into a binary class. All non-zero values were mapped to 1 (indicating the presence of heart disease), while zero remained as 0 (indicating no heart disease). This conversion supports the development of a binary classifier that focuses on distinguishing between healthy and diseased individuals, which is often more practical for early screening applications.



```
      2    3    8      9     11   15   18     31   37   39   40   43   50   57
0  40.0  1.0  2.0  140.0  289.0  0.0  0.0  172.0  0.0  0.0  NaN  NaN  NaN  0.0
1  49.0  0.0  3.0  160.0  180.0  0.0  0.0  156.0  0.0  1.0  2.0  NaN  NaN  1.0
2  37.0  1.0  2.0  130.0  283.0  0.0  1.0   98.0  0.0  0.0  NaN  NaN  NaN  0.0
3  48.0  0.0  4.0  138.0  214.0  0.0  0.0  108.0  1.0  1.5  2.0  NaN  NaN  3.0
4  54.0  1.0  3.0  150.0    NaN  0.0  0.0  122.0  0.0  0.0  NaN  NaN  NaN  0.0
```

(a) Raw data sorted by head (top)

```
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       294 non-null    float64
 1   sex       294 non-null    float64
 2   cp        294 non-null    float64
 3   trestbps  293 non-null    float64
 4   chol      271 non-null    float64
 5   fbs       286 non-null    float64
 6   restecg   293 non-null    float64
 7   thalach   293 non-null    float64
 8   exang     293 non-null    float64
 9   oldpeak   294 non-null    float64
 10  slope     104 non-null    float64
 11  ca        4 non-null      float64
 12  thal      28 non-null     float64
 13  target    294 non-null    float64
```

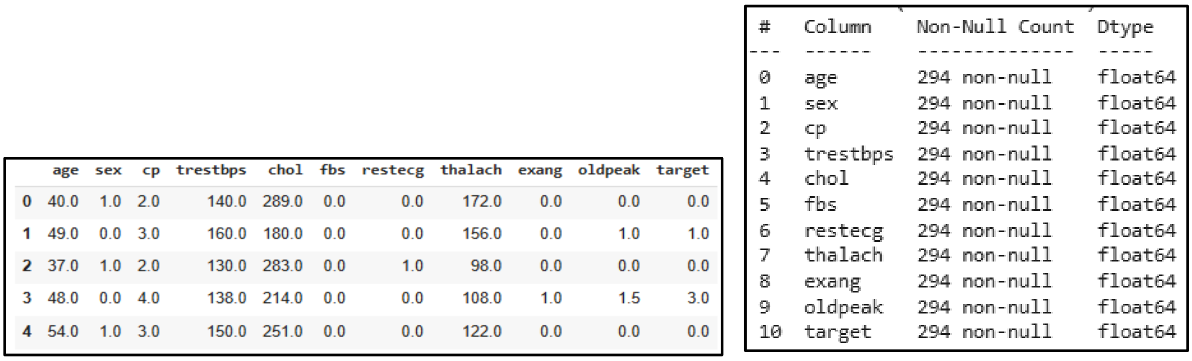
(b) Raw data information

Figure 2. Data collection (raw data)

2.3. Cleaning Data

Data cleaning is an integral part of any machine learning pipeline, especially given that medical datasets are often accompanied by missing records [18]. The dataset used for this study had many features with varying levels of missing values. In the analysis, the slope, ca, and thal features had missing values for most of their entries, indicating 190, 290, and 266 missing values, respectively, out of the total 294 entries [19], [20]. Since such features had well over 65% missing values, imputation would be illogical for them, hence it became necessary that they be dropped from the dataset so as not to compromise the quality of the data. This action is aimed at avoiding bias or distortion of the outcome that otherwise would have been caused by imputation of features with a very high incompleteness level.

After the aforementioned columns, there were only a few missing values for the other features. In specific, the features trestbps (one missing value), chol (twenty-three missing values), fbs (eight missing values), restecg (one missing value), thalach (one missing value), and exang (one missing value) were recognized as incomplete [21]. To solve for this problem, mean imputation is utilized, where missing values are replaced by the mean of the respective feature. The method, being simple, is sufficient for numerical medical data, especially for cases where missing inputs are infrequent, and the distribution of the data is assumed to be near-normal. Then, the mean of each feature is determined after converting values of the column into float, followed by averaging the same values, and then the values are rounded off up to the nearest integer for comparison purposes with the data structure.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	target
0	40.0	1.0	2.0	140.0	289.0	0.0	0.0	172.0	0.0	0.0	0.0
1	49.0	0.0	3.0	160.0	180.0	0.0	0.0	156.0	0.0	1.0	1.0
2	37.0	1.0	2.0	130.0	283.0	0.0	1.0	98.0	0.0	0.0	0.0
3	48.0	0.0	4.0	138.0	214.0	0.0	0.0	108.0	1.0	1.5	3.0
4	54.0	1.0	3.0	150.0	251.0	0.0	0.0	122.0	0.0	0.0	0.0

(a) Clean data sorted by head (top)

```
#   Column    Non-Null Count  Dtype
---  ------    --------------  -----
0   age       294 non-null    float64
1   sex       294 non-null    float64
2   cp        294 non-null    float64
3   trestbps  294 non-null    float64
4   chol      294 non-null    float64
5   fbs       294 non-null    float64
6   restecg   294 non-null    float64
7   thalach   294 non-null    float64
8   exang     294 non-null    float64
9   oldpeak   294 non-null    float64
10  target    294 non-null    float64
```

(b) Clean data information

Figure 3. Data collection (clean data)

2.4. Synthetic Minority Over-sampling Technique (SMOTE)

Synthetic Minority Over-sampling Technique (SMOTE) is used in this study to handle class integration on target attributes [22]. Based on the initial distribution of target data, it was found that class 0 dominates the amount of data, while other classes such as 1, 2, 3, and 4 have much smaller numbers. This imbalance has the potential to cause the prediction model to be biased towards the majority class, so a data balancing strategy is needed before the model training process.

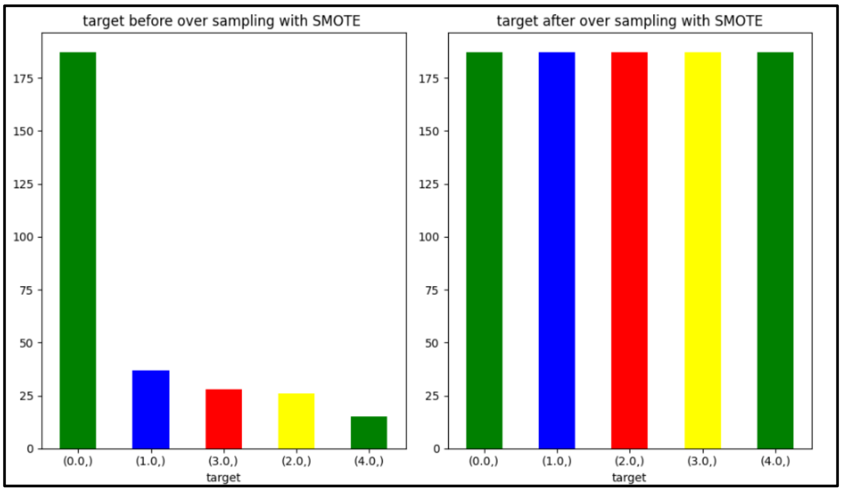


Figure 4. Results of SMOTE processing

SMOTE works by creating synthetic data on the minority class through interpolation between existing data in the feature space. In its application, this technique uses SMOTE from the imblearn library with the random_state=42 parameter to ensure reproducible results. This oversampling process successfully equalizes the amount of data in each target class, which is shown through the visualization of the bar chart before and after SMOTE [23]. After this process, all target classes have a balanced amount of data, so that the model training process can be carried out with equivalent class representation.

2.5. Xtreme Gradient Boosting (XGBoost) Model

After the data balancing process is carried out with SMOTE, the next stage is to build a classification model using the Xtreme Gradient Boosting (XGBoost) algorithm [24]. XGBoost is an ensemble algorithm based on boosting techniques, with advantages in terms of speed, accuracy, and the ability to handle unbalanced data and complex features. This algorithm works by building a decision tree sequentially, where each new tree will try to correct the errors of the previous tree by minimizing the loss function iteratively. In this study, XGBoost is used to study the pattern of feature data against multi-class targets (five classes) of heart disease prediction results [25]. The model is trained using oversampling data, so it is expected to be able to provide more balanced and accurate predictions for all target classes. Equationlly, the XGBoost model builds predictions additively through the following objective function as seen in Eq (1).

$$L(\phi) = \sum_{i=1}^{n} l\left(yi, yi^{(t)}\right) + \sum_{k=1}^{t} \Omega(fk) \tag{1}$$

Where, $L(\phi)$ is the overall objective function, $l\left(yi, yi^{(t)}\right)$ is the loss function measuring the difference between the predicted value $yi^{(t)}$ and the actual label yi at iteration t, $\Omega(fk)$ is the regularization term that penalizes the complexity of each tree fk. The regularization function of $\Omega(fk)$ given by eq (2).

$$\Omega(fk) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{2}$$

Where, T is the number of leaves in the tree, w_j is the output score on leaf j, γ and λ are regularization parameters to control overfitting. The final prediction at iteration t is obtained by summing the predictions of all previous trees in Eq (3). With F being the space of all regression trees. This formulation allows XGBoost to efficiently model complex, nonlinear relationships in the data, making it a go-to algorithm in many real-world machine learning competitions and research.

$$yi^{t} = \sum_{k=1}^{t} fk(xi), fk \in F \tag{3}$$

2.6. Confusion Matrix Evaluation

To assess the performance of a multiclass classification model, a confusion matrix is often used as a visualization tool that illustrates how well the model distinguishes between different output classes [26], [27]. Rather than providing a single scalar value, a confusion matrix breaks down predictions into a tabular format consisting of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each class. This matrix facilitates a more in-depth analysis of the model's behavior, especially for imbalanced datasets, where the overall accuracy may be misleading [28].

In this study, a confusion matrix is used to extract key evaluation metrics, including precision, recall, and F1 score for each class [29], [30]. Precision refers to the proportion of correct predictions among all positive predictions for a class. Recall (also known as sensitivity) measures the model's ability to identify all relevant class instances. F1 score, the harmonic mean of precision and recall, serves as a balancing measure in situations with class imbalance. This metric allows for a nuanced evaluation of model performance, highlighting not only the overall predictive ability but also how well each class is handled. The equation can be seen in Eq (4) – (7).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

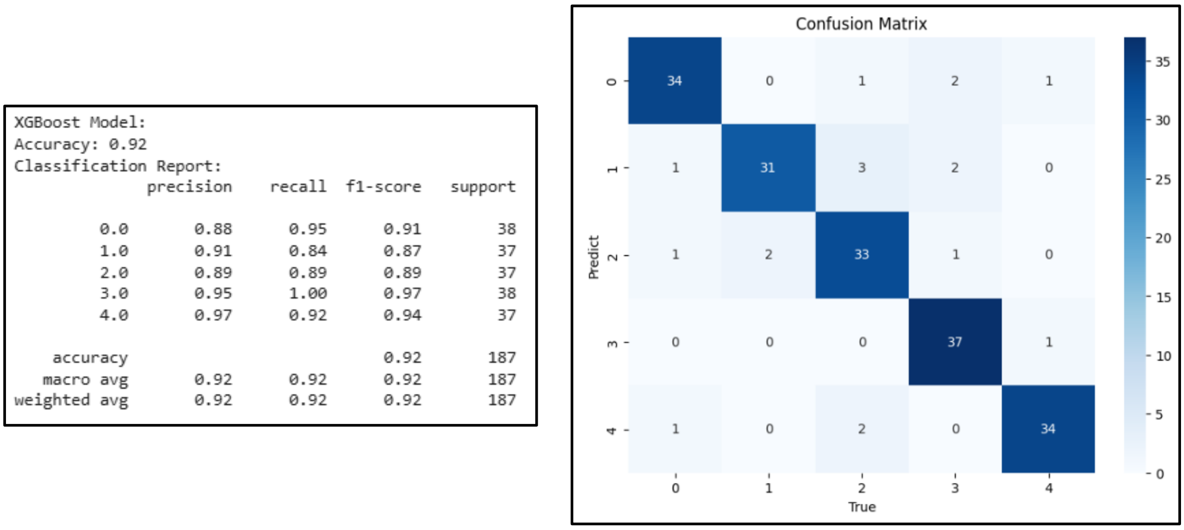
$$Recall = \frac{TP}{TP+FN} \tag{6}$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision \times Recall} \tag{7}$$

3. RESULT

All the computational experiments undertaken in this research were carried out in the Google Colaboratory (Colab) environment, which supports cloud-based GPU acceleration. The model training execution environment set up was configured for GPU hardware to maximize both the training time and the efficiency of the models. Although the Colab environment was the computational basis, the validations and tests of the codes were carried out on a locally running machine with an Intel Core i3 12th Generation processor, an NVIDIA RTX 2060 GPU, 16 GB of memory, and a 2 TB SSD. The use of this combined cloud and local infrastructure made the training of deep models both flexible and optimal for efficiently processing large datasets.

The training procedure involves developing a predictive model from a preprocessed and balanced dataset. After applying SMOTE for class imbalance correction, the dataset is split between training and testing datasets based on an 80:20 ratio. This split allows the model to have enough volume for the learning procedure while leaving a dedicated portion for evaluating performance. The Xtreme Gradient Boosting (XGBoost) algorithm is then trained on 80% of the training set, leveraging advanced ensemble learning philosophy and gradient optimization to efficiently handle structured data and understand complex feature interactions. Results of this phase can be seen in Figure 5.

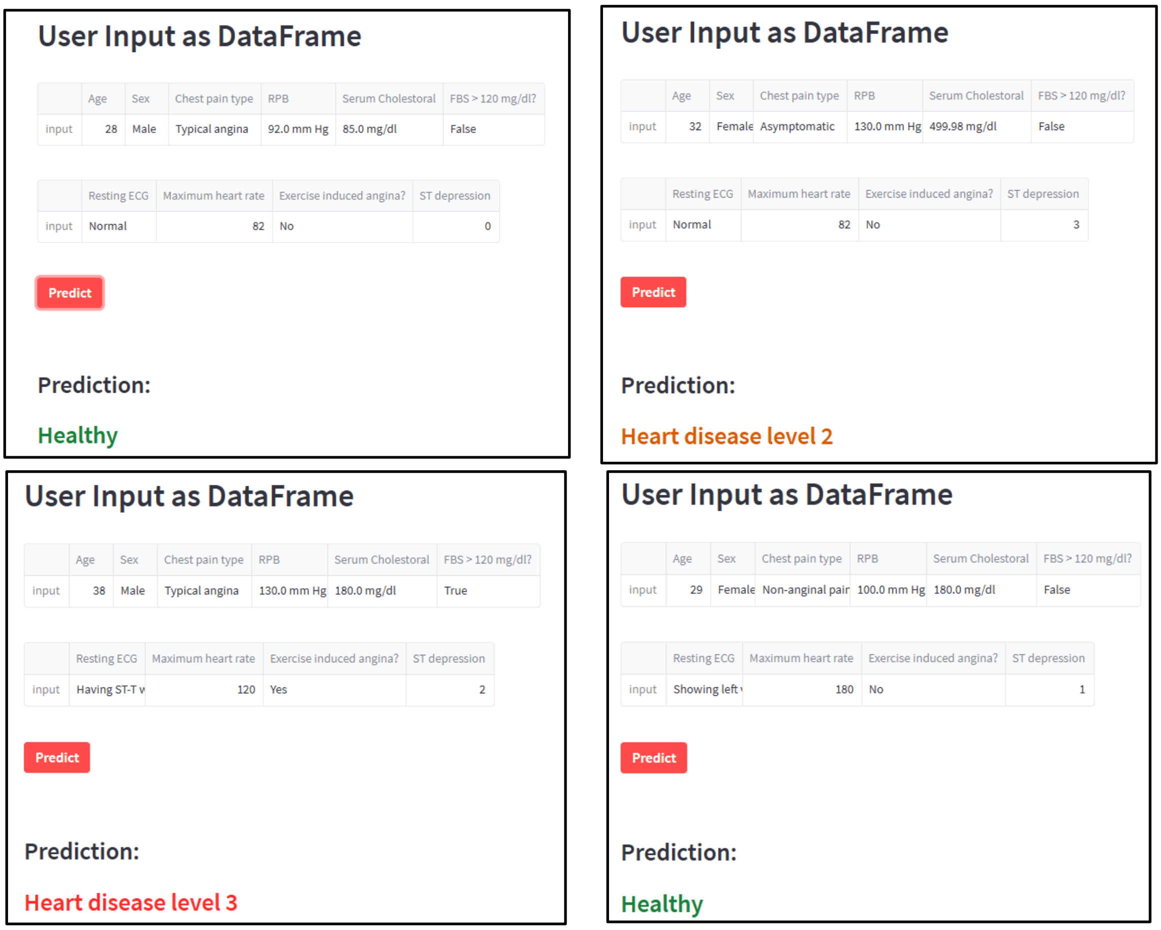


```
XGBoost Model:
Accuracy: 0.92
Classification Report:
              precision    recall  f1-score   support

         0.0       0.88      0.95      0.91        38
         1.0       0.91      0.84      0.87        37
         2.0       0.89      0.89      0.89        37
         3.0       0.95      1.00      0.97        38
         4.0       0.97      0.92      0.94        37

    accuracy                           0.92       187
   macro avg       0.92      0.92      0.92       187
weighted avg       0.92      0.92      0.92       187
```

(a) Performance matrix (b) Tabel of confusion matrix by 20% testing data

Figure 5. Performance of XGBoost model

For best performance, the model is trained on optimized hyperparameters, allowing it to learn complex decision boundaries between different classes of heart disease. The remaining twenty percent of the dataset is used to evaluate the predictivity of the model.



User Input as DataFrame

	Age	Sex	Chest pain type	RPB	Serum Cholestoral	FBS > 120 mg/dl?
input	28	Male	Typical angina	92.0 mm Hg	85.0 mg/dl	False

	Resting ECG	Maximum heart rate	Exercise induced angina?	ST depression
input	Normal	82	No	0

Prediction:

Healthy

User Input as DataFrame

	Age	Sex	Chest pain type	RPB	Serum Cholestoral	FBS > 120 mg/dl?
input	32	Female	Asymptomatic	130.0 mm Hg	499.98 mg/dl	False

	Resting ECG	Maximum heart rate	Exercise induced angina?	ST depression
input	Normal	82	No	3

Prediction:

Heart disease level 2

User Input as DataFrame

	Age	Sex	Chest pain type	RPB	Serum Cholestoral	FBS > 120 mg/dl?
input	38	Male	Typical angina	130.0 mm Hg	180.0 mg/dl	True

	Resting ECG	Maximum heart rate	Exercise induced angina?	ST depression
input	Having ST-T w	120	Yes	2

Prediction:

Heart disease level 3

User Input as DataFrame

	Age	Sex	Chest pain type	RPB	Serum Cholestoral	FBS > 120 mg/dl?
input	29	Female	Non-anginal pain	100.0 mm Hg	180.0 mg/dl	False

	Resting ECG	Maximum heart rate	Exercise induced angina?	ST depression
input	Showing left	180	No	1

Prediction:

Healthy

Figure 6. Streamlit model prediction

To optimize performance, the model underwent hyperparameter tuning, enabling it to effectively capture complex, non-linear relationships between clinical features and heart disease severity. The final model was evaluated on a 20% hold-out test set, ensuring that no data leakage occurred during training. The confusion matrix derived from the test data indicates high classification accuracy across both positive and negative classes.

Notably, the model achieves balanced precision and recall values (92%)—a clear indication that the oversampling strategy via SMOTE successfully mitigated the issue of class imbalance. Without SMOTE, the model exhibited a tendency to underpredict minority (positive) cases, as commonly observed in imbalanced datasets. However, after SMOTE was applied, the synthetic augmentation of underrepresented classes allowed XGBoost to better generalize decision boundaries, particularly for early-stage disease instances that typically present more subtle clinical signatures.

Moreover, the F1-score, as the harmonic mean of precision and recall, further confirms the model's robustness. This metric is especially relevant in medical diagnostics, where both false positives (which may cause unnecessary alarm) and false negatives (which may delay treatment) carry serious implications. The XGBoost+SMOTE model maintained this metric at 92%, highlighting a reliable balance between sensitivity and specificity. Figure 5 and Table 1 consolidate these results, showing that the proposed approach consistently outperforms other models in the literature.

4. DISCUSSIONS

Table 1. Comparison of machine learning models

Researcher	Model	Performance Evaluation

		Accuracy	Precision	Recall	F1-Score
Ref [15]	Decision Tree with SMOTE	91%	91%	91%	90%
Ref [16]	Random Forest with SMOTE	90.53%	N/A	85.71%	N/A
Ref [17]	K-Means with SMOTE	85.34%	N/A	N/A	85.15%
Our	XGBoost with SMOTE	92%	92%	92%	92%

Based on Table 1, the proposed model using XGBoost combined with SMOTE outperformed other referenced models across all performance evaluation metrics—achieving 92% accuracy, precision, recall, and F1-score. Compared to the Decision Tree by [15], which attained 91% accuracy and a slightly lower F1-score of 90%, our model shows improved consistency and balance across all metrics. Furthermore, while Random Forest by [16] reached 90.53% accuracy, the lack of reported precision and F1-score indicates limited evaluation transparency. The K-Means model by [17] performed notably worse, with only 85.34% accuracy and an F1-score of 85.15%.

This comparison highlights the robustness and generalization ability of our XGBoost model in handling class imbalance through SMOTE, particularly in multi-class heart disease prediction. The integration of a powerful ensemble-based learner (XGBoost) and effective oversampling significantly reduces bias toward majority classes, resulting in more stable and reliable classification across all classes. While decision trees and random forests remain strong contenders, they may be more prone to overfitting or fail to optimize recall and F1-score equally. In contrast, our model achieves balanced performance without sacrificing any evaluation aspect, thus positioning it as a superior approach in terms of both predictive accuracy and comprehensive classification capability.

In addition to the numerical superiority of the XGBoost with SMOTE approach, a deeper analysis highlights the underlying factors behind this performance gain. Unlike traditional classifiers such as Decision Trees [15] or even ensemble models like Random Forest [16], XGBoost incorporates gradient boosting with regularization, which enables it to reduce both bias and variance more effectively. This characteristic is critical in medical datasets, where overfitting is a common issue due to limited and noisy data.

Furthermore, the superior F1-score achieved by our model indicates not only improved prediction accuracy but also consistency in handling both the positive (diseased) and negative (healthy) classes. This is essential in clinical practice where misclassification of minority cases (false negatives) can have severe consequences. Models like K-Means SMOTE [17], although effective in certain specialized applications such as Acute Coronary Syndrome (ACS), lack the generalizability and robustness needed to handle broader heart disease classifications.

Another factor contributing to the performance of our proposed model is the preprocessing strategy, including rigorous feature selection and noise reduction, which were either absent or underexplored in previous works [15], [16]. These preprocessing steps play a crucial role in enhancing the interpretability and clinical relevance of the model. In summary, the combined use of SMOTE and XGBoost, augmented by careful data preparation and a real-time deployment interface, positions this research not just as a numerical improvement but also as a step forward in practical, clinically applicable heart disease prediction tools.

5. CONCLUSION

The study successfully puts the XGBoost algorithm together with SMOTE into action for the improvement of heart disease prediction classification quality. By precise preprocessing of the dataset, handling missing values, removing incomplete features, and using SMOTE for balancing the dataset. the model could learn effectively from the given input features. It divided the dataset so that 80% of it could be used for training purposes and 20% for testing purposes, with the training being carried out on the Google Colab environment, which provided GPU support for efficient computation and minimized execution time. The test phase verifies that the model of XGBoost, as trained, had excellent performance, as shown through an overall accuracy, precision, recall, and F1-score of 92% in average. These performances outdo those of other benchmark models such as Decision Tree, Random Forest, and K-Means where combined with SMOTE, as shown through the comparison table. The balanced and high performance metrics clearly prove the stability of the model as well as its ability to generalize well across the classes of heart diseases. For future research, the system can be further developed into a multi-input prediction platform using Streamlit, allowing simultaneous assessment of multiple patients or batch predictions. Moreover, exploring and comparing other advanced machine learning models, such as LightGBM, CatBoost, or neural network-based approaches may lead to performance improvements and deeper insights in heart disease classification tasks.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

REFERENCES

[1] A. V. Poznyak, L. Litvinova, P. Poggio, V. N. Sukhorukov, and A. N. Orekhov, “Effect of Glucose Levels on Cardiovascular Risk,” Cells, vol. 11, no. 19, p. 3034, Sep. 2022, doi: 10.3390/cells11193034.
[2] Y. Wang and D. J. Magliano, “Special Issue: ‘New Trends in Diabetes, Hypertension, and Cardiovascular Diseases,’” Mar. 01, 2024, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/ijms25052711.
[3] Md. A. Rahman, S. Cronmiller, Y. Shanjana, M. A. Bhuiyan, and Md. R. Islam, “The WHO announced COVID-19 is no longer a global public health emergency amid the spreading of arcturus variant: a correspondence evaluating this decision,” International Journal of Surgery, May 2023, doi: 10.1097/JS9.0000000000000522.
[4] World Health Organization, “World Heart Day: Cardiovascular diseases claim 3.9 million lives in the WHO South-East Asia Region every year.” Accessed: Apr. 17, 2025. [Online]. Available: https://www.who.int/southeastasia/news/detail/29-09-2024-world-heart-day
[5] R. G. Russo et al., “COVID-19, Social Determinants of Health, and Opportunities for Preventing Cardiovascular Disease: A Conceptual Framework,” J Am Heart Assoc, vol. 10, no. 24, p. e022721, Dec. 2021, doi: 10.1161/JAHA.121.022721.
[6] F. Sapna et al., “Advancements in Heart Failure Management: A Comprehensive Narrative Review of Emerging Therapies,” Cureus, Oct. 2023, doi: 10.7759/cureus.46486.
[7] N. Chandrasekhar and S. Peddakrishna, “Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization,” Processes, vol. 11, no. 4, Apr. 2023, doi: 10.3390/pr11041210.
[8] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, “Effective Heart Disease Prediction Using Machine Learning Techniques,” Algorithms, vol. 16, no. 2, Feb. 2023, doi: 10.3390/a16020088.
[9] B. Arjmand et al., “Machine Learning: A New Prospect in Multi-Omics Data Analysis of Cancer,” Jan. 27, 2022, Frontiers Media S.A. doi: 10.3389/fgene.2022.824451.
[10] J. Wen, Z. Zhang, Y. Lan, Z. Cui, J. Cai, and W. Zhang, “A survey on federated learning: challenges and applications,” International Journal of Machine Learning and Cybernetics, vol. 14, no. 2, pp. 513–535, Feb. 2023, doi: 10.1007/s13042-022-01647-y.
[11] N. R. D. Cahyo, C. A. Sari, E. H. Rachmawanto, C. Jatmoko, R. R. A. Al-Jawry, and M. A. Alkhafaji, “A Comparison of Multi Class Support Vector Machine vs Deep Convolutional Neural Network for Brain Tumor Classification,” in 2023 International Seminar on Application for Technology of Information and Communication (iSemantic), IEEE, Sep. 2023, pp. 358–363. doi: 10.1109/iSemantic59612.2023.10295336.
[12] I. P. Kamila, C. A. Sari, E. H. Rachmawanto, and N. R. D. Cahyo, “A Good Evaluation Based on Confusion Matrix for Lung Diseases Classification using Convolutional Neural Networks,” Advance Sustainable Science, Engineering and Technology, vol. 6, no. 1, p. 0240102, Dec.

2023, doi: 10.26877/asset.v6i1.17330.

[13] X. Jiang and Z. Ge, "Data augmentation classifier for imbalanced fault classification," IEEE Transactions on Automation Science and Engineering, vol. 18, no. 3, pp. 1206–1217, 2020.

[14] S. J. Basha, S. R. Madala, K. Vivek, E. S. Kumar, and T. Ammannamma, "A Review on Imbalanced Data Classification Techniques," in 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA), IEEE, Mar. 2022, pp. 1–6. doi: 10.1109/ICACTA54488.2022.9753392.

[15] A. J. Albert, R. Murugan, and T. Sripriya, "Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology," Research on Biomedical Engineering, vol. 39, no. 1, pp. 99–113, Dec. 2022, doi: 10.1007/s42600-022-00253-9.

[16] U. Hasanah, A. M. Soleh, and K. Sadik, "Effect of Random Under sampling, Oversampling, and SMOTE on the Performance of Cardiovascular Disease Prediction Models," Jurnal Matematika, Statistika dan Komputasi, vol. 21, no. 1, pp. 88–102, Sep. 2024, doi: 10.20956/j.v21i1.35552.

[17] M. F. Muzakki, R. D. Prayogo, and M. A. Rizky A, "Handling Imbalanced Data for Acute Coronary Syndrome Classification Based on Ensemble and K-Means SMOTE Method," JOIV : International Journal on Informatics Visualization, vol. 7, no. 3–2, p. 1989, Nov. 2023, doi: 10.30630/joiv.7.3-2.1429.

[18] P. Li, X. Rao, J. Blase, Y. Zhang, X. Chu, and C. Zhang, "CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks," in 2021 IEEE 37th International Conference on Data Engineering (ICDE), IEEE, Apr. 2021, pp. 13–24. doi: 10.1109/ICDE51399.2021.00009.

[19] A. Palanivinayagam and R. Damaševičius, "Effective Handling of Missing Values in Datasets for Classification Using Machine Learning Methods," Information, vol. 14, no. 2, p. 92, Feb. 2023, doi: 10.3390/info14020092.

[20] Z. Abidin, T. Tamrin, V. Harsono, D. N. Aziza, and I. Kansania, "OPTIMALISASI DIAGNOSIS STROKE DENGAN ALGORITMA C4.5 DAN STRATEGI IMPUTASI k-NN UNTUK MENGATASI MISSING VALUE," Jurnal Disprotek, vol. 15, no. 2, pp. 152–160, Aug. 2024, doi: 10.34001/jdpt.v15i2.6701.

[21] P. Dabhade, R. Agarwal, K. P. Alameen, A. T. Fathima, R. Sridharan, and G. Gopakumar, "Educational data mining for predicting students' academic performance using machine learning algorithms," Mater Today Proc, vol. 47, pp. 5260–5267, 2021, doi: 10.1016/j.matpr.2021.05.646.

[22] F. R. Adi Pratama and S. I. Oktora, "Synthetic Minority Over-sampling Technique (SMOTE) for handling imbalanced data in poverty classification," Stat J IAOS, vol. 39, no. 1, pp. 233–239, Mar. 2023, doi: 10.3233/SJI-220080.

[23] A. D. Amirruddin, F. M. Muharam, M. H. Ismail, N. P. Tan, and M. F. Ismail, "Synthetic Minority Over-sampling TEchnique (SMOTE) and Logistic Model Tree (LMT)-Adaptive Boosting algorithms for classifying imbalanced datasets of nutrient and chlorophyll sufficiency levels of oil palm (Elaeis guineensis) using spectroradiometers and unmanned aerial vehicles," Comput Electron Agric, vol. 193, p. 106646, Feb. 2022, doi: 10.1016/j.compag.2021.106646.

[24] A. Farzipour, R. Elmi, and H. Nasiri, "Detection of Monkeypox Cases Based on Symptoms Using XGBoost and Shapley Additive Explanations Methods," Diagnostics, vol. 13, no. 14, p. 2391, Jul. 2023, doi: 10.3390/diagnostics13142391.

[25] J. Ou et al., "Coupling UAV Hyperspectral and LiDAR Data for Mangrove Classification Using XGBoost in China's Pinglu Canal Estuary," Forests, vol. 14, no. 9, p. 1838, Sep. 2023, doi: 10.3390/f14091838.

[26] N. R. D. Cahyo and M. M. I. Al-Ghiffary, "An Image Processing Study: Image Enhancement, Image Segmentation, and Image Classification using Milkfish Freshness Images," IJECAR) International Journal of Engineering Computing Advanced Research, vol. 1, no. 1, pp. 11–22, 2024.

[27] F. Farhan, C. A. Sari, E. H. Rachmawanto, and N. R. D. Cahyo, "Mangrove Tree Species Classification Based on Leaf, Stem, and Seed Characteristics Using Convolutional Neural

Networks with K-Folds Cross Validation Optimalization," Advance Sustainable Science Engineering and Technology, vol. 5, no. 3, p. 02303011, Oct. 2023, doi: 10.26877/asset.v5i3.17188.

[28] M. M. I. Al-Ghiffary, C. A. Sari, E. H. Rachmawanto, N. M. Yacoob, N. R. D. Cahyo, and R. R. Ali, "Milkfish Freshness Classification Using Convolutional Neural Networks Based on Resnet50 Architecture," Advance Sustainable Science Engineering and Technology, vol. 5, no. 3, p. 0230304, Oct. 2023, doi: 10.26877/asset.v5i3.17017.

[29] R. R. Ali et al., "Learning Architecture for Brain Tumor Classification Based on Deep Convolutional Neural Network: Classic and ResNet50," Diagnostics, vol. 15, no. 5, p. 624, Mar. 2025, doi: 10.3390/diagnostics15050624.

[30] E. H. Rachmawanto, C. A. Sari, and F. O. Isinkaye, "A good result of brain tumor classification based on simple convolutional neural network architecture," Telkomnika (Telecommunication Computing Electronics and Control), vol. 22, no. 3, pp. 711–719, Jun. 2024, doi: 10.12928/TELKOMNIKA.v22i3.25863.