Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2633-2645
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5626

Performance Comparison Of K-Nearest Neighbors And Decision Tree Algorithms With Random Oversampling For Imbalanced Heart Disease Classification

Dita Yustianisa[*1], Farid Wajidi[2], Wawan Firgiawan[3], Adinda Gama Sholeha[4]

[1,2,3]Informatics, Universitas Sulawesi Barat, Indonesia
[4]Computer Science, Albukhary International University, Malaysia

Email: [1]yustianisadita@gmail.com

Received : Jan 6, 2026; Revised : Jan 6, 2026; Accepted : Jan 18, 2026; Published : Jun 15, 2026

Abstract

Heart disease remains one of the leading causes of mortality worldwide, including in Indonesia, where delayed detection continues to be a serious challenge. In medical data mining, class imbalance often degrades classification performance by reducing sensitivity toward minority class cases. This study aims to compare the performance of the K-Nearest Neighbors (KNN) and Decision Tree algorithms for heart disease classification and to evaluate the effectiveness of random oversampling in handling imbalanced data. This research uses a heart disease dataset consisting of 10,000 medical records obtained from Kaggle. Data preprocessing includes categorical transformation, missing value imputation using KNN Imputer, and Min–Max normalization. Random oversampling is applied to increase minority class representation. Model evaluation is performed using stratified 10-fold cross-validation with accuracy, precision, recall, F1-score, and Receiver Operating Characteristic–Area Under the Curve (ROC–AUC) as performance metrics. Experimental results show that after random oversampling, the KNN model achieves the best performance with an accuracy of 94%, precision of 96%, recall of 90%, F1-score of 92%, and ROC–AUC of 90.2%. In comparison, the Decision Tree model records an accuracy of 80%, precision of 78%, recall of 81%, F1-score of 79%, and ROC–AUC of 81.5%. These findings demonstrate that random oversampling significantly improves minority class detection, particularly for KNN. This study contributes to Informatics by providing empirical evidence that simple and efficient data mining strategies can effectively address class imbalance in large-scale medical datasets, supporting the development of accurate, interpretable, and accessible AI-based diagnostic systems for early heart disease detection.

Keywords : *Classification, Data Mining, Decision Tree, Heart Disease, K-Nearest Neighbors, Random Oversampling.*

1. INTRODUCTION

Globally, including in Indonesia, heart disease continues to be a leading cause of death. According to a report by the World Health Organization (WHO) in 2021, deaths caused by heart disease reached approximately 17.8 million cases, accounting for one out of every three deaths worldwide each year [1]. Heart diseases or disorders encompass various types and classifications, including cardiovascular disease, coronary heart disease, and heart attacks [2]. Cardiovascular disease remains the primary cause of mortality worldwide to this day [3], [4], [5].

The heart is a vital organ in the human body that functions to pump and circulate blood carrying oxygen to all parts of the body [6], [7]. Heart disease is classified as a dangerous condition because it can disrupt these vital functions [8]. Heart disease is generally caused by various risk factors associated with unhealthy lifestyles, such as lack of physical activity, consumption of foods high in fat and salt, smoking habits, excessive alcohol consumption, and high stress levels [9], [10]. One of the main challenges in managing heart disease is the delay in early detection, as symptoms often do not appear

Vol. 7, No. 3, June 2026, Page. 2633-2645
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5626

clearly at the initial stages [8]. In the medical diagnostic process, the involvement of competent medical experts is required. However, with the advancement of technology, data mining approaches can be utilized to assist the disease diagnosis process based on patient medical record data, including in the case of heart disease [11], [12].

In the field of data mining, several methods are commonly applied, including association, clustering, classification, and regression. Among these methods, classification is one of the most widely used approaches [13], [14], [3].Classification is the process of assigning a class or category to each data sample by utilizing sample attributes as inputs to the classification model, while the sample class serves as the output [15]. One of the common challenges in the classification process is class imbalance, which arises when the data distribution among classes is uneven, with the minority class containing fewer samples than the majority class [16], [17]. Therefore, this study applies a random oversampling technique to balance the dataset by randomly oversampling the minority class (Heart Disease). This approach aims to equalize the data distribution by increasing the number of minority class samples through random duplication, as described by [18]. This method is chosen due to its simplicity and its ability to improve minority class representation without removing the original data contained in the dataset [19].

Various algorithms can be applied in the classification process, including Decision Tree, Naïve Bayes, K-Nearest Neighbor, and other algorithms [20]. In this study, the Decision Tree and K-Nearest Neighbor algorithms are employed to perform heart disease classification. The Decision Tree algorithm offers advantages in terms of interpretability, as its classification results can be visualized in the form of a decision tree structure [21]. This approach is also effective for data exploration and for uncovering latent relationships between multiple input variables and the target variable [22]. In contrast, the K-Nearest Neighbor (KNN) algorithm is considered a straightforward yet powerful technique, especially when applied to large-scale datasets. The algorithm performs classification by evaluating the distance between a new data instance and its closest neighbors. The number of neighbors involved in this process is denoted by the parameter K, which has a significant influence on the decision-making mechanism of the KNN algorithm [9], [23].

In contrast to the study by [24] which utilized a relatively limited dataset and did not specifically address class imbalance issues, potentially leading to biased model performance, this study employs a large-scale heart disease dataset consisting of 10,000 records and applies a random oversampling technique to enhance minority class representation. As a result, the proposed approach achieves more balanced and stable classification performance. Furthermore, a study by [25] focused on the application of oversampling in supervised learning algorithms and reported a decrease in accuracy after the oversampling process. In contrast to that approach, the present study evaluates the effect of simple oversampling on non-ensemble algorithms, namely Decision Tree and K-Nearest Neighbors (KNN). This approach provides a clearer understanding of the behavior and performance of commonly used classification algorithms in data mining under imbalanced data conditions. Meanwhile, study [26] entitled “Comparative Analysis of XGBoost, KNN, and SVM Algorithms for Heart Disease Prediction Using SMOTE-Tomek Balancing” demonstrates that complex models such as XGBoost combined with the SMOTE-Tomek technique are able to achieve high accuracy and AUC values. However, this approach requires greater computational resources and results in models with lower interpretability, making it less practical for real-world medical datasets with limited resources. Therefore, this study adopts a different approach by analyzing the performance of the K-Nearest Neighbors and Decision Tree algorithms, which are lighter and more interpretable, and by evaluating the effectiveness of random oversampling as a simple and practical solution for handling class imbalance in resource-constrained clinical decision support systems.

Vol. 7, No. 3, June 2026, Page. 2633-2645
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5626

2. METHOD

This study employs a data mining approach using classification methods to identify heart disease based on patients' medical record data. The research stages are systematically and structurally organized, starting from a literature review, followed by data collection and preprocessing, model development using classification algorithms, and concluding with the evaluation of classification performance. The overall research workflow is illustrated in Figure 1.

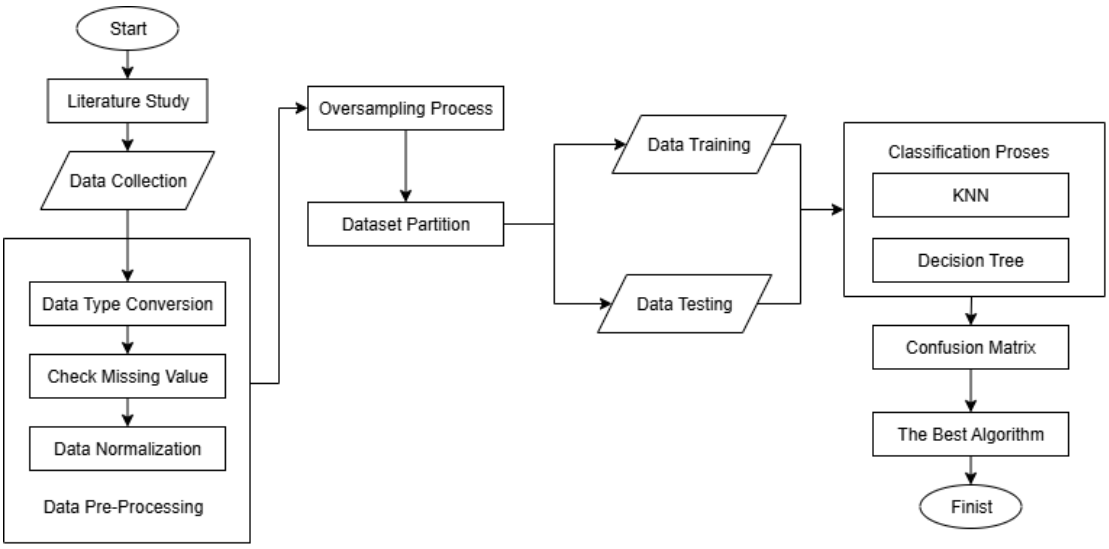


Figure 1. Research Flowchart

Figure 1 illustrates the research flowchart workflow, which begins with literature review and data collection. The acquired data then undergo preprocessing stages, including data type conversion, missing value checking, and data normalization. Subsequently, oversampling is performed to address class imbalance in the dataset. The balanced data are then divided into training and testing sets. In the next stage, classification is conducted using the K-Nearest Neighbors (KNN) and Decision Tree algorithms. The classification results are evaluated using a confusion matrix, and based on this evaluation, the best-performing algorithm is determined as the final output of the study

2.1. Literature Study

At this stage, a comprehensive review of various sources relevant to the research is conducted. The literature review aims to collect supporting reference materials derived from previous scientific journals, books, and other related sources that can serve as a foundation for completing the study [27]. Subsequently, the research problem is formulated to determine how the K-Nearest Neighbors (KNN) and Decision Tree algorithms can be implemented and compared in classifying heart disease using patient medical record data, as well as how the random oversampling technique can be applied to address class imbalance in order to enhance the performance and accuracy of the classification models.

2.2. Data Collection

In this study, the dataset used is obtained from a secondary data source downloaded from the Kaggle website entitled "Heart Disease", which consists of 10,000 entries. Healthcare practitioners, academics, and data analysts can use this information to look at heart disease trends, find risk factors, and perform other health-related investigations. Consisting of 21 attributes that include demographic data, lifestyle factors, clinical conditions, and biochemical indicators. Demographic attributes include age and gender, while lifestyle factors comprise exercise habits, smoking status, alcohol consumption, stress level, sleep duration, and sugar intake. Individual health conditions are represented by attributes

Vol. 7, No. 3, June 2026, Page. 2633-2645
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5626

such as family history of heart disease, diabetes, high blood pressure, body mass index (BMI), cholesterol levels (HDL and LDL), triglyceride levels, and fasting blood glucose. In addition, biomarker attributes such as CRP level and homocysteine level are used to describe inflammatory conditions and vascular health. The heart disease status attribute serves as the target variable, indicating the presence or absence of heart disease in an individual.

2.3. Data Pre-processing

In this study, the data pre-processing stage consists of data type conversion, missing value handling, and data normalization to ensure optimal model performance. Data type conversion is performed using the label encoding technique, where ordinal attributes are mapped to values ranging from 0 to 2 (Low, Medium, High), while binary attributes are encoded as 0 and 1 (No and Yes), enabling all features to be processed numerically during the modeling stage. Subsequently, missing value checking is conducted to identify incomplete data within each attribute, as missing values can negatively affect model accuracy and overall performance [11], [22]. To address this issue, the KNN Imputer method with five nearest neighbors (n_neighbors = 5) is applied, which imputes missing values based on the similarity of neighboring data instances. Finally, data normalization is carried out using the Min–Max normalization technique [31] to standardize feature values within the range of 0 to 1, thereby ensuring that all attributes contribute proportionally to the learning process.

2.4. Imbalance Dataset

In this study, class imbalance is handled using the random oversampling method. In this study, the size of the minority class is expanded to reach 50% of the majority class. The use of the random_state = 42 parameter ensures the reproducibility of the experimental results [4]. Following the oversampling procedure, the class distribution becomes more balanced, enabling the classification models to learn more effectively from both classes and minimizing bias toward the majority class.

2.5. Dataset Partition

As stated in [28], training data consist of labeled instances that are utilized by the model to learn data characteristics and to construct classification patterns or models. In contrast, testing data are labeled instances used to assess the accuracy of the trained model when classifying previously unseen data. In this study, model accuracy is evaluated using the percentage split approach, where the dataset is divided into 75% training data and 25% testing data [31].

2.6. K-Nearest Neighbor (KNN) Model

In the K-Nearest Neighbor (KNN) algorithm, the classification process is carried out by calculating the distance between a data instance to be classified and the instances within the existing dataset. The algorithm then identifies a number of closest data points, referred to as k neighbors, which are used to determine the class of the new instance. [29].

The procedural steps of the K-Nearest Neighbor method are described as follows [30]:

a. Determination of the Value of K

In this study, the Elbow method is utilized to identify the optimal K value, as it is widely acknowledged for its simplicity and effectiveness in visually indicating the most suitable parameter value [31]. The Elbow method helps interpret and validate consistency in cluster analysis and select the optimal number of clusters by adjusting the model to a range of K values. The way to determine the comparison is by knowing the SSE (Sum of Squared Error) value of each class/group. The more dominant the number of K in the class automatically the lower the SSE value [32].

Vol. 7, No. 3, June 2026, Page. 2633-2645
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5626

b. Distance Calculation Between Training and Testing Data.

$$d_i = \sqrt{\sum i = (x_i - y_i)^2} \quad (1)$$

Where :
d = Distance between data points
I = Number of data features
x_i= Testing Data
y_i = Training Data

c. Sorting of Distance Calculation Results
d. Class Determination

2.7. Decision Tree Model

Although classification has been extensively studied in the past, many of the proposed classification techniques have not been effectively applied to large-scale datasets. However, the use of large datasets is important for improving classification accuracy [33]. Decision Tree is one of the most widely used classification algorithms in data mining due to its ability to produce models that are easy to understand and visually interpretable [34]. The decision tree process transforms tabular data into a tree-based model, then converts the tree into a set of rules, and subsequently simplifies these rules. The objective of this model is to decompose complex decision-making processes into simpler ones, thereby producing more accurate decisions as solutions to the problem [6].

The pseudocode for Decision-Tree algorithm is as follows :

a. Keep the best feature of the input attributes at the root portion of the tree.
b. Then make a splitting of training dataset into subsections.
c. These splitted subsets can be done by making the each subset with data having the similar value for a input attribute.
d. Now repeat the step 1, 2 and step 3 on each subset till the leaf portion in every branches of the tree is found.

2.8. Confusion Matrix Evaluation

The classification results are evaluated to measure the accuracy of the analyzed algorithms and to determine the feasibility of the developed classification models. The evaluation process involves visualizing the model's accuracy and loss to detect potential issues related to overfitting or underfitting. Model performance is assessed using the confusion matrix, which provides a comparative representation between predicted classification outcomes and actual data in the form of a matrix.

The performance of the proposed model is evaluated using accuracy, precision, recall, F1-score, and ROC–AUC metrics. Accuracy measures the proportion of correctly classified instances among all predictions, while precision indicates the reliability of positive predictions by comparing True Positives to the total predicted positives. Recall evaluates the model's sensitivity in identifying actual positive cases by measuring the ratio of True Positives to all actual positives. The F1-score, defined as the harmonic mean of precision and recall, provides a balanced evaluation, particularly for imbalanced datasets. The corresponding equations for accuracy, precision, recall, and F1-score are presented in Equations (4)–(7). Additionally, the ROC curve and AUC are used to assess the model's ability to distinguish between positive and negative classes, where an AUC value closer to 1 indicates better classification performance.

$$Accuracy = \frac{TP}{TN+FP+FN+TP} \quad (4)$$

Vol. 7, No. 3, June 2026, Page. 2633-2645
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5626

$$Precision = \frac{TP}{FP+TP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1 - Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \quad (7)$$

2.9. The Best Algorithm

The best results in this study are obtained through a performance comparison of two classification algorithms, namely K-Nearest Neighbors (KNN) and decision tree (dt), which are tested on the same dataset using a quantitative evaluation approach. Following a proportional distribution of training and testing data, each algorithm's performance is rigorously assessed via accuracy, precision, recall, and F1-score. High results in these metrics serve as primary indicators of the model's efficacy, signifying its capacity for generating precise and consistent classifications [35].

3. RESULT

In this study, the dataset was obtained from a secondary data source on Kaggle titled 'Heart Disease,' consisting of 10,000 records and a total of 21 attributes. This dataset includes a combination of numerical and categorical attributes representing health conditions, lifestyle, and heart disease status.A comprehensive pre-processing stage was conducted to ensure data quality prior to the modeling process. Categorical attributes were converted into numerical values using label encoding techniques, where binary attributes such as gender were encoded as 0 and 1, while attributes with more than two categories were converted into a discrete numerical scale. This process aimed to preserve categorical information in a format compatible with classification algorithms.Furthermore, missing values were handled using the KNN Imputer method with five nearest neighbors (n_neighbors = 5). Following the imputation process, all attributes in the dataset were complete with no null values. The dataset was then normalized using the Min–Max scaling method, mapping all feature values to a range between 0 and 1. This normalization plays a crucial role in enhancing data stability and optimizing the performance of distance-based algorithms such as K-Nearest Neighbors (KNN), while maintaining consistent class separation in Decision Tree algorithms.

To address the class imbalance issue present in the initial dataset, an oversampling strategy was implemented on the minority class using the Random Oversampling technique. This process involved increasing the number of minority samples to 50% of the majority class. A parameter of random_state = 42 was applied to control the randomization process, ensuring that the oversampling results are consistent and reproducible. Figure 2 presents the class distribution results following the oversampling process.

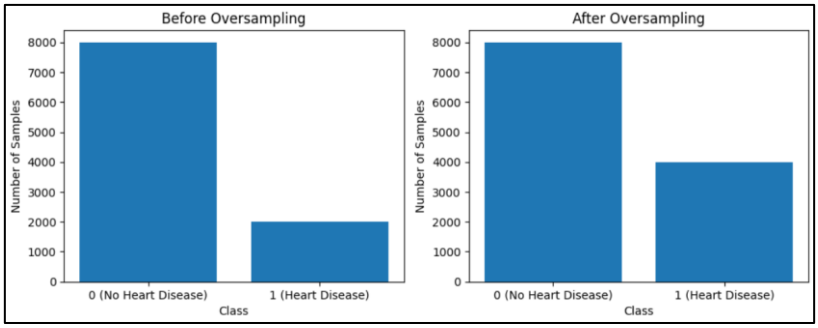


Figure 2. Result of Random Oversampling Processing

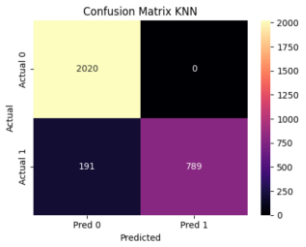
As illustrated in Figure 2, the implementation of the oversampling process resulted in a substantial increase in the number of samples within the minority class compared to the initial data distribution.

Vol. 7, No. 3, June 2026, Page. 2633-2645
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5626

This increase indicates that the class imbalance issue has been successfully mitigated, allowing the classification model to better learn patterns from the minority class and, consequently, enhancing the overall classification performance.

The balanced dataset was subsequently partitioned into training and testing sets with a ratio of 75:25. The results of this split show that the training data consists of 9,000 samples and the testing data comprises 3,000 samples, each with 20 attributes. This division ensures that the model performance evaluation can be conducted objectively and representatively.

The determination of optimal parameters for the KNN algorithm was conducted using the Elbow method, which identified an optimal K-value of 155 with a maximum accuracy of 0.9363. This value was selected because further increases in K beyond this point did not yield significant improvements in accuracy. The KNN model was tested under two scenarios: without oversampling and with the application of random oversampling. Under the condition without oversampling, the KNN model exhibited a severe bias toward the majority class, where all heart disease cases failed to be detected. Although the overall accuracy reached 80%, the recall and F1-score for the minority class were 0%, indicating that this accuracy is misleading and does not reflect the model's ability to detect heart disease. Conversely, after implementing random oversampling as shown in Figure 3, the performance of the KNN model improved significantly.

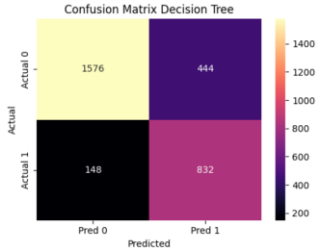



```
==== CLASSIFICATION REPORT ====
              precision    recall  f1-score   support

         0.0       0.91      1.00      0.95      2020
         1.0       1.00      0.81      0.89       980

    accuracy                           0.94      3000
   macro avg       0.96      0.90      0.92      3000
weighted avg       0.94      0.94      0.93      3000
```

Figure 3. Performance of the K-Nearest Neighbors model with random oversampling technique

Based on Figure 3, the model successfully classified 789 heart disease cases correctly, with an overall accuracy of 94%. The recall and F1-score for the heart disease class increased to 81% and 89%, respectively, indicating an improvement in the model's sensitivity toward the minority class. These results confirm that random oversampling is effective in reducing class bias and producing a more balanced and reliable classification performance for the KNN algorithm.

In the Decision Tree algorithm, the selection of the max_depth parameter shows that model performance improves as the tree depth increases until it reaches an optimal point at max_depth = 30. Increasing the tree depth beyond this value does not result in a significant performance improvement. In the evaluation without oversampling, the Decision Tree model achieved an accuracy of only 68% and demonstrated poor capability in detecting the heart disease class, with very low precision, recall, and F1-score for the minority class. This condition indicates that the model tends to favor the majority class in an imbalanced dataset. After applying random oversampling, the performance of the Decision Tree model improved substantially, as shown in Figure 4.




```
==== CLASSIFICATION REPORT DECISION TREE ====
              precision    recall  f1-score   support

         0.0       0.91      0.78      0.84      2020
         1.0       0.65      0.85      0.74       980

    accuracy                           0.80      3000
   macro avg       0.78      0.81      0.79      3000
weighted avg       0.83      0.80      0.81      3000
```

Figure 4. Performance of the Decision Tree model using Random Oversampling technique

Vol. 7, No. 3, June 2026, Page. 2633-2645
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5626

Based on the confusion matrix and classification report in Figure 12, the model achieved an accuracy of 80%, with a precision of 65%, recall of 85%, and an F1-score of 74% for the heart disease class. Although some misclassifications persist, these results indicate that random oversampling effectively enhances the model's sensitivity toward the minority class and balances the overall performance of the Decision Tree in heart disease risk classification. The final evaluation using the ROC-AUC curve in Figure 5 demonstrates that the KNN model exhibits superior performance compared to the Decision Tree model.

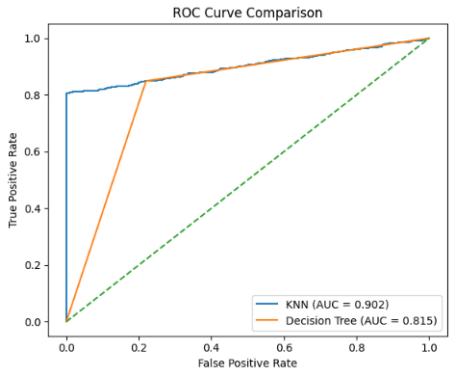


Figure 5. ROC Curce AUC Comparison

Based on Figure 5, the KNN model achieved an AUC value of 0.902, while the Decision Tree obtained an AUC of 0.815. The AUC value for KNN, which is closer to 1, indicates a very strong discriminative ability in distinguishing between patients with and without heart disease. Consequently, based on the overall experimental results, it can be concluded that the KNN algorithm, integrated with random oversampling, provides the most optimal and reliable classification performance for heart disease prediction on the utilized dataset.

To determine the best-performing algorithm, the performance evaluation of each model was conducted based on the classification metrics presented in Table 1. This comparison enables the identification of the algorithm most capable of accurately and balancedly distinguishing heart disease classes, particularly under imbalanced dataset conditions. The evaluation results serve as the basis for establishing the most optimal algorithm for heart disease prediction.

Table 1. The evaluation of each model

Oversampling	Accuracy	Class	Precision	Recall	F1-Score	ROC- AUC
K-Nearest Neighbors with random oversampling technique	94%	0	91%	100%	95%	90.2%
		1	100%	81%	89%	
		avg	96%	90%	92%	
Decision Tree with random oversampling technique	80%	0	91%	78%	84%	81.5%
		1	65%	85%	74%	
		avg	78%	81%	79%	

Based on Table 1, it can be concluded that the K-Nearest Neighbors (KNN) algorithm with the application of random oversampling is the best-performing model in this study. The model achieves the highest accuracy of 94%, with a precision of 100%, recall of 81%, and an F1-score of 89% for the heart disease class, as well as a macro-average F1-score of 92%. In addition, the ROC curve AUC value of 90.2% indicates a very strong discriminative ability in distinguishing between patients with and without heart disease. In contrast, although the Decision Tree algorithm with oversampling shows an improvement in performance compared to the model without oversampling, its overall performance remains lower than that of KNN, with an accuracy of 80% and an AUC value of 81.5%. Therefore, it

Vol. 7, No. 3, June 2026, Page. 2633-2645
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5626

can be concluded that KNN with random oversampling provides the most optimal, balanced, and reliable classification performance, and is thus recommended as the best algorithm for heart disease risk classification in this study.

4. DISCUSSIONS

Table 2. Comparison of Related Studies on Heart Disease Classification

Researcher	Dataset	Model	Performance Evaluation			
			Accuracy	Precision	Recall	F1-Score
Ref [4]	UCI – 294 Data	XGBoost	92%	92%	92%	92%
Ref [25]	CDC	C4.5	70%	70%	N/A	N/A
		Random Forest	87%	87%	N/A	N/A
		SVM	52%	56%	N/A	N/A
		Logistic Regression	73%	73%	N/A	N/A
		KNN	86%	87%	N/A	N/A
		Naïve Bayes	70%	70%	N/A	N/A
Ref [26]	CDC	XGBoost	94%	98%	99%	90%
	BRFSS	KNN	87%	80%	98%	88%
	2015 survey	SVM	79%	76%	84%	80%
Our	Kaggle –	KNN	94%	96%	90%	92%
	10.000 Data	Decision Tree	80%	78%	81%	79%

Based on the model performance comparison table 2, it is evident that the performance of heart disease classification algorithms is strongly influenced by the type of dataset, data size, and the class imbalance handling approach employed. In Ref. [4], which used a relatively small UCI dataset (294 records), the XGBoost model achieved an accuracy of 92% with balanced precision, recall, and F1-score values. These results highlight the strength of ensemble models in maximizing performance on small and well-controlled datasets; however, they do not sufficiently address model complexity and practical limitations for deployment in real-world clinical environments.

Furthermore [25], which utilized the CDC dataset, reported considerable performance variation among conventional algorithms such as C4.5, Random Forest, SVM, Logistic Regression, KNN, and Naïve Bayes. Random Forest and KNN achieved relatively higher accuracies (87% and 86%, respectively), whereas SVM exhibited the lowest accuracy at 52%. A major limitation of this study is the absence of recall and F1-score reporting, which prevents a thorough evaluation of the models' ability to detect the minority class (patients with heart disease). This omission represents a critical weakness, as failing to detect high-risk patients (false negatives) can have serious clinical consequences.

In [26], which employed the CDC BRFSS 2015 survey dataset, XGBoost again demonstrated superior performance, achieving an accuracy of 94% and a recall of 99%, followed by KNN with an accuracy of 87% and a very high recall of 98%. These findings indicate that KNN possesses strong sensitivity toward the minority class, although its accuracy remains slightly lower than that of XGBoost. Nevertheless, the use of ensemble models such as XGBoost introduces challenges related to interpretability and higher computational requirements.

In contrast to previous studies, the present research employs a larger and more representative Kaggle dataset consisting of 10,000 records, reflecting real-world medical data conditions. Experimental results show that after applying random oversampling, the KNN algorithm achieved an accuracy of 94%, precision of 96%, recall of 90%, and an F1-score of 92%. This improvement demonstrates that KNN benefits significantly from a more balanced data distribution, as the algorithm is distance-based and

Vol. 7, No. 3, June 2026, Page. 2633-2645
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5626

relatively robust to duplicated samples generated through oversampling. By increasing the representation of the minority class, KNN is able to form more balanced and less biased classification decisions.

Meanwhile, the Decision Tree algorithm in this study exhibited lower performance compared to KNN, with an accuracy of 80%. Nevertheless, the performance of Decision Tree remained relatively stable and did not experience a drastic decline after oversampling. This behavior can be attributed to the inherent characteristics of Decision Tree models, which rely on selecting the most informative attributes at each node; thus, the addition of oversampled data does not necessarily alter the tree structure significantly. This stability indicates that Decision Tree remains a reliable and highly interpretable model, although it is less optimal than KNN in exploiting the balanced data distribution produced by oversampling.

When compared with [25], which reported inconsistent or degraded performance after oversampling for certain algorithms, the findings of this study indicate that oversampling does not inherently have a negative impact. This difference suggests that the effectiveness of oversampling is highly dependent on both the algorithm characteristics and the data structure. In the case of KNN, oversampling improves minority class representation without substantially increasing noise, leading to positive gains in accuracy and, more importantly, recall.

From a clinical and health informatics perspective, these findings highlight that simple random oversampling is a practical solution for handling class imbalance in real-world medical datasets. The use of interpretable non-ensemble models such as KNN and Decision Tree enhances clinician trust while maintaining computational efficiency. This study demonstrates that improving minority-class recall—critical for heart disease prediction—can be achieved without complex models, supporting the development of lightweight and applicable clinical decision support systems in resource-limited healthcare settings.

5. CONCLUSION

This study concludes that the K-Nearest Neighbors (KNN) algorithm outperforms the Decision Tree model in heart disease classification when combined with random oversampling. KNN achieves superior performance, with an accuracy of 94%, precision of 96%, recall of 90%, F1-score of 92%, and a ROC–AUC of 90.2%, indicating strong discriminative capability. In contrast, the Decision Tree model attains lower results, with an accuracy of 80%, F1-score of 79%, and ROC–AUC of 81.5%. These findings confirm that class imbalance significantly impacts medical classification tasks and that simple random oversampling can effectively enhance minority class detection without complex ensemble models. From an Informatics perspective, this study demonstrates that lightweight and interpretable algorithms can deliver high predictive performance on large imbalanced medical datasets, supporting the development of accessible and efficient AI-based diagnostic systems.

ACKNOWLEDGEMENT

As the lead author, I would like to express my deepest gratitude to my supervisors, Mr. Farid Wajidi, S.Kom., M.T., and Mr. Wawan Firgiawan, S.T., M.Kom., for their invaluable mentorship and scholarly guidance throughout this study. Their constructive insights were fundamental to the successful preparation of this manuscript. I am also immensely grateful to the Faculty of Engineering and Universitas Sulawesi Barat for providing the institutional support and facilities required to conduct this research.

REFERENCES

[1] Kemenkes, “Satu dari Tiga Kematian Disebabkan oleh Jantung, Ayo Cegah serangan jantung,” Unit Pelayanan Kesehatan. [Online]. Available: https://upk.kemkes.go.id/new/satu-dari-tiga-

Vol. 7, No. 3, June 2026, Page. 2633-2645
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5626

kematian-disebabkan-oleh-jantung-ayo-cegah-serangan-jantung

[2] D. P. Utomo, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," *J. Media Inform. Budidarma*, vol. 4, no. April, pp. 437–444, 2020, doi: 10.30865/mib.v4i2.2080.

[3] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Comput. Biol. Med.*, vol. 136, no. May, p. 104672, 2021, doi: 10.1016/j.compbiomed.2021.104672.

[4] D. Rohmayani, C. A. Sugianto, R. S. Perdana, and M. Mansoor, "Improving Extreme Gradient Boosting Model for Heart Disease Prediction Using SMOTE for Class Imbalance," *J. Tek. Inform.*, vol. 6, no. 4, pp. 1717–1728, 2025, doi: https://doi.org/10.52436/1.jutif.2025.6.4.4753.

[5] T. S. Sriya, "Heart Disease Prediction Using KNN," *Int. J. Res. Eng. Sci. Manag.*, vol. 7, no. 6, pp. 156–157, 2024, doi: https://journal.ijresm.com/index.php/ijresm/article/view/3097.

[6] J. D. Muthohhar and A. Prihanto, "Analisis Perbandingan Algoritma Klasifikasi untuk Penyakit Jantung," *J. Informatics Comput. Sci.*, vol. 04, pp. 298–304, 2023, doi: 10.26740/jinacs.v4n03.p298-304.

[7] A. A. Surya and Y. Yamasari, "Penerapan Algoritma Naïve Bayes (NB) untuk Klasifikasi Penyakit Jantung," *J. Informatics Comput. Sci.*, vol. 5, no. 03, pp. 447–455, 2024, doi: 10.26740/jinacs.v5n03.p447-455.

[8] A. Sepharni, I. E. Hendrawan, and C. Rozikin, "Klasifikasi Penyakit Jantung dengan Menggunakan Algoritma C4.5," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.*, vol. 7, no. 2, p. 117, 2022, doi: 10.30998/string.v7i2.12012.

[9] A. Yogianto, A. Homaidi, and Z. Fatah, "Implementasi Metode K-Nearest Neighbors (KNN) untuk Klasifikasi Penyakit Jantung," *G-Tech J. Teknol. Terap.*, vol. 8, no. 3, pp. 1720–1728, 2024, doi: 10.33379/gtech.v8i3.4495.

[10] H. Hidayat, A. Sunyoto, and H. Al Fatta, "Klasifikasi Penyakit Jantung Menggunakan Random Forest Clasifier," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 7, no. 1, pp. 31–40, 2023, doi: 10.47970/siskom-kb.v7i1.464.

[11] S. R. Azizah, R. Herteno, A. Farmadi, D. Kartini, and I. Budiman, "Kombinasi Seleksi Fitur Berbasis Filter Dan Wrapper Combinations Of Feature Selection Based On Filter And Wrapper," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 6, pp. 1361–1368, 2023, doi: 10.25126/jtiik.2023107467.

[12] P. A. Jusia, A. Rahim, H. Yani, and J. Jasmir, "Improving Performance of KNN and C4.5 using Particle Swarm Optimization in Classification of Heart Disease," *J. Resti*, vol. 5, no. 158, pp. 1–6, 2026, doi: 10.29207/resti.v8i3.5710.

[13] E. F. Laili *et al.*, "Komparasi Algoritma Decision Tree Dan Support Vector Machine (Svm) Dalam," *J. Sist. Inf. dan Inform.*, vol. 8, no. 1, pp. 67–76, 2025, doi: https://doi.org/10.47080/simika.v8i1.3683.

[14] I. W. Gamadarenda, I. Waspada, U. Diponegoro, P. Korespondensi, S. Atribut, and A. Backward, "Implementasi Data Mining Untuk Deteksi Penyakit Ginjal Kronis (Pgk) Menggunakan K-Nearest Neighbor (Knn) Dengan Backward Data Mining Implementation For Detection Of Chronic Kidney (Ckd) Using K-Nearest Neighbor (Knn) With Backward Elimination," *J. Teknol. dan ilmu Komput.*, vol. 7, no. 2, pp. 417–426, 2020, doi: 10.25126/jtiik.202071896.

[15] A. C. Wibowo, S. A. Lestari, S. Informasi, F. I. Komputer, U. Duta, and B. Surakarta, "Analisis Penggunaan Machine Learning Dalam Klasifikasi Penentuan Penyakit Jantung," *J. Sist. Inf. DAN Tek. Komput.*, vol. 9, no. 2, pp. 9–13, 2024, doi: https://doi.org/10.51876/simtek.v9i2.395.

[16] C. Kaope and Y. Pristyanto, "The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance," *J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 22, no. 2, pp. 227–238, 2023, doi: 10.30812/matrik.v22i2.2515.

[17] A. N. Kasanah, Muladi, and U. Pujianto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam," *J. Rekayasa Sist. dan Teknol. Inf.*, vol. 1, no. 10, 2021, doi: 10.29207/resti.v3i2.945.

[18] R. Amelia, Indahwati, A. Fitrianto, and A. Rizki, "Komparasi Teknik Undersampling Dan Oversampling Pada Regresi," *J. TIMES (Technology Informatics Comput. Syst.*, vol. X, no. 2,

Vol. 7, No. 3, June 2026, Page. 2633-2645
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5626

pp. 1–11, 2024, doi: https://doi.org/10.51351/jtm.10.2.2021652.

[19] Z. Abidin, T. Suratno, and M. F. Putri, "Penerapan Random Oversampling Dan Principal Component Analysis Untuk Meningkatkan Akurasi Prediksi Kebangkrutan Application Of Random Oversampling And Principal Component Analysis To Enhance The Accuracy Of Bankruptcy Prediction For," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 12, no. 5, pp. 1209–1220, 2025, doi: https://doi.org/10.25126/jtiik.2025125.

[20] A. F. Riany and G. Testiana, "Penerapan Data Mining untuk Klasifikasi Penyakit Jantung Koroner Menggunakan Algoritma Naïve Bayes," *MDP Student Conf.*, vol. 2, no. 1, pp. 297–305, 2023, doi: 10.35957/mdp-sc.v2i1.4388.

[21] M. A. Fais *et al.*, "Implementasi Algoritma Decision Tree untuk Klasifikasi Serangan Jantung," *J. Sist. Inf. dan Ilmu Komput.*, vol. 1, no. 4, pp. 207–212, 2023, doi: https://doi.org/10.59581/jusiik-widyakarya.v1i4.1895.

[22] A. S. Arifianto, K. D. Safitri, K. Agustianto, and I. G. Wiryawan, "Pengaruh Prediksi Missing Value Pada The Effect Of Missing Value Prediction On," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 4, pp. 779–786, 2022, doi: 10.25126/jtiik.202294778.

[23] Ainurrohmah and D. T. Wijayanti, "Analisis Performa Algoritma Decision Tree , Naïve Bayes , K- Nearest Neighbor Untuk Klasifikasi Zona Daerah Risiko Covid-19 Di Indonesia Performance Analysis Of Decision Tree , Naïve Bayes , K-Nearest Neighbor Algorithm For Covid-19 Risk Zone Classificati," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 1, pp. 115–122, 2023, doi: 10.25126/jtiik.2023105935.

[24] P. V. Bhamare, S. R. Chikhale, N. S. Sawakare, A. Y. Kurkunde, and M. S. Autade, "Heart Disease Prediction Using Machine Learning Algorithms," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. April, pp. 559–564, 2024, doi: https://doi.org/10.22214/ijraset.2022.44895.

[25] A. F. Masruriyah, H. Y. Novita, C. E. Sukmawati, A. R. Ramadhan, S. N. N. Arif, and B. A. Dermawan, "Pengukuran Kinerja Model Klasifikasi dengan Data Oversampling pada Algoritma Supervised Learning untuk Penyakit Jantung," *Comput. Sci.*, vol. 4, no. 1, pp. 62–70, 2024, doi: 10.31294/coscience.v4i1.2389.

[26] Yuliana, Robet, and L. Hoki, "Comparative Analysis of XGBoost , KNN , and SVM Algorithms for Heart Disease Prediction Using SMOTE-Tomek Balancing," vol. 10, no. 1, pp. 305–316, 2026, doi: https://doi.org/10.33395/sinkron.v10i1.15469 e-ISSN.

[27] I. N. Abrar, A. Abdullah, and Sucipto, "Klasifikasi Penyakit Liver Menggunakan Metode Elbow Untuk Menentukan K Optimal pada Algoritma K-Nearest Neighbor (K-NN)," *J. SISFOKOM (Sistem Inf. dan Komputer)*, vol. 12, no. 1, pp. 218–228, 2023, doi: 10.32736/sisfokom.v12i2.1643.

[28] B. Aribowo, B. Tjahjono, G. Firmasnyah, and A. M. Widodo, "Prediksi Peringkat Akreditasi BAN PT Program Studi Sarjana Rumpun Ilmu Komputer Menggunakan Klasifikasi Machine Learning," *J. Al-azhar Indones. Seri Sains dan Teknol.*, vol. 10, no. 2, pp. 122–127, 2025, doi: http://dx.doi.org/10.36722/sst.v10i2.3089.

[29] S. Sudianto, J. A. Masheli, N. Nugroho, R. R. W. Ananda, and A. Zarkasih, "Comparison of Support Vector Machines and K-Nearest Neighbor Algorithm Analysis of Spam Comments on Youtube Covid Omicron," *J. Tek. Inform.*, vol. 15, no. 2, pp. 110–118, 2022, doi: https://doi.org/10.15408/jti.v15i2.24996.

[30] J. Yos, S. No, K. Lubuklinggau, and S. Selatan, "Perbandingan Tingkat Akurasi Metode KNN Dan Decision Tree Dalam Memprediksi Lama Studi Mahasiswa," vol. 03, no. 97, pp. 6–14, 2021.

[31] C. A. Sinaga and A. K. Ginting, "Implementasi Algoritma K-Nearest Neighbors Dengan Pendekatan Elbow Method Untuk Klasifikasi Status Ketahanan Pangan Provinsi Di Indonesia," *Kumpul. Artik. Ilm. Fak. Ilmu Komput.*, vol. 07, no. 01, pp. 27–34, 2025, doi: https://doi.org/10.54367/kakifikom.v7i1.4949.

[32] I. Maulana and R. Roestam, "Optimizing KNN Algorithm Using Elbow Method Predicting Voter Participation Using Fixed Voter List Data (DPT)," *J. Sos. dan Teknol.*, vol. 4, pp. 441–451, 2024, doi: 10.59188/jurnalsostech.v4i7.1308.

[33] Z. Cetinkaya and F. Horasan, "Decision Trees in Large Data Sets," *J. Eng. Res. Dev.*, vol. 13, pp. 140–151, 2021, doi: 10.29137/umagd.763490.

[34] N. A. Sivi, I. Mualim, and C. A. Lestari, "Data Mining Menggunakan Decision Tree untuk

Vol. 7, No. 3, June 2026, Page. 2633-2645
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5626

Prediksi Nilai Akhir Siswa," *J. Ilm. Tek. Inform. dan Komun.*, vol. 4, no. November, pp. 26–36, 2024, doi: https://doi.org/10.55606/juitik.v4i3.1824.

[35] L. Hakim, A. Sobri, L. Sunardi, and D. Nurdiansyah, "Prediksi penyakit jantung berbasis mesin learning dengan menggunakan metode k-nn," *J. Digit. Teknol. Inf.*, vol. 07, no. 02, pp. 14–20, 2025, doi: https://doi.org/10.32502/digital.v7i2.9429.