Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1753−1760
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8471

Revisiting SMOTE in Balanced Medical Data: A Comparative Evaluation of SVM, Random Forest, and KNN

Kelvin Leonardi Kohsasih[1,*], Joni[1], Herman[2], Pieter Octaviandy[2]

[1] Teknik Informatika, STMIK TIME, Medan, Indonesia
[2] Sistem Infromasi, STMIK TIME, Medan, Indonesia
Email: [1,*]kelvinleonardi@stmik-time-ac.id, [2]joni.hgw@gmail.com, [3]hrman_ang@yahoo.com, [4]pieter.lecture@gmail.com
Email Penulis Korespondensi: kelvinleonardi@stmik-time-ac.id
Submitted: **06/10/2025**; Accepted: **16/12/2025**; Published**: 16/12/2025**

Abstract−Heart disease is one of the leading causes of death worldwide, making data-driven early detection crucial for supporting medical decision-making systems. A major challenge in developing heart disease prediction models is dataset quality, including the often imbalanced class distribution, which can impact the performance of classification algorithms. This study aims to analyze the effect of the Synthetic Minority Oversampling Technique (SMOTE) on the performance of three classification algorithms: Support Vector Classifier (SVC), Random Forest (RF), and K-Nearest Neighbor (KNN). The dataset used is heart_disease50.csv with 4,001 patient data consisting of 21 predictor attributes and one target variable (heart disease status: "Yes" or "No") with a relatively balanced class distribution. The research process includes data preprocessing (data cleaning, normalization, and encoding), data partitioning using Stratified K-Fold Cross Validation (k=5), applying SMOTE to training data, building a classification model, and evaluation using accuracy, precision, recall, F1-score, and AUC-ROC metrics. The results showed that applying SMOTE did not always improve performance. The SVC model with SMOTE experienced a decrease in accuracy (0.4819) compared to the one without SMOTE (0.5106), while Random Forest remained relatively stable with insignificant differences (0.4669 without SMOTE and 0.4644 with SMOTE). KNN with SMOTE emerged as the best model with an accuracy of 0.5268 and a precision of 0.5271, although the AUC-ROC remained the same as KNN without SMOTE (0.5135). Overall, these results confirm that the effectiveness of SMOTE is highly dependent on dataset conditions, and in cases with relatively balanced data, SMOTE does not provide significant benefits. Therefore, improving the performance of heart disease prediction classification is recommended through hyperparameter optimization strategies, relevant feature selection, or the use of more sophisticated algorithms such as Gradient Boosting or Neural Networks.

Keywords: Heart Disease; Classification; SMOTE; Support Vector Machine; Random Forest; K-Nearest Neighbor

1. INTRODUCTION

Cardiovascular disease remains one of the leading causes of mortality worldwide. According to the World Health Organization (WHO), more than 17 million individuals die annually from cardiovascular diseases, with coronary heart disease accounting for the largest proportion [1], [2], [3].. Early detection of heart disease is a crucial factor in both prevention and treatment efforts, making the development of data-driven and artificial intelligence (AI)-based prediction methods increasingly relevant.

In the era of big data and machine learning, the ability to classify patients at risk of heart disease provides strategic value for both the medical community and society at large [4], [5], [6]. One of the major challenges in developing predictive models for heart disease lies in the issue of imbalanced datasets. This condition occurs when the number of patients with heart disease (minority class) is significantly lower than the number of healthy patients (majority class) [7], [8], [9]. Imbalanced datasets often lead to prediction models that are biased toward the majority class, resulting in deceptively high accuracy while producing poor performance in detecting the minority class (low recall/sensitivity). In the medical context, failing to correctly identify patients with heart disease (false negatives) can be fatal. Therefore, specific strategies are required to address data imbalance in order to develop predictive models that are both reliable and fair across classes [10], [11].

In the case of heart disease prediction, a variety of classification models have been utilized to develop medical decision support systems. Among the most widely applied models are Support Vector Machine (SVM) [12], [13], [14], which excels in finding optimal separating hyperplanes in high-dimensional data; Random Forest (RF) [15], [16], which combines multiple decision trees to produce stable predictions and effectively handle complex data; and K-Nearest Neighbor (KNN) [17], [18], [19], a simple similarity-based algorithm that classifies instances according to their nearest neighbors. Beyond these three primary algorithms, alternative approaches such as Logistic Regression, which is commonly applied in medical analysis due to its interpretability, and Neural Networks, which can capture complex patterns albeit requiring larger datasets and higher computational resources, are also available. The choice of classification model in this study largely depends on dataset characteristics, required accuracy, and interpretability of results, thereby necessitating comparative performance analysis to identify the most suitable approach [20], [21].

Previous research discussed the implementation of the RS-SMOTE-GBT method, a combination of Random Search (RS), Synthetic Minority Oversampling Technique (SMOTE), and Gradient Boosting Tree (GBT), for classifying rock discontinuity traces from three-dimensional point cloud data [22]. This approach successfully addressed the data imbalance problem (trace vs. non-trace), achieved the highest F-score among 16 benchmark models, and demonstrated superior generalization ability, particularly when training data were diverse. Additionally, the study showed that discontinuity features played a more critical role than basic features, thereby improving classification reliability. However, limitations were identified, including persistent misclassifications in fractured and

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1753−1760
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8471

irregularly layered rocks, reliance on the quality and accuracy of expert-labeled data, and increased computational cost during parameter optimization. Overall, the method offered an innovative solution for rock trace mapping, although its effectiveness remained constrained by data quality and geological complexity.

Hairani (2023) focused on improving the performance of Random Forest in classifying imbalanced diabetes data through the Smote-Tomeklink technique [23]. This method successfully mitigated the dominance of the majority class (negative diabetes), thereby enhancing the detection of the minority class (positive diabetes). Results demonstrated improved performance, with accuracy reaching 86.4%, sensitivity 88.2%, precision 83.3%, and F1-score 85.1%, surpassing both the standard Random Forest and the SMOTE-only approach. Furthermore, the combination of oversampling and undersampling not only reduced data noise but also minimized overfitting, yielding more stable classification results. Nevertheless, its drawbacks included increased process complexity, as the integration of SMOTE and Tomeklink required additional time and computational resources, and limited validation since it had not yet been tested on multiclass or larger, more heterogeneous medical datasets. Overall, this approach demonstrated clear improvements in early diabetes diagnosis, yet further research is needed for broader applicability.

Among various approaches, the Synthetic Minority Oversampling Technique (SMOTE) has been widely adopted. Unlike simple duplication, SMOTE generates synthetic minority samples through interpolation between neighboring minority instances [7], [24], [25]. This method has proven effective in many studies, improving classification performance on imbalanced datasets across domains such as healthcare, finance, and anomaly detection. However, SMOTE must be applied carefully, as improper use may introduce noise or generate patterns unrepresentative of the underlying data [11], [26], [27].

In addition to data balancing techniques, the selection of classification algorithms plays a key role in developing optimal prediction models. Popular algorithms in medical data classification include Support Vector Machine (SVM) [[28], [29], Random Forest (RF) [30], [31], [32], and K-Nearest Neighbor (KNN) [33], [34], [35]. SVM operates by identifying the optimal hyperplane that maximizes the margin between classes, making it suitable for high-dimensional data, although its performance is sensitive to kernel parameters and regularization.

RF, an ensemble learning method that integrates multiple decision trees, is recognized for its robustness, ability to handle heterogeneous data, and feature importance estimation, which makes it widely used in medical applications. Conversely, KNN, an instance-based learning algorithm, classifies data based on majority voting among nearest neighbors. While conceptually simple, its performance is heavily influenced by the choice of k and preprocessing requirements such as normalization.

Several studies have demonstrated that combining oversampling techniques (such as SMOTE) with classification algorithms enhances model performance. For instance, research in cancer prediction, diabetes detection, and heart failure classification reported improvements in recall and F1-score when SMOTE was applied [36], [37]. Nonetheless, outcomes often depended on dataset characteristics, chosen algorithms, and parameter configurations. This suggests that no universal conclusion exists regarding the most optimal algorithm when combined with SMOTE, particularly for heart disease datasets [36], [37], [38].

Based on these considerations, the present study aims to analyze and compare the performance of three classification algorithms—SVM, Random Forest, and KNN—with and without the application of SMOTE, using a heart disease dataset. The focus is to assess the extent to which SMOTE addresses class imbalance (if present) and its impact on key evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. The findings are expected to contribute to the development of machine learning-based medical decision support systems, particularly for early detection of heart disease, while also enriching the literature on the use of SMOTE in healthcare domains.

2. RESEARCH METHODOLOGY

This research is a quantitative study with an experimental approach. The focus of the study is to evaluate the effect of using the Synthetic Minority Oversampling Technique (SMOTE) data balancing technique on the performance of three classification algorithms: Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbor (KNN) on a heart disease dataset. The goal is to obtain an optimal prediction model and compare the performance of the algorithms with and without SMOTE.

2.1 Research Data

Contains an explanation of the stages of research (MANDATORY IN THE ARTICLE) which describes the sequence/stages in conducting research, how the stages of applying the method in research and testing methods in obtaining research results are in accordance with the expectations and description of the research. It is better if there are pictures and tables, they must be presented with the names of the tables and figures accompanied by serial numbers. The dataset used is heart_disease50.csv, which consists of 4001 rows of data with 21 attributes (independent variables) and one target variable, Heart Disease Status, with two classes: "Yes" (positive for heart disease) and "No" (negative for heart disease). The class distribution in this dataset is relatively balanced, with 2000 positive patient data and 2001 negative patient data. The variables in the dataset include demographic information, medical history, medical examination results, and heart disease risk factors. The dataset can be seen in Table 1 below.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1753−1760
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8471

Table 1. Sample Research Dataset

Age	Gender	Blood Pressure	Cholesterol Level	Exercise Habits	Smoking	…. Etc	Heart Disease Status
56	Male	153	155	High	Yes		No
69	Female	146	286	High	No		No
46	Male	126	216	Low	No		No
32	Female	122	293	High	Yes		No
60	Male	166	242	Low	Yes		No
…	…	…	…	…	…	…	…etc
25	Female	136	243	Medium	Yes		Yes
38	Male	172	154	Medium	No		Yes
73	Male	152	201	High	Yes		Yes
23	Male	142	299	Low	Yes		Yes
38	Female	128	193	Medium	Yes		Yes

Before use, the dataset underwent a data cleaning process, including checking for missing values, normalizing numeric variables, and encoding categorical variables using one-hot encoding to facilitate processing by the classification algorithm. Available attributes included demographic factors such as age and gender, clinical variables such as resting blood pressure, cholesterol levels, and maximum heart rate, as well as medical examination results such as electrocardiogram recordings, ST-segment depression, chest pain type, and history of angina during exercise. Some attributes were numeric with varying value ranges and therefore required normalization, while categorical attributes such as gender, chest pain type, and ST-segment slope required encoding to facilitate processing by the classification algorithm. The initial dataset showed no missing values, making it relatively clean and ready for processing. This dataset was selected based on its relatively large size compared to classic cardiac datasets, the availability of medically relevant clinical attributes, and its balanced class distribution. This allowed this study not only to experiment with data balancing techniques like SMOTE but also to produce more representative and interpretable results in the context of a medical decision support system.

2.2 Research Design

In order to ensure methodological rigor and reproducibility, this study adopted a systematic research workflow for developing and evaluating classification models on the heart disease dataset. The workflow integrates data preprocessing, sampling strategies, model development, and evaluation, thereby providing a structured approach to assess the impact of SMOTE on balanced data. The research design also incorporates cross-validation and hyperparameter tuning to minimize bias and optimize performance. The overall process is illustrated in Figure 1.

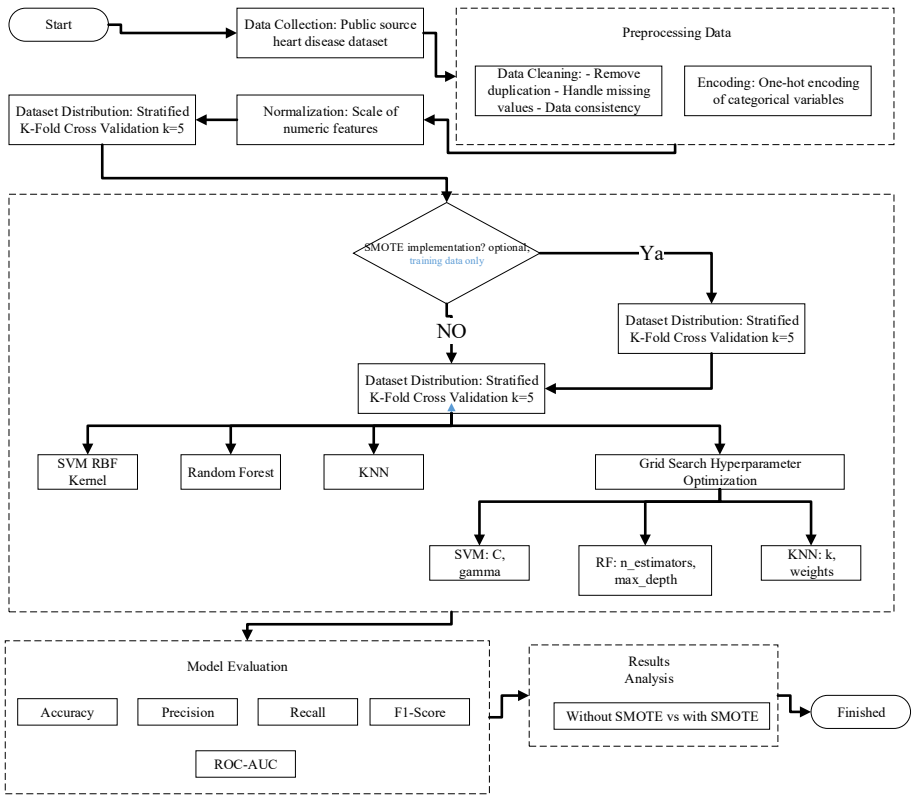


Figure 1. Research Design

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1753−1760
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8471

Figure 1 illustrates the research workflow adopted in this study. The process begins with the collection of the publicly available heart_disease50.csv dataset, which then undergoes a series of preprocessing steps. These steps include data cleaning to remove duplicate records, handle missing values, and ensure consistency, normalization to scale numerical attributes into uniform ranges and prevent bias in distance-based algorithms, and categorical encoding using one-hot encoding so that non-numeric variables can be properly processed by machine learning models. Once preprocessing is complete, the dataset is partitioned using Stratified K-Fold Cross Validation (k=5) to maintain class balance across training and testing folds. At this stage, a branching decision is made regarding whether to apply the Synthetic Minority Oversampling Technique (SMOTE). If selected, SMOTE is applied exclusively to the training data to avoid potential data leakage into the testing set.

Model development is then conducted using three well-established algorithms: Support Vector Machine (SVM) with an RBF kernel, Random Forest (RF), and K-Nearest Neighbor (KNN). In the SMOTE-applied scenario, each model undergoes hyperparameter optimization through Grid Search, where SVM parameters such as C and gamma, Random Forest parameters such as the number of estimators and maximum tree depth, and KNN parameters such as the number of neighbors (k) and weighting schemes are systematically adjusted. The resulting models are subsequently evaluated using five performance metrics: accuracy, precision, recall, F1-score, and ROC-AUC. This evaluation provides a comprehensive measure of both predictive accuracy and the ability to distinguish between positive and negative classes. Finally, the results are analyzed by comparing models trained with SMOTE against those trained without it, enabling this study to assess the effectiveness of oversampling in the context of a nearly balanced medical dataset..

2.3 Comparison of Six Classification Models

This study compared six classification models obtained from a combination of three algorithms: Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbor (KNN). Each model was tested under two scenarios: without and with SMOTE. The comparison was conducted to determine the extent to which the SMOTE technique affects the performance of the classification algorithms on the heart disease dataset used. The evaluation was conducted using five main metrics: accuracy, precision, recall, F1-score, and ROC-AUC.

In general, the experimental results show that applying SMOTE does not always improve performance on this dataset, given that the actual class distribution is relatively balanced (50:50). In the SVM model, applying SMOTE actually caused a decrease in accuracy and AUC compared to the model without SMOTE, indicating that synthetic oversampling does not provide a significant benefit. In the Random Forest model, both with and without SMOTE, performance was relatively stable with minor differences, indicating that this ensemble-based algorithm is quite robust to class distribution. Meanwhile, in the KNN model, the experimental results showed that using SMOTE did not provide a significant difference, with accuracy and AUC values being relatively similar in both scenarios.

Based on the comparison of the six models, it can be concluded that the effectiveness of SMOTE is significantly influenced by the class distribution in the dataset. In the case of this relatively balanced heart disease dataset, SMOTE did not provide a significant advantage, and in some algorithms, it even tended to decrease performance. Nevertheless, these results are still important to emphasize that the use of oversampling techniques should consider the characteristics of the dataset, rather than simply being applied automatically. In other words, the selection of algorithms and data handling strategies must be done contextually according to the needs and conditions of the dataset being analyzed.

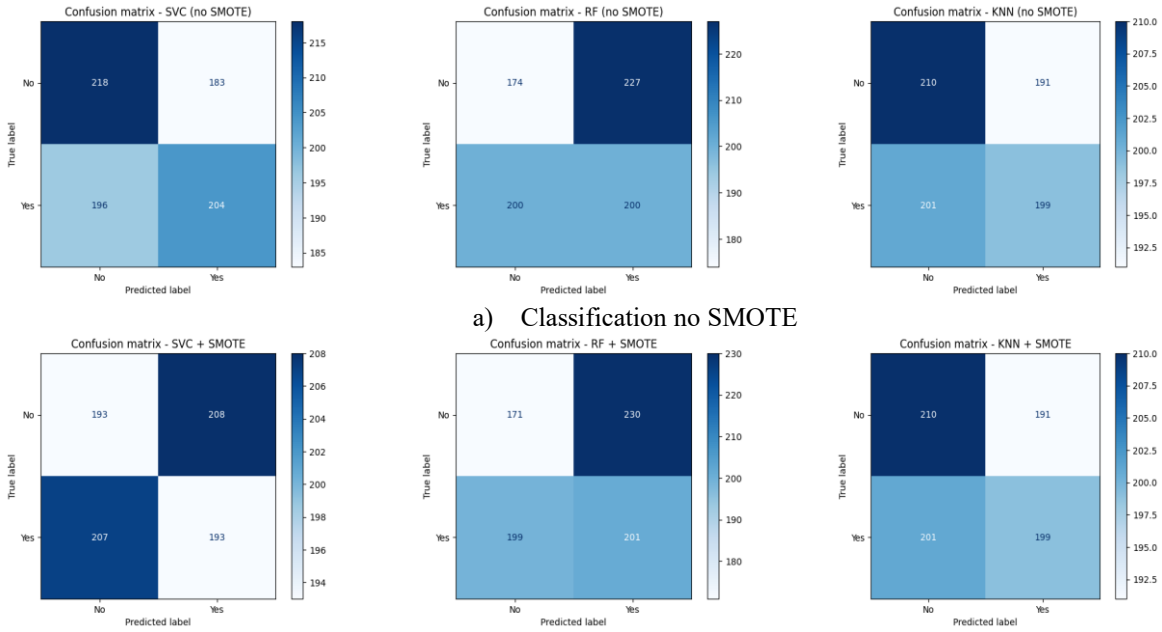
3. RESULT AND DISCUSSION

This section presents the results of the experiments conducted, along with an in-depth analysis of the performance of each classification model. The six models tested were combinations of three algorithms: Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbor (KNN). Each was tested in two scenarios: without and with SMOTE. The results of this study are not only presented in the form of evaluation figures such as accuracy and ROC-AUC, but also analyzed to understand the pattern of performance differences between the models. Through this discussion, it is hoped that a comprehensive overview of the effectiveness of the SMOTE technique on a relatively balanced heart disease dataset will be obtained, as well as its implications for selecting the most appropriate classification algorithm to support medical prediction systems.

3.1 Experimental Results

This study tested six classification models: combinations of three popular algorithms: Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbor (KNN), in two scenarios: without and with SMOTE. The evaluation was conducted using the Accuracy and ROC-AUC metrics, which represent the model's overall ability to predict and distinguish between positive and negative classes. This study compares six classification models: SVC, Random Forest (RF), and K-Nearest Neighbor (KNN) in two scenarios: without SMOTE and with SMOTE. The evaluation was conducted using five key performance metrics: accuracy, precision, recall, F1-score, and AUC-ROC.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1753−1760
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8471

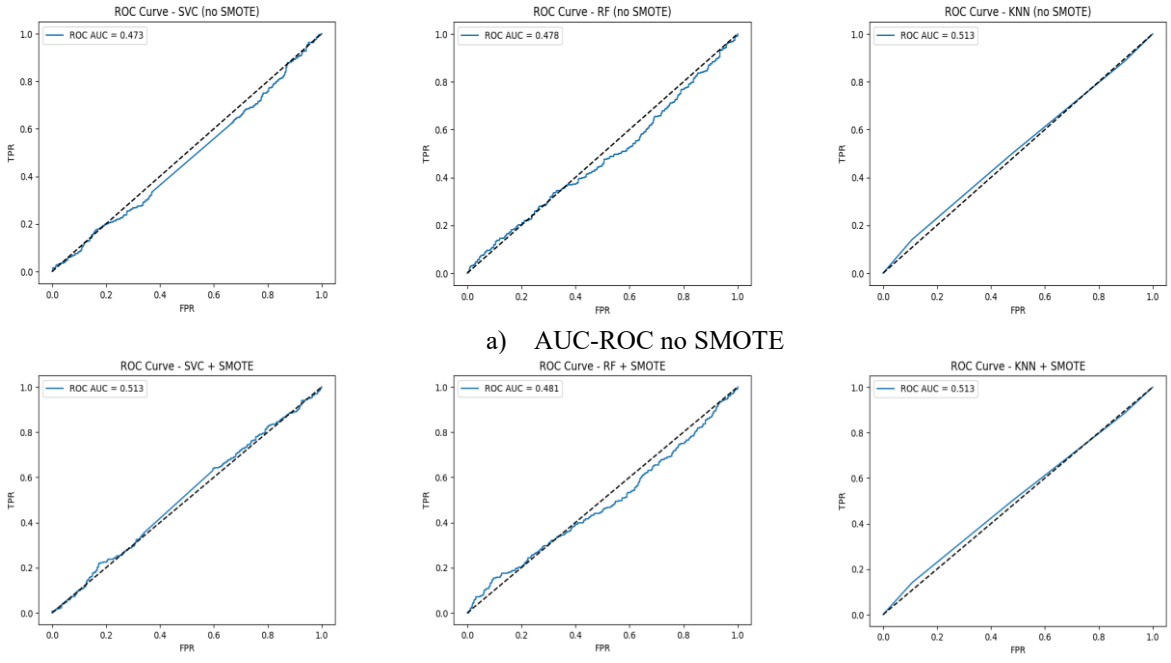


a) Classification no SMOTE

b) Classification + SMOTE

Figure 2. Comparison of Confusion Matrices for Classification without SMOTE and Classification with SMOTE

Based on Figure 2, the confusion matrix evaluation results show that all six classification models produced relatively balanced predictions between positive and negative classes, but still had a relatively high rate of misclassification. In the SVC and KNN models without SMOTE, the recall value approached 0.50, meaning only about half of patients with heart disease were correctly detected, while the rest were incorrectly classified as healthy (false negatives). The Random Forest model, both with and without SMOTE, showed a similar trend, with slightly higher recall (around 0.50) but lower precision, resulting in a high number of false positives. Meanwhile, the KNN model with SMOTE performed relatively better with higher accuracy and precision, indicating a more balanced distribution of predictions between true positives and true negatives. Overall, the confusion matrix illustrates that although all models performed basic classification, the prediction error rate was still quite high, especially for positive cases of heart disease. Therefore, further optimization strategies are needed to reduce false negatives, which are critical in medical contexts. The AUC-ROC results can be seen in Figure 3 below.



a) AUC-ROC no SMOTE

b) AUC-ROC + SMOTE

Figure 3. Comparison of AUC-ROC for Classification without SMOTE and Classification with SMOTE

Figure 3 displays the Receiver Operating Characteristic (ROC) curves and AUC (Area Under Curve) values for six classification models, each in two scenarios: without SMOTE (a) and with SMOTE (b). In general, all ROC

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1753−1760
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8471

curves tend to approach the diagonal line, indicating that the class separation ability of all models is still limited. In the scenario without SMOTE, the AUC value for SVC was 0.473, RF was 0.478, and KNN reached 0.513, the highest among the three. After applying SMOTE, the AUC value for SVC increased to 0.513, RF slightly increased to 0.481, while KNN remained at 0.513. This indicates that SMOTE did not significantly improve model performance, except for slight improvements in SVC and RF. Overall, this graph illustrates that both with and without SMOTE, the model's performance in distinguishing patients with heart disease from healthy patients remains close to random predictions, requiring further optimization to achieve more reliable classification results.

3.2 Results Analysis

3.2.1 Support Vector Classifier (SVC)

The SVC without SMOTE produced an accuracy of 0.5106, with relatively balanced precision and recall (0.5103 and 0.4975). The F1-score of 0.5184 indicates a harmonious combination of precision and recall, although the AUC-ROC is low (0.4730), indicating limited class separation ability. When SMOTE is applied, the accuracy decreases to 0.4819, and although recall increases slightly (0.4825), the F1-score actually drops to 0.4819. The AUC-ROC increases slightly (0.5130), but overall the SVC performance with SMOTE is not superior to that without SMOTE. This indicates that adding synthetic data actually decreases the SVC classification accuracy.

3.2.2 Random Forest (RF)

RF without SMOTE recorded an accuracy of 0.4669 with a recall of 0.5000, indicating the model was relatively capable of recognizing the positive class, but had low precision (0.4684). The F1-score of 0.4837 supports this finding, and the AUC-ROC of 0.4782 indicates suboptimal class separation. After applying SMOTE, accuracy decreased slightly to 0.4644, but recall increased slightly to 0.5025. Despite the increase in recall, overall performance remained low, with an AUC-ROC of 0.4814, which was not significantly different from the scenario without SMOTE. This indicates that RF is relatively stable to the addition of synthetic data, but remains suboptimal for this dataset.

3.2.3 K-Nearest Neighbor (KNN)

KNN without SMOTE provided an accuracy of 0.5106, a precision of 0.5103, a recall of 0.4975, and an F1-score of 0.5038. The AUC-ROC value of 0.5135 was the highest among the models without SMOTE, indicating relatively better class separation ability. When SMOTE was applied, KNN's performance actually improved, reaching an accuracy of 0.5268 and a precision of 0.5271, while the recall increased slightly to 0.5100. The AUC-ROC value remained stable at 0.5135, but the increase in accuracy indicates that KNN is more responsive to the addition of synthetic samples than SVC and RF.

3.3 Discussion

To evaluate the comparative performance of the classification algorithms, five standard metrics were used: accuracy, precision, recall, F1-score, and AUC-ROC. These metrics provide a comprehensive assessment of both predictive accuracy and the discriminative power of the models. The results of the experiments for Support Vector Classifier (SVC), Random Forest (RF), and K-Nearest Neighbor (KNN), with and without the application of SMOTE, are presented in Table 2.

Table 2. Classification Model Evaluation Results

Model	Akurasi	Presisi	Recall	F1 Score	AUC-ROC
SVC	0,51060	0,51030	0,49750	0,51840	0,47300
RF	0,46690	0,46840	0,50000	0,48370	0,47820
KNN	0,51060	0,51030	0,49750	0,50380	0,51350
SVC + SMOTE	0,48190	0,48130	0,48250	0,48190	0,51300
RF + SMOTE	0,46440	0,46640	0,50250	0,48380	0,48140
KNN + SMOTE	0,52680	0,52710	0,51000	0,50380	0,51350

Table 2 highlights that the impact of SMOTE on model performance is not consistent across algorithms. Without SMOTE, both SVC and KNN produced similar accuracy values of 0.5106, while RF recorded a lower accuracy of 0.4669. Precision and recall values followed a comparable pattern, with KNN achieving the highest AUC-ROC score of 0.5135, indicating slightly stronger discriminative capability compared to SVC (0.4730) and RF (0.4782). When SMOTE was applied, the results varied. For SVC, performance decreased, with accuracy dropping from 0.5106 to 0.4819 and F1-score also declining. Random Forest remained relatively stable across conditions, showing only minor fluctuations with an accuracy of 0.4644 and AUC-ROC of 0.4814. In contrast, KNN demonstrated a modest improvement, with accuracy increasing to 0.5268 and precision to 0.5271. However, its F1-score and AUC-ROC values showed little to no change, suggesting that the improvements in accuracy and precision did not translate into a stronger discriminative ability. These findings provide an important observation: SMOTE does not consistently enhance classification performance when applied to balanced medical data. In fact, it may even reduce performance,

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1753−1760
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8471

as observed with SVC. This result underscores the importance of considering dataset characteristics before applying oversampling techniques, as their effectiveness is not universal.

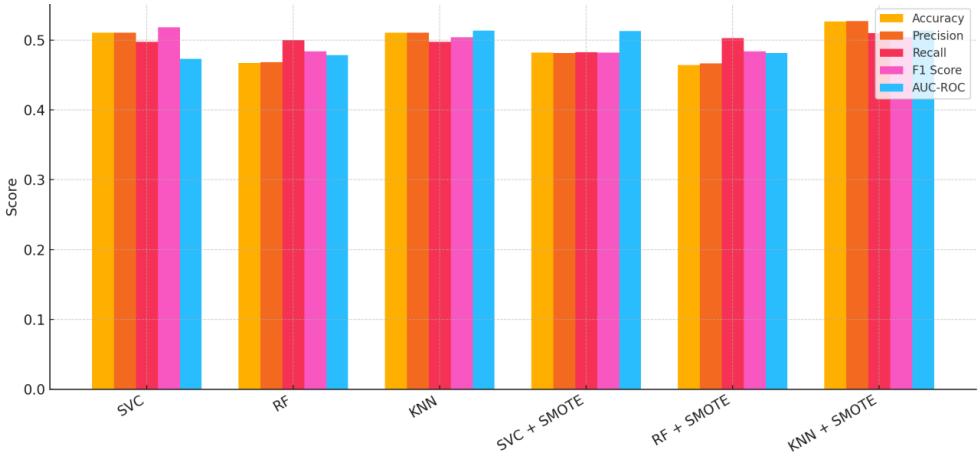


Figure 4. Comparison of Model Performance with and without SMOTE

Figure 4 provides a visual comparison that reinforces the tabular results. It clearly shows that applying SMOTE did not universally enhance performance. Instead, SVC suffered from decreased accuracy and F1-score, RF maintained stable but suboptimal results, while KNN benefited modestly in accuracy and precision but without significant improvement in discriminative ability (AUC-ROC). These results emphasize that SMOTE is not a one-size-fits-all solution and its application should be tailored to dataset characteristics.

4. CONCLUSION

This study revisited the role of SMOTE in the context of a nearly balanced heart disease dataset by comparing three popular classification algorithms: SVM, Random Forest, and KNN. The results demonstrated that SMOTE did not consistently improve model performance. In particular, SVM suffered a decrease in accuracy and F1-score after oversampling, while Random Forest remained relatively stable but still suboptimal. Only KNN showed a modest gain in accuracy and precision, although its AUC-ROC value did not improve. These findings highlight an important insight: SMOTE, while effective for imbalanced learning, may not provide tangible benefits and can even reduce performance when applied to balanced datasets. This contribution challenges the common assumption in machine learning that oversampling is universally advantageous, particularly in medical data classification. For future research, performance improvements are better pursued through hyperparameter optimization, feature selection, and the application of more advanced methods such as Gradient Boosting or Neural Networks. Furthermore, exploring alternative strategies such as cost-sensitive learning or hybrid data augmentation approaches may provide more robust solutions for medical prediction tasks..

REFERENCES

[1] M. Said, Y. Omar, S. Safwat, and A. Salem, “Explainable Artificial Intelligence Powered Model for Explainable Detection of Stroke Disease BT - Proceedings of the 8th International Conference on Advanced Intelligent Systems and Informatics 2022,” A. E. Hassanien, V. Snášel, M. Tang, T.-W. Sung, and K.-C. Chang, Eds., Cham: Springer International Publishing, 2023, pp. 211–223.

[2] P. R. Kumar, “Identification of noteworthy features and data mining techniques for heart disease prediction,” *Int. J. Model. Simulation, Sci. Comput.*, 2024, doi: 10.1142/S1793962325500102.

[3] N. A. M. Zaini and M. K. Awang, “Performance Comparison between Meta-classifier Algorithms for Heart Disease Classification,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 10, pp. 323–328, 2022, doi: 10.14569/IJACSA.2022.0131039.

[4] S. Sowmya and D. Jose, “Contemplate on ECG signals and classification of arrhythmia signals using CNN-LSTM deep learning model,” *Meas. Sensors*, vol. 24, no. October, p. 100558, 2022, doi: 10.1016/j.measen.2022.100558.

[5] P. Alkhairi, A. P. Windarto, and M. M. Efendi, “Optimasi LSTM Mengurangi Overfitting untuk Klasifikasi Teks Menggunakan Kumpulan Data Ulasan Film Kaggle IMDB,” vol. 6, no. 2, pp. 1142–1150, 2024, doi: 10.47065/bits.v6i2.5850.

[6] J. Lee *et al.*, “Unsupervised machine learning for identifying important visual features through bag-of-words using histopathology data from chronic kidney disease,” *Sci. Rep.*, vol. 12, no. 1, pp. 1–13, 2022, doi: 10.1038/s41598-022-08974-8.

[7] S. A. Alex, “Classification of Imbalanced Data Using SMOTE and AutoEncoder Based Deep Convolutional Neural Network,” *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, vol. 31, no. 3, pp. 437–469, 2023, doi: 10.1142/S0218488523500228.

[8] A. Desiani, “Handling the imbalanced data with missing value elimination smote in the classification of the relevance education background with graduates employment,” *Iaes Int. J. Artif. Intell.*, vol. 10, no. 2, pp. 346–354, 2021, doi: 10.11591/ijai.v10.i2.pp346-354.

[9] J. A. Adisa, “The Effect of Imbalanced Data and Parameter Selection via Genetic Algorithm Long Short-Term Memory

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1753−1760
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8471

(LSTM) for Financial Distress Prediction," *IAENG Int. J. Appl. Math.*, vol. 53, no. 3, 2023.
[10] Asniar, "SMOTE-LOF for noise identification in imbalanced data classification," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 6, pp. 3413–3423, 2022, doi: 10.1016/j.jksuci.2021.01.014.
[11] J. Park, "A study on improving turnover intention forecasting by solving imbalanced data problems: focusing on SMOTE and generative adversarial networks," *J. Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00715-6.
[12] K. V. Rani, "Lung Lesion Classification Scheme Using Optimization Techniques and Hybrid (KNN-SVM) Classifier," *IETE J. Res.*, vol. 68, no. 2, pp. 1485–1499, 2022, doi: 10.1080/03772063.2019.1654935.
[13] K. M. Sunnetci, "Face Mask Detection Using GoogLeNet CNN-Based SVM Classifiers," *Gazi Univ. J. Sci.*, vol. 36, no. 2, pp. 645–658, 2023, doi: 10.35378/gujs.1009359.
[14] S. DEMİR and E. K. ŞAHİN, "Evaluation of Oversampling Methods (OVER, SMOTE, and ROSE) in Classifying Soil Liquefaction Dataset based on SVM, RF, and Naïve Bayes," *Eur. J. Sci. Technol.*, no. 34, pp. 142–147, 2022, doi: 10.31590/ejosat.1077867.
[15] L. Liu, "Average AoI Minimization in UAV-Assisted Data Collection With RF Wireless Power Transfer: A Deep Reinforcement Learning Scheme," *IEEE Internet Things J.*, vol. 9, no. 7, pp. 5216–5228, 2022, doi: 10.1109/JIOT.2021.3110138.
[16] J. Huang, "Deciphering decision-making mechanisms for the susceptibility of different slope geohazards: A case study on a SMOTE-RF-SHAP hybrid model," *J. Rock Mech. Geotech. Eng.*, vol. 17, no. 3, pp. 1612–1630, 2025, doi: 10.1016/j.jrmge.2024.03.008.
[17] Z. A. Sejuti and M. S. Islam, "A hybrid CNN–KNN approach for identification of COVID-19 with 5-fold cross validation," *Sensors Int.*, vol. 4, no. November 2022, p. 100229, 2023, doi: 10.1016/j.sintl.2023.100229.
[18] W. Sinhashthita, "Improving KNN Algorithm Based on Weighted Attributes by Pearson Correlation Coefficient and PSO Fine Tuning," 2020.
[19] T. Benil, "Efficient data pruning using optimal KNN for weather forecasting in cloud computing," *Int. J. Glob. Warm.*, vol. 30, no. 2, pp. 137–151, 2023, doi: 10.1504/IJGW.2023.130984.
[20] İ. Atik, "Pneumonia detection on chest x-ray images using residual convolutional neural network," *J. Fac. Eng. Archit. Gazi Univ.*, vol. 39, no. 3, pp. 1719–1731, 2024, doi: 10.17341/gazimmfd.1271385.
[21] D. I. Swasono, "Classification of Air-Cured Tobacco Leaf Pests Using Pruning Convolutional Neural Networks and Transfer Learning," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 12, no. 3, pp. 1229–1235, 2022, doi: 10.18517/ijaseit.12.3.15950.
[22] J. Chen, H. Huang, A. G. Cohn, D. Zhang, and M. Zhou, "Machine learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning," *Int. J. Min. Sci. Technol.*, vol. 32, no. 2, pp. 309–322, 2022, doi: 10.1016/j.ijmst.2021.08.004.
[23] H. Hairani, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link," *Int. J. Informatics Vis.*, vol. 7, no. 1, pp. 258–264, 2023, doi: 10.30630/joiv.7.1.1069.
[24] T. Wu, "Intrusion detection system combined enhanced random forest with SMOTE algorithm," *EURASIP J. Adv. Signal Process.*, vol. 2022, no. 1, 2022, doi: 10.1186/s13634-022-00871-6.
[25] N. G. Ramadhan, "MODIFIED SMOTE AND ENSEMBLE LEARNING BASED ON EXPERT JUDGMENT FOR CHRONIC DISEASES PREDICTION," *Int. J. Innov. Comput. Inf. Control*, vol. 21, no. 4, pp. 1003–1023, 2025, doi: 10.24507/ijicic.21.04.1003.
[26] M. Mishra, "DTCDWT-SMOTE-XGBoost-Based Islanding Detection for Distributed Generation Systems: An Approach of Class-Imbalanced Issue," *IEEE Syst. J.*, vol. 16, no. 2, pp. 2008–2019, 2022, doi: 10.1109/JSYST.2021.3086298.
[27] H. Naz, "SMOTE-SMO-based expert system for type II diabetes detection using PIMA dataset," *Int. J. Diabetes Dev. Ctries.*, vol. 42, no. 2, pp. 245–253, 2022, doi: 10.1007/s13410-021-00969-x.
[28] A. R. Vaka, B. Soni, and S. R. K., "Breast cancer detection by leveraging Machine Learning," *ICT Express*, vol. 6, no. 4, pp. 320–324, 2020, doi: 10.1016/j.icte.2020.04.009.
[29] A. Rahmat, H. Hardi, F. A. Syam, Z. Zamzami, B. Febriadi, and A. P. Windarto, "Utilization of the field of data mining in mapping the area of the Human Development Index (HDI) in Indonesia," *J. Phys. Conf. Ser.*, vol. 1783, no. 1, 2021, doi: 10.1088/1742-6596/1783/1/012035.
[30] H. Wang, "Research on the Application of Random Forest-based Feature Selection Algorithm in Data Mining Experiments," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 10, pp. 505–518, 2023, doi: 10.14569/IJACSA.2023.0141054.
[31] M. Irshada, "SMOTE and ExtraTreesRegressor based random forest technique for predicting Australian rainfall," *Int. J. Inf. Technol. Singapore*, vol. 15, no. 3, pp. 1679–1687, 2023, doi: 10.1007/s41870-023-01185-y.
[32] A. E. K. Gunawan, "Stock Price Movement Classification Using Ensembled Model of Long Short-Term Memory (LSTM) and Random Forest (RF)," *Int. J. Informatics Vis.*, vol. 7, no. 4, pp. 2255–2262, 2023, doi: 10.30630/joiv.7.4.1640.
[33] C. Y. Lee, K. Y. Huang, Y. X. Shen, and Y. C. Lee, "Improved weighted k-nearest neighbor based on PSO for wind power system state recognition," *Energies*, vol. 13, no. 20, 2020, doi: 10.3390/en13205520.
[34] S. Pande, A. Khamparia, and D. Gupta, "Feature selection and comparison of classification algorithms for wireless sensor networks," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 3, pp. 1977–1989, 2023, doi: 10.1007/s12652-021-03411-6.
[35] A. P. Ayudhitama and U. Pujianto, "Analisa 4 Algoritma Dalam Klasifikasi Liver Menggunakan Rapidminer," *J. Inform. Polinema*, vol. 6, no. 2, pp. 1–9, 2020, doi: 10.33795/jip.v6i2.274.
[36] H. Sun, "Multi-variety monitoring of potato late blight severity using UAV data with improved SMOTE-CS for small sample modeling and deep feature learning," *Eur. J. Agron.*, vol. 169, 2025, doi: 10.1016/j.eja.2025.127702.
[37] A. R. Doni, "Artificial Intelligence Driven Skin Cancer Detection Using R-FCN Enhanced Deep Convolutional Neural Networks with SMOTE Balancing," *Int. J. Eng. Sci. Inf. Technol.*, vol. 5, no. 4, pp. 28–36, 2025, doi: 10.52088/ijesty.v5i4.1196.
[38] L. Azoulay-Younes, "Zinc electrode manufacturing quality control with machine learning: Using SMOTE & image augmentation to prevent overfitting," *J. Eng. Res. Kuwait*, 2025, doi: 10.1016/j.jer.2025.01.002.