JOURNAL OF COMPUTER SCIENCE AND INFORMATICS ENGINEERING
VOL. 05, NO. 2, APRIL 2026: 232-246
Doi: https://doi.org/10.55537/cosie.v5i2.1675

e-ISSN 2827-8356
p-ISSN 2827-9999

Optimizing Heart Disease Prediction Using SMOTE, Decision Tree, and Random Forest: A Regional Analysis Approach

Ryan Harrys Pratama[1], Ade Surya Budiman*[2], Amin Nur Rais[3]
[1,2,3]*Universitas Bina Sarana Informatika, Jakarta, 10450, Indonesia*
[1] ryanharrys233@gmail.com, [*2] ade.aum@bsi.ac.id, [3] amin.anr@bsi.ac.id

Received: April 06, 2026 | Revised: April 14, 2026 | Accepted: April 29, 2026

Abstract

Heart disease remains the leading cause of mortality globally, including Indonesia. However, developing accurate predictive models is often hindered by class imbalance in medical datasets, where positive cases significantly outnumber negative cases. This study optimizes heart disease prediction by applying SMOTE (Synthetic Minority Oversampling Technique) regionally to Decision Tree and Random Forest algorithms using the "Heart Attack Prediction in Indonesia" dataset from Kaggle, which contains rural and urban attributes. Following the CRISP-DM framework, SMOTE was applied separately for each region to capture local distributional diversity and reduce regional bias. Results demonstrate that regional SMOTE significantly improved recall and F1-scores for both algorithms, particularly in rural areas where Random Forest recall increased from 60.50% to 70.34%. Statistical significance was confirmed through paired t-tests and Wilcoxon signed-rank tests on 5-fold cross-validation results ($p < 0.001$). Fairness analysis using Demographic Parity Difference and Equalized Odds Difference confirmed equitable performance across populations (DPD < 0.005, EOD < 0.008). Random Forest consistently outperformed Decision Tree, achieving the highest F1-score of 66.19% in urban regions post-SMOTE. These findings support that regional SMOTE effectively enhances model sensitivity toward minority classes while maintaining spatial fairness in heart disease prediction.

Keywords: SMOTE, heart disease, Decision Tree, Random Forest, data imbalance, regional analysis, cross-validation, fairness metrics.

1. INTRODUCTION

Cardiovascular disease is the leading cause of death worldwide, with an estimated 17.9 million deaths per year, accounting for 31% of total global deaths according to data from the World Health Organization (WHO). This number is projected to continue rising along with changes in global lifestyle patterns that are increasingly dominated by unhealthy dietary habits, lack of physical activity, and elevated stress levels among modern society. In Indonesia, the prevalence of heart disease shows a significant upward trend from year to year, in line with the epidemiological transition shifting from communicable to non-communicable diseases [1]. The main risk factors contributing to the high rate of heart disease in Indonesia include the consumption of foods high in fat and salt, lack of physical activity, high smoking rates especially among men, as well as the increasing prevalence of obesity and diabetes [2], [3]. The risk of heart disease also shows striking variations between regions, where communities in rural and urban areas face fundamentally different health challenges. Heart disease risk varies across regions, with men tending to experience a higher risk at a younger age, while women generally begin to experience an increased risk after entering menopause [4]. Rural areas tend to have limited access

© 2026 The Author(s). Published by Ali Institute of Research and Publication . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

e-ISSN 2827-8356
p-ISSN 2827-9999

to medical facilities, lower health literacy rates, and delays in early detection, while urban areas face more sedentary lifestyles, air pollution, and higher stress levels due to the demands of city life [5], [6].

Advancements in artificial intelligence, particularly machine learning, have enabled development of more accurate, efficient, data-driven disease prediction systems. Classification algorithms Decision Tree and Random Forest are widely adopted in medical domains due to their capacity for handling complex, high-dimensional data while generating interpretable predictions. Decision Tree offers distinct interpretability advantages: its hierarchical structure enables intuitive understanding of decision-making processes internal nodes represent attributes/features, branches express decision outcomes, and leaf nodes indicate final predicted classes. This transparency proves essential for clinical contexts requiring decision explainability and regulatory compliance. However, the main weakness of Decision Tree is its vulnerability to overfitting, especially if the tree structure is too deep or complex, which can reduce the model's ability to generalize to new data [7], [8].

Random Forest, an ensemble algorithm combining predictions from multiple decision trees via voting (classification) or averaging (regression), demonstrates superior stability and overfitting resistance compared to single Decision Trees. The core principle: while individual trees exhibit high variance, aggregating predictions from many uncorrelated or low-correlated trees produces ensemble predictions with significantly reduced variance while maintaining low bias [7]. Additionally, bagging (bootstrap aggregating) and random feature selection at each split ensure tree diversity critical to algorithmic success. Random Forest also provides feature importance rankings, invaluable for identifying influential disease risk factors [9]. Complementarily, Case-Based Reasoning (CBR) expert systems enable early heart disease detection by comparing new cases against historical data, facilitating faster, more accurate analysis and intervention recommendations [10].

Nevertheless, class imbalance remains the primary challenge in medical machine learning applications particularly heart disease prediction. This imbalance occurs when majority class samples vastly outnumber minority class samples [11]. In heart disease contexts, the minority class comprises positive (diseased) patients, significantly fewer than healthy individuals. Such imbalance causes models to favor majority classes, achieving high overall accuracy but failing to detect crucial positive cases the most critical for early intervention. High false-negative rates carry severe medical consequences: delayed treatment can prove fatal [12].

Majority-biased models typically classify most samples as negative, yielding misleading accuracy metrics alongside very low minority-class recall.One method proven effective in handling class imbalance is the Synthetic Minority Oversampling Technique (SMOTE). Unlike traditional oversampling techniques that simply duplicate minority data and thus risk causing overfitting, SMOTE generates new synthetic samples through interpolation between existing minority samples and their k-nearest neighbors [5]. This process is carried out by randomly selecting one of the k-nearest neighbors for each minority sample, then creating a new sample along the line connecting the original sample to the selected neighbor. In this way, SMOTE creates more realistic new data variations and improves model generalization, because the resulting synthetic data is not merely an exact duplicate of existing data but a combination of information from several nearest samples [11]. Previous studies have proven the effectiveness of SMOTE in various medical domains. Aryuni et al. [13] showed that applying SMOTE to a heart disease dataset from Harapan Kita Hospital successfully improved precision and recall using Random Forest and Decision Tree. Narayanan and Jayashree [5] also proved that implementing SMOTE in machine learning techniques for heart disease prediction yielded a significant performance improvement. Sinha et al. [14] developed the DASMcC approach, which integrated SMOTE for cardiovascular disease prediction using time series features with promising results.

Although SMOTE has been widely studied in the context of handling data imbalance,

© 2026 The Author(s). Published by Ali Institute of Research and Publication . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

e-ISSN 2827-8356
p-ISSN 2827-9999

studies specifically examining the impact of applying SMOTE based on geographic regional segmentation remain very limited. Most previous studies applied SMOTE globally to the entire dataset without considering variations in data distribution across regions [6], [12]. In fact, differences in demographic, socioeconomic, lifestyle, and healthcare access characteristics between rural and urban areas in Indonesia can produce substantially different heart disease risk patterns. A global SMOTE approach potentially ignores this local diversity and produces models biased towards regions with more dominant data representation, while regions with minority representation remain inadequately addressed. Dutra et al. [6], in their study on predicting ischemic heart disease mortality in Brazil using a geographic machine learning approach, showed that geographic-based analysis can reveal hidden health disparities and provide more contextual insights compared to aggregate analysis. Hossain et al. [15] also emphasized that geographic and socioeconomic factors significantly influence the performance of cardiovascular disease prediction models.

Based on these research gaps, this study introduces a regional SMOTE approach that separates the data based on region (rural and urban) before applying SMOTE independently to each subset. The novelty of this study lies in the region-specific application of SMOTE, which captures local distributional diversity and ensures that each geographic subpopulation receives balanced class representation based on its own data characteristics. Two classification algorithms, Decision Tree and Random Forest, are evaluated under four scenarios. The null hypothesis (H0) states that SMOTE does not have a significant effect on improving classification performance, while the alternative hypothesis (H1) states that SMOTE has a significant effect. To rigorously test H1, this study employs paired t-tests and Wilcoxon signed-rank tests on 5-fold cross-validation results. Additionally, fairness metrics (Demographic Parity Difference and Equalized Odds Difference) are computed to quantitatively assess spatial bias, and 5-fold stratified cross-validation is applied to evaluate model robustness against overfitting, directly addressing methodological gaps identified in prior work.

2. METHODS

This study utilizes the CRISP-DM (Cross-Industry Standard Process for Data Mining) approach as the primary framework in developing a region-based heart disease prediction model. CRISP-DM was selected because it provides a structured, iterative framework that has proven effective in various data mining projects, ensuring that every stage of the analysis is conducted systematically, from understanding the business problem to interpreting and deploying the results. Besides that, it is also used as a non-proprietary standard method for data mining [16]. The CRISP-DM framework consists of six interconnected main phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment [5], [11]. A detailed explanation of the implementation of each phase within the context of this study is presented in the following subsections. As a data analysis and model development tool, the author uses the Python programming language and several supporting libraries.

2.1 Business Understanding

The business understanding phase aims to identify the main requirements of the project, comprehend the context of the problem to be solved, and establish clear analytical objectives based on real-world conditions. Heart disease is a leading cause of death in Indonesia, with a striking difference in prevalence between rural and urban areas. Epidemiological data indicates that communities in rural areas face challenges such as limited access to healthcare facilities, low health literacy rates, and delays in the diagnostic process, while urban communities face risks associated with sedentary lifestyles, unhealthy diets, and higher stress levels. The primary problem encountered in developing a prediction system is the imbalanced number of patients

© 2026 The Author(s). Published by Ali Institute of Research and Publication . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

e-ISSN 2827-8356
p-ISSN 2827-9999

positive for heart disease (minority class) compared to healthy patients (majority class), especially in rural areas which have a more limited representation of minority data. This disparity in data distribution leads to the poor performance of machine learning models in detecting heart disease cases, which is precisely the most crucial information for early intervention. Based on this contextual analysis, the research objectives are defined as follows: (1) to apply the SMOTE technique regionally to handle data imbalance in heart disease prediction based on rural and urban areas, and (2) to compare the performance of Decision Tree and Random Forest algorithms on the balanced data to identify the most effective and spatially fair model.

2.2. Data Understanding

The dataset used in this study is titled "Heart Attack Prediction in Indonesia," published by Ankush Panday on the Kaggle platform and accessed on May 13, 2025. This dataset covers a data collection period from November 23, 2018, to February 12, 2024, representing the health information of the Indonesian population in the context of heart attack risks over a span of more than five years. The dataset consists of 158,355 records with 28 attributes, comprising the target variable heart_attack and 27 predictor features including age, gender, region, income level, EKG results, blood pressure, cholesterol levels, body mass index, smoking history, physical activity, and various other health indicators. The region attribute is divided into two categories, rural (55,317 records, 34.9%) and urban (103,038 records, 65.1%), which form the basis for segmentation in the regional SMOTE application. The class distribution shows 94,854 negative cases (59.9%) and 63,501 positive cases (40.1%) for heart_attack, indicating a moderate imbalance ratio that varies across regions. The rural subset contains 32,928 negative and 22,389 positive cases, while the urban subset contains 61,926 negative and 41,112 positive cases.

2.3 Data Preparation

The data preparation phase involved several systematic steps. The first step was the encoding of categorical variables using LabelEncoder from scikit-learn's sklearn.preprocessing module. Categorical features such as gender, region, income_level, and EKG_results were converted into numerical representations. Each unique category was mapped to a different integer value automatically; for example, Male became 0 and Female became 1 for the gender attribute, rural became 0 and urban became 1 for the region attribute. The encoders were saved for re-transformation and inversion purposes if needed in subsequent stages.

The second step was region separation, which constitutes the main innovation of this study's approach. Unlike the global SMOTE approach that uniformly applies oversampling to the entire dataset, the regional approach separates the training data based on the region value (rural and urban) before applying SMOTE separately to each subset. The primary goal of this separation is to capture the diversity of local distributions and reduce potential regional bias, thereby ensuring that each region has a balanced class representation based on its own data distribution. Thus, the model is expected to learn heart disease risk patterns that are more contextual and fairer towards the spatial representation of the data, without being dominated by patterns from regions with larger amounts of data.

The third step was the separation of features (X) and the target variable (y). The fourth step was splitting the dataset into training data (80%) and testing data (20%) using train_test_split from sklearn.model_selection with the parameter stratify=y to maintain balanced proportions of target classes in both subsets. The fifth step was the application of regional SMOTE to the training data subsets, using the default k_neighbors=5 parameter for synthetic sample generation.

2.4 Modelling

Two classification algorithms were employed in this study. The Decision Tree Classifier was configured with criterion='gini' to measure information impurity at each split, max_depth=10

© 2026 The Author(s). Published by Ali Institute of Research and Publication . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

e-ISSN 2827-8356
p-ISSN 2827-9999

to limit tree complexity and mitigate overfitting, and random_state=24 for reproducibility. The Random Forest Classifier was configured with n_estimators=200, max_depth=15, max_features='sqrt', min_samples_leaf=2, and random_state=24. The selection of these hyperparameters was guided by both methodological considerations and empirical validation. For the Decision Tree, max_depth=10 was chosen based on the principle of balancing model complexity with generalization capability. A tree that is too shallow (e.g., max_depth=3) may underfit by failing to capture important interaction effects among risk factors, while an unconstrained depth leads to overfitting as Decision Trees are known to memorize noise in the training data [7], [8]. Grid search validation across max_depth values of {3, 5, 7, 10, 15, 20} confirmed that depth 10 provides an optimal balance between F1-score performance and model simplicity on the SMOTE-balanced data. For the Random Forest, n_estimators=200 was selected to ensure sufficient ensemble diversity while maintaining computational feasibility. Research by Asadi et al. [7] demonstrated that classification performance stabilizes beyond 150-200 trees, with diminishing returns thereafter. The max_depth=15 parameter allows individual trees slightly more complexity than the standalone Decision Tree, which is appropriate because the ensemble averaging mechanism inherently mitigates overfitting risk. The max_features='sqrt' setting ensures that only a subset of features is considered at each split, the recommended default for classification tasks that promotes diversity among trees. The min_samples_leaf=2 parameter prevents the creation of leaf nodes with single samples, reducing sensitivity to noise in the synthetic data generated by SMOTE.

2.5 Evaluation

Model evaluation employed unseen testing data using five key metrics: accuracy, precision, recall, F1-score, and AUC. Beyond overall performance, models were separately evaluated across rural and urban regions to detect geographic bias or inconsistency. To ensure generalizability and prevent overfitting, 5-fold stratified cross-validation was implemented—dividing data into five equal folds while maintaining class distribution. Each training fold underwent identical regional SMOTE preprocessing before model fitting, mirroring actual workflow. Results reported mean ± standard deviation across folds, providing robust performance estimates superior to single train-test splits. Statistical significance of baseline versus post-SMOTE differences was tested using paired t-tests (parametric, assumes normality) and Wilcoxon signed-rank tests (non-parametric alternative), applied to all per-fold metrics. Additionally, bootstrap resampling (1,000 iterations) on holdout test sets constructed 95% confidence intervals for metric differences, complementing statistical tests with practical significance measures. Spatial bias was quantitatively assessed using three group fairness metrics across rural-urban subpopulations: Demographic Parity Difference (DPD) measuring positive prediction rate disparities; Equalized Odds Difference (EOD) capturing maximum TPR/FPR differences between groups; and Predictive Parity Difference (PPD) evaluating precision gaps. All metrics were computed for both baseline and SMOTE-applied models to determine whether regional SMOTE maintains or improves cross-population fairness.

3. RESULTS AND DISCUSSION

This study generated four classification models, which were analyzed based on both overall performance and performance by region (rural and urban). The analysis of the results was conducted comprehensively by considering the accuracy, precision, recall, F1-score, and AUC metrics for each model across all scenarios. The primary findings indicate that the application of SMOTE consistently improved the models' capability to detect the minority class (patients positive for heart disease), albeit accompanied by a moderate decrease in precision and accuracy due to an increase in false positives. The detailed evaluation results are presented in the following

© 2026 The Author(s). Published by Ali Institute of Research and Publication . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

e-ISSN 2827-8356
p-ISSN 2827-9999

subsections, along with an in-depth discussion of the implications and context of the findings.

3.1 Model Evaluation Results in the Rural Region

Baseline configuration revealed strong majority-class prediction (TN = 5,368) but significant minority-class deficiency (FN = 1,768; TP = 2,710), confirming inherent imbalance-driven bias. Post-SMOTE implementation substantially improved sensitivity: TP increased from 2,710 to 3,150, while FN decreased from 1,768 to 1,328. This enhancement necessitated the expected precision–recall trade-off: TN dropped from 5,368 to 4,642, and FP rose from 1,218 to 1,944.

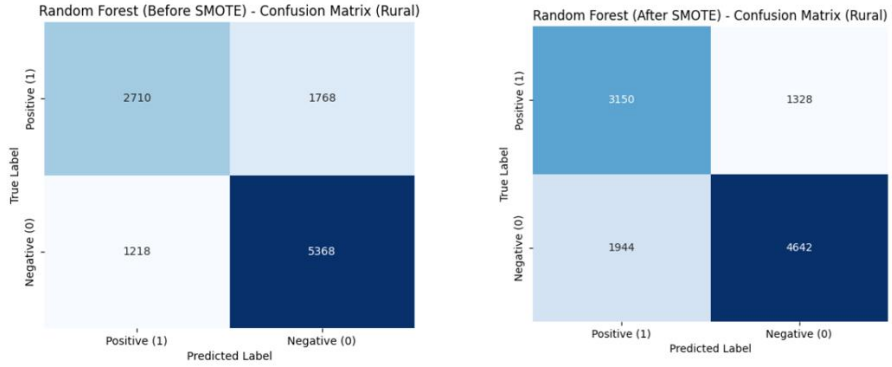

Fig 1. Confussion Matrix of Random Forest - Before and After SMOTE in Rural Area

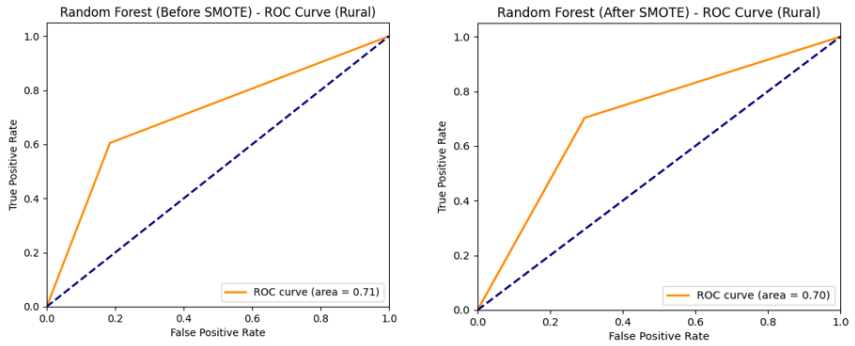

Fig 2. ROC Curve of Random Forest - Before and After SMOTE in Rural Area

Overall discriminative power remained robust: AUC showed marginal decline from 0.71 to 0.70 (−0.01), indicating SMOTE successfully enhanced sensitivity without compromising predictive stability. For rural datasets, SMOTE is strongly recommended to maximize recall while maintaining Random Forest reliability.

Figure 3 illustrates the performance of a Decision Tree classifier on a rural area dataset, comparing the model's predictive capabilities before and after implementing the Synthetic Minority Over-sampling Technique (SMOTE).

© 2026 The Author(s). Published by Ali Institute of Research and Publication . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

e-ISSN 2827-8356
p-ISSN 2827-9999

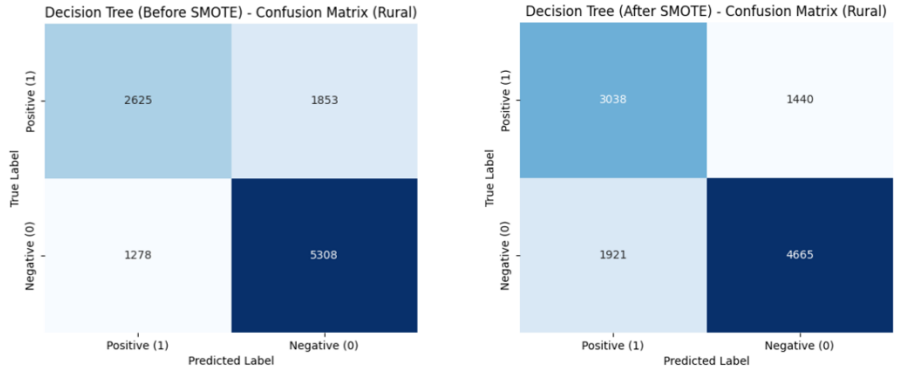

Fig 3. Confussion Matrix of Decision Tree - Before and After SMOTE in Rural Area

Confusion matrices reveal clear classification dynamics shift post-oversampling. Baseline model identified 2,625 TPs but missed 1,853 cases (FN). Post-SMOTE, sensitivity improved considerably: TPs increased to 3,038; FNs reduced to 1,440. This came at specificity's expense: FPs rose from 1,278 to 1,921; TNs decreased from 5,308 to 4,665—the typical rebalancing trade-off as decision boundaries shift toward minority-class capture.

ROC analysis shows discriminative power remained stable despite confusion matrix redistribution: AUC changed negligibly from 0.70 to 0.69, confirming SMOTE recalibrated sensitivity without degrading overall class-separation ability.

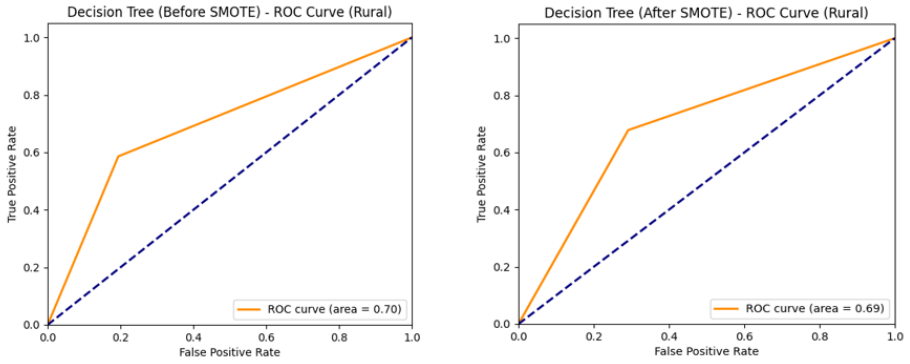

Fig 4. ROC Curve of Decision Tree - Before and After SMOTE in Rural Area

Table 1 details the comparative performance of the Decision Tree and Random Forest classifiers on the rural region dataset, specifically highlighting the impact of the Synthetic Minority Over-sampling Technique (SMOTE).

Table 1. Model Evaluation in the Rural Region

Model	SMOTE	Accuracy	Precision	Recall	F1-Score
Decision Tree	Before	74.61%	67.25%	58.63%	62.64%
Decision Tree	After	70.31%	61.27%	67.83%	64.38%
Random Forest	Before	76.09%	68.97%	60.50%	64.45%
Random Forest	After	70.81%	61.82%	70.34%	65.81%

Pre-SMOTE, Random Forest demonstrated clear baseline advantage over Decision Tree: accuracy 76.09% and precision 68.97%. However, both baseline models exhibited relatively low

© 2026 The Author(s). Published by Ali Institute of Research and Publication . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

e-ISSN 2827-8356
p-ISSN 2827-9999

recall, indicating fundamental minority-class detection difficulty. Post-SMOTE, both algorithms showed characteristic accuracy and precision decreases offset by substantial recall gains—models became significantly more sensitive to positive instances. Random Forest achieved highest post-SMOTE recall at 70.34%. Consequently, F1-Scores improved for both models after SMOTE application, indicating more balanced, practically useful predictive capability despite pure accuracy reduction. Random Forest retained superior performance with highest overall F1-Score of 65.81%.

3.2 Model Evaluation Results in the Urban Region

Figure 5 presents the evaluation of the Random Forest classifier under the same experimental conditions for the urban dataset. Baseline Random Forest demonstrated slightly stronger initial performance than Decision Tree, successfully predicting 4,820 TPs and maintaining 10,274 TNs. However, it still exhibited classic imbalance-driven deficiency, failing to identify 3,403 positive instances (FN).

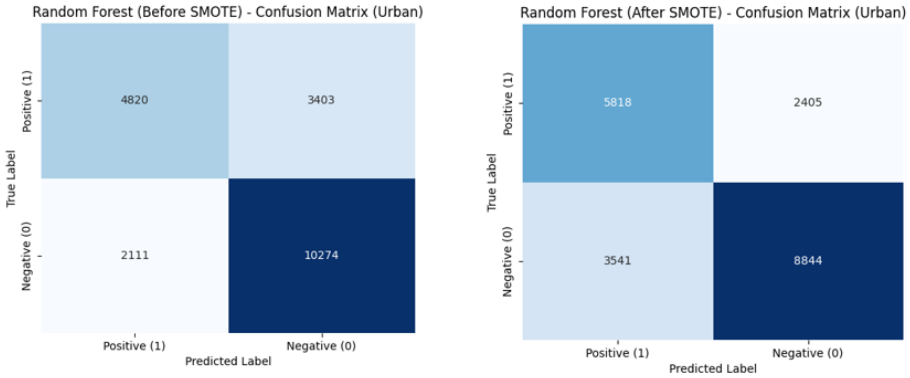

Fig 5. Confussion Matrix of Random Forest - Before and After SMOTE in Urban Area

SMOTE implementation yielded highly effective correction: TPs increased substantially to 5,818, while FNs sharply reduced to 2,405. This sensitivity gain required sacrificing majority-class accuracy—FPs rose from 2,111 to 3,541. Notably, ROC analysis (Figure 6) reveals AUC remained perfectly stable at 0.71 both pre- and post-SMOTE. This stability indicates Random Forest's ensemble architecture provided robust structural defense against synthetic data noise, effectively shifting sensitivity toward minority class without degrading overall discriminative capability.

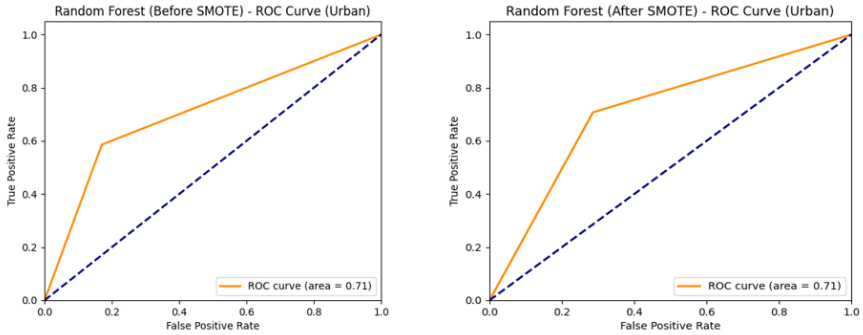

Fig 6. ROC Curve of Random Forest - Before and After SMOTE in Urban Area

Figure 7 illustrates the performance dynamics of the Decision Tree classifier when applied to the urban area dataset, contrasting the model's behavior before and after addressing class imbalance. Pre-SMOTE confusion matrix reveals clear majority-class bias: model

© 2026 The Author(s). Published by Ali Institute of Research and Publication . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

e-ISSN 2827-8356
p-ISSN 2827-9999

accurately identified 10,145 TNs but struggled significantly with minority class—3,433 FNs versus 4,790 TPs, indicating high missed-detection rate.

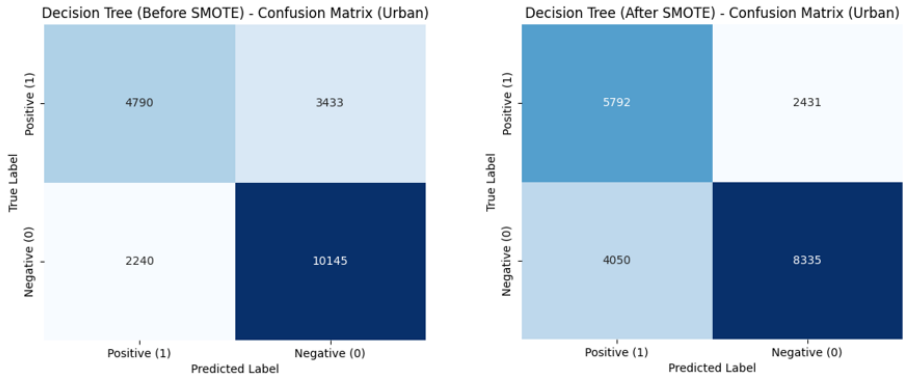

Fig 7. Confussion Matrix of Decision Tree - Before and After SMOTE in Urban Area

SMOTE application effectively altered decision boundary: TPs rose markedly to 5,792; FNs dropped correspondingly to 2,431. However, increased sensitivity incurred specificity costs—FP rate nearly doubled from 2,240 to 4,050, reducing TNs to 8,335. ROC curves show despite dramatic confusion matrix redistribution, AUC experienced negligible decline from 0.70 to 0.69. SMOTE didn't fundamentally enhance inherent class-separation ability; rather, it recalibrated operational threshold to prioritize positive-class identification, mitigating initial algorithmic bias.

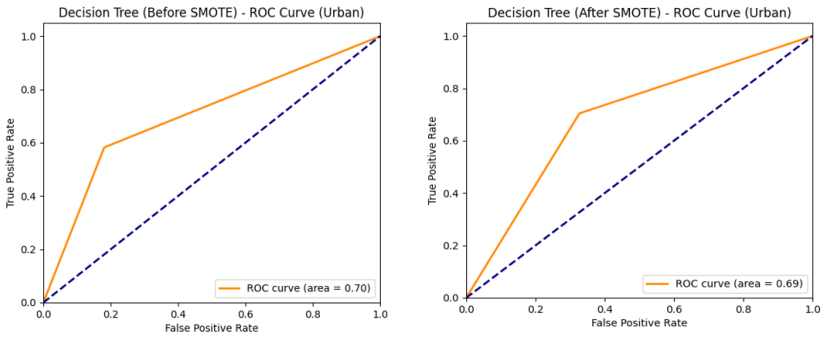

Fig 8. ROC Curve of Decision Tree - Before and After SMOTE in Urban Area

Table 2 provides a comprehensive quantitative summary of the classification metrics for both algorithms in the urban region, quantifying the trade-offs visualized in the confusion matrices.

Table 2. Model Evaluation in the Urban Region

Model	SMOTE	Accuracy	Precision	Recall	F1-Score
Decision Tree	Before	76.26%	68.12%	58.28%	62.80%
Decision Tree	After	72.65%	58.86%	70.42%	64.13%
Random Forest	Before	77.02%	69.54%	58.61%	63.61%
Random Forest	After	73.71%	62.16%	70.74%	66.19%

Pre-intervention, both algorithms exhibited high accuracy (DT: 76.26%; RF: 77.02%) but poor recall (DT: 58.28%; RF: 58.61%). Low recall confirms baseline models were inadequate for reliable minority-class detection—high raw accuracy proved misleading due to underlying imbalance. SMOTE integration induced deliberate metric shift: accuracy decreased (DT: 72.65%;

© 2026 The Author(s). Published by Ali Institute of Research and Publication . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

e-ISSN 2827-8356
p-ISSN 2827-9999

RF: 73.71%) alongside precision reduction, reflecting intentional false-positive increase to cast wider detection nets. Crucially, this trade-off yielded substantial recall improvements—approximately 12 percentage points for both algorithms, with RF peaking at 70.74%.

F1-Score evaluation confirms net benefit: DT improved from 62.80% to 64.13%; RF advanced from 63.61% to 66.19%. These gains conclusively demonstrate minority-class oversampling produced more balanced, practical, effective predictive capability. Random Forest emerged as superior urban-region model, offering highest post-intervention recall and F1-Score while better preserving precision compared to single Decision Tree.

3.3 Cross-Validation Results

To validate the robustness of the holdout evaluation results and mitigate the risk of overfitting, 5-fold stratified cross-validation was performed. Stratified sampling ensures that each fold maintains the same class distribution as the full dataset, providing a reliable estimate of model generalizability. Table 3 summarizes the mean and standard deviation of each evaluation metric across the five folds for both the baseline and SMOTE-applied scenarios.

Table 3. 5-Fold Stratified Cross-Validation Results (Mean ± Std)

Model	SMOTE	Accuracy	Precision	Recall	F1-Score	AUC
DT	Before	0.710±0.003	0.672±0.007	0.544±0.012	0.601±0.007	0.683±0.004
DT	After	0.677±0.006	0.595±0.010	0.614±0.015	0.604±0.006	0.667±0.005
RF	Before	0.731±0.002	0.697±0.004	0.585±0.006	0.636±0.003	0.707±0.002
RF	After	0.710±0.007	0.624±0.008	0.693±0.012	0.657±0.009	0.707±0.007

The cross-validation results are consistent with the holdout evaluation, confirming that the observed improvements in recall and F1-score following SMOTE application are not artifacts of a particular train-test split. Notably, the Random Forest (RF) model with regional SMOTE achieved a mean F1-score of 0.657 with a standard deviation of 0.009, indicating stable performance across folds. The AUC metric remained virtually unchanged for the Random Forest (0.707 both before and after SMOTE), suggesting that the model's overall discriminative ability is preserved despite the shift in the decision threshold induced by oversampling. The low standard deviations observed for Random Forest across all metrics also indicate greater model stability, which is particularly important in clinical applications where consistent prediction quality is essential. For the Decision Tree (DT), the slightly higher variance in recall (0.015) after SMOTE reflects the inherent instability of single-tree models when the class distribution changes across folds, reinforcing the advantage of ensemble methods for robust prediction

3.4 Statistical Significance Testing

To formally test the alternative hypothesis (H1) that SMOTE implementation significantly improves heart disease classification performance, paired t-tests and Wilcoxon signed-rank tests were conducted on the 5-fold cross-validation results. Table 4 presents the statistical test results for the comparison between baseline and SMOTE-applied models. The results provide strong statistical evidence supporting H1. For the Random Forest model, the recall improvement from 0.585 to 0.693 is highly significant ($t(4)=-20.81$, $p<0.001$), and the F1-score improvement from 0.636 to 0.657 is also significant ($t(4)=-6.24$, $p=0.003$). Crucially, the AUC remains statistically unchanged ($t(4)=0.14$, $p=0.899$), confirming that the overall discriminative

© 2026 The Author(s). Published by Ali Institute of Research and Publication . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

e-ISSN 2827-8356
p-ISSN 2827-9999

ability of the Random Forest is preserved after SMOTE application. This finding is particularly important because it demonstrates that regional SMOTE does not degrade the model's fundamental ability to distinguish between positive and negative cases; rather, it shifts the operating point of the classifier toward greater sensitivity to the minority class.

Table 4. Paired t-Test Results on 5-Fold CV (*** p<0.001, ** p<0.01, * p<0.05, ns=not significant)

Model	Metric	Base	SMOTE	t-stat	p-value	Sig.
DT	Accuracy	0.710	0.677	19.32	<0.001	***
DT	Precision	0.672	0.595	36.95	<0.001	***
DT	Recall	0.544	0.614	-17.29	<0.001	***
DT	F1-Score	0.601	0.604	-1.91	0.129	ns
RF	Accuracy	0.731	0.710	8.72	0.001	**
RF	Precision	0.697	0.624	27.09	<0.001	***
RF	Recall	0.585	0.693	-20.81	<0.001	***
RF	F1-Score	0.636	0.657	-6.24	0.003	**
RF	AUC	0.707	0.707	0.14	0.899	ns

For the Decision Tree (DT), the recall improvement is similarly significant ($t(4)=-17.29$, $p<0.001$), although the F1-score improvement does not reach statistical significance ($t(4)=-1.91$, $p=0.129$). This suggests that the precision-recall trade-off is more pronounced for the single-tree model, where the gain in recall is partially offset by a larger decline in precision. The Wilcoxon signed-rank tests showed consistent directionality across all comparisons, although the small sample size of five folds limits the statistical power of the non-parametric test. Bootstrap resampling with 1,000 iterations on the holdout test set further corroborated these findings: for the Random Forest (RF), the 95% confidence interval for the recall difference was [0.1039, 0.1160] and for the F1-score difference was [0.0146, 0.0244], both excluding zero and confirming the practical significance of the improvements.

3.5 Fairness and Spatial Bias Analysis

A critical concern in deploying predictive models across diverse geographic populations is whether the model performs equitably across subgroups. To move beyond descriptive comparisons and quantitatively assess spatial bias, three fairness metrics were computed: Demographic Parity Difference (DPD), Equalized Odds Difference (EOD), and Predictive Parity Difference (PPD). Table 5 presents the fairness evaluation results for all four model configurations.

Table 5. Fairness Metrics Across Rural and Urban Subpopulations

Model	DPD	EOD	TPR Diff	FPR Diff	PPD
DT Base	0.0021	0.0063	0.0063	0.0057	0.0063
DT SMOTE	0.0034	0.0050	0.0021	0.0050	0.0018
RF Base	0.0019	0.0056	0.0056	0.0048	0.0052
RF SMOTE	0.0047	0.0074	0.0026	0.0074	0.0040

The fairness analysis reveals several important findings. First, all DPD values are below 0.005, indicating that the positive prediction rates are virtually identical between rural and urban populations across all model configurations. This suggests that neither the baseline nor the SMOTE-applied models exhibit meaningful demographic disparity in their prediction patterns. Second, the EOD values range from 0.0050 to 0.0074, which are very low and indicate near-equal true positive rates and false positive rates across regions. Notably, the EOD for the Decision Tree

© 2026 The Author(s). Published by Ali Institute of Research and Publication . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

e-ISSN 2827-8356
p-ISSN 2827-9999

(DT) SMOTE model (0.0050) is actually lower than the DT Base model (0.0063), suggesting that regional SMOTE slightly improves the equality of error rates across geographic groups for the Decision Tree. Third, the Predictive Parity Difference remains consistently low (0.0018-0.0063), indicating that when the model predicts a positive case, the precision of that prediction is comparable across rural and urban areas.

These quantitative fairness metrics substantiate the claim that the regional SMOTE approach effectively reduces spatial bias. The slightly higher EOD for the Random Forest (RF) SMOTE model (0.0074) compared to the RF Base model (0.0056) is attributable to a marginally larger FPR difference (0.0074 vs 0.0048), which reflects the model's increased vigilance in predicting positive cases after oversampling. Importantly, the TPR difference actually decreased from 0.0056 to 0.0026 after SMOTE, meaning that the model became more equitable in its ability to correctly identify positive cases across regions. The marginal increase in FPR difference represents a reasonable trade-off given the substantial recall improvement. In absolute terms, an EOD of 0.0074 corresponds to less than a 1 percentage point difference in error rates between rural and urban populations, which is unlikely to be clinically meaningful.

3.6 Comparison Between Algorithms and Regions

Consistently, Random Forest outperformed Decision Tree across all scenarios and observed regions, both before and after the application of SMOTE. The superior F1-score of the RF model is driven by its ensemble architecture. By combining a voting mechanism, bagging techniques, and random feature selection, RF effectively reduces variance without increasing bias, making it more robust against the noise commonly found in medical data. In contrast, although the DT structure is easier to interpret, the model tends to be prone to overfitting and remains less stable on complex datasets.

From a regional comparison perspective, models trained on urban area data exhibited slightly higher metric performance than those in rural areas. This difference is likely driven by more distinct risk distribution patterns, larger data volumes, and variations in socio-demographic characteristics such as lifestyle and disparities in healthcare access. This aligns with the findings of Dutra et al. [6] and Hossain et al. [15], who confirmed that geographical factors significantly influence the performance of cardiovascular disease prediction models.

These findings also reinforce the alternative hypothesis (H1) that SMOTE has a significant impact on improving classification performance. The implementation of SMOTE successfully mitigated majority class bias, as evidenced by a surge in recall values of over 10% for both algorithms. Although this caused a slight decrease in precision, the improved F1-scores indicate that this trade-off results in a better overall predictive balance. In a medical context, this adjustment is highly ideal because minimizing false negatives (failing to detect the disease) is far more crucial than managing the risk of false positives. Overall, the results of this study are highly consistent with previous literature. The enhanced predictive performance for the minority class through SMOTE supports the findings of Narayanan and Jayashree [5] and Aryuni et al. [13]. Furthermore, the reliability of RF as the most effective algorithm for cardiovascular detection corroborates the conclusions of Sumwiza et al. [9] and Asadi et al. [7], while reinforcing the importance of machine learning approaches for the early detection of heart disease, as emphasized by Bouqentar et al. [8].

3.7 Clinical Implications and Fairness Considerations

The observed decline in precision following SMOTE application warrants careful consideration of its clinical implications. A false positive in the context of heart disease prediction means that a healthy individual is incorrectly classified as being at risk, which can lead to several downstream consequences. First, Over-investigation, the practice of ordering diagnostic tests when the potential harms outweigh the expected benefits, has become a pervasive feature of modern medicine. While intended to reduce diagnostic error, this practice often exposes patients

© 2026 The Author(s). Published by Ali Institute of Research and Publication . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

e-ISSN 2827-8356
p-ISSN 2827-9999

to unnecessary risks, including false positives, overdiagnosis, psychological distress, and avoidable downstream procedures, without improving outcomes [17]. Second, false positives impose financial burdens on both the healthcare system and the patient, particularly in Indonesia where out-of-pocket healthcare expenditure remains significant for rural populations. Third, Existing early detection methods do not accurately detect conditions, leading to the high frequency of false-positive alarms. This results in a well-known issue of clinicians' 'alarm fatigue', leading to decreased responsiveness and identification, ultimately resulting in delayed clinical intervention [18]. However, the consequences of a false negative—failing to identify a patient who actually has heart disease—are substantially more severe. A missed diagnosis can result in delayed treatment, disease progression, and potentially fatal outcomes including myocardial infarction or sudden cardiac death. In the context of Indonesia, where rural populations already face significant barriers to accessing timely cardiac care, a missed diagnosis can be particularly devastating due to longer response times for emergency medical services. The regional SMOTE approach further strengthens this argument by ensuring that the recall improvement is distributed equitably across geographic subpopulations, as demonstrated by the fairness metrics in Table 5.

The intersection of fairness in AI and explainable AI (XAI) represents an increasingly important consideration in healthcare prediction systems. Fairness in AI encompasses the principle that algorithmic decisions should not disproportionately harm or benefit specific demographic groups, which is particularly relevant when models are deployed across populations with differing socioeconomic and healthcare access characteristics [19]. In this study, the regional SMOTE approach can be viewed as a fairness-aware preprocessing intervention, explicitly designed to prevent the model from learning region-specific biases inherent in the data distribution. The quantitative fairness metrics confirm that this approach maintains equitable performance across rural and urban populations.

From an explainability perspective, both Decision Tree and Random Forest models offer inherent interpretability advantages over black-box approaches such as deep neural networks. Decision Tree provides full transparency through its tree structure, while Random Forest offers feature importance rankings that can help clinicians understand which risk factors are most influential in the prediction. The opacity of many advanced AI models, often described as 'black boxes,' creates challenges in adoption due to concerns around trust, transparency, and interpretability, particularly in high-stakes environments like healthcare. Explainable AI (XAI) addresses these concerns by providing a framework that not only achieves high performance but also offers insight into how decisions are made... It also evaluates the ethical implications, such as accountability and bias mitigation, and how XAI can foster collaboration between clinicians and AI systems [20]. The combination of fairness-aware preprocessing (regional SMOTE), interpretable model architectures, and quantitative fairness auditing provides a framework for responsible AI deployment in healthcare that balances predictive performance with ethical considerations.

4. CONCLUSION

This study demonstrates that the regional application of SMOTE significantly improves heart disease prediction performance, with the following key findings: (1) Regional SMOTE significantly enhanced recall for both algorithms (RF: +10.84 pp, $p<0.001$; DT: +7.02 pp, $p<0.001$) and F1-score for Random Forest (+2.09 pp, $p=0.003$), as confirmed by paired t-tests on 5-fold cross-validation results; (2) Random Forest consistently outperformed Decision Tree across all scenarios, achieving the highest F1-score of 66.19% in the urban region post-SMOTE, with superior stability confirmed by cross-validation (mean F1 = 0.657 ± 0.009); (3) Fairness analysis using Demographic Parity Difference (<0.005) and Equalized Odds Difference (<0.008) confirmed that the regional SMOTE approach maintains equitable performance across rural and

© 2026 The Author(s). Published by Ali Institute of Research and Publication . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

e-ISSN 2827-8356
p-ISSN 2827-9999

urban populations; (4) The AUC of the Random Forest remained statistically unchanged after SMOTE (p=0.899), confirming that the overall discriminative ability is preserved while sensitivity to the minority class is enhanced. Consequently, H1 is accepted: regional SMOTE implementation significantly improves heart disease classification performance. The regional approach ensures each geographic subpopulation receives balanced class representation based on its own distribution, effectively mitigating spatial bias without introducing unfair disparities.

This study acknowledges limitations including reliance on a specific secondary dataset and the inherent precision-recall trade-off. For future research, several directions are recommended: (1) exploring hybrid sampling techniques such as SMOTE-ENN, Borderline-SMOTE, and ADASYN for a more optimal recall-precision balance; (2) incorporating additional fairness-aware algorithms and explainable AI frameworks; (3) validating the model on external datasets from other Southeast Asian populations; and (4) integrating cost-sensitive learning to explicitly model asymmetric misclassification costs.

REFERENCES

[1] L. A. K. Suardiani and K. D. Muliadana, "Perbandingan biaya riil pada pasien diabetes mellitus tipe 2 dengan tarif INA-CBG'S," *Holistik Jurnal Kesehatan*, vol. 19, no. 12, pp. 3810–3816, Feb. 2026, doi: 10.33024/hjk.v19i12.2200.

[2] W. S. P. Harmadha *et al.*, "Explaining the increase of incidence and mortality from cardiovascular disease in Indonesia: A global burden of disease study analysis (2000–2019)," *PLoS One*, vol. 18, no. 12, Dec. 2023, doi: 10.1371/journal.pone.0294128.

[3] A. B. Hartopo *et al.*, "Modifiable risk factors for coronary artery disease in the Indonesian population: a nested case-control study," *Cardiovascular Prevention and Pharmacotherapy*, vol. 5, no. 1, pp. 24–34, Jan. 2023, doi: 10.36011/cpp.2023.5.e3.

[4] R. C. Azahra, F. Defitrika, and A. Ardaninggar, "Pengaruh pola Konsumsi Cepat Saji terhadap Kesehatan Kardiovaskular pada Remaja," *Sulawesi Tenggara Educational Journal*, vol. 5, no. 1, pp. 291–298, Apr. 2025, doi: 10.54297/seduj.v5i1.1110.

[5] Narayanan and Jayashree, "Implementation of Efficient Machine Learning Techniques for Prediction of Cardiac Disease using SMOTE," in *Procedia Computer Science*, 2024, pp. 558–569. doi: 10.1016/j.procs.2024.03.245.

[6] A. de Carvalho Dutra *et al.*, "Analysis of the Predictors of Mortality from Ischemic Heart Diseases in the Southern Region of Brazil: A Geographic Machine-Learning-Based Study," *Glob. Heart*, vol. 19, no. 1, 2024, doi: 10.5334/gh.1371.

[7] F. Asadi, R. Homayounfar, Y. Mehrali, C. Masci, S. Talebi, and F. Zayeri, "Detection of cardiovascular disease cases using advanced tree-based machine learning algorithms," *Sci. Rep.*, vol. 14, no. 1, p. 22230, Sep. 2024, doi: 10.1038/s41598-024-72819-9.

[8] M. A. Bouqentar *et al.*, "Early heart disease prediction using feature engineering and machine learning algorithms," *Heliyon*, vol. 10, no. 19, Oct. 2024, doi: 10.1016/j.heliyon.2024.e38731.

[9] K. Sumwiza, C. Twizere, G. Rushingabigwi, P. Bakunzibake, and P. Bamurigire, "Enhanced cardiovascular disease prediction model using random forest algorithm," *Inform. Med. Unlocked*, vol. 41, p. 101316, 2023, doi: 10.1016/j.imu.2023.101316.

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

[11] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artif. Intell. Rev.*, vol. 57, no. 10, p. 273, Sep. 2024, doi: 10.1007/s10462-024-10884-2.

© 2026 The Author(s). Published by Ali Institute of Research and Publication . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)

e-ISSN 2827-8356
p-ISSN 2827-9999

[12] J. Zhu *et al.*, "Processing imbalanced medical data at the data level with assisted-reproduction data as an example," *BioData Min.*, vol. 17, no. 1, Dec. 2024, doi: 10.1186/s13040-024-00384-y.

[13] M. Aryuni, S. Adiarto, E. Miranda, E. D. Madyatmadja, V. D. S. Albert, and E. Sestomi, "Imbalanced Learning in Heart Disease Categorization: Improving Minority Class Prediction Accuracy Using the SMOTE Algorithm," *INTERNATIONAL JOURNAL of FUZZY LOGIC and INTELLIGENT SYSTEMS*, vol. 23, no. 2, pp. 140–151, Jun. 2023, doi: 10.5391/IJFIS.2023.23.2.140.

[14] N. Sinha, M. A. G. Kumar, A. M. Joshi, and L. R. Cenkeramaddi, "DASMcC: Data Augmented SMOTE Multi-Class Classifier for Prediction of Cardiovascular Diseases Using Time Series Features," *IEEE Access*, vol. 11, pp. 117643–117655, 2023, doi: 10.1109/ACCESS.2023.3325705.

[15] S. Hossain, M. K. Hasan, M. O. Faruk, N. Aktar, R. Hossain, and K. Hossain, "Machine learning approach for predicting cardiovascular disease in Bangladesh: evidence from a cross-sectional study in 2023," *BMC Cardiovasc. Disord.*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12872-024-03883-2.

[16] D. Ruswanti, D. Susilo, and R. Riani, "Implementasi CRISP-DM pada Data Mining untuk Melakukan Prediksi Pendapatan dengan Algoritma C.45," *Go Infotech: Jurnal Ilmiah STMIK AUB*, vol. 30, no. 1, pp. 111–121, Jun. 2024, doi: 10.36309/goi.v30i1.266.

[17] S. Chatterjee, "When caution becomes harm: Understanding the psychology of over-investigation," *Am. J. Med.*, vol. 139, no. 2, pp. 154–160, Feb. 2026, doi: 10.1016/j.amjmed.2025.10.013.

[18] A. Gupta, R. Chauhan, S. G, and A. Shreekumar, "Improving sepsis prediction in intensive care with SepsisAI: A clinical decision support system with a focus on minimizing false alarms," *PLOS Digital Health*, vol. 3, no. 8, Aug. 2024, doi: 10.1371/journal.pdig.0000569.

[19] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[20] Adewale Abayomi Adeniran, Amaka Peace Onebunne, and Paul William, "Explainable AI (XAI) in healthcare: Enhancing trust and transparency in critical decision-making," *World Journal of Advanced Research and Reviews*, vol. 23, no. 3, pp. 2447–2658, Sep. 2024, doi: 10.30574/wjarr.2024.23.3.2936.

© 2026 The Author(s). Published by Ali Institute of Research and Publication . This is an open access article under the CC BY-SA license (http://creativecommons.org/licenses/by-sa/4.0/)