Optimizing KNN: Impact of Distance Metrics and SMOTE on Heart Disease Classification

Windra Swastika[1], *

[1]Informatics Engineering, Faculty of Technology and Design, Universitas Ma Chung, Malang, Indonesia

Corresponding author: Windra Swastika (e-mail: windra.swastika@machung.ac.id).

ABSTRACT Heart disease remains the leading cause of mortality worldwide, accounting for approximately 32% of all global deaths. The development of accurate and clinically reliable machine learning-based prediction systems is therefore essential for supporting early clinical decision-making. This study proposes a comprehensive optimization framework for the K-Nearest Neighbor (KNN) algorithm applied to heart disease classification using the UCI Cleveland Heart Disease Dataset. While prior work has addressed class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE) and feature scaling via Min-Max Normalization, no study has simultaneously investigated the effect of distance metric selection and systematic K value optimization in the context of preprocessed imbalanced medical data. This paper makes three contributions: (1) a comparative analysis of three distance metrics, Euclidean, Manhattan, and Minkowski (p=3), applied to KNN after preprocessing; (2) systematic optimal-K identification using Grid Search with Stratified 10-Fold Cross-Validation across all metric-scenario combinations; and (3) a structured ablation study across four preprocessing scenarios to quantify the individual and combined contributions of SMOTE and Min-Max Normalization. Experiments were conducted on 297 samples with 13 clinical features. Results show that the best clinically oriented model (Scenario C: SMOTE + Manhattan, K=9) achieves 81.67% accuracy, 82.14% recall, and 80.70% F1-score. The Minkowski metric in the fully combined scenario (D) achieves the highest AUC of 92.47%, with optimal K=21, a markedly different configuration than Euclidean and Manhattan, which converge at K=1. These findings demonstrate that distance metric choice and K optimization interact significantly, offering practical configuration guidelines for KNN in medical classification tasks.

KEYWORDS K-Nearest Neighbour, SMOTE, heart disease prediction, distance metric optimization

I. INTRODUCTION

Cardiovascular disease, particularly coronary heart disease, continues to represent the most significant burden on global public health. The World Health Organization estimates that approximately 17.9 million lives are lost annually to cardiovascular events, constituting 32% of all global deaths [1]. In Indonesia, national health surveys (Riskesdas) have consistently reported increasing prevalence of heart disease, placing it among the primary contributors to national mortality and healthcare expenditure [2]. Early and accurate identification of at-risk patients is critical in reducing these figures; however, traditional diagnostic pathways rely heavily on specialist expertise, multi-step laboratory investigations, and clinical experience, resources that may not be uniformly available across healthcare settings.

The emergence of machine learning (ML) and data mining techniques has opened a promising avenue for developing automated clinical decision support systems. Among the most widely deployed classification algorithms, K-Nearest Neighbor (KNN) stands out for its interpretability, minimal training complexity, and competitive performance on moderately sized datasets, characteristics particularly well-suited to clinical tabular data [3]. KNN classifies a new data point by identifying its K nearest neighbors in the feature space and assigning the majority class label. Its non-parametric nature means it makes no distributional assumptions, which is advantageous when working with heterogeneous medical features.

Despite its advantages, KNN faces two well-documented challenges in medical classification settings. First, real-world clinical datasets are frequently imbalanced: the number of disease-positive samples is typically far smaller than disease-negative samples. This class imbalance causes the learning algorithm to bias toward the majority class, resulting in high overall accuracy but poor sensitivity toward the clinically critical minority class (true disease cases). Second, features in

medical datasets often span vastly different numerical scales, for instance, cholesterol values may range between 100 and 600 mg/dL, while binary-coded features such as fasting blood sugar occupy only {0,1}. Without proper scaling, features with large magnitudes disproportionately dominate the distance computation in KNN. To address these issues, the Synthetic Minority Over-sampling Technique (SMOTE) [4] and Min-Max Normalization are widely employed as preprocessing steps.

While the combination of KNN, SMOTE, and Min-Max Normalization has been explored in various heart disease prediction studies [5], [6], [7], a consistent gap remains in the literature: the majority of studies apply Euclidean distance as the default KNN metric without empirically evaluating alternative distance functions, and the value of K is typically fixed heuristically (e.g., K=3 or K=5) without systematic search on the preprocessed, over-sampled data distribution. Crucially, SMOTE modifies the spatial distribution of training data by introducing synthetic minority-class points; the K value optimal for the original data may no longer be optimal for the SMOTE-augmented distribution. Similarly, Min-Max Normalization compresses features into [0,1], which may alter the relative discriminative power of different distance metrics.

Importantly, the present study recognizes that the UCI Cleveland dataset exhibits only mild class imbalance (ratio 1.17:1). Rather than treating SMOTE as a solution to a severe imbalance problem, this study includes SMOTE as a component of a controlled ablation design. The ablation framework is structured precisely to empirically quantify whether SMOTE is beneficial, neutral, or detrimental when applied to mildly imbalanced data, a question that has important practical implications for practitioners who routinely apply SMOTE without first examining their dataset's imbalance severity. This framing is consistent with guidance from Fernández et al. [8], who note that SMOTE's effectiveness is conditional on the degree of imbalance and the representativeness of the minority class manifold.

This paper addresses the identified gaps through three explicit contributions. First, we conduct a comparative analysis of three KNN distance metrics, Euclidean (p=2), Manhattan (p=1), and Minkowski (p=3), evaluated after each preprocessing configuration. Second, we apply Grid Search with Stratified 10-Fold Cross-Validation to identify the optimal K value for each distance metric across all preprocessing scenarios, treating K optimization as a downstream task that follows data balancing. Third, we design a four-scenario ablation study that isolates and quantifies the individual contributions of SMOTE and Min-Max Normalization to model performance. The experiments are conducted on the UCI Cleveland Heart Disease Dataset [9], the most widely used benchmark for heart disease classification, enabling direct comparison with prior art.

II. *RESEARCH METHOD*

A. KNN for Medical Classification

KNN has been applied extensively to medical classification problems owing to its simplicity and transparency. In the context of heart disease prediction, several comparative studies have benchmarked KNN against other algorithms such as Support Vector Machine (SVM), Decision Tree, Random Forest, and Logistic Regression. Rahmat et al. reported KNN performance of around 89% accuracy on the UCI Cleveland dataset without specialized preprocessing [10]. Absar et al. [11] evaluated KNN alongside Random Forest, Decision Tree, and AdaBoost on the combined UCI Cleveland dataset, reporting KNN accuracy of 97.83% on the single Cleveland split, a result that reflects the high variance inherent in KNN without systematic hyperparameter tuning. A key weakness of both studies is the absence of distance metric comparison and systematic K selection: Absar et al. used Manhattan distance (p=1) only, while Rahmat et al. did not report the metric used, limiting reproducibility.

B. SMOTE for Imbalanced Medical Data

SMOTE, introduced by Chawla et al. [4], generates synthetic minority-class samples by interpolating between existing minority instances and their K nearest neighbors in the feature space. This approach has been widely adopted to address class imbalance in medical datasets. Khan et al. [5] applied SMOTE alongside Modified Artificial Bee Colony (M-ABC) feature selection and KNN for heart disease prediction on the UCI dataset, reporting that SMOTE effectively balanced the training distribution; however, K was not re-optimized after oversampling and only Euclidean distance was used, two methodological gaps addressed directly by the present study.

Irawan et al. [6] compared KNN, Logistic Regression, and Random Forest with SMOTE on the UCI Heart Failure dataset, finding that Random Forest consistently outperformed KNN, partly because KNN's hyperparameters were not tuned post-balancing. Wei and Shi [12] applied SMOTE-ENN combined with PCA and five classifiers on a coronary heart disease dataset, demonstrating that KNN recall improved dramatically after balancing, yet no distance metric comparison or ablation analysis was performed. Navita et al. [13] used SMOTE-ENN with a stacking ensemble (including KNN as a base learner) and achieved high accuracy and AUC, but KNN was not the focal classifier, and no ablation study was conducted to quantify its individual contribution.

Critically, the effectiveness of SMOTE is not unconditional. Fernández et al. [8] reviewed 15 years of SMOTE research and concluded that its benefit diminishes significantly when the imbalance ratio falls below 1.5:1. At such mild imbalance levels, synthetic samples generated by SMOTE may introduce noise into feature regions already well-represented by real data, potentially degrading classifier

performance rather than improving it. This theoretical position is directly tested in the present study's ablation design.

C. Normalization in KNN and Its Interaction with Distance Metrics

Feature normalization is a prerequisite for distance-based classifiers, and several studies have shown that Min-Max Normalization consistently improves KNN performance on tabular medical data [14]. However, the interaction between normalization and distance metric selection has received limited attention. An empirical study by Cha [15] across fifteen datasets showed that Manhattan distance often achieves performance comparable to or better than Euclidean distance, particularly for data with heterogeneous feature scales, a common characteristic of clinical tabular data.

Existing research emphasizes that one of SMOTE's main challenges is its sensitivity to noise and the potential to exacerbate class overlap by generating synthetic samples indiscriminately near noisy or borderline instances[16]. Several SMOTE variants address this by carefully selecting neighbours or filtering noise prior to oversampling, often by integrating k-nearest neighbours with noise detection or cleaning methods [17], [18]. However, most standard implementations rely on Euclidean distance for neighbour identification, which can be suboptimal in certain feature spaces or when feature distributions are anisotropic.

TABLE I
COMPARISON OF RELATED STUDIES

Study	Algorithm	Dataset	Distance Metric	K-Optimization	Ablation Study
Absar et al. (2022) [11]	KNN + RF + DT + AdaBoost	UCI Cleveland	Manhattan (p=1)	No	No
Khan et al. (2024) [5]	M-ABC + KNN + SMOTE	UCI Heart	Euclidean only	No	No
Agustina et al. (2025) [6]	KNN + LR + RF + SMOTE	UCI Heart Failure	Not Stated	No	No
Wei & Shi (2025) [12]	SMOTE-ENN + KNN + PCA	CHD dataset	No Stated	Grid Search	No
Navita et al. (2025) [13]	SMOTE-ENN + Stacking + KNN	Multi-dataset	Euclidean only	No	No

D. Research Gap

Table I summarizes the positioning of this work relative to closely related studies. As shown, no prior study has simultaneously addressed all three of the following: (1) comparative distance metric analysis, (2) systematic K optimization on SMOTE-augmented data, and (3) component-level ablation across preprocessing configurations. Prior studies tend to fix one or two of these dimensions while leaving others unexplored. The present study is the first to address all three simultaneously on the UCI Cleveland dataset, while also explicitly framing SMOTE inclusion as an empirical test of its utility under mild imbalance rather than as an assumed improvement.

III. RESEARCH METHOD

A. Research Design

The proposed research framework follows the pipeline illustrated in Figure 1, proceeding from raw data ingestion through four ablation scenarios, metric-aware K optimization, and final evaluation on a held-out test set.

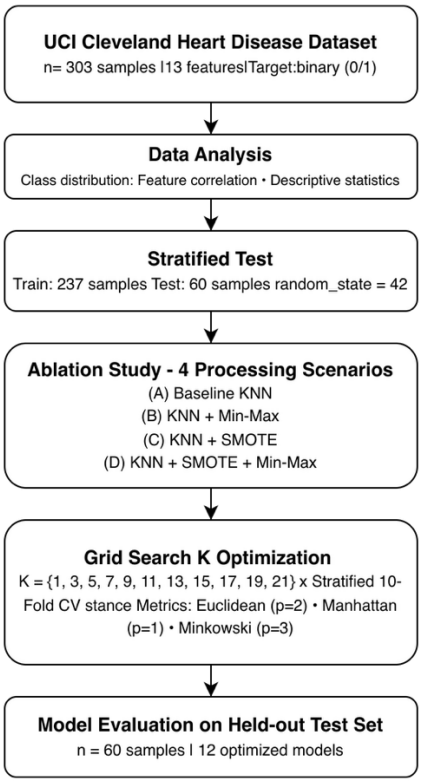


Figure 1. End-to-end research pipeline showing data flow across four ablation scenarios, Grid Search cross-validation, and test set evaluation.

The key design principle is that SMOTE is applied exclusively within the training fold to prevent data leakage: synthetic samples generated from test-set instances would artificially inflate performance metrics. Similarly, Min-Max normalization parameters (minimum and maximum values per feature) are computed solely from the training data and applied consistently to both training and test sets.

B. Dataset

Experiments were conducted on the **Cleveland Heart Disease Dataset** from the UCI Machine Learning Repository

[9]. This dataset is the most widely used benchmark for heart disease classification, enabling direct comparison with prior studies. The dataset characteristics is shown in Table II.

TABLE II
DATASET CHARACTERISTICS

Property	Value
Original samples	303
Samples after removing missing values	297
Features (input)	13 clinical attributes
Target	Binary: 0 = No disease, 1 = Disease
Class 0 (no disease)	160 samples (53.9%)
Class 1 (disease)	137 samples (46.1%)
Class imbalance ratio	1.17:1 (mild)
Train / Test split	80% / 20% (stratified)
Training samples	237 (128 class-0, 109 class-1)
Test samples	60 (32 class-0, 28 class-1)

Six samples with missing values in the *ca* and *thal* features were removed prior to analysis, resulting in a final working dataset of 297 samples. The class imbalance ratio of 1.17:1 is considered mild by the literature standard (ratios below 1.5:1), a factor with important implications for interpreting SMOTE's contribution, discussed in Section 5.

The 13 input features include both numerical and categorical attributes: age, sex, chest pain type (cp), resting blood pressure (trestbps), serum cholesterol (chol), fasting blood sugar (fbs), resting ECG result (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression (oldpeak), slope of peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), and thalassemia type (thal).

C. Min-Max Normalization

Min-Max Normalization was applied to continuous numerical features (age, trestbps, chol, thalach, oldpeak) to map all values to the [0,1] range (1).

$$x'_i = \frac{x_i - \min(X_{train})}{\max(X_{train}) - \min(X_{train})} \quad (1)$$

where $\min(X_{train})$and $\max(X_{train})$ are computed exclusively from the training subset. Binary and ordinal categorical features were left unchanged, as they already reside within an appropriate bounded range.

D. SMOTE

SMOTE was applied to the training data to address the mild class imbalance. For each minority sample x_i, a synthetic sample is generated as (2).

$$x_{new} = x_i + \lambda \cdot (\hat{x} - x_i), \quad \lambda \in [0,1] \quad (2)$$

where $\hat{x}$ is a randomly selected neighbours from the k_{SMOTE} nearest minority-class neighbours of x_i. The configuration used is: $k_{SMOTE} = 5$, sampling_strategy = 'auto' (balancing classes to equal size), and random_state = 42. After SMOTE, the training set contains 128 samples per class (256 total in Scenarios C and D).

E. KNN Configuration and Optimal K Selection

Three distance metrics were evaluated as (3-5)

- **Euclidean** $(p = 2)$: $d_E(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ (3)
- **Manhattan** $(p = 1)$: $d_M(x,y) = \sum_{i=1}^{n}|x_i - y_i|$ (4)
- **Minkowski** $(p = 3)$: $d_{Mk}(x,y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{1/p}$ (5)

For each combination of scenario and distance metric, the optimal K was determined through Grid Search over K ∈ {1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21} using Stratified 10-Fold Cross-Validation on the training data, with mean F1-score as the selection criterion. This yields 4 scenarios × 3 metrics × 11 K values = **132 configurations** evaluated in total, producing **12 optimized models** (one per scenario–metric pair).

F. Ablation Study Design

The four scenarios are defined as shown in Table III:

TABLE III
ABLATION SCENARIO DEFINITIONS

Scenario	SMOTE	Min-Max	Purpose
A	No	No	Baseline: standard KNN
B	No	Yes	Isolate effect of normalization
C	Yes	No	Isolate effect of oversampling
D	Yes	Yes	Full combination (proposed)

This design enables quantification of each component's individual contribution and their synergistic effect when combined.

G. Evaluation Metrics

Performance was assessed using accuracy (6), precision (7), Recall (8), Specificity (9), F1-Score (10) derived from the confusion matrix (TP = true positive, TN = true negative, FP = false positive, FN = false negative):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (8)$$

$$\text{Specificity} = \frac{TN}{TN+FP}, \quad (9)$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

The Area Under the ROC Curve (AUC) was also computed as a threshold-independent measure of discriminative ability. In the clinical context of heart disease prediction, Recall (sensitivity) is treated as the primary metric of clinical significance, since false negatives, failing to detect at-risk patients, carry the highest clinical cost.

IV. RESULTS AND DISCUSSION

A. Dataset Analysis

Figure 2 shows the class distribution of the dataset. The class imbalance ratio of 1.17:1 is mild, with 160 negative cases (no disease) and 137 positive cases (disease). After applying SMOTE to the training fold, the minority class (disease) is augmented to match the majority class, resulting in 128 samples per class within the training set.

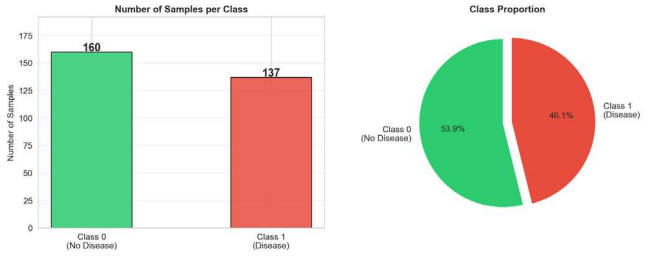


Figure 2. Class distribution of the UCI Cleveland Heart Disease Dataset before and after SMOTE oversampling

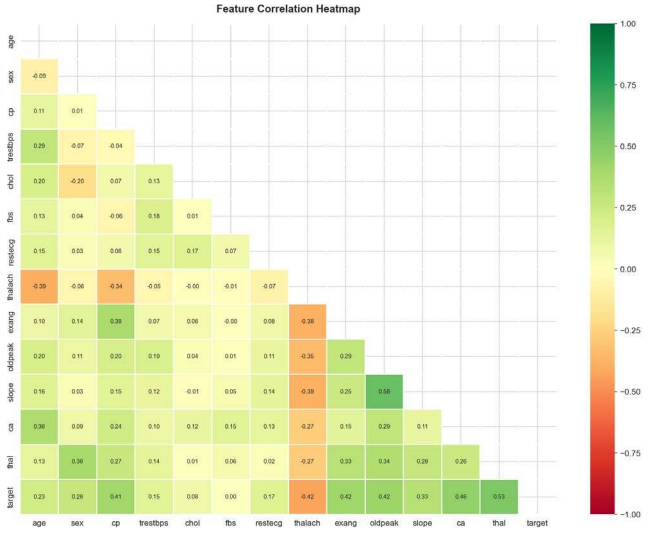


Figure 3. Correlation heatmap of all features including the target variable.

Figure 3 presents the correlation heatmap of all features against the target variable. The features most positively correlated with heart disease are *thal* (0.53), *ca* (0.46), *oldpeak* (0.42), and *exang* (0.42), while *thalach* shows the strongest negative correlation (−0.42), indicating that lower maximum heart rate is associated with higher disease likelihood.

The feature distributions by class (Figure 4) confirm that several continuous features, particularly *thalach*, *oldpeak*, and *age*, exhibit meaningful class separation, providing informative signals for distance-based classification.

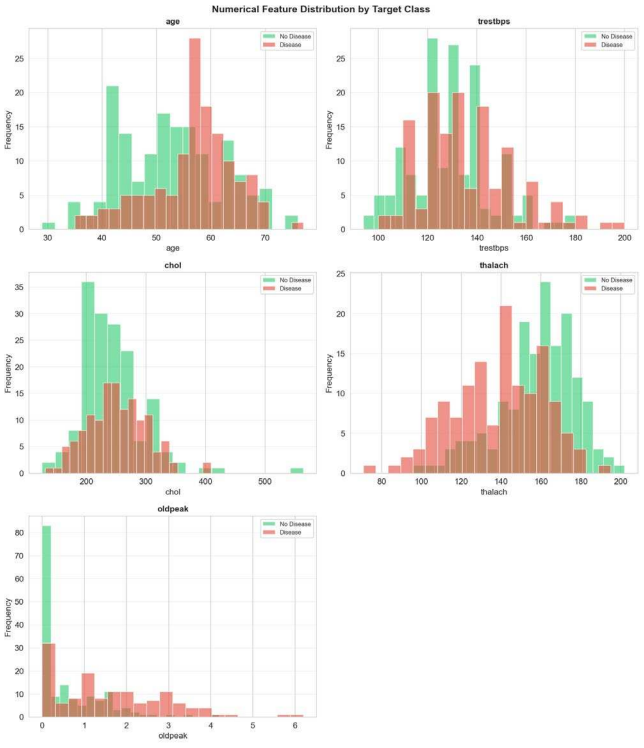


Figure 4. Distribution of continuous features partitioned by class label (0 = no disease, 1 = disease).*

B. Ablation Study Results

Table IV reports the best-performing model per scenario (selected by highest F1-score across all distance metric–K combinations). Figure 5 visualizes the comparative performance across the four scenarios.

Several noteworthy observations arise from Table IV. Scenario C (KNN + SMOTE, no normalization) achieves the highest Recall (82.1%) and F1-Score (80.7%), outperforming the fully combined Scenario D. Scenario B (KNN + Min-Max, no SMOTE) achieves the highest AUC (94.4%), suggesting that normalization most strongly improves the model's probability calibration and overall discriminative ability. The baseline Scenario A achieves the highest Specificity (90.6%) and Precision (86.9%), indicating that without oversampling, the model is more conservative, it correctly identifies most negative cases but misses more positive cases (Recall = 71.4%).

TABLE IV
PERFORMANCE TESTING RESULTS

Scenario	Best Metric	K	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Specificity (%)
A: Baseline KNN	Manhattan	7	81.7	86.9	71.4	78.4	90.6
B: KNN + Min-Max	Manhattan	11	81.7	84.0	75.0	79.2	87.5
C: KNN + SMOTE	Manhattan	9	81.7	79.3	82.1	80.7	81.2
D: KNN + SMOTE+MinMax	Minkowski (p=3)	21	80.0	83.3	71.4	76.9	87.5

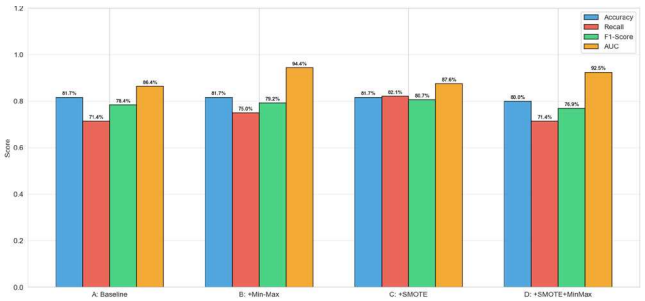


Figure 5. **Bar chart comparison of Accuracy, Recall, F1-Score, and AUC across the four ablation scenarios (best configuration per scenario).**

C. K Optimization Results

Figure 6 shows the cross-validation F1-score as a function of K for each distance metric in Scenario D. The optimal K values vary substantially across metrics.

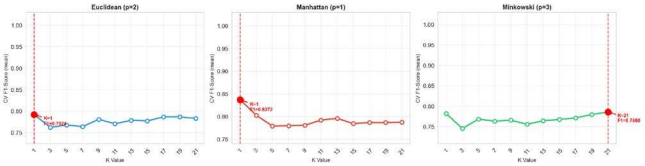


Figure 6. **Grid Search cross-validation F1-score across K values for Euclidean, Manhattan, and Minkowski (p=3) distance metrics in Scenario D (KNN + SMOTE + Min-Max).**

Optimal K value depends on both the preprocessing scenario and the distance metric, validating the motivation for scenario-specific K search rather than using a universal default. In Scenario D specifically, Euclidean and Manhattan converge at K=1 with high CV-F1 values (0.7924 and 0.8372 respectively), while Minkowski (p=3) selects K=21, the maximum tested value.

D. Distance Metric Comparison

Figure 7 present the comparison of all three distance metrics under Scenario D (the proposed combination), each evaluated at its own optimal K.

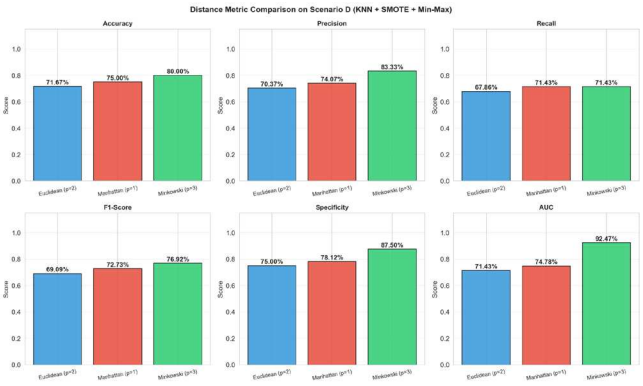
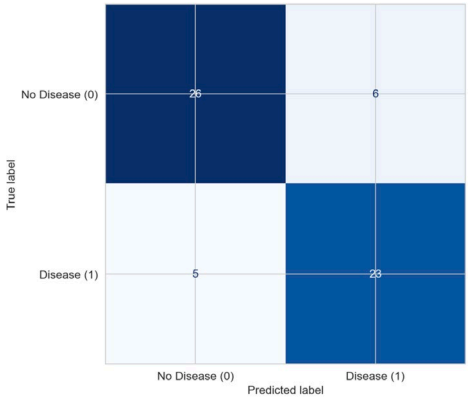
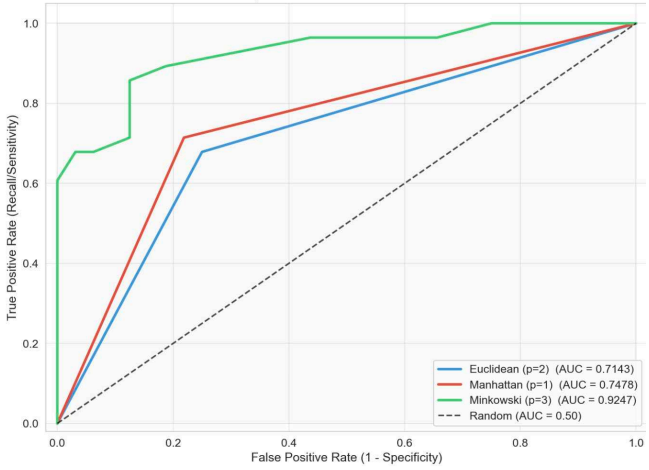


Figure 7. **Comparison of evaluation metrics across three KNN distance metrics under Scenario D (KNN + SMOTE + Min-Max).**



(a)



(b)

Figure 8. **Figure 8. (a) Confusion matrix of the best clinical model (Scenario C, Manhattan, K=9). (b) ROC curves for all three-distance metrics under Scenario D.**

Minkowski (p=3) with K=21 dominates on Accuracy, Precision, F1-Score, Specificity, and AUC. The most striking difference is in AUC: Minkowski achieves 92.47% compared to 71.43% for Euclidean, a gap of 21 percentage points. Manhattan outperforms Euclidean across all metrics but falls

substantially short of Minkowski in AUC (74.78% vs. 92.47%).

E. Best Model, Confusion Matrix and ROC Curve

The clinically recommended best model, Scenario C with Manhattan distance and K=9, is evaluated on the 60-sample test set. Its confusion matrix is shown in Figure 8(a), and the ROC curves for all Scenario D metric variants are shown in Figure 8(b).

From the confusion matrix: TP=23, TN=26, FP=6, FN=5. This yields Accuracy=81.67%, Recall=82.14%, Precision=79.31%, Specificity=81.25%, and F1-Score=80.70%. The model correctly identified 23 out of 28 true heart disease cases (5 false negatives) and correctly classified 26 out of 32 healthy patients (6 false positives).

The ROC curves in Figure 8(b) illustrate the substantial AUC advantage of Minkowski (p=3) with K=21 over the other two metrics in Scenario D. The smooth, staircase-free shape of the Minkowski ROC curve is attributable to its larger K (21 neighbors), which provides more gradual class probability estimates compared to the binary-like probabilities from K=1.

F. Comparison with Prior Art

To contextualize these results, Table V places the best-performing model from the present study alongside KNN-specific results reported in prior studies on the same UCI Cleveland dataset.

TABLE V. COMPARISON OF KNN CLASSIFICATION RESULTS ON UCI CLEVELAND HEART DISEASE DATASET

Scenario	Method	Acc (%)	Recall (%)	F1	Distance Metric	K Optimized?
Khan et al. (2024) [5]	M-ABC + KNN + SMOTE	~83*	Not reported	Not reported	Euclidean only	No
Absar et al. (2022) [11]	KNN on Cleveland only	97.83*	Not reported	Not reported	Manhattan	No
This study (Scenario C)	KNN + SMOTE + Manhattan, K=9	81.67	82.14	80.70	Manhattan	Yes (Grid Search)
This study (Scenario D)	KNN + SMOTE + MinMax + Minkowski, K=21	80.00	71.43	76.92	Minkowski (p=3)	Yes (Grid Search)

** Approximate or derived from reported confusion matrix.*

Direct comparison is constrained by differences in train-test splits, feature preprocessing, and evaluation protocols across studies. The result reported by Absar et al. (97.83%) on the single Cleveland split appears inflated relative to all other studies and is likely sensitive to the specific random split used, particularly given the small dataset size (n=297). The present study's accuracy of 81.67% (Scenario C) is consistent with the range reported by multiple independent studies using standard preprocessing on the same dataset, lending credibility to the result. The key advance of this study over prior KNN work on this dataset is not a higher accuracy number per se, but the empirical demonstration of how metric choice and K optimization interact, findings that are not captured by any single-metric, fixed-K study.

V. *DISCUSSION*

The most counterintuitive finding of this study is that Scenario C (SMOTE without normalization) achieves higher Recall and F1-Score than the fully combined Scenario D (SMOTE + Min-Max Normalization). This result is explained by the mild nature of the class imbalance in the Cleveland dataset (ratio = 1.17:1). When imbalance is severe (e.g., >5:1), SMOTE's benefit is substantial and normalization can further stabilize the distance space. However, at mild imbalance ratios, the synthetic samples generated by SMOTE occupy a feature space that already closely mirrors the real data distribution. The subsequent application of Min-Max Normalization compresses feature ranges uniformly, inadvertently reducing the distance separability between minority synthetic points and the majority class in the normalized space. This finding aligns with the theoretical position of Fernández et al. [8] and reinforces a practical guideline: SMOTE and normalization should not be applied as a blanket pipeline without first examining the dataset's imbalance ratio.

The second principal finding, that Minkowski (p=3) with K=21 achieves the highest AUC (92.47%) under Scenario D, is explained by two concurring factors. First, with p=3, the distance metric places intermediate weighting on feature differences, providing a balanced discriminative profile after Min-Max normalization. Second, K=21 produces smoother, more calibrated class probability estimates (reflecting the votes of 21 neighbours), yielding a smooth ROC curve with many operating points. This contrasts with K=1, which produces binary-like probabilities and deflates AUC. This analysis implies that Minkowski (p=3) with K=21 is the superior choice when the application requires probabilistic risk stratification rather than binary classification.

From a clinical perspective, Recall (sensitivity) is the most consequential metric, as false negatives lead to delayed treatment and adverse outcomes. The model recommended for clinical deployment is Scenario C with Manhattan distance and K=9, achieving the highest Recall (82.14%) and balanced F1-Score (80.70%). For applications requiring risk stratification rather than binary diagnosis, for instance,

prioritizing patients for cardiology consultation, the Scenario D with Minkowski (p=3) and K=21 configuration is preferred, as it provides a well-calibrated probability score reflected in its AUC of 92.47%.

Third, Manhattan distance consistently outperforms Euclidean across all four scenarios, suggesting it as a more appropriate default for KNN in clinical tabular classification. This finding is consistent with prior empirical evidence [19] and extends it to a pre-processed, over-sampled setting.

This study is subject to the following limitations: (1) experiments were conducted on a single dataset with mild imbalance and a moderate sample size (n=297), and conclusions may differ on more severely imbalanced or higher-dimensional datasets; (2) the test set size (n=60) limits the statistical precision of point estimates; and (3) SMOTE variants (Borderline-SMOTE, ADASYN) were not evaluated. Future work should validate these findings across multiple UCI heart disease sub-datasets (Hungarian, Swiss, VA Long Beach), investigate SMOTE variants in combination with different distance metrics, and explore SHAP-based explainability to provide interpretable predictions for clinical use.

VI. *CONCLUSION*

This study presented a systematic optimization framework for KNN applied to heart disease prediction, incorporating SMOTE for class imbalance handling and Min-Max Normalization for feature scaling, with explicit comparative analysis of distance metrics and data-driven K selection. Three principal findings emerge from experiments on the UCI Cleveland dataset.

First, the optimal distance metric and K value interact non-trivially: combining SMOTE and Min-Max Normalization leads Euclidean and Manhattan to converge at K=1 (indicating SMOTE-induced overfitting under normalization), while Minkowski (p=3) selects K=21, yielding substantially better AUC (92.47% vs. 71.43%). This finding empirically validates the necessity of metric-specific K optimization.

Second, on mildly imbalanced data (ratio 1.17:1), SMOTE alone (Scenario C) outperforms the fully combined pipeline (Scenario D) in terms of Recall and F1-Score. This challenges the assumption that stacking more preprocessing steps always improves performance and underscores the importance of empirically validating each component's contribution.

Third, Manhattan distance consistently outperforms Euclidean across all scenarios on this clinical dataset, suggesting it as a more appropriate default for KNN in medical tabular classification. These findings offer concrete, actionable configuration guidelines for practitioners deploying KNN in medical classification tasks.

AUTHORS CONTRIBUTION

Windra Swastika: Conceptualization, methodology design, project supervision, system architecture review, validation of results, review and editing of manuscript.

COPYRIGHT

REFERENCES

[1] Y. Li, G. Cao, W. Jing, J. Liu, and M. Liu, "Global trends and regional differences in incidence and mortality of cardiovascular disease, 1990–2019: findings from 2019 global burden of disease study," *Eur. J. Prev. Cardiol.*, vol. 30, no. 3, pp. 276–286, Feb. 2023, doi: 10.1093/eurjpc/zwac285.

[2] T. Suryati and S. Suyitno, "Prevalence and Risk Factors of the Ischemic Heart Disease in Indonesia: A Data Analysis of Indonesia Basic Health Research (RISKESDAS) 2013," *Public Health of Indonesia*, vol. 6, no. 4, pp. 138–144, Dec. 2020, doi: 10.36685/phi.v6i4.366.

[3] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[5] M. A. Khan *et al.*, "Optimal feature selection for heart disease prediction using modified Artificial Bee colony (M-ABC) and K-nearest neighbors (KNN)," *Sci. Rep.*, vol. 14, no. 1, p. 26241, Oct. 2024, doi: 10.1038/s41598-024-78021-1.

[6] C. Irawan, L. Erawan, C. Jatmoko, D. Sinaga, H. Lestiawan, and others, "Optimization of Heart Failure Classification on Imbalanced Data Using a Supervised Learning Approach Based on Logistic Regression, Random Forest, and K-Nearest Neighbor: Optimalisasi Klasifikasi Gagal Jantung pada Data Imbalanced Menggunakan Pendekatan Supervised Learning Berbasis Regresi Logistik, Random Forest, dan K-Nearest Neighbor," *Jurnal Informatika Polinema*, vol. 12, no. 1, pp. 38–45, 2025.

[7] R. Detrano *et al.*, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *Am. J. Cardiol.*, vol. 64, no. 5, pp. 304–310, Aug. 1989, doi: 10.1016/0002-9149(89)90524-9.

[8] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, Apr. 2018, doi: 10.1613/jair.1.11192.

[9] S. W. P. M. Janosi Andras and R. Detrano, "Heart Disease," 1989.

[10] D. Rahmat, A. A. Putra, Hamrin, and A. W. Setiawan, "Heart Disease Prediction Using K-Nearest Neighbor," in *2021 International Conference on Electrical Engineering and Informatics (ICEEI)*, IEEE, Oct. 2021, pp. 1–6. doi: 10.1109/ICEEI52609.2021.9611110.

[11] N. Absar *et al.*, "The Efficacy of Machine-Learning-Supported Smart System for Heart Disease Prediction," *Healthcare*, vol. 10, no. 6, p. 1137, Jun. 2022, doi: 10.3390/healthcare10061137.

[12] X. Wei and B. Shi, "Reducing bias in coronary heart disease prediction using Smote-ENN and PCA," *PLoS One*, vol. 20, no. 8, p. e0327569, Aug. 2025, doi: 10.1371/journal.pone.0327569.

[13] Navita *et al.*, "Advanced Hybrid Machine Learning Model for Accurate Detection of Cardiovascular Disease," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, p. 51, Mar. 2025, doi: 10.1007/s44196-025-00771-1.

[14] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: a review of classification and combining techniques," *Artif. Intell. Rev.*, vol. 26, no. 3, pp. 159–190, Nov. 2006, doi: 10.1007/s10462-007-9052-3.

[15] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *City*, vol. 1, no. 2, p. 1, 2007.
[16] N. A. Azhar, M. S. Mohd Pozi, A. Mohamed Din, and A. Jatowt, "An Investigation of SMOTE based Methods for Imbalanced Datasets with Data Complexity Analysis," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2022, doi: 10.1109/TKDE.2022.3179381.
[17] X. Yi, Y. Xu, Q. Hu, S. Krishnamoorthy, W. Li, and Z. Tang, "ASN-SMOTE: a synthetic minority oversampling method with adaptive qualified synthesizer selection," *Complex & Intelligent Systems*, vol. 8, no. 3, pp. 2247–2272, Jun. 2022, doi: 10.1007/s40747-021-00638-w.
[18] K. Bashir, S. Abdelwahab Ghorashi, A. Ahmed, and A. Khader, "ABL-SMOTE: A Novel Resampling Method by Handling Noisy and Borderline Challenge for Imbalanced Dataset for Software Defect Prediction," *IEEE Access*, vol. 13, pp. 78781–78794, 2025, doi: 10.1109/ACCESS.2025.3564183.
[19] H. A. Abu Alfeilat *et al.*, "Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review," *Big Data*, vol. 7, no. 4, pp. 221–248, Dec. 2019, doi: 10.1089/big.2018.0175.