Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 5, October 2025, Page. 3958-3970
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5189

Evaluating Synthetic Minority Oversampling Technique Strategies for Diabetes Mellitus Classification using K-Nearest Neighbors Algorithm

Imam Riadi*[1], Anton Yudhana[2], Gusti Chandra Kurniawan[3]

[1]Departement of Information System, Universitas Ahmad Dahlan Yogyakarta, Indonesia
[2]Departement of Electrical Engineering, Universitas Ahmad Dahlan Yogyakarta, Indonesia
[3]Master Program of Informatics, Universitas Ahmad Dahlan Yogyakarta, Indonesia

Email: [1]imam.riadi@is.uad.ac.id

Received : Jul 29, 2025; Revised : Aug 16, 2025; Accepted : Aug 16, 2025; Published : Oct 23, 2025

Abstract

Data-driven classification of Diabetes Mellitus is a crucial strategy in developing medical decision support systems that are both accurate and efficient. A major challenge in this classification task is the imbalanced class distribution, which tends to reduce the model's sensitivity to positive cases. This research utilizes a dataset of 1,000 patient medical records obtained from the Mendeley Data repository, containing clinical attributes relevant to diabetes diagnosis. This research examines the impact of various K values on the K-Nearest Neighbors (KNN) algorithm when it is combined with the SMOTE oversampling technique to enhance classification performance. The experiment employs a 10-Fold Cross-Validation methodology with five principal assessment metrics: accuracy, precision, recall, F1-score, and Area Under Curve (AUC). Compared to prior studies, this work advances the methodology by applying SMOTE within each fold of the cross-validation process, effectively preventing data leakage and improving model generalizability. Results indicate that the K=3 configuration yields the highest F1-score of 95.13% and recall of 91.83%, while the highest AUC of 96.40% is achieved at K=9 with lower sensitivity. Applying SMOTE within each fold of the cross-validation process preserves evaluation integrity and prevents potential data leakage. The model demonstrates the ability to detect positive cases more effectively while maintaining high precision. These findings highlight that combining KNN with SMOTE and proper validation strategy is a promising approach for developing a reliable early detection system for Diabetes Mellitus that is adaptive to imbalanced clinical data.

Keywords : *Cross-Validation, Diabetes Mellitus, K-Nearest Neighbors, Medical Classification, SMOTE.*

1. INTRODUCTION

The categorization of chronic illnesses like Diabetes Mellitus (DM) is pivotal in the advancement of artificial intelligence systems using medical data, especially for early identification and the prevention of enduring consequences. Diabetes is a non-communicable illness whose incidence is rising globally, posing a significant public health challenge. The early identification of diabetes not only helps prevent severe consequences in essential organs, including the heart, kidneys, and retina, but also facilitates more effective and prompt medical management [1], [2]. A machine learning-based technique serves as a strategic option for automating the illness identification and classification procedure with more precision and measurability.

The imbalance in class distribution within medical datasets, particularly between the number of patients diagnosed as positive and negative for diabetes, is a major constraint in developing reliable classification models. Disproportionate data leads algorithms to tend to ignore minority classes that are crucial to identify, such as patients with pre-diabetes or early-stage diabetes. Model performance often shows high overall accuracy values but fails to detect minority cases that have a significant impact in a clinical context [3], [4]. This challenge is driving researchers to seek solutions that can improve the representation of minority classes without disrupting the overall data distribution.

A prevalent way of mitigating class imbalance is through data synthesis-based resampling techniques, with the Synthetic Minority Oversampling Technique (SMOTE) being the most renowned option. SMOTE operates by creating synthetic data derived from the difference vectors between minority data points and their neighbors, thereby enhancing the diversity of the minority class's representation while adhering to a legitimate feature space [5]. SMOTE has demonstrated efficacy in enhancing model sensitivity towards the minority class and equilibrating learning proportions during model training [6].

The K-Nearest Neighbor (KNN) algorithm is one of the classification methods consistently used in various studies related to disease diagnosis, including diabetes. Its distance-based nature makes it highly intuitive in classifying new data but also makes it very sensitive to uneven data distribution [5]. When the training data is dominated by the majority class, the model's predictions tend to favor that class even if there are unique characteristics of the minority class [7], [8]. Balancing strategies like SMOTE become relevant for improving KNN performance in the context of imbalanced datasets.

Much earlier research has shown that using SMOTE can greatly enhance how well the KNN model performs, particularly in recall and F1-score [9]. These two metrics are crucial in evaluating medical classification because they reflect the model's ability to recognize true positive cases and balance precision and sensitivity [10], [11]. The combination of SMOTE and KNN is one of the promising approaches for automatic and data-driven early detection of diabetes.

Previous research has looked at different ways to balance data, like SMOTE, Borderline-SMOTE, and ADASYN, and used them with classification methods such as Support Vector Machine (SVM), Random Forest, and KNN. The results show that balancing the data is crucial in the classification process, especially for diseases with uneven class distributions like diabetes [12], [13]. These studies provide a strong foundation for developing an SMOTE-based approach as an effective pre-processing technique.

As medical data becomes more complicated, different methods have been created to make classification models work better, including using SMOTE along with optimization techniques like Particle Swarm Optimization (PSO) and Grey Wolf Optimization (GWO). This combination aims to balance the data and optimize feature selection to make the model more accurate and efficient in diagnosing diabetes [6], [14]. This approach reflects ongoing efforts to produce a statistically effective and practically implementable classification system.

While SMOTE provides tangible benefits in improving the representation of minority classes, it still presents technical challenges that need careful consideration. Some common challenges include the risk of overfitting, blurring of class boundaries, and increased sensitivity to noise, which can hinder model performance. This challenge becomes increasingly complex when SMOTE is combined with distance-based algorithms like KNN, which are highly dependent on the consistency of data structure [15], [16]. Therefore, a comprehensive evaluation of this combination of the two methods is highly necessary.

This research seeks to thoroughly assess the SMOTE oversampling method for classifying Diabetes Mellitus with the K-Nearest Neighbor algorithm. The main focus of the evaluation lies in measuring performance metrics such as accuracy, precision, recall, and F1-score, as well as how the balancing process affects the model's ability to accurately recognize minority classes. Evaluation was performed on various values of k to identify the optimal configuration [17], [18]. The main contribution of this research is to show real-world proof that SMOTE helps KNN-based classification work better with unbalanced medical datasets. This research also presents a comprehensive analysis of the limitations and potential for applying the model in medical decision support systems. The research results are expected to support the creation of a machine learning system that can detect diabetes early, making it more flexible, effective, and easier to use in real life.

2. METHOD

This research follows the Knowledge Discovery in Databases (KDD) framework, consisting of data understanding, preprocessing, class balancing, modeling, and evaluation [19]. Figure 1 illustrates the research flow. The entire experimental process was implemented using RapidMiner Studio 10.1, which allows for the modular and visual design of the experimental flow, including integration between data balancing, cross-validation, and classification processes.

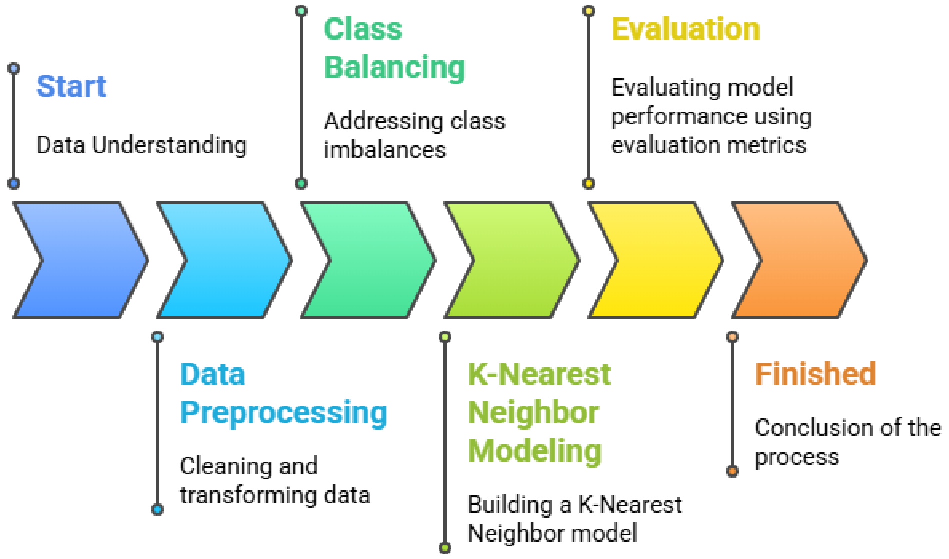
Start
Data Understanding
Data Preprocessing
Cleaning and transforming data
Class Balancing
Addressing class imbalances
K-Nearest Neighbor Modeling
Building a K-Nearest Neighbor model
Evaluation
Evaluating model performance using evaluation metrics
Finished
Conclusion of the process

Figure 1. Research Workflow Based on the KDD Framework, Showing Sequential Stages From Data Understanding to Model Evaluation

As seen in Figure 1, the workflow starts with data comprehension to ascertain attribute features and probable anomalies within the dataset. The preprocessing phase tackles absent values and rectifies anomalies to guarantee data integrity. The class balancing phase utilizes the Synthetic Minority Oversampling Technique (SMOTE) to address uneven class distribution. Modeling is conducted utilizing the K-Nearest Neighbors (KNN) technique with diverse setups of the K parameter. The assessment phase employs a 10-Fold Cross-Validation method to assess performance with five metrics: accuracy, precision, recall, F1-score, and AUC. This systematic technique guarantees methodological uniformity, mitigates data loss, and improves the dependability of experimental outcomes.

2.1. Dataset

This research utilizes a dataset obtained from the Mendeley Data platform titled "Diabetes Dataset," published by the University of Information Technology and Communications, Iraq [20]. The dataset consists of 1,000 patient medical record entries reflecting clinical characteristics relevant to the diagnosis of Diabetes Mellitus. Overall, there are 14 attributes in this dataset, including one target attribute (diagnosis class). Input attributes include physiological information and laboratory test results, such as age, blood glucose levels, blood pressure, body mass index, and other supporting variables commonly used in diabetes risk analysis.

The target attribute in this dataset is divided into three classes: Y (Yes) with 844 samples, N (No) with 103 samples, and P (Probable) with 53 samples. This distribution shows a significant class imbalance, with class Y dominating as the representation of patients diagnosed with diabetes. This condition has the potential to cause bias in the model training process, as algorithms tend to prioritize predictions on the majority class [21].

To simplify the class structure and improve the model's stability in recognizing minorities, classes P and N were merged into a single class representing the non-diabetic category. This approach refers to the technique of aggregating minority classes, which is commonly applied in medical classification research with the aim of minimizing noise and reducing the risk of underfitting in minority classes [22], [23]. After the merging process was carried out, the dataset was converted to a binary format with the final distribution: Y (Diabetes) with 844 samples and N (Non-diabetes) with 156 samples. Despite class consolidation, the imbalance in distribution between the majority and minority classes remains high. Data balancing methods based on oversampling, such as SMOTE, will be applied.

2.2. Data Preprocessing

The data preprocessing stage is carried out systematically to ensure the quality, consistency, and reliability of the dataset before it is used in training the classification model [24], [25]. This process includes two main stages: handling missing values and correcting outliers. These two stages are designed to improve data representation to ensure it remains valid in a medical context and to minimize potential distortions to model performance.

The process of handling missing values was carried out because the initial dataset showed missing values in several attributes. To handle this, an imputation strategy based on data type was applied. Missing values in numerical attributes were filled with the mean, while the mode was used for categorical attributes. This process is performed automatically using the Replace Missing Values operator available in the RapidMiner software. This approach was chosen to maintain the statistical integrity of the data distribution without the need to remove data rows that contain informative values, which, if removed, could actually reduce the complexity of the clinical representation.

Attention is also directed at extreme values that could potentially be outliers, especially for clinical attributes such as urea, creatinine, HbA1c, cholesterol, and triglycerides. Outlier detection is performed using the interquartile method, which identifies data points that fall below the first quartile (Q1) or above the third quartile (Q3). To correct these extreme values, the winsorizing technique is used, which involves limiting values outside the threshold to keep them within a clinically relevant range [26].

Some examples of outlier correction implementation include creatinine values exceeding 133 µmol/L being adjusted to that threshold, HbA1c values less than 3% being replaced with the median value from all data, and urea values below 2.0 mmol/L being corrected to 4.6 mmol/L (median), while values above 20.0 mmol/L are adjusted to that upper limit, reflecting severe kidney failure. The winsorizing approach was chosen because it preserves the original distribution structure of the data better than the deletion method, and it is also widely used in medical classification studies to maintain model stability and improve its ability to generalize to new data [27], [28].

2.3. Class Balancing

Unbalanced class distribution is a crucial challenge in medical classification systems, including cases of Diabetes Mellitus. In the dataset used, the positive class (diabetics) dominates approximately 84% of the total samples, while the negative class (non-diabetics) only accounts for 16%. This imbalance in proportions risks introducing prediction bias towards the majority class and reducing the model's sensitivity to the minority class.

The Synthetic Minority Over-sampling Technique (SMOTE) was utilized to equilibrate the class distribution in the training dataset. This approach produces synthetic data by interpolating between examples of the minority class and their neighboring instances, rather than merely duplicating them.

Enhanced minority representation can fortify the parameters of model judgments. Prior research indicates that SMOTE can enhance sensitivity and AUC without compromising model accuracy [7], [29].

SMOTE was applied exclusively to the training subset within the 10-Fold Cross-Validation process to balance the class distribution while preserving the original structure of the test subset. This design prevents information leakage and maintains evaluation validity. Figure 2 illustrates the placement of SMOTE in the overall experimental workflow [10], [30]. Figure 2 illustrates the technical structure of SMOTE placement. This visualization shows that the balancing process is performed before the model is constructed, and then the results are passed to the testing process without being contaminated by synthetic data.

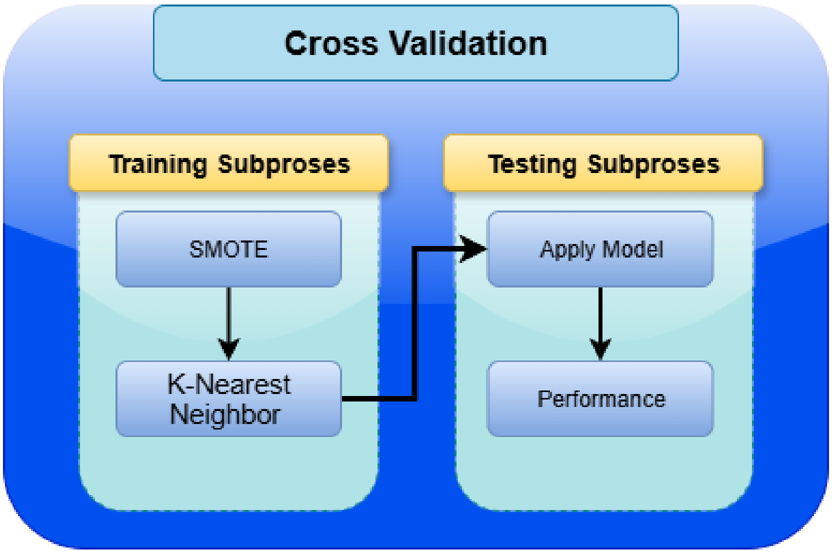

Figure 2. Diagram of SMOTE Placement in Cross-Validation Scheme

As illustrated in Figure 2, each fold begins by splitting the dataset into training and test subsets. The training subset is balanced using the Synthetic Minority Oversampling Technique (SMOTE) before model training with the K-Nearest Neighbors (KNN) algorithm. The trained model is subsequently evaluated on the untouched test subset. This process is repeated for all folds, and the resulting performance metrics are averaged to obtain a robust and unbiased estimate.

2.4. Implementation of the K-Nearest Neighbor Algorithm

The K-Nearest Neighbor (KNN) algorithm is a type of classification method that predicts the category of new data by examining the most common category among the K closest data points. Unlike other algorithms that create a specific model during training, KNN uses all the training data as a guide for making class predictions, making it a lazy learner and a non-parametric method [31], [32].

In this research, KNN was applied with three variations of the parameter k value, namely 3, 5, 7, and 9. The purpose of testing these four values is to determine the most optimal configuration k in the context of Diabetes Mellitus classification. A value of k that is too small can cause the model to be overly sensitive to noise, while a value of k that is too large can lead to overgeneralization (underfitting). Evaluation is conducted to balance the model's sensitivity and its ability to generalize to new data.

To measure the proximity between data, the Euclidean Distance metric, which is commonly used in numerical classification, is employed. This distance is considered effective in capturing similarity patterns in multidimensional feature space, especially since all medical attributes in the dataset have undergone a normalization process. The classification system is binary, meaning it has two categories:

Y (Diabetes) and N (Non-diabetes). This matches the results from the preprocessing stage, where the label "P" (probable) was combined with the negative class to make the classification simpler.

The choice of the KNN algorithm in this research is based on its intuitive, simple, and efficient characteristics for low- to medium-dimensional datasets. KNN is also known to have competitive performance in medical diagnosis tasks, especially when combined with balancing techniques such as SMOTE [11], [5]. Additionally, this algorithm is relatively robust to noise after undergoing preprocessing steps that include normalization and extreme value correction, making it a viable option for diabetes classification cases.

2.5. Cross Validation

This research implements the 10-fold cross-validation methodology. This method divides the dataset into 10 approximately equal-sized subgroups, known as folds. The training dataset is comprised of the remaining nine folds, while one fold serves as the test dataset in each cycle. This approach is performed ten times to guarantee that each subset has an equal probability of being utilized as test data. The conclusive evaluation results are based on the mean of all performance indicators calculated during each cycle.

The selection of the 10-Fold Cross Validation method over more rudimentary validation techniques, such as holdout (e.g., 80:20 or 70:30 splits), is predicated on many methodological factors. This method mitigates the danger of overfitting by validating the model against diverse combinations of test data. This method offers a more reliable and representative assessment, particularly for datasets characterized by restricted size and imbalance. Third, 10-Fold Cross Validation optimizes data use since each instance is employed for both training and testing over various iterations [33], [34].

This validation technique is widely recognized as the standard in medical classification studies and has been shown to improve the reliability of evaluations, particularly when combined with class balancing methods such as SMOTE [3], [5]. The application of cross-validation in this experiment provides a more realistic picture of model performance, reduces the risk of bias, and ensures the stability of the evaluation. This strategy is crucial for maintaining consistent performance in the medical classification context when applying the model to new patient data.

2.6. Evaluate Model Performance

The classification model's performance was assessed using five primary metrics: Accuracy, Precision, Recall, F1-Score, and Area Under Curve (AUC). The choice of these measures takes into account the attributes of the unbalanced dataset and the significance of the medical setting, where inaccuracies in identifying positive cases (diabetics) might have severe consequences for clinical care [29].

Accuracy measures the proportion of correct total predictions against the entire sample and is calculated using Equation (1).

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \qquad (1)$$

Precision indicates the proportion of positive predictions that are actually positive cases, and it is important to avoid misdiagnosis in healthy individuals. It is calculated using equation (2).

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (2)$$

Recall, sometimes referred to as Sensitivity or True Positive Rate, assesses the model's capacity to identify all positive instances and is computed using equation (3). This statistic is essential for applications where the identification of positive instances is prioritized over total accuracy, such as in medical diagnosis. A high recall signifies that the model is proficient in reducing false negatives, ensuring that the majority of positives are accurate.

$$Recall \;= \frac{TP}{TP+FN} \qquad (3)$$

The F1-Score is the harmonic mean of accuracy and recall, pertinent when a balance between the two is required, and it is computed using equation (4). This statistic is especially beneficial in scenarios with uneven class distribution, since it offers a more thorough assessment of a model's performance. The F1-Score emphasizes both false positives and false negatives, ensuring the model's accuracy and reliability in recognizing pertinent events.

$$F1 - Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \qquad (4)$$

The Area Under the Curve (AUC) quantifies a model's efficacy in differentiating between positive and negative classes. AUC values closer to 1 indicate superior discriminating capability of the model between the two classes. These five criteria are utilized collectively to provide a comprehensive evaluation of the classification model's efficacy, especially for unbalanced medical data, as shown by Diabetes Mellitus.

Description:

- True Positive (TP) : A positive instance that is accurately identified.
- False Positive (FP) : A negative instance erroneously classified as positive.
- True Negative (TN) : An accurately forecasted negative instance.
- False Negative (FN) : A positive instance erroneously classified as negative.

3. RESULT

Experiments were performed to assess the efficacy of the K-Nearest Neighbor (KNN) algorithm in identifying Diabetes Mellitus using unbalanced data. The Synthetic Minority Oversampling Technique (SMOTE) is employed to address the imbalance, implemented within a 10-Fold Cross Validation framework to avert data leakage. The evaluated values for the parameter K were 3, 5, 7, and 9, utilizing five primary assessment metrics: accuracy, precision, recall, F1-score, and Area Under Curve (AUC). Table 1 displays the comprehensive results of the tests performed for each value of K.

Table 1. Performance Evaluation of KNN Model with SMOTE and 10-Fold Cross Validation

	Accuracy %	Precision %	Recall %	F1-Score %	AUC %
K=3	92.10	98.75	91.83	95.13	94.40
K=5	91.80	98.88	91.35	94.93	95.50
K=7	91.40	98.51	91.24	94.70	96.20
K=9	90.70	98.73	90.17	94.22	96.40

All of the k value combinations in Table 1 demonstrate relatively consistent accuracy (ranging from 90.70% to 92.10%) and high precision (exceeding 98%), indicating a generally strong classification capability. Notable variations are observed in recall and F1-score, suggesting that the model’s sensitivity to the positive class and the balance between precision and recall are influenced by the selected k value. These fluctuations highlight that tuning the k parameter can significantly affect the model’s ability to detect minority cases, which is critical in medical applications. This indicates the need for targeted optimization, such as grid search or refined cross-validation, to identify parameter configurations that maximize recall without sacrificing overall performance. Understanding these trade-offs allows researchers to align parameter selection with the specific diagnostic priorities of the application domain.

The configuration with k = 3 yields the highest recall and F1-score compared to other configurations, indicating its ability to recognize minority data more effectively. Although the AUC value increased gradually and reached a maximum at k = 9, this improvement was not accompanied by a rise in recall or F1-score. The configuration k = 3 can be considered optimal because it maintains a balance between precision and sensitivity while showing stable classification performance across all tested configurations. This performance pattern can be attributed to the inherent characteristics of the K-Nearest Neighbors (KNN) algorithm. Lower k values make the model more responsive to local patterns in the feature space, enabling better detection of minority class instances and resulting in higher recall and F1-scores. This sensitivity also increases susceptibility to noise. Higher k values aggregate decisions over broader neighborhoods, smoothing class boundaries and improving overall class separation as reflected in higher AUC scores. While this broader generalization benefits discrimination across all classes, it can reduce sensitivity to rare or borderline minority cases.

This pattern indicates that a global increase in discriminatory ability does not always correspond to improved detection of minority groups. Model configuration should not rely solely on AUC but should also consider metrics more sensitive to data imbalance, such as recall and F1-score. This comprehensive approach ensures that the model remains both accurate and equitable across different classes. Prioritizing these additional metrics allows for a more complex understanding of model performance and its real-world impact. In healthcare, reduced sensitivity can lead to missed diagnoses, resulting in delayed treatment and higher risks of complications. Therefore, implementing strategies to address these disparities can enhance the overall effectiveness and fairness of predictive models. To complement the numerical results in Table 1, Figure 3 presents a comparative visualization in the form of a line graph, enabling a clearer view of the performance trade-offs across different k values.

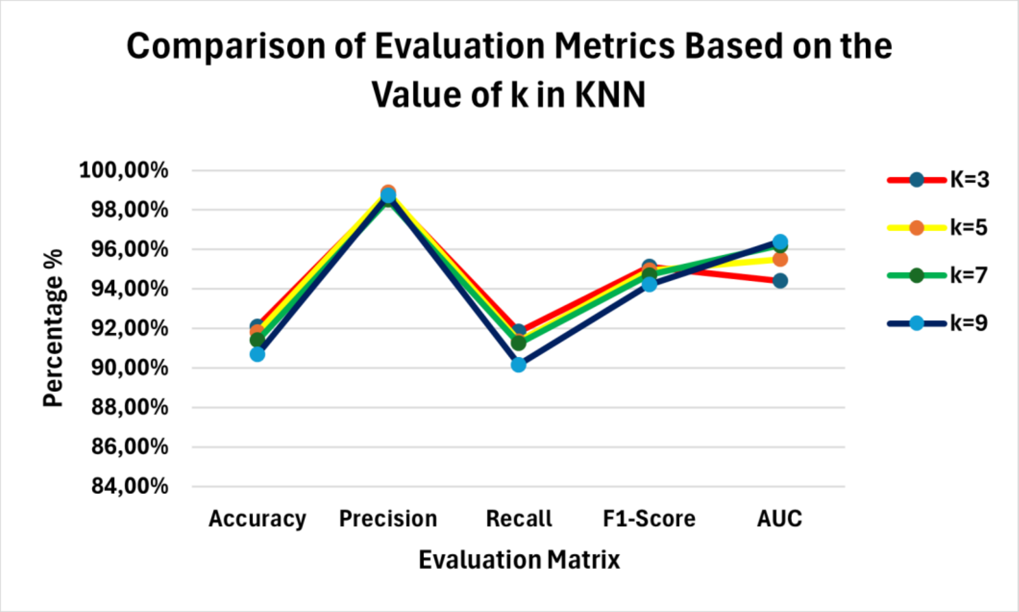

Figure 3. Comparison of Evaluation Metrics Based on the Value of k in KNN

The visual trends in Figure 3 reinforce the trade-offs identified in the numerical results. Lower k values are positioned higher on recall and F1-score, reflecting superior sensitivity to the positive class, which is critical for minimizing false negatives in medical diagnosis. Higher k values, while offering incremental gains in AUC, show a gradual decline in recall, indicating a reduced ability to detect minority cases. This visual interpretation highlights the importance of aligning parameter selection with diagnostic priorities, ensuring that the chosen configuration supports both accuracy and equitable class representation. Figure 4 below displays a confusion matrix that illustrates the quantity of accurate and erroneous predictions for each class, offering a comprehensive overview of the classification results distribution in the optimal configuration, K=3.

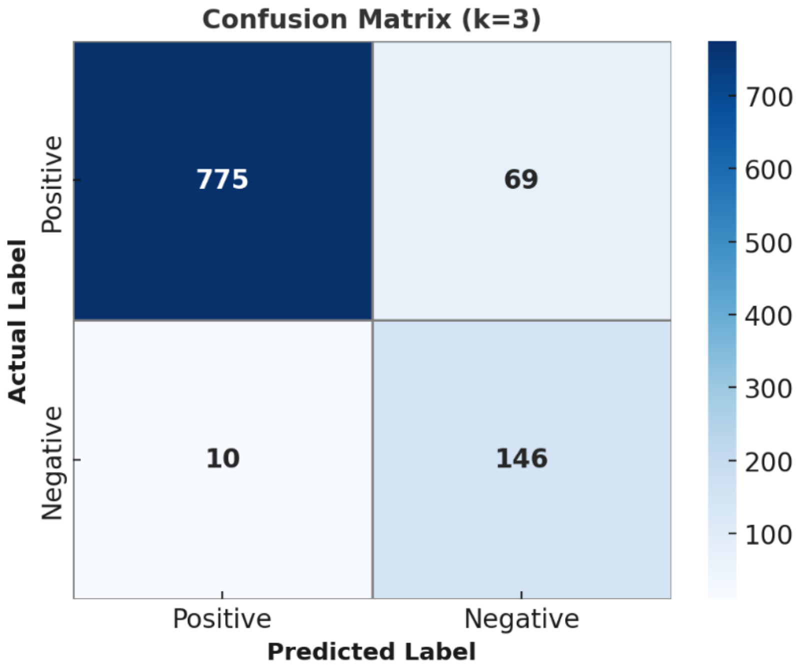

Figure 4. KNN confusion matrix for K=3 configuration value

Figure 4 presents the confusion matrix for the best configuration, which is the KNN model with a K value of 3. This matrix shows the number of true and false predictions for two main classes, Positive and Negative. The model successfully classified 775 positive data points correctly (True Positives), while 69 positive data points were misclassified as negative (False Negatives). There were 146 negative data points that were correctly identified (True Negatives), and only 10 negative data points that were misclassified as positive (False Positives).

The color distribution on this graph also illustrates the quantitative proportion of each matrix cell, with darker blue indicating a larger number. The high number of True Positives and True Negatives confirms that the model at K=3 is not only sensitive to positive cases but also has outstanding specificity in avoiding misclassification errors in the negative class. This image provides a clear visual representation of the model's classification performance while also reinforcing the rationale for selecting the K=3 configuration as the most optimal.

4. DISCUSSIONS

This research's primary conclusion reveals that altering the parameter K in the K-Nearest Neighbor (KNN) algorithm subsequent to the use of the Synthetic Minority Oversampling Technique (SMOTE) significantly affects diabetes classification performance. Configuration k = 3 had the maximum performance in the F1-score (95.13%) and recall (91.83%) measures, signifying that the model exhibits more sensitivity in identifying positive instances. Configuration k = 9 exhibited the best AUC (96.40%), but with a reduction in recall, signifying a trade-off between the model's discriminative capacity and sensitivity to the minority class.

The performance differences between configurations confirm the role of the k value as a controlling factor in balancing precision and recall in KNN-based classification. A smaller k value allows the model to be more responsive to local patterns but can increase vulnerability to noise. Conversely, a larger k value results in more conservative classifications but risks overlooking minority cases. In the context of diagnosing chronic diseases like diabetes, this condition becomes extremely crucial because misclassifying a false negative can have serious implications for delayed treatment and an increased risk of complications.

These results substantially expand upon the findings reported by Susanto and Ismanto (2025), who found that KNN without balancing only yielded a recall of 68.7% in diabetes classification. Data imbalance causes the model to be biased towards the majority class, leading it to fail to recognize most positive cases. This research confirms that integrating SMOTE methodologically, i.e., only on a subset of the training data within each cross-validation fold, is an effective approach for improving sensitivity without sacrificing the validity of model evaluation [35].

The model's performance in this study was also consistently higher than the results reported by Mohammed (2024), who noted a recall of 82% and an AUC of 88% after applying SMOTE. The absence of cross-validation in previous experimental designs could potentially lead to overoptimistic evaluations. This study implements stratified cross-validation combined with internal SMOTE, thus avoiding data leakage and producing more generalizable evaluations against real-world data [36].

In the context of the experimental structure, the approach applied aligns with the principle suggested by Abushahla and Pala (2024), which emphasizes the importance of harmoniously integrating balancing and cross-validation in the design of medical classification. Although the model they proposed is based on deep learning, the principle of caution regarding class distribution and data independence remains a foundation applicable across algorithms. This research not only applies that principle but also provides a replicable framework with a high degree of experimental control [7].

Pratap and Singh (2023) claim that SMOTE can improve classification performance in diabetes diagnosis. However, it is unclear whether this balancing method is used internally within the cross-validation process or applied to the entire dataset before data splitting. Without this clarification, the potential for information leakage from training data to test data remains a risk. This research systematically addresses this issue, demonstrating that controlling the learning process through the appropriate placement of SMOTE significantly impacts the reliability of the evaluation [11].

The empirical contribution of this research is strengthened by the research by Arsyadani and Purwinarko (2023), which showed that applying SMOTE was able to improve the F1-score and recall in diabetes classification. However, the study has not systematically explored the influence of varying the value of k on KNN. This research fills that gap and shows that parameter selection in the basic algorithm can directly affect model sensitivity [5].

This research makes a conceptual and methodological contribution by integrating balancing strategies and model parameter selection within a disciplined cross-validation framework. The evaluation was conducted on pure test data that was not contaminated by synthetic data from the SMOTE process. The results obtained show that the KNN model can remain competitive in the context of imbalanced data if accompanied by the appropriate parameter configuration and validation design.

The implications of these findings are not only relevant within an academic framework but also have the potential to be adopted into clinical decision support systems. Models with high sensitivity, maintained precision, and a solid validation process provide a strong foundation for integration into machine learning-based diagnostic applications that are safe, accurate, and feasible for real-world implementation.

This research is subject to numerous constraints, despite the promising outcomes.The dataset is limited, consisting of 1,000 patient records from a singular source, thereby constraining the model's generalizability to wider and more heterogeneous populations. The utilized characteristics are confined to those included in the dataset, perhaps excluding additional pertinent clinical signs. The present assessment exclusively addresses binary classification, hence reducing the intricacy of actual clinical diagnoses where multi-class situations and comorbidities frequently occur. Incorporating these criteria in future studies may enhance the model's applicability and reliability across different healthcare settings. Future endeavors will concentrate on augmenting the dataset with multi-center and longitudinal data, investigating supplementary machine learning algorithms, and using sophisticated feature selection techniques to further improve classification efficacy and resilience in practical clinical environments.

5. CONCLUSION

This research makes a conceptual and methodological contribution by integrating balancing strategies and model parameter selection within a disciplined cross-validation framework. The evaluation was conducted on pure test data that was not contaminated by synthetic data from the SMOTE process. The results obtained show that the KNN model can remain competitive in the context of imbalanced data if accompanied by the appropriate parameter configuration and validation design. These findings are relevant in academia and could be used in clinical decision support systems. A model with high sensitivity, maintained precision, and a solid validation process provides a strong foundation for integration into secure, accurate, and real-world deployable machine learning-based diagnostic applications. This research indicates that the selection of the K value in the K-Nearest Neighbor (KNN) algorithm significantly impacts the performance of Diabetes Mellitus classification, particularly when combined with the SMOTE balancing technique. The configuration K=3 yields the best results based on an F1-score of 95.13% and recall of 91.83%, reflecting the model's ability to detect positive classes in a balanced and accurate manner. Although the highest AUC value of 96.40% was obtained at K=9, this configuration was not accompanied by an increase in sensitivity, indicating a trade-off that needs to be considered. The use of SMOTE in-fold during cross-validation has also been shown to effectively maintain evaluation validity and prevent data leakage, which makes this approach suitable for implementation in data-driven clinical decision support systems. For future work, expanding the dataset with multi-center and longitudinal data, incorporating additional clinical features, and exploring other classification algorithms or hybrid approaches are recommended to further improve model robustness, generalizability, and applicability in diverse healthcare environments.

REFERENCES

[1] A. I. ElSeddawy, F. K. Karim, A. M. Hussein, and D. S. Khafaga, “Predictive Analysis of Diabetes-Risk with Class Imbalance,” *Computational Intelligence and Neuroscience.*, vol. 2022, pp. 1–16, Oct. 2022, doi: 10.1155/2022/3078025.

[2] M. N. Abdullah and Y. B. Wah, “Improving Diabetes Mellitus Prediction with MICE and SMOTE for Imbalanced Data,” in *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, IEEE, Sep. 2022, pp. 209–214. doi: 0.1109/AiDAS56890.2022.9918773.

[3] A. Wibowo, A. F. N. Masruriyah, and S. Rahmawati, “Refining Diabetes Diagnosis Models: The Impact of SMOTE on SVM, Logistic Regression, and Naïve Bayes,” *Journal of Electronics, Electromedical Engineering, and Medical Informatics.*, vol. 7, no. 1, pp. 197–207, Jan. 2025, doi: 10.35882/jeeemi.v7i1.596.

[4] A. J. Mohammed, M. M. Hassan, and D. H. Kadir, “Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method,” *International Journal of Advanced Trends in Computer Science and Engineering.*, vol. 9, no. 3, pp. 3161–3172, Jun. 2020, doi: 10.30534/ijatcse/2020/104932020.

[5] F. Arsyadani and A. Purwinarko, “Implementation of Synthetic Minority Oversampling Technique and Two-phase Mutation Grey Wolf Optimization on Early Diagnosis of Diabetes using K-Nearest Neighbors,” *Recursive Journal of Informatics*, vol. 1, no. 1, pp. 9–17, Mar.

2023, doi: 10.15294/rji.v1i1.64406.

[6] D. R. Damayanti and A. Purwinarko, "Application of C4.5 Algorithm Using Synthetic Minority Oversampling Technique (SMOTE) and Particle Swarm Optimization (PSO) for Diabetes Prediction," *Recursive Journal of Informatics*, vol. 2, no. 1, pp. 18–27, Mar. 2024, doi: 10.15294/rji.v2i1.64928.

[7] K. H. Abushahla and M. A. Pala, "Optimizing Diabetes Prediction: Addressing Data Imbalance with Machine Learning Algorithms," *ADBA Computer Science.*, Jul. 2024, doi: 10.69882/adba.cs.2024075.

[8] R. Taher, S. H. Basha, and A. Abdalla, "Improving Machine Learning Techniques with Imbalanced Data Treatment for Predicting Diabetes," in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 184, Springer Science and Business Media Deutschland GmbH, 2023, pp. 380–391. doi: 10.1007/978-3-031-43247-7_34.

[9] I. Leguen-de-Varona, J. Madera, H. Gonzalez, L. Tubex, and T. Verdonck, "Oversampling Method Based Covariance Matrix Estimation in High-Dimensional Imbalanced Classification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14335 LNCS, Springer Science and Business Media Deutschland GmbH, 2024, pp. 16–23. doi: 10.1007/978-3-031-49552-6_2.

[10] A. Hashmi, M. T. Nafis, S. Naaz, and I. Hussain, "Comparative Analysis of Resampling Techniques and Machine Learning Classifiers in Multiclass Classification of Diabetes Mellitus," in *2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, IEEE, Oct. 2023, pp. 230–238. doi: 10.1109/ICSSAS57918.2023.10331822.

[11] V. Pratap and A. P. Singh, "A Comparative Analysis of Classification Methods Using Oversampling Methods for Diabetes Dataset," in *2023 3rd International Conference on Advancement in Electronics & Communication Engineering (AECE)*, IEEE, Nov. 2023, pp. 921–926. doi: 10.1109/AECE59614.2023.10428438.

[12] A. Prastyo, S. Sutikno, and K. Khadijah, "Improving support vector machine and backpropagation performance for diabetes mellitus classification," *Computer Science and Information Technology.*, vol. 5, no. 2, pp. 140–149, Jul. 2024, doi: 10.11591/csit.v5i2.p140-149.

[13] N. M. Nayan, A. Islam, M. U. Islam, E. Ahmed, M. M. Hossain, and M. Z. Alam, "SMOTE Oversampling and Near Miss Undersampling Based Diabetes Diagnosis from Imbalanced Dataset with XAI Visualization," in *2023 IEEE Symposium on Computers and Communications (ISCC)*, IEEE, Jul. 2023, pp. 1–6. doi: 10.1109/ISCC58397.2023.10218281.

[14] Q. Dong and W. Lu, "Imbalance Data Classification Method Based on Improved SMOTE Algorithm and Granular Computing," in *2022 41st Chinese Control Conference (CCC)*, IEEE, Jul. 2022, pp. 3196–3201. doi: 10.23919/CCC55666.2022.9902406.

[15] H. A. Gameng, B. B. Gerardo, and R. P. Medina, "Modified Adaptive Synthetic SMOTE to Improve Classification Performance in Imbalanced Datasets," in *2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, IEEE, Dec. 2019, pp. 1–5. doi: 10.1109/ICETAS48360.2019.9117287.

[16] N. Cahyana, S. Khomsah, and agus sasmito Aribowo, "Improving Imbalanced Dataset Classification Using Oversampling and Gradient Boosting," in *2019 5th International Conference on Science in Information Technology (ICSITech)*, IEEE, Oct. 2019, pp. 217–222. doi: 10.1109/ICSITech46713.2019.8987499.

[17] F. Mesquita, J. Mauricio, and G. Marques, "Oversampling Techniques for Diabetes Classification: a Comparative Study," in *2021 International Conference on e-Health and Bioengineering (EHB)*, IEEE, Nov. 2021, pp. 1–6. doi: 10.1109/EHB52898.2021.9657542.

[18] N. Sigeef, "An Oversampling Algorithm combining SMOTE and RF for Imbalanced Medical Data," *International Journal of Research in Applied Science and Engineering Technology.*, vol. 11, no. 6, pp. 2429–2434, Jun. 2023, doi: 10.22214/ijraset.2023.54074.

[19] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *Journal of Engineering and Applied Sciences.*, vol. 12, no. 16, pp. 4102–4107, 2017, doi: 10.3923/jeasci.2017.4102.4107.

[20] A. Rashid, "Diabetes Dataset," vol. 1, 2020, doi: 10.17632/WJ9RWKP9C2.1.

[21] A. P. Monika, F. E. P. Risti, I. Binanto, and N. F. Sianipar, "Analisis Perbandingan Algoritma Knn, Gaussian Naive Bayes, Random Forest Untuk Data Tidak Seimbang Dan Data Yang Diseimbangkan Dengan Metode Tomek Link Undersampling Pada Dataset Lcms Tanaman Keladi Tikus," *Pros. Sains Nas. dan Teknol.*, vol. 13, no. 1, p. 156, 2023, doi: 10.36499/psnst.v13i1.9002.

[22] K. Natarajan, D. Baskaran, and S. Kamalanathan, "An adaptive ensemble feature selection technique for model-agnostic diabetes prediction," *Sci. Rep.*, vol. 15, no. 1, pp. 1–12, 2025, doi: 10.1038/s41598-025-91282-8.

[23] A. Rakhmadi, A. Yudhana, and S. Sunardi, "A Study Of Worldwide Patterns In Alphabet Sign Language Recognition Using Convolutional And Recurrent Neural Networks," *Jurnal Teknik Informatika.*, vol. 6, no. 1, pp. 187–204, Feb. 2025, doi: 10.52436/1.jutif.2025.6.1.4202.

[24] A. Yudhana, R. Umar, and S. Saputra, "Fish Freshness Identification Using Machine Learning: Performance Comparison of k-NN and Naïve Bayes Classifier," *Journal of Computer Science and Engineering.*, vol. 16, no. 3, pp. 153–164, 2022, doi: 10.5626/JCSE.2022.16.3.153.

[25] Sunardi, A. Yudhana, and A. R. W. Putri, "Optimization of Breast Cancer Classification Using Faster R-CNN," *Revue d'Intelligence Artificielle.*, vol. 37, no. 1, pp. 39–45, Feb. 2023, doi: 10.18280/ria.370106.

[26] J. Chukwura and J. Chukwura Obi, "A comparative study of several classification metrics and their performances on data," *https://wjaets.com/sites/default/files/WJAETS-2023-0054.pdf*, vol. 8, no. 1, pp. 308–314, Feb. 2023, doi: 10.30574/WJAETS.2023.8.1.0054.

[27] J. L. Speiser, "A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data," *Journal of Biomedical Informatics.*, vol. 117, p. 103763, May 2021, doi: 10.1016/j.jbi.2021.103763.

[28] D. R. Rajan, G. V. Sena, R. K, and M. K. Faizan, "Disease Prediction using Machine Learning," *BOHR International Journal of Computer Science.*, vol. 1, no. 1, pp. 69–72, Jul. 2022, doi: 10.54646/bijcs.2022.11.

[29] K. F. Habie, M. Murinto, and S. Sunardi, "Impact of Optimizer Selection on MobileNetV1 Performance for Skin Disease Detection Using Digital Images," *Jurnal Teknik Informatika.*, vol. 6, no. 3, pp. 1589–1604, Jul. 2025, doi: 10.52436/1.jutif.2025.6.3.4685.

[30] M. Abdelaoui, "Analysis of the diabetes dataset using a SMOTE machine learning approach," *Stud. Eng. Exact Sci.*, vol. 5, no. 2, p. e12076, Dec. 2024, doi: 10.54021/seesv5n2-772.

[31] I. Riadi, R. Umar, and R. Anggara, "Prediksi Kelulusan Tepat Waktu Berdasarkan Riwayat Akademik Menggunakan Metode K-Nearest Neighbor," *Jurnal Teknologi Informasi dan Ilmu Komputer.*, vol. 11, no. 2, pp. 249–256, Apr. 2024, doi: 10.25126/jtiik.20241127330.

[32] S. Helmiyah, I. Riadi, R. Umar, A. Hanif, A. Yudhana, and A. Fadlil, "Identifikasi Emosi Manusia Berdasarkan Ucapan Menggunakan Metode Ekstraksi Ciri LPC dan Metode Euclidean Distance," *Jurnal Teknologi Informasi dan Ilmu Komputer.*, vol. 7, no. 6, p. 1177, Dec. 2020, doi: 10.25126/jtiik.2020722693.

[33] G. A. Ansari, S. S. Bhat, and M. D. Ansari, "Machine Learning Techniques for Diabetes Mellitus Based on Lifestyle Predictors," *Recent Advances in Electrical and Electronic Engineering. (Formerly Recent Patents Electr. Electron. Eng.*, vol. 18, no. 7, pp. 1060–1071, Aug. 2025, doi: 10.2174/0123520965291435240508111712.

[34] P. Talari *et al.*, "Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2," *PLoS One*, vol. 19, no. 1 January, Jan. 2024, doi: 10.1371/journal.pone.0292100.

[35] E. B. Susanto, A. N. Anzila, and B. Ismanto, "Comparison Of The Effectiveness Of K-Nearest Neighbor (KNN) And Naive Bayes Algorithms In Identifying Diabetes Patients," *Journal of Artificial Intelligence and Software Engineering.*, vol. 5, no. 1, p. 22, Mar. 2025, doi: 10.30811/jaise.v5i1.6275.

[36] A. R. Mohammed, "Enhancing Diabetes Mellitus Onset Prediction through Advanced Ensemble Learning Techniques," *Journal of Statistical Modeling and Analytics.*, vol. 6, no. 2, pp. 1–18, Dec. 2024, doi: 10.22452/josma.vol6no2.2.