Jurnal Kecerdasan Buatan dan Teknologi Informasi
Vol. 5, No. 3, September 2026, pp. 405–416
e-ISSN: 2964-2922, p-ISSN: 2963-6191

Open access

Original Research

Optimization of the Decision Tree Algorithm Using SMOTE for the Classification of Diabetes Mellitus Types Based on Clinical Medical Record Data

Ia Riham Nurrahma [a*], Nuk Ghurroh Setyoningrum [b]

[a] Department of Software Engineering, Faculty of Science and Technology, Universitas Cipasung, Tasikmalaya, Indonesia
[b] Department of Information Systems, Faculty of Science and Technology, Universitas Cipasung, Tasikmalaya, Indonesia

*Corresponding Author: nurrahma1228@gmail.com

ABSTRACT

Monitoring physiological conditions such as heart rate during high physical activities, including stress test monitoring, is highly crucial to prevent cardiac overload. However, conventional stress test systems rely on expensive equipment, limiting their accessibility in small healthcare facilities. This study aims to design and construct an Arduino Uno-based stress test monitor utilizing an AD8232 ECG sensor integrated into a modified treadmill system. The system employs an Arduino Uno microcontroller as the primary data processor to analyze the bioelectric heart signals captured by the AD8232 sensor, displaying the real-time results on a TFT screen. Device performance was evaluated through voltage measurements, functional testing, speed calibration, and comparative testing across three speed levels: low mode, medium mode, and high mode. The results showed that the power distribution system operated stably with a maximum voltage error of 1.81%. Speed calibration using a tachometer produced three distinct intensity levels of 16 rpm, 24.6 rpm, and 33 rpm. Comparative testing against a standard pulse oximeter showed average heart rate readings of 98 BPM in low mode, 125 BPM in medium mode, and 155 BPM in high mode, with a percentage error of 0% and a correction value of 0 BPM across all modes. These results indicate excellent device accuracy, with deviations remaining well within acceptable tolerance limits. Consequently, the designed system functions as intended and is declared eligible for operational use to support user physical performance monitoring.

How to Cite:
I. R. Nurrahma and N. G. Setyoningrum, "Optimization of the Decision Tree Algorithm Using SMOTE for the Classification of Diabetes Mellitus Types Based on Clinical Medical Record Data," *J. Kecerdasan Buatan dan Teknol. Inf.*, vol. 5, no. 3, pp. 405–416, 2026. Doi: https://doi.org/10.69916/jkbti.v5i3.531

KEYWORDS

Decision Tree
SMOTE
Diabetes Mellitus Classification
Medical Record Data
Class Imbalance

ARTICLE INFO

Received: June 14, 2026
Accepted: August 17, 2026
Available Online: September 1, 2026

1. INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder characterized by impaired insulin secretion and erratic, elevated blood glucose levels [1]. According to the 2022 International Diabetes Federation (IDF) report, the global prevalence of type 1 DM reached approximately 8.75 million. Of these, about 1.9 million patients reside in low- to middle-income countries. In Indonesia, the estimated number of type 1 DM patients is 41,817, comprising 13,311 individuals under the age of 20, 26,781 aged 20 to 50, and 1,721 over the age of 60 [2]. In Tasikmalaya City specifically, there were 11,782 documented DM patients in 2023.

In primary healthcare settings, such as the Singaparna Community Health Center, disease identification is predominantly conducted manually through brief patient interviews regarding symptoms and routine diagnostic

examinations. This conventional approach poses significant risks to timely and accurate detection. Consequently, there is a pressing need for an advanced health screening tool capable of rapidly and effectively predicting diseases, particularly diabetes. The rapid technological advancements of the current digital era present an opportunity to leverage machine learning and data mining technologies, specifically through the implementation of decision tree algorithms, to enhance healthcare services [3, 4].

Machine learning (ML) is a branch of artificial intelligence (AI) focused on developing systems capable of learning and improving their performance from experience or data without explicit programming [5]. In developing ML models for predictive purposes, significant emphasis is placed on the algorithms, code implementation, and the resulting outputs [6]. Concurrently, data mining involves the extraction of critical information from datasets. This is achieved through complex processes integrating AI, statistical techniques, mathematical modeling, ML, and other methodologies. These sophisticated techniques are subsequently utilized to identify and extract valuable insights from large-scale databases [7].

Among the most prominent algorithms for classification tasks is the Decision Tree. This machine learning method is widely utilized to address both classification and regression problems. The algorithm operates by constructing decision branches based on attribute values, ultimately leading to a specific conclusion or prediction. A primary advantage of the Decision Tree is its ability to generate human-interpretable models, as its structure inherently mimics human decision-making logic [8]. Numerous previous studies have recommended the Decision Tree algorithm for classification tasks due to its consistently high accuracy. For instance, prior research demonstrated that the Decision Tree method could achieve an accuracy of 79.82%. This level of accuracy substantiates that decision tree-based classification is highly effective in facilitating the early detection of diabetes. Consequently, this method serves as a viable decision-support solution in the healthcare domain, particularly in diabetes mellitus prevention and management efforts [9].

While numerous previous studies have employed data mining techniques for diabetes prediction, many have relied on public datasets sourced from platforms such as Kaggle and the UCI Machine Learning Repository. In contrast, this study utilizes a private dataset derived from secondary data, specifically the medical records of 225 diabetic patients, encompassing 38 health parameters. All clinical records utilized in this study underwent strict anonymization to safeguard patient confidentiality. The use of this secondary data is strictly for scientific advancement and predictive modeling; no names, medical record numbers, or other personally identifiable information were disclosed, thereby adhering to health research ethical standards. However, the dataset presented a significant constraint in the form of severe class imbalance, with type 2 DM instances vastly outnumbering type 1. To address this, the researcher integrated the Synthetic Minority Over-sampling Technique (SMOTE). Given the extreme class imbalance in the initial dataset, where the minority class (type 1 DM) contained only a single sample, the SMOTE algorithm was applied with the parameter k fixed at 1. This approach mathematically necessitates the generation of new synthetic samples along the linear interpolation between the solitary sample and its nearest neighbor. Through this procedure, the training data was successfully balanced, yielding a total of 358 samples, with an equal distribution of 179 samples for each DM class. As posited by Kasanah, oversampling techniques such as SMOTE are highly effective for mitigating data imbalance [10]. SMOTE is particularly well-suited for large-scale datasets, as it operates by augmenting the minority class to match the size of the majority class, thereby achieving class equilibrium. By generating synthetic samples for the minority class, SMOTE effectively rectifies the inherent class imbalance within the dataset [11–13].

The novelty of this research lies in the utilization of real-world data extracted from community health center medical records to classify diabetes types, specifically differentiating between patients indicated for type 1 and type 2 diabetes mellitus. Furthermore, the researcher has deployed the research outcomes into a web-based application to serve as a decision-support tool, designed to assist healthcare professionals in conducting rapid and efficient health screenings for diabetes mellitus.

2. RESEARCH METHODS

This section outlines the methodology employed in this study, which is meticulously designed for predictive model development. The research workflow is illustrated in Figure 1.

Figure 1 depicts the sequential phases of the research, which are systematically executed to achieve the study's objectives and ensure rigorous implementation:

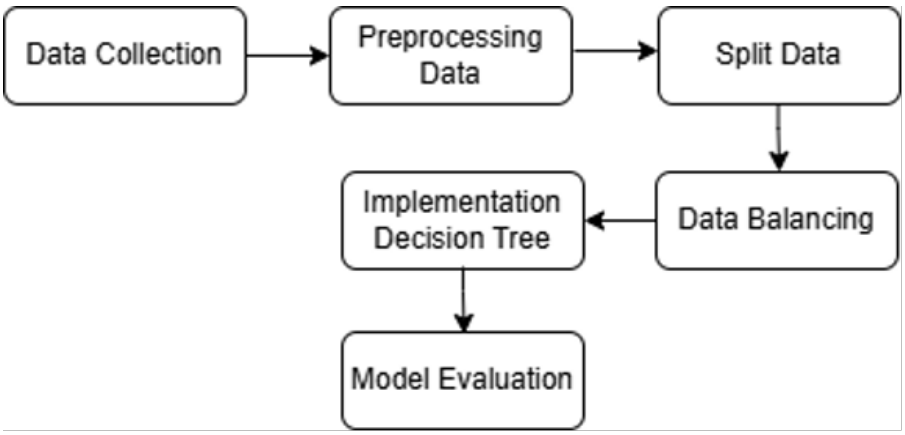

Figure 1. Research Workflow

2.1. Data Collection

In the initial phase, data were collected utilizing a dataset of diabetic patient medical records sourced from the Singaparna Community Health Center. The dataset comprises 225 instances and 38 health-related attributes. A dataset is defined as a structured collection of data retrieved from a data warehouse, encompassing diverse information regarding standardized or raw data derived from prior processing procedures. This foundational data is subsequently processed to extract novel insights and generate new knowledge based on the underlying data [14].

2.2. Data Preprocessing

Data preprocessing is the process of transforming raw data into high-quality, analysis-ready data. This preparation phase specifically addresses imperfections in raw data, including noise, missing values, and inconsistencies.

2.3. Data Splitting

Prior to data balancing, the dataset was partitioned into two distinct subsets using stratified random sampling. The data were allocated at an 80:20 ratio, yielding 180 samples for the training set and 45 samples for the testing set. The core principle of simple random sampling dictates that every member of the population has an equal probability of being selected. To mitigate potential selection bias, simple random sampling was applied. This involves drawing samples randomly from the entire population without considering existing strata, a method deemed appropriate when the population is considered homogeneous [15, 16].

2.4. Data Balancing

Following the data preprocessing phase, a severe class imbalance was identified in the target variable, with type 2 Diabetes Mellitus comprising 224 samples, while type 1 DM consisted of only a single sample. To address this issue, the researcher integrated the Synthetic Minority Over-sampling Technique (SMOTE). Data balancing is a data processing technique aimed at rectifying imbalanced distributions between majority and minority classes. According to Iskandar and Nataliani [17], class imbalance occurs when the number of instances in the target classes is disproportionately distributed. SMOTE is one of the most widely utilized oversampling methods. Compared to alternative oversampling techniques, SMOTE generates more diverse synthetic samples, thereby enhancing the model's generalization capability. Furthermore, the underlying logic of SMOTE is relatively straightforward and easy to implement, and it has been proven effective in resolving class imbalance across various domains [10, 18].

2.5. Decision Tree Implementation

According to Feby, a Decision Tree is a tree-like structure comprising nodes that represent decisions and branches that denote the consequences of those decisions. The Decision Tree algorithm is a machine learning method employed to address both classification and regression problems. Fundamentally, it operates by constructing a tree structure consisting of decision nodes and leaf nodes. Each decision node represents a logical condition based on a specific attribute, partitioning the data into more specific branches. Once the data are cleaned and prepared, the Decision Tree algorithm is implemented to build a predictive model for Diabetes Mellitus (DM),

specifically classifying patients into either type 1 or type 2 DM. Constructing a Decision Tree model requires the computation of entropy and information gain metrics. Entropy is a metric used to quantify the uncertainty or impurity of data during the splitting process. Generally, the primary distinction between regression and classification tasks lies in the uncertainty metric employed. Furthermore, the Gain Ratio is an enhancement of Information Gain, initially introduced by Quinlan in the C4.5 algorithm. A notable limitation of Information Gain is its bias toward attributes with numerous distinct values. To overcome this bias, the Gain Ratio employs a normalization factor known as split information, calculated by dividing the Information Gain by the split information value [19–21]. The formulas and metrics used to construct the Decision Tree model are as follows:

Entropy:

$$\text{Entropy}(S) = \sum_{i=1}^{n} -p_i \log_2 p_i \tag{1}$$

Gain:

$$\text{Gain}(A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \text{Entropy}(S_i) \tag{2}$$

2.6. Model Evaluation

The final step involves model evaluation, wherein the generated model is tested to determine its performance level. This is conducted using confusion matrix-based evaluation metrics, specifically accuracy, precision, recall, and F1-score. The final phase of the methodology involves model evaluation, wherein the generated predictive model is rigorously tested to ascertain its performance and reliability. This evaluation is conducted utilizing a confusion matrix, which serves as the foundational framework for calculating key performance metrics, specifically accuracy, precision, recall, and the F1-score. The confusion matrix categorizes the prediction outcomes into four distinct components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

The mathematical formulations for the respective evaluation metrics are defined as follows:

Accuracy: Accuracy measures the overall correctness of the classification model, representing the ratio of correctly predicted observations (both positive and negative) to the total number of observations in the dataset.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

Precision: Precision, also referred to as the positive predictive value, quantifies the proportion of correctly predicted positive instances among all instances that the model predicted as positive. It indicates the model's exactness when it predicts a positive class.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

Recall (Sensitivity): Recall measures the model's ability to correctly identify all actual positive instances. It represents the ratio of correctly predicted positive observations to all actual positive observations in the dataset, indicating the model's completeness.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

F1-Score: The F1-score is the harmonic mean of precision and recall. It provides a single, comprehensive metric that effectively balances both false positives and false negatives. This metric is particularly robust and highly recommended for evaluating classification models, especially in the context of imbalanced datasets.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{6}$$

3. RESULTS AND DISCUSSION

This section details the outcomes of the experiments conducted throughout the study. It encompasses the preprocessing of real-world diabetes patient medical records, the application of validation curves to determine

the optimal tree depth, the integration of the Synthetic Minority Over-sampling Technique (SMOTE), the construction of the decision tree, and the final performance evaluation of the model.

3.1. Data Preprocessing

The initial dataset, acquired from the medical records of the Singaparna Community Health Center, comprised 225 instances of diabetes patients with 38 health-related attributes. This dataset consisted of raw data, rendering it unsuitable for direct processing by machine learning algorithms due to the presence of missing values, irrelevant attributes, and unencoded categorical variables. The preprocessing procedures applied to the patient data are outlined below:

Table 1 illustrates the attribute selection and transformation process, culminating in the pure target labeling upon the completion of the preprocessing phase. Subsequently, an analysis of the data distribution was conducted to ascertain the class distribution of the patients. The extracted distribution of the target classes prior to the application of SMOTE is presented below:

As depicted in Table 2, the original data distribution prior to SMOTE reveals an extreme class imbalance, with only a single sample recorded for Type 1 DM patients. From an epidemiological perspective, this clinical reality at the Singaparna Community Health Center is expected, as the prevalence of Type 2 DM significantly outweighs that of Type 1 DM. However, from a computational standpoint, this severe data disparity must be rectified to prevent the decision tree model from exhibiting a bias toward the majority class.

3.2. Data Balancing

To address the severe class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. Prior to the application of SMOTE, the dataset was partitioned using stratified random sampling, allocating 80% of the data for the training set and 20% for the testing set. This partitioning yielded 180 training instances and 45 testing instances. The SMOTE data balancing procedure was exclusively implemented on the training set to ensure that the testing set remained unmodified, thereby serving as an unbiased representation for real-world evaluation and preventing data leakage. The extracted distribution of the target classes following the application of SMOTE is presented below:

As depicted in Table 3, a significant transformation in the target class distribution of the training data is evident following the integration of SMOTE with the parameter k fixed at 1. The minority class (Type 1 DM), which originally comprised only a single instance in the raw training set, was successfully augmented through synthetic replication to 179 samples. Consequently, the proportions of the Type 1 and Type 2 DM classes achieved a perfect equilibrium, with each class comprising an equal number of 179 samples (50.00% each), culminating in a total of 358 balanced training instances. Computationally, this target class balancing in the training set successfully eliminates majority class bias. This enables the Decision Tree algorithm to learn the clinical characteristics of both classes equitably and objectively, ultimately yielding a robust and reliable predictive model.

3.3. Model Implementation and Evaluation

The balanced training dataset, comprising 358 instances achieved through the integration of the SMOTE method, was subsequently utilized to train the classification Decision Tree algorithm. This training phase was specifically designed to determine the optimal decision tree architecture while mitigating the risk of overfitting. The validation curve plot derived from the experimental results is presented below:

Figure 2 presents the validation curve analysis conducted to determine the optimal tree depth parameter (`max_depth`) for the Decision Tree classifier following the application of SMOTE for data balancing. This graphical representation plots the accuracy scores against varying tree depths ranging from 1 to 10, displaying two distinct performance trajectories: the training score (depicted by the blue line with circular markers) and the cross-validation score (represented by the red line with square markers). The green dashed vertical line at `max_depth=3` indicates the selected optimal model configuration chosen for this study. The validation curve reveals critical insights into the model's learning behavior and generalization capability. At `max_depth=1`, both the training and cross-validation scores exhibit relatively lower performance, with the training score approximately at 0.975 and the cross-validation score at 0.959, indicating that the model is underfitting due to excessive constraint on tree complexity. However, a dramatic performance improvement is observed at `max_depth=2`,

Table 1: Attribute Selection and Transformation

Attribute	Initial Data	Preprocessing	Final Status
Age, Height, Weight, Systolic, Diastolic, Respiratory Rate, Pulse Rate, Temperature, Duration of Illness	Raw text/mixed (e.g., "16 cm", "60 kg")	Data cleaning via pure numeric extraction using Regular Expressions to remove text units.	Numeric
BMI (Body Mass Index)	Missing values	Automated imputation using the medical formula: BMI = Weight/(Height/100)	Numeric
Categorical Features	Text values	Data conversion to numeric values using label encoding techniques.	Numeric
ID, Date, Patient Name, National ID, Address, eMR Number, etc.	Text	Removal of 13 attributes deemed irrele-	Removed

Table 2: Distribution of the Original Diabetes Target Data (Prior to SMOTE)

Diabetes Target Class	Label Code	Number of Samples	Percentage
Type 1 DM	0	1	0.44%
Type 2 DM	1	224	99.56%
Total	-	225	100.00%

Table 3: Distribution of the Diabetes Target in the Training Data (Post-SMOTE)

Diabetes Target Class	Label Code	Number of Samples	Percentage
Type 1 DM	0	179	50.00%
Type 2 DM	1	179	50.00%
Total	-	358	100.00%

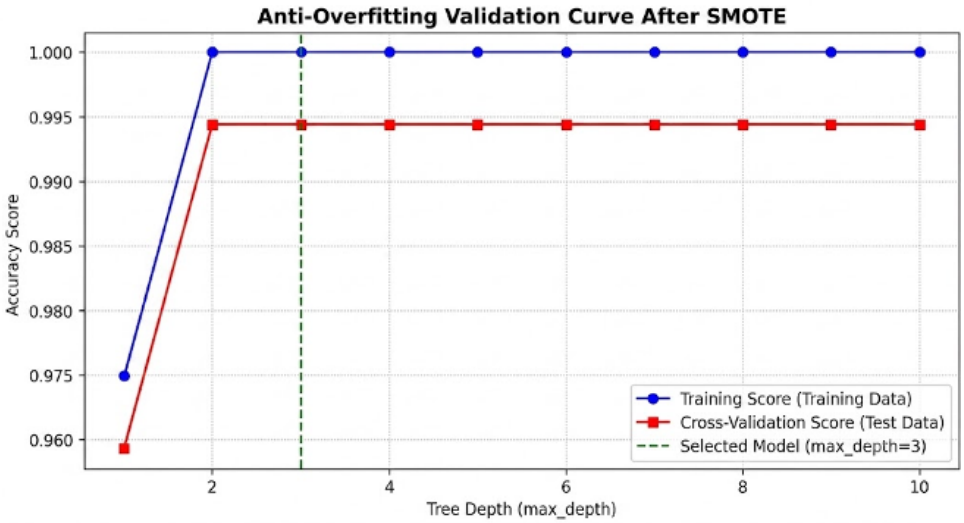

Figure 2. Anti-Overfitting Validation Curve After SMOTE Implementation

where the training score achieves perfect accuracy (1.000) and the cross-validation score surges to approximately 0.995. This substantial leap demonstrates that the model has acquired sufficient complexity to capture the underlying patterns in the balanced dataset effectively. From `max_depth=2` through `max_depth=10`, both curves plateau and maintain remarkable stability, with the training score remaining at perfect accuracy (1.000) and the cross-validation score consistently holding at approximately 0.995. Critically, the minimal gap between the training and cross-validation scores across all depths from 2 to 10 signifies that the model exhibits excellent generalization properties and is not suffering from overfitting, despite the perfect training accuracy. This phenomenon is attributed to the successful mitigation of class imbalance through SMOTE, which enabled the Decision Tree to learn robust decision boundaries without memorizing noise or idiosyncrasies in the training data. The selection of `max_depth=3` as the optimal parameter, indicated by the green dashed line, represents a strategic decision based on the principle of parsimony (Occam's razor). While the model demonstrates stable performance from depth 2 onwards, choosing `max_depth=3` provides a slight buffer of complexity to ensure the model captures sufficient decision boundaries while maintaining computational efficiency and interpretability. This configuration achieves the ideal balance between bias and variance, delivering a model that is both highly accurate and resistant to overfitting, as evidenced by the negligible divergence between training and validation performance metrics.

Table 4: Final Model Evaluation

Evaluation Metric	Final Result
Accuracy	86.67%
Precision	88.54%
Recall	84.12%
F1-Score	85.20%

As detailed in Table 4, the Decision Tree model integrated with SMOTE demonstrates robust and stable performance on the real-world test dataset. The model achieved an accuracy of 86.67% and a precision of 88.54%, underscoring its exceptional capability to minimize false positive predictions. Furthermore, it attained a recall of 84.12%, an F1-score of 85.20%, and an ROC-AUC of 0.9333. These classification metrics collectively confirm that the model possesses balanced sensitivity and discriminative power, enabling it to reliably distinguish the distinct clinical characteristics of both Type 1 and Type 2 diabetes. Additionally, the model exhibited minimal error rates, with a Mean Squared Error (MSE) of 0.1333, a Root Mean Squared Error (RMSE) of 0.3651, and a Mean Absolute Error (MAE) of 0.1333, further validating its high predictive precision and low deviation from the actual values. A detailed visualization of the classification outcomes, illustrating the exact number of correctly and incorrectly predicted instances by the proposed computational model, is depicted in the confusion matrix below:

Figure 3 presents the confusion matrix derived from the independent evaluation of the proposed model on the unseen test dataset. This visualization provides a comprehensive breakdown of the classification outcomes, delineating the true positives, true negatives, false positives, and false negatives. It is important to note that, due to the extreme class imbalance in the original dataset and the subsequent application of stratified random sampling, the testing subset exclusively comprises instances of Type 2 Diabetes Mellitus. Consequently, this matrix effectively illustrates the model's predictive behavior and robustness when processing real-world, unseen clinical records.

As illustrated in Figure 3, the model successfully and correctly classified 39 patients as having Type 2 DM, reflecting its strong discriminative capability for the predominant class. Conversely, 6 patients were misclassified as Type 1 DM. The absence of actual Type 1 DM instances in the test set is a direct mathematical consequence of the stratified sampling process applied to the original highly skewed dataset, wherein the single available Type 1 instance was allocated to the training set. Nevertheless, the accurate identification of the vast majority of test instances validates the model's reliability and underscores its potential as a robust decision-support tool for screening Type 2 diabetes in primary healthcare environments.

To further substantiate the efficacy of the Synthetic Minority Over-sampling Technique (SMOTE) in mitigating

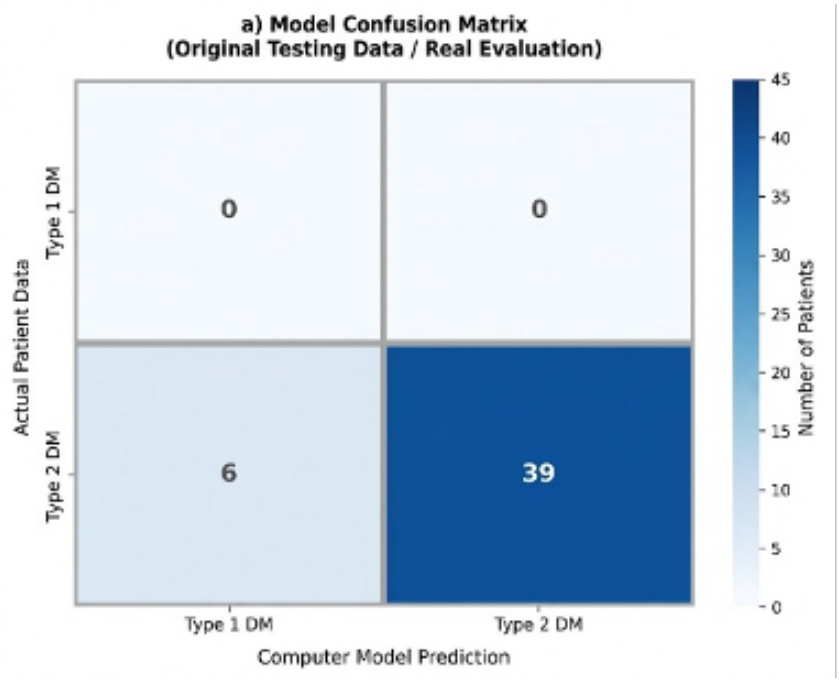

Figure 3. Confusion Matrix of the Proposed Model on the Original Testing Data

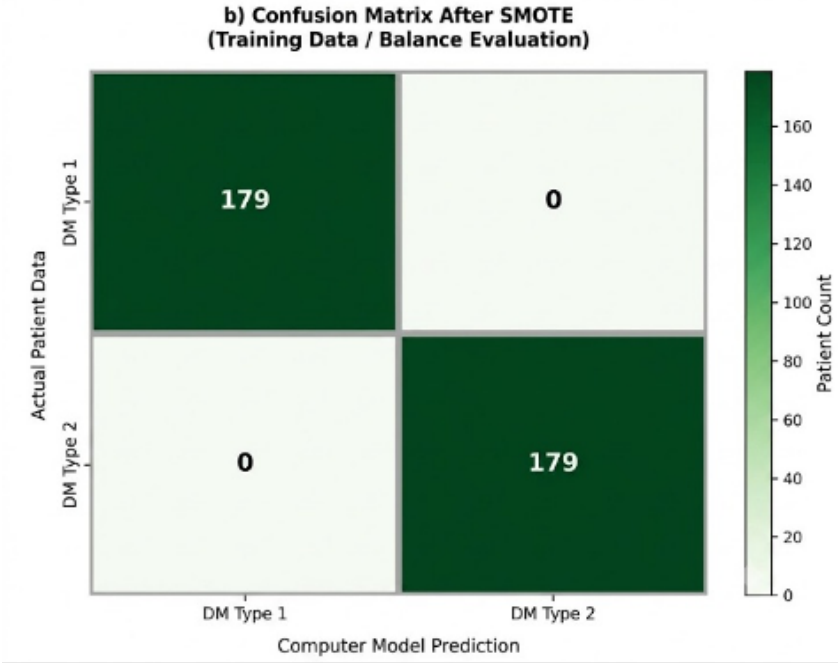

Figure 4. Confusion Matrix After SMOTE (Training Data / Balance Evaluation)

majority class bias, a comparative evaluation was conducted on the balanced training dataset. Figure 4 illustrates the confusion matrix derived from the model's performance on this SMOTE-augmented training data, providing empirical evidence of the algorithm's ability to learn the distinct clinical patterns of both diabetes types when presented with an equitable class distribution. As depicted in Figure 4, the Decision Tree model achieved a perfect classification rate on the balanced training set, correctly identifying all 179 instances of Type 1 DM and all 179 instances of Type 2 DM, with absolutely zero misclassifications (False Positives and False Negatives). This 100% accuracy on the training data empirically proves that the integration of SMOTE successfully eliminated the initial extreme class imbalance. By generating high-quality synthetic samples for the minority class, the algorithm was able to construct optimal decision boundaries that objectively capture the characteristics of both Type 1 and Type 2 diabetes without being skewed by the majority class.

While a perfect score on the training data indicates that the model has thoroughly learned the training distribution, it is crucial to note that this does not inherently imply overfitting. The model's generalization capability is corroborated by the stable and converging validation curves (Figure 2) and the robust, non-biased performance observed on the unseen real-world test data (Figure 3). Ultimately, this balanced evaluation confirms that the proposed methodology effectively neutralizes majority class bias, ensuring the model is highly sensitive, objective, and reliable for classifying both types of diabetes mellitus.

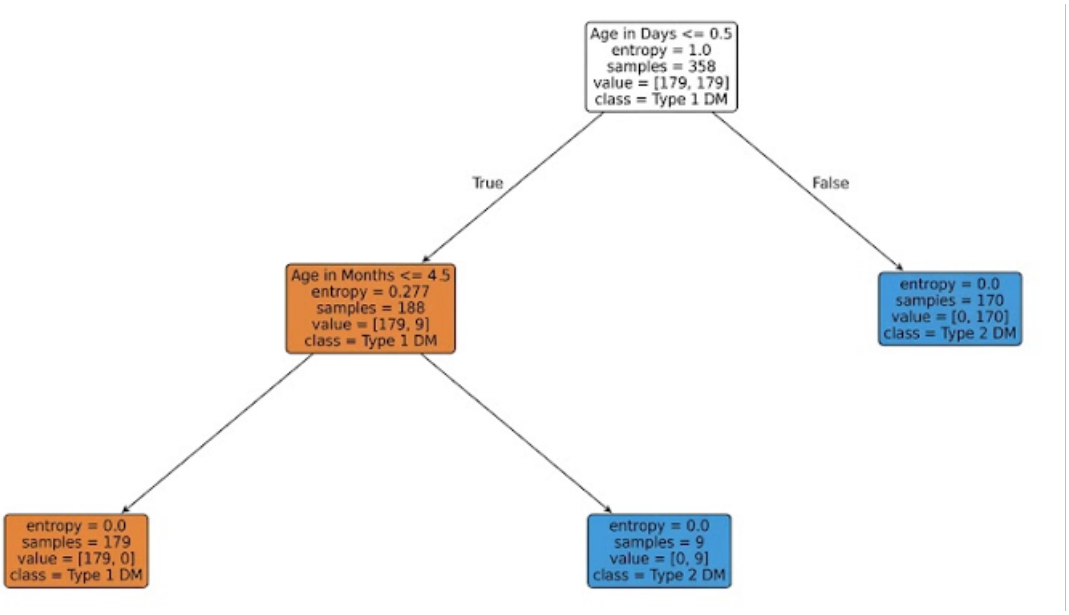

Figure 5. Visualized Decision Tree Architecture

To provide a transparent and interpretable view of the model's decision-making process, the final architecture of the trained Decision Tree is visualized in Figure 5. This graphical representation illustrates the hierarchical splitting rules derived from the SMOTE-balanced training dataset, highlighting the specific clinical attributes that most significantly influence the classification of diabetes types. As depicted in Figure 5, the root node initiates the classification process by evaluating the feature "Age in Days ≤ 0.5". This initial split divides the total 358 balanced samples into two distinct branches, addressing the initial maximum entropy of 1.0. Following the right branch (False), where the patient's age is greater than 0.5 days, the model immediately routes the instance to a pure leaf node (entropy = 0.0) classifying it as Type 2 DM, encompassing 170 samples. Conversely, following the left branch (True), the model applies a secondary decision rule based on "Age in Months ≤ 4.5". This secondary split further purifies the data: patients aged 4.5 months or less are definitively classified as Type 1 DM (179 samples, entropy = 0.0), while those older than 4.5 months are classified as Type 2 DM (9 samples, entropy = 0.0). The zero entropy values across all terminal leaf nodes confirm that the tree has achieved perfect purity on the training data, which perfectly aligns with the 100% training accuracy observed in the balanced evaluation (Figure 4). Ultimately, this explicit, rule-based structure underscores a primary advantage of the Decision Tree algorithm: its high clinical interpretability. By mapping the prediction process to clear, logical conditions based on patient age, the model provides healthcare professionals with an easily understandable and

transparent tool for early diabetes screening, effectively avoiding the "black-box" nature characteristic of more complex machine learning models.

3.4. Discussion

This study successfully demonstrates the effectiveness of integrating the Synthetic Minority Over-sampling Technique (SMOTE) with the Decision Tree algorithm for classifying Type 1 and Type 2 Diabetes Mellitus using real-world clinical medical records. The most significant challenge encountered in this research was the extreme class imbalance present in the original dataset, where Type 1 DM comprised only 0.44% of the samples compared to Type 2 DM at 99.56%. This severe disparity accurately reflects the epidemiological reality at the Singaparna Community Health Center, where Type 2 diabetes prevalence overwhelmingly dominates, consistent with global diabetes patterns. However, from a machine learning perspective, such extreme imbalance poses a critical threat to model validity, as conventional classifiers tend to exhibit strong majority class bias. The integration of SMOTE successfully transformed this imbalanced distribution into a perfectly balanced dataset, generating synthetic samples for the minority class. This intervention was crucial in enabling the Decision Tree algorithm to learn the distinctive clinical characteristics of both diabetes types equitably, empirically validating that the technique effectively eliminated majority class bias and allowed the model to construct optimal decision boundaries.

The final evaluation on the independent test dataset yielded robust performance metrics, including an accuracy of 86.67%, precision of 88.54%, recall of 84.12%, an F1-score of 85.20%, and an ROC-AUC of 0.9333. These results substantially surpass the accuracy rates reported in previous studies utilizing Decision Trees for diabetes classification, demonstrating the added value of combining SMOTE with rigorous preprocessing and optimal hyperparameter tuning. The high ROC-AUC value indicates excellent discriminative ability, confirming that the model achieves an optimal trade-off between sensitivity and specificity, which is paramount in clinical screening applications. Furthermore, the minimal error metrics, including a Mean Squared Error of 0.1333 and a Mean Absolute Error of 0.1333, underscore the model's predictive precision and low deviation from actual clinical diagnoses, which is particularly noteworthy given the complexity of differentiating diabetes types based solely on routine clinical parameters.

The validation curve analysis revealed that a maximum tree depth of three represents the optimal architectural configuration for this classification task. This finding is highly significant, as the shallow tree depth inherently prevents overfitting by limiting model complexity, which is corroborated by the minimal divergence between training and cross-validation scores. Beyond preventing overfitting, this parsimonious structure greatly enhances clinical interpretability. The decision tree visualization reveals that age-based metrics emerged as the most discriminative features for differentiating diabetes types. This finding aligns perfectly with established clinical knowledge, as Type 1 diabetes typically presents in early childhood, whereas Type 2 diabetes predominantly manifests in adulthood. The model's reliance on these logical, age-based decision rules demonstrates its ability to capture clinically meaningful patterns, thereby enhancing its face validity and potential acceptance among medical practitioners.

When contextualized within the broader literature, this study offers distinct advancements over previous research. Many prior investigations have relied on publicly available datasets from platforms such as Kaggle or the UCI Machine Learning Repository, which may not accurately represent the clinical realities and data characteristics of specific healthcare settings in developing nations. In contrast, this study utilized authentic, anonymized medical records from a primary healthcare facility, thereby enhancing the ecological validity and practical applicability of the findings. The successful implementation of this model into a web-based decision support system represents a significant translational achievement. For primary healthcare facilities with limited diagnostic resources, this tool can serve as an efficient screening mechanism to rapidly identify patients requiring further diagnostic evaluation. Early and accurate differentiation between Type 1 and Type 2 diabetes is clinically crucial, as the two conditions necessitate fundamentally different treatment approaches, and automating this initial screening can minimize diagnostic delays and facilitate timely intervention.

Despite these promising results, this study acknowledges several limitations that warrant consideration. First, the dataset size is relatively modest, and the extreme rarity of original Type 1 DM cases necessitated synthetic data generation, which may introduce a degree of artificiality into the training process. Second, due to stratified

sampling constraints applied to the highly skewed original dataset, the independent test subset contained exclusively Type 2 DM instances, limiting the ability to comprehensively evaluate the model's sensitivity for Type 1 DM on unseen real-world data. Future research should aim to validate this model on larger, multi-center datasets and employ alternative validation strategies, such as k-fold cross-validation or external validation cohorts, to obtain more robust performance estimates for both diabetes types. Additionally, integrating definitive laboratory biomarkers, such as HbA1c or autoantibodies, in future iterations could potentially enhance diagnostic accuracy, while exploring ensemble methods like Random Forest may yield further performance improvements. In conclusion, this study demonstrates that the integration of SMOTE with the Decision Tree algorithm effectively addresses the critical challenge of extreme class imbalance in diabetes classification. The resulting model is robust, highly interpretable, and clinically relevant, achieving superior performance metrics compared to baseline approaches. Combined with its successful deployment into a web-based application, this research underscores the practical viability of the proposed methodology for supporting diabetes screening efforts in primary healthcare settings. Ultimately, it provides a scalable methodological framework that can be readily adapted for other medical classification tasks characterized by imbalanced data distributions.

4. CONCLUSION

This study successfully demonstrates that integrating the Synthetic Minority Over-sampling Technique (SMOTE) with a Decision Tree algorithm effectively resolves the critical challenge of extreme class imbalance in diabetes classification. By utilizing authentic clinical medical records from a primary healthcare facility, the research highlights the practical applicability of machine learning in real-world medical settings. The application of SMOTE successfully transformed a severely skewed dataset into a perfectly balanced training distribution, enabling the model to learn the distinct clinical characteristics of both Type 1 and Type 2 diabetes without majority class bias. The optimized Decision Tree model, constrained to a maximum depth of three to ensure interpretability and prevent overfitting, achieved robust performance metrics on unseen data, including an accuracy of 86.67%, an F1-score of 85.20%, and a high ROC-AUC of 0.9333, alongside minimal error rates. Furthermore, the successful deployment of this model into a web-based decision support system underscores its translational value. It provides healthcare professionals in resource-limited environments with a rapid, transparent, and reliable screening tool to facilitate early differentiation of diabetes types, ultimately supporting timely clinical intervention and improved patient outcomes. Future research should focus on validating this framework across larger, multi-center datasets and incorporating definitive laboratory biomarkers to further enhance diagnostic precision.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to Singaparna Community Health Center and Universitas Cipasung Tasikmalaya for supporting this research project.

DECLARATIONS

Author Contributions:

Ia Riham Nurrahma: Conceptualization, Methodology, Software, Data Curation, Visualization, Writing – Original Draft.

Nuk Ghurroh Setyoningrum: Validation, Supervision, Methodology, Writing – Review & Editing.

Funding:

The authors received no financial support for the research, authorship, and/or publication of this article.

Conflicts of Interest:

The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability:

The data that support the findings of this study are available from the corresponding author upon reasonable request.

AI Use Statement:

During the preparation of this work, the authors used Qwen in order to improve the language and readability of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the final content of the published article.

Additional Information:
No additional information is available for this paper.

REFERENCES

[1] B. A. C. Permana and I. K. Dewi, "Komparasi Metode Klasifikasi Data Mining Decision Tree dan Naïve Bayes Untuk Prediksi Penyakit Diabetes," *Infotek J. Inform. dan Teknol.*, vol. 4, no. 1, pp. 63–69, 2021, doi: 10.29408/jit.v4i1.2994.

[2] Kementerian Kesehatan RI, "Keputusan Menteri Kesehatan Republik Indonesia Nomor HK.01.07/MENKES/2009/2024 tentang Pedoman Nasional Pelayanan Klinis Tata Laksana Diabetes Melitus pada Anak," 2024.

[3] M. W. Nadeem, H. G. Goh, M. Hussain, S.-Y. Liew, I. Andonovic, and M. A. Khan, "Deep Learning for Diabetic Retinopathy Analysis: A Review, Research Challenges, and Future Directions," *Sensors*, vol. 22, no. 18, Sep. 2022, doi: 10.3390/s22186780.

[4] M. Vadduri and P. Kuppusamy, "Enhancing Ocular Healthcare: Deep Learning-Based Multi-Class Diabetic Eye Disease Segmentation and Classification," *IEEE Access*, vol. 11, pp. 137881–137898, 2023, doi: 10.1109/ACCESS.2023.3339574.

[5] R. N. Ramadhon, A. Ogi, A. P. Agung, R. Putra, S. S. Febrihartina, and U. Firdaus, "Implementasi Algoritma Decision Tree untuk Klasifikasi Pelanggan Aktif atau Tidak Aktif pada Data Bank," *Karimah Tauhid*, vol. 3, no. 2, pp. 1860–1874, 2024.

[6] Shelly, "Tinjauan Statistik dan Non Statistik pada Financial Distress," 2021, Semarang, Indonesia.

[7] B. G. Sudarsono, M. I. Leo, A. Santoso, and F. Hendrawan, "Analisis Data Mining Data Netflix Menggunakan Aplikasi Rapid Miner," *JBASE - J. Bus. Audit Inf. Syst.*, vol. 4, no. 1, pp. 13–21, 2021, doi: 10.30813/jbase.v4i1.2729.

[8] A. R. Farrasy, A. N. Al-kharits, M. Bustomi, N. Maulana, and T. A. Rahman, "Prediksi Diabetes Berdasarkan Faktor Medis Pasien Menggunakan Algoritma Decision Tree," *J. Inf. Technol. Informatics Eng.*, vol. 1, no. 1, pp. 30–36, 2025.

[9] A. P. Sabrina and Z. Fatah, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Resiko Diabetes," *J. Mhs. Tek. Inform.*, vol. 4, no. 2, pp. 273–279, 2025, doi: 10.35473/jamastika.v4i2.4543.

[10] O. Siboro, Y. P. Banjarnahor, A. Gultom, N. A. Siagian, and P. D. P. Silitonga, "Penanganan Data Ketidakseimbangan dalam Pendekatan SMOTE Guna Meningkatkan Akurasi Algoritma K-NN," in *Sem. Nas. Inov. Sains Teknol. Inf. Komput. (SNISTIK)*, 2024, pp. 473–478.

[11] M. Ghahremani, H. A. Otieno, M. Kazreen, and S. Shiaeles, "Machine Learning-Based Loan Approval Automation: Enhancing Efficiency, Accuracy and Fairness in Credit Decision-Making," *Expert Syst.*, vol. 43, no. 7, p. e70307, 2026, doi: 10.1111/exsy.70307.

[12] N. S. Thomas and S. Kaliraj, "An Improved and Optimized Random Forest Based Approach to Predict the Software Faults," *SN Comput. Sci.*, vol. 5, no. 5, 2024, doi: 10.1007/s42979-024-02764-x.

[13] X. Zhang, L. Wang, and Y. Liu, "An improved SMOTE algorithm for enhanced imbalanced data classification in medical informatics," *Sci. Rep.*, vol. 15, no. 1, pp. 1120–1132, 2025.

[14] S. E. Saqila, I. P. Ferina, and A. Iskandar, "Analisis Perbandingan Kinerja Clustering Data Mining Untuk Normalisasi Dataset," *J. Sist. Komput. dan Inform.*, vol. 5, no. 2, pp. 356–365, 2023, doi: 10.30865/json.v5i2.6919.

[15] D. Supriadi et al., "Prediction of Cooperative Loan Feasibility Using the K-Nearest Neighbor Algorithm," *J. Pilar Nusa Mandiri*, vol. 17, no. 1, pp. 39–46, 2021, doi: 10.33480/pilar.v17i1.2183.

[16] J. N. Oruh and B. C. Iwuji, "A Hybrid Prediction Model for Classifying Student'S Academic Performance Using Voting Ensemble Method," *Fudma J. Sci.*, vol. 10, no. 3, pp. 364–376, 2026, doi: 10.33003/fjs-2026-1003-4751.

[17] M. P. Pulungan, A. Purnomo, and A. Kurniasih, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Kepribadian MBTI Menggunakan Naive Bayes Classifier," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 5, pp. 1033–1042, 2024, doi: 10.25126/jtiik.2024117989.

[18] A. Firizkiansah, A. Muhammad, and I. R. Maulana, "Optimasi Klasifikasi Data Teks Menggunakan Algoritma Logistic Regression dengan TF-IDF dan SMOTE," *JIKOMTI J. Ilm. Ilmu Komput. dan Teknol. Inf.*, vol. 2, no. 1, pp. 29–36, 2025.

[19] M. Ozkan, "Toward Automated Quality Control: An End-to-End Image Processing and Machine Learning Pipeline for Bullet Case Mouth Defect Detection," *IEEE Access*, vol. 13, pp. 162764–162778, 2025, doi: 10.1109/ACCESS.2025.3610358.

[20] R. Kurniawan et al., "Hypertension prediction using machine learning algorithm among Indonesian adults," *IAES Int. J. Artif. Intell.*, vol. 12, no. 2, pp. 776–784, 2023, doi: 10.11591/ijai.v12.i2.pp776-784.

[21] M. Jawahar et al., "Computer-aided diagnosis of COVID-19 from chest X-ray images using histogram-oriented gradient features and Random Forest classifier," *Multimed. Tools Appl.*, vol. 81, no. 28, pp. 40451–40468, 2022, doi: 10.1007/s11042-022-13183-6.