RESEARCH ARTICLE OPEN ACCESS

Manuscript received October 13, 2024; revised December 20, 2024; date of publication January 28, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i1.596

How to cite: Arief Wibowo, Anis Fitri Nur Masruriyah, Selly Rahmawati, "Refining Diabetes Diagnosis Models: The Impact of SMOTE on SVM, Logistic Regression, and Naïve Bayes for Imbalanced Datasets", Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 7, no. 1, pp. 197-207, January 2025.

Refining Diabetes Diagnosis Models: The Impact of SMOTE on SVM, Logistic Regression, and Naïve Bayes for Imbalanced Datasets

Arief Wibowo[1], Anis Fitri Nur Masruriyah[2], Selly Rahmawati[1]

[1] Department of Computer Science, Universitas Budi Luhur, Jakarta, Indonesia.
[2] Department of Informatics, Universitas Pembangunan Nasional "Veteran" Jakarta, Indonesia.

Corresponding author: Arief Wibowo (e-mail: arief.wibowo@budiluhur.ac.id).

This research was funded by the Directorate of Higher Education, Ministry of Education, Research, and Technology, Republic of Indonesia, through grant number 6556 based on Decree No. 0667/E5/AL.04/2024.

ABSTRACT Accurate diabetes classification is a significant challenge in medical diagnostics, especially in imbalanced datasets. This study addresses this issue by introducing A New Modified Weighted SMOTE (ANMWS), integrated with Priority of Attribute by Expert Judgement (PAEJ) framework, to enhance the performance of machine learning models for imbalanced data. PAEJ categorizes attributes into three levels—high, medium and low priority—based on expert knowledge, while ANMWS applies weighted oversampling using these priority levels to generate synthetic data more representative of real-world cases. The proposed method was evaluated using three algorithms: Support Vector Machine (SVM), Logistic Regression, and Naïve Bayes. Results indicate that applying ANMWS algorithm with PAEJ framework significantly improved predictive performance, with AUC values increasing to 0.995 for SVM, 0.993 for Logistic Regression, and 0.990 for Naïve Bayes, compared to 0.980, 0.978, and 0.975, respectively, using standard SMOTE. Additionally, precision and recall for SVM improved by 5% and 7%, respectively. These findings demonstrate the critical role of ANMWS algorithm and PAEJ framework in addressing class imbalance, providing a reliable method for early diabetes diagnosis and informed clinical decision-making.

INDEX TERMS Diabetes Mellitus, Logistic Regression, Naïve Bayes, Support Vector Machine, A New Modified Weighted SMOTE (ANMWS) algorithm, Priority of Attribute by Expert Judgement (PAEJ).

I. INTRODUCTION

The global prevalence of diabetes mellitus, a chronic metabolic disorder marked by persistent hyperglycemia, continues to escalate, creating significant challenges for healthcare systems worldwide. Early and precise diagnosis is essential for effective management and treatment of this condition. Machine learning (ML) algorithms have demonstrated substantial potential in improving diagnostic accuracy across various medical domains, including diabetes. However, the inherent class imbalance in medical datasets—where diabetic cases are frequently outnumbered by non-diabetic cases—poses a critical challenge to the performance of ML models.

In 2020, a study [1] employing Python programming within the Jupyter Notebook application reported that the Naïve Bayes algorithm achieved superior predictive results compared to ID3. Similarly, another study [2] evaluated the Naïve Bayes method against Support Vector Machine (SVM) for diabetes classification using the WEKA tool, concluding that the SVM algorithm with a polynomial kernel outperformed Naïve Bayes. In 2021, research by Mulyo Widodo et al. [3] compared the performance of K-Nearest Neighbors (KNN), J48, Naïve Bayes, and Logistic Regression for diabetes classification, identifying KNN as the most accurate, achieving 98% accuracy. In 2022, Karo et al. [4] utilized SVM, Decision Tree, and Naïve Bayes to classify diabetes, further exploring the application of machine learning in this domain. The study indicated that the SVM algorithm was the most effective in performance. In the same year, Ginanjar [5] utilized the Adaboost Classifier for diabetes classification, incorporating mean and median input techniques in the testing process. The results showed that the highest accuracy of 80.09% was achieved using the mean input technique. In 2022, Putry and Sari [6] compared the KNN and Naïve Bayes algorithms for diabetes classification. Their findings indicated that Naïve Bayes had the highest accuracy compared to KNN.

In order to deal with the imbalance of data, Hairani et al. [7] stated that addressing the class imbalance in diabetes datasets can be effectively achieved through oversampling methods like SMOTE before classification. Their study demonstrated improved performance metrics such as accuracy, sensitivity, and specificity. Specifically, combinations like K-Means - SMOTE and SVM - SMOTE achieved higher accuracy and sensitivity rates of 82% and 77%, respectively, while Naïve Bayes achieved a sensitivity of 89%.

Modifications of the SMOTE technique have been extensively explored in various studies to address data imbalance. One notable modification is Weighted-SMOTE, which introduces weighting based on the distance between minority class samples to enhance the distribution of synthetic data [11]. Another approach, Feature Weighted-SMOTE, leverages feature weights to account for the significance of each attribute during the oversampling process [12]. Additionally, adaptive methods like Adaptive Weighting-SMOTE have been proposed, where weights are updated iteratively based on their contribution to evaluation outcomes [13]. These methods have shown significant improvements in handling imbalanced data by refining the oversampling process, either through selective sampling or dynamic weighting strategies. However, none explicitly consider the prioritization of attributes determined by expert knowledge, particularly in medical datasets where domain expertise plays a crucial role.

This study introduces a novel approach to oversampling with weighting, termed A New Modified Weighted SMOTE (ANMWS), which integrates the Priority of Attribute by Expert Judgement (PAEJ) framework. In this method, dataset attributes are assigned one of three weights—high, medium and low—based on their clinical relevance as determined by the internist doctor. High-weight attributes represent features categorized as critical for diabetes diagnosis (Level 1), medium-weight attributes correspond to factors with moderate importance (Level 2), and low-weight attributes are those considered supplementary or less influential (Level 3).These weights are incorporated into the oversampling process to ensure that synthetic data better reflects the clinical significance of each attribute. By prioritizing key attributes, ANMWS and PAEJ offer a novel approach that bridges data-driven algorithms and domain expertise, addressing class imbalance in medical datasets and improving the reliability of predictive models for early diabetes diagnosis.

Existing studies have proven the benefits of classification models for healthcare professionals in diabetes prevention and treatment. However, there remains a gap in analyzing lifestyle-related factors influencing diabetes diagnosis, and prior research has not sufficiently addressed the ethical implications of integrating machine learning into healthcare. This study aims to enhance health analytics by evaluating the impact of SMOTE on the performance of SVM, Logistic Regression, and Naïve Bayes, addressing class imbalance to improve diagnostic accuracy.

The contributions of this study are as follows: 1) It introduces and evaluates A New Modified Weighted SMOTE (ANMWS) integrated with the Priority of Attribute by Expert Judgement (PAEJ) framework, which prioritizes attributes into high, medium, and low importance based on expert knowledge from internists to improve synthetic data generation. 2) It provides a comparative analysis of machine learning algorithms (SVM, Logistic Regression, and Naïve Bayes) in diabetes classification before and after applying SMOTE, demonstrating significant improvements in addressing class imbalance. 3) It offers insights into the ethical implications of using machine learning in healthcare, emphasizing the importance of equitable and responsible deployment of diagnostic tools.

II. DATA AND METHOD

The CRISP-DM framework (Cross-Industry Standard Process for Data Mining) was utilized in this study to provide a structured approach for developing diabetes prediction models. In the Business Understanding phase, the primary objective was to support healthcare practitioners in improving early diabetes diagnosis and treatment strategies. Specifically, the goal was to develop models that could aid in identifying at-risk patients, allowing for timely intervention and reducing long-term healthcare costs associated with diabetes complications. During the Data Understanding phase, the dataset was thoroughly examined to identify relevant features such as age, BMI, and blood sugar levels, while uncovering potential issues like missing data and imbalanced class distribution, which could affect model performance.

The Data Preparation phase involved cleaning the dataset, normalizing features, and applying the SMOTE technique to address class imbalance, ensuring the models would not be biased toward the majority class. In the Modeling phase, three machine learning algorithms—SVM, Logistic Regression, and Naïve Bayes—were applied to evaluate their effectiveness both before and after SMOTE application as shown in FIGURE 1.

The Evaluation phase involved assessing the models using metrics like AUC, accuracy, precision, and recall to ensure that the best model could accurately predict diabetes risk across various patient profiles. By following the CRISP-DM process, this study aligns predictive modeling with healthcare goals, providing clinicians with tools to make more informed decisions and ultimately improve patient outcomes.

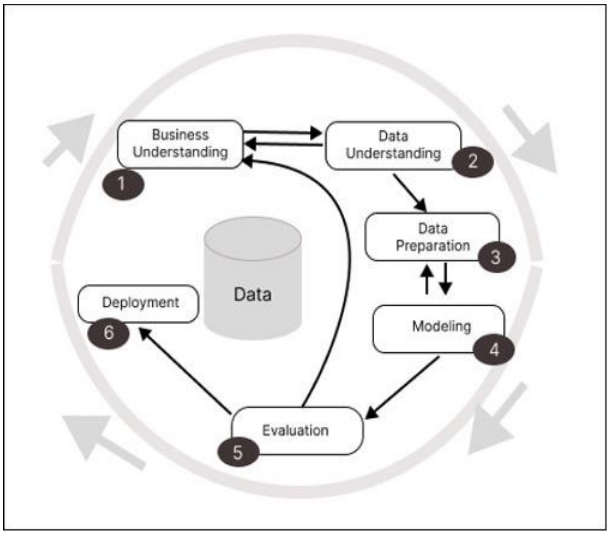

FIGURE 1. Stages of CRISP-DM

TABLE 1
Attributes of Datasets

Attribute	Category	Detail
Age	Medium	Age of the patient in years at the time of data collection.
Gender	Low	Classification of individuals as male or female.
Family History	Medium	Presence of diabetes in the patient's immediate family.
Body Mass Index	High	Measure of body fat based on height and weight.
Blood Presure	High	Measurement of the force of blood against the walls of arteries
Blood Sugar Levels	High	Measurement of glucose concentration in the blood
Pregnancy Status	Medium	Indicates whether the patient is currently pregnant
Smoking Habits	Low	Indicates whether the patient smokes tobacco products
Physical Activities	Medium	Level of regular physical exercise or activity
Sleep Patterns	Low	Duration and quality of sleep patterns

A. BUSINESS UNDERSTANDING

The first stage, Business Understanding, focuses on clarifying project objectives and requirements from a business perspective [9]. This involves defining goals, establishing success criteria, and aligning data mining efforts with business objectives to ensure relevance and impact. In this process, it is crucial to conduct a comprehensive analysis of business requirements and map out how data mining can provide effective solutions. This encompasses identifying the challenges encountered and the potential opportunities that can be leveraged through the application of data mining.

B. DATA UNDERSTANDING

Data Understanding stage involves initial data collection and exploration to gain insights into the dataset [10]. It includes assessing data quality, identifying potential issues, and forming initial hypotheses about relationships within the data [9]. The dataset used in this study consists of diabetes patient records collected from a public hospital in West Java, Indonesia, covering the period from January 2022 to December 2023, with a total of 657 patients.

The data was gathered through direct patient surveys, ensuring comprehensive collection of key health metrics. The attributes include age, gender, family history of diabetes, Body Mass Index (BMI), blood pressure, blood sugar levels, pregnancy status, smoking habits, physical activity, and sleep patterns. Each attribute offers specific insights into the patient’s health profile, enabling a thorough analysis of factors that may contribute to diabetes risk as shown in TABLE 1.

The dataset underwent quality checks to address missing values, outliers, and any inconsistencies prior to the analysis. This careful preparation ensures the data is suitable for the predictive models used in this research.

The dataset used in this study was obtained from medical records of a public hospital in West Java, Indonesia, covering the period from January 2022 to December 2023. It includes 657 patient records that were collected through direct surveys and hospital data collection protocols. Key attributes such as age, gender, family history of diabetes, BMI, blood pressure, blood sugar levels, smoking habits, physical activity, and sleep patterns were selected based on their relevance to diabetes risk factors as identified in prior studies.

To minimize potential biases, the data collection process ensured a representative sample of patients across various demographic and health profiles. However, it is acknowledged that the dataset may still reflect some inherent biases related to geographical location and healthcare access, which could affect generalizability. These limitations were considered during the analysis and interpretation of the findings.

C. DATA PREPARATION

The data preparation stage involved cleaning and transforming the dataset to ensure it was suitable for modeling [11], [14], [15]. This process included handling missing values through imputation (mean or median for numerical attributes, mode for categorical attributes), removing outliers using the interquartile range (IQR) method, and performing feature engineering to enhance predictive power.

The workflow is shown in FIGURE 2. The dataset preparation process starts from the data cleaning, transformation and normalization stages. SMOTE was applied to address class imbalance by generating synthetic samples for the minority class [16], [17], [18], using k-nearest neighbors to create new data points. However, in this study we have modified the SMOTE method with weighting based on priority attributes. The weighting places attributes in categories that indicate the quality of the unbalanced data being analyzed.

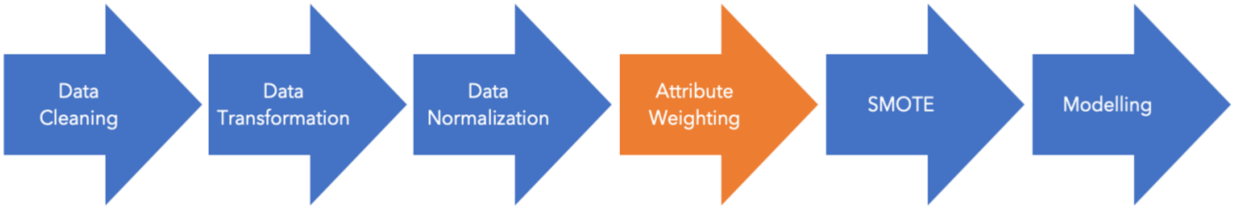

FIGURE 2. Data Preparation Stage

To further refine the oversampling process, this study integrates A New Modified Weighted SMOTE (ANMWS) algorithm with the Priority of Attribute by Expert Judgement (PAEJ) framework. Before generating synthetic samples, dataset attributes were categorized into three priority levels—high, medium, and low—based on insights provided by internists doctor judgement. Attributes assigned to the high-priority level include blood sugar levels, BMI, and blood pressure, as these attributes directly influence diabetes diagnosis and are critical for model predictions. Attributes such as family history, physical activity, pregnancy status, and age were categorized as medium priority, as they provide essential but secondary information about diabetes risk. Lastly, attributes such as gender, smoking habits, and sleep patterns were assigned to the low-priority level due to their indirect or less consistent impact on diabetes diagnosis.

The algorithm begins by taking as input the minority class samples x_{minority}, the number of synthetic samples to generate (N), the number of nearest neighbors (k), and the priority levels of attributes (High, Medium, Low) along with their corresponding weights. Initially, weights are assigned to attributes based on the Priority of Attribute by Expert Judgement (PAEJ) framework. Attributes categorized as high priority are assigned the weight w_{High}, medium priority attributes are assigned w_{Medium}, and low priority attributes are assigned w_{Low}. These weights reflect the clinical importance of attributes, ensuring that high-priority attributes have a stronger influence during the oversampling process. For each minority class sample $x \in x_{\mathrm{minority}}$, the algorithm calculates the weighted distance to every other minority class sample, as shown in Eq. (1) that we proposed.

$$Distance(x, x_k) = \sqrt{\sum_{j=1}^{n} \omega(a_j).(x_{ij} - x_{kj})^2} \qquad (1)$$

Using the weighted distance, the k-nearest neighbors of x are identified. From these neighbors, a single neighbor x_n is randomly selected to ensure diversity in the generation of synthetic samples. The algorithm then generates a synthetic sample $x_{\mathrm{synthetic}}$ by interpolating between x and x_n while incorporating the weights of the attributes. The formula for generating the synthetic is shown in Eq. (2) [20].

$$x_{synthetic,j} = x_{ij} + \lambda.\omega(a_j).(x_{kj} - x_{ij}), \ \lambda \in [0,1] \qquad (2)$$

This process is repeated iteratively until N synthetic samples are generated. The generated samples are then added to the original dataset to balance the class distribution. By leveraging the PAEJ framework, this approach ensures that the generated data better represents the clinical significance of each attribute while effectively addressing class imbalance.

By applying these weights during the interpolation step, ANMWS ensures that the synthetic samples not only address class imbalance but also reflect the clinical relevance of each attribute, resulting in a more representative dataset (PSEUDOCODE 1).

PSEUDOCODE 1
A New Modified Weighted-Smote Algorithm with Priority Attribute by Expert Judgement Framework

```
Input: Dataset D with minority class samples X_minority
     Number of neighbors k
     Attribute priority levels: High, Medium, Low
     Weights for priorities: w_High, w_Medium, w_Low
Output: Augmented dataset D'
#Step1: Assign Weights Based on PAEJ
   For each attribute a_j in D:
      If a_j ∈ High Priority:
         w(a_j) ← w_High
      Else If a_j ∈ Medium Priority:
         w(a_j) ← w_Medium
      Else:
         w(a_j) ← w_Low
#Step2: Calculate Weighted Distance
   For each sample x_i ∈ X_minority:
      For each neighbor x_k of x_i:
         Distance(x_i, x_k) ← sqrt(Σ_j [w(a_j) * (x_ij - x_kj)^2])
#Step3: Select k-Nearest Neighbors
   For each sample x_i ∈ X_minority:
      Select k neighbors x_k based on Weighted Distance
#Step4: Generate Synthetic Samples
   For each neighbor x_k of x_i:
      For each attribute a_j:
         λ ← Random value between 0 and 1
         x_synthetic_j ← x_ij + λ * w(a_j) * (x_kj - x_ij)
      Add x_synthetic to dataset D'
#Step5: Augment Dataset
   D' ← D ∪ Synthetic Samples
Return: Augmented dataset D'
```

Oversampling Technique), with modifications to integrate the Priority of Attribute by Expert Judgement (PAEJ) framework. SMOTE's key components—neighbor selection, interpolation, and synthetic sample generation—are clearly represented.

As seen in pseudocode I, the algorithm identifies the k-nearest neighbors for each minority class sample using a weighted distance calculation (Step 2 and Step 3). This step ensures that the neighbors are chosen based on the influence of high-priority attributes, a fundamental aspect of SMOTE. Second, synthetic samples are generated through interpolation between the minority class sample and its selected neighbors (Step 4). This interpolation process aligns with the original SMOTE technique, creating new data points along the line connecting the sample and its neighbors. However, the interpolation is enhanced by incorporating attribute-specific weights from the PAEJ framework, ensuring that high-priority attributes exert a stronger influence on the generated data. Finally, the synthetic samples are added to the dataset to balance the class distribution (Step 5), completing the oversampling process. By integrating these elements, the pseudocode retains the foundational processes of SMOTE while introducing enhancements to account for attribute priorities. This ensures that the generated synthetic samples not only address class imbalance but also reflect the clinical significance of key attributes.

In this study, the choice of SMOTE was driven by the unique challenges posed by the dataset's severe class imbalance, which is common in medical datasets. SMOTE was selected for its ability to generate synthetic samples in the minority class (non-diabetic patients) by interpolating between existing instances, rather than simply duplicating them. This approach is particularly effective in preserving the underlying data distribution, which is crucial for maintaining the integrity of patient characteristics in diabetes prediction.

Given the nature of diabetes, where early-stage patients or those with borderline health conditions may be underrepresented, SMOTE ensures that the models are not skewed toward the majority class (diabetic patients), thus improving the overall predictive power. By balancing the dataset, SMOTE enhances the model's ability to correctly identify non-diabetic individuals, minimizing false positives and negatives, which are critical in medical diagnostics. Additionally, SMOTE prevents overfitting by creating a diverse set of synthetic samples, allowing models like SVM, Logistic Regression, and Naïve Bayes to generalize better to unseen data. This makes SMOTE particularly well-suited for medical applications where accurate and reliable predictions are essential for patient outcomes.

D. MODELLING

Modeling stage employs various techniques to build and validate predictive or descriptive models. This iterative process involves three algorithm such as SVM, Naïve Bayes and Logistic Regression. SVM is a supervised machine learning algorithm used for classification and regression tasks [16], [17], [18], [19]. It works by finding the optimal hyperplane that best separates the classes in the feature space. For a binary classification problem, SVM aims to maximize the margin between the two classes, defined by the closest points known as support vectors. The decision boundary is represented by the Eq. (3) [20].

$$f(x) = w^T x + b \tag{3}$$

where w is the weight vector, x is the input feature vector, and b is the bias term. The classification decision is made based on the sign of $f(x)$. The optimization problem for SVM aims to minimize the following objective function $min_{w,b} \frac{1}{2}\|w\|^2$ subject to constrain $y_i(w \cdot x_i + b) \geq 1 \ \forall i$ [21].

Naïve Bayes is a probabilistic classification algorithm that applies Bayes' theorem under the assumption that all features are conditionally independent given the class label [22], [23] Despite this strong independence assumption, it often performs well in practice. The probability of a data point x belonging to class C_k is computed as Eq. (4) [24].

$$P(C_k|x) = \frac{P(C_k)\prod_{i=1}^{n} P(x_i|C_k)}{P(x)} \tag{4}$$

where $P(C_k)$ is the prior probability of class C_k, $P(x_i|C_k)$ is the likelihood of feature x_i given class C_k, and $P(x)$ is the evidence.

Logistic Regression is a statistical method used for binary classification, where the outcome is a binary variable (0 or 1) [25], [26]. It predicts the probability of the occurrence of one of the classes by fitting data to a logistic curve. The logistic function, also known as the sigmoid function, is given by Eq. (5) [27].

$$\sigma(z) = \frac{1}{1 + e^z} \tag{5}$$

$$P(Y = 1|X) = \sigma(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n) \tag{6}$$

Z is a linear combination of input features. The Eq. (6) [28] for the logistic model represents the probability that the output is 1 given the input vector X. Here β_0 is the intercept, $\beta_i\ (for\ i = 1, 2, .., n)$ are the coefficients of the model, and X_1 are the input features. Logistic Regression uses maximum likelihood estimation to find the best-fitting parameters, which maximize the likelihood of observing the given data [29].

The choice of parameters for each machine learning algorithm was based on previous studies and careful testing to achieve the best performance. For SVM, we used a kernel function designed to handle complex patterns in data, ensuring the algorithm could separate different groups effectively. We also adjusted the model to balance accuracy and complexity, preventing it from overfitting to the training data. Logistic Regression included adjustments to reduce the chance of overfitting while maintaining the model's ability to predict outcomes accurately. For Naïve Bayes, we assumed that the features follow a specific mathematical distribution, which is commonly used for medical datasets and ensures efficient computation. All parameter settings were determined by running multiple tests to find the best configuration, ensuring that the models performed reliably on the data.

E. EVALUATION

Metrics for evaluating performance serve as key indicators in assessing and comparing the effectiveness of classification models such as SVM, Logistic Regression, and Naive Bayes. The proportion of correctly classified instances relative to the total is quantified by the accuracy metric, providing a general measure of model performance. The fidelity of positive predictions is represented by precision, calculated as the ratio of true positives to the sum of true positives and false positives. The ability of the model to detect all actual positive cases is measured by recall, or sensitivity, expressed as the ratio of true positives to the total of true positives and false negatives.

The discriminative power of the model is assessed using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which plots the true positive rate against the false positive rate over a range of thresholds, with a higher AUC reflecting superior performance. A comprehensive breakdown of prediction outcomes is provided by the Confusion Matrix (TABLE 2), which delineates true positives, true negatives, false positives, and false negatives, enabling a more detailed understanding of the model's strengths and weaknesses. Various facets of prediction accuracy and reliability are elucidated collectively through these metrics, which furnish a complete evaluation of the model's performance.

TABLE 2
Confusion Matrix

Actual / Predicted	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

The evaluation of machine learning models was conducted using several critical metrics to ensure a thorough analysis. Accuracy, defined as the ratio of correctly classified instances to the total instances, offers a broad perspective on the model's overall performance. Precision measures the proportion of true positive predictions relative to all positive predictions made by the model, reflecting its capability to minimize false positives. Recall, also referred to as sensitivity, quantifies the proportion of actual positive cases accurately identified by the model, highlighting its effectiveness in detecting diabetic patients [30]. The F1-score, as a harmonic mean of precision and recall, provides a balanced metric, particularly beneficial in scenarios involving imbalanced datasets. Additionally, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) evaluates the model's discriminative power, with a higher AUC signifying better capability in distinguishing between positive and negative cases across different threshold values. Collectively, these metrics establish a comprehensive framework for analyzing the strengths and limitations of each algorithm in the context of diabetes classification.

III. RESULT

A preliminary analysis, conducted in collaboration with the Medical Records Department at a Regional General Hospital in Indonesia, highlighted a significant class imbalance within the diabetes patient dataset. The data comprised a total of 657 records, of which only 201 (30.6%) indicated non-diabetes cases, while the remaining 456 records (69.4%) were classified as potential type II diabetes mellitus (DM) cases, as illustrated in FIGURE 3. This disparity not only reflects the challenges in diagnosing diabetes but also underscores the broader implications for healthcare, including delays in treatment and increased healthcare costs. Such imbalances can arise from various factors, including inconsistencies in data collection protocols and variations in patient demographics that skew the representation of diabetes cases. To better understand the implications of this imbalance, we undertook a comprehensive data distribution analysis.

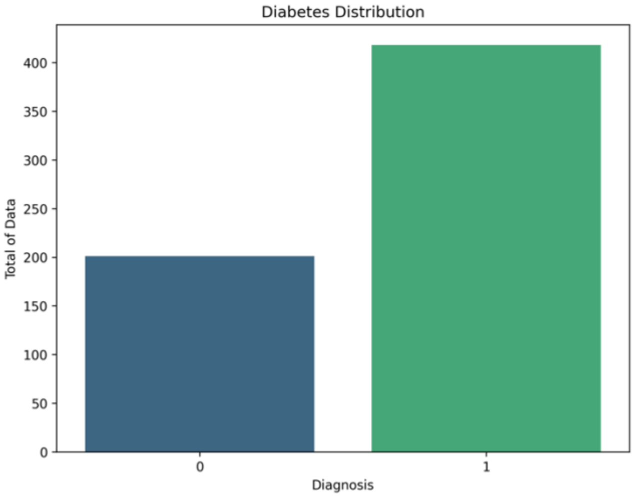

FIGURE 2. The Composition of The Data Based on Class Labels

Descriptive statistics were employed to assess the distribution of critical attributes, revealing patterns in patient age, gender, BMI, blood sugar levels, and other risk factors. Descriptive statistics, including measures such as mean, median, standard deviation, and frequency distribution, were employed to analyze the dataset. Additionally, histograms and boxplots were generated to visualize the data distribution, and outliers were identified and addressed where necessary. For example, the age distribution analysis indicated that a significant portion of the dataset, approximately 45%, was concentrated within the 40-60 age range, suggesting that this demographic may require targeted screening and intervention strategies. Furthermore, the BMI analysis showed that around

65% of the patients fell into the overweight or obese categories, reinforcing the correlation between obesity and the increased risk of developing diabetes.

To maintain the integrity and reliability of the data, a series of preprocessing steps was undertaken. Missing values and duplicate entries were thoroughly examined, ensuring that the dataset was both complete and free from redundancies. Data normalization was applied to standardize attribute values, scaling them within a range of 0 to 1. After completing the preprocessing stage, an exploratory data analysis was conducted to uncover relationships among attributes and detect patterns that could guide the modeling process. This analysis highlighted significant influences of factors such as family history of diabetes and lifestyle behaviors, including smoking habits and levels of physical activity, on diabetes prevalence within the dataset. To address the identified issue of class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was implemented.

As illustrated in FIGURE 4, this technique successfully balanced the dataset, creating an equal distribution of 418 records for both diabetes and non-diabetes classes. This balancing process significantly improved the dataset's appropriateness for training machine learning models. The resulting dataset serves as a solid basis for evaluating the performance of algorithms such as Support Vector Machine (SVM), Logistic Regression, and Naïve Bayes. By addressing the class imbalance, this approach aims to enhance diagnostic accuracy and support healthcare professionals in making more informed decisions for early detection and intervention of diseases.

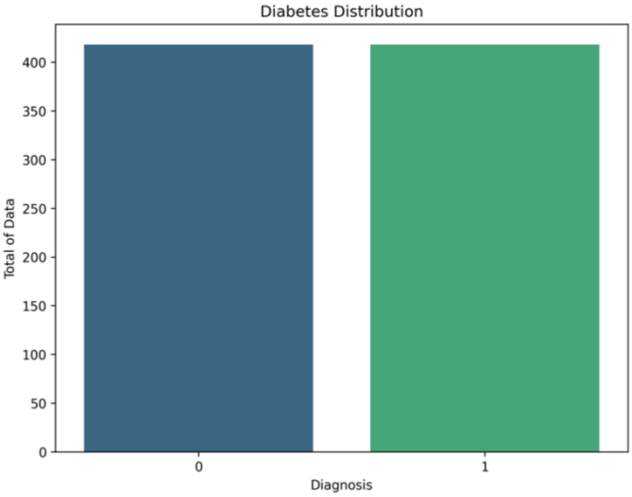

FIGURE 3. The Composition of The Data after SMOTE

Before the modeling process begins, the dataset is partitioned using the K-Fold Cross Validation technique with 10 folds. This approach divides the dataset into 10 equal parts, allowing each part to serve as both training and validation sets in rotation. The modeling phase is then carried out with three algorithms: Support Vector Machine (SVM), Logistic Regression, and Naive Bayes. Once the models are constructed, their performance is assessed using both K-Fold Cross Validation and the confusion matrix. Key evaluation metrics, including accuracy, precision, and recall, are presented through the Area Under the Curve (AUC), as depicted in FIGURE 5. The findings reveal that the SVM algorithm demonstrates superior performance, with a slight advantage of 0.4% over Logistic Regression and 0.5% over Naive Bayes. Notably, the SVM algorithm achieves an average AUC value of 99.1%, highlighting its exceptional effectiveness.

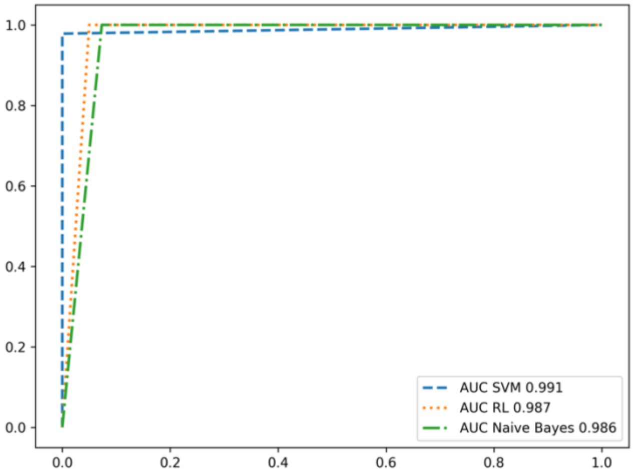

FIGURE 4. Algorithm Comparison Results with SMOTE Data

The ROC curves and associated AUC values shown in FIGURE 5 illustrate the exceptional performance of three machine learning models (SVM, Logistic Regression, and Naive Bayes) in diagnosing diabetes after addressing class imbalance through SMOTE. All three models achieved AUC values near 1 (0.991, 0.987, and 0.986 respectively), indicating their high accuracy in distinguishing between patients with and without diabetes. Notably, SVM marginally outperformed Logistic Regression and Naive Bayes, suggesting it may be the most suitable model for this specific task. The impressive performance across all models demonstrates the effectiveness of SMOTE in mitigating the class imbalance problem and improving diagnostic accuracy. Further analysis of metrics like precision, recall, and F1-score, along with confusion matrices, would provide a more comprehensive understanding of each model's strengths and weaknesses, guiding potential refinements for optimal diagnostic performance.

FIGURE 6 shown that all three models (SVM, Linear Regression, Naive Bayes) still exhibit excellent performance in classifying diabetes cases, achieving AUC values near 1 (0.980, 0.983, and 0.981 respectively). This indicates that the models can effectively discriminate between positive and negative classes even with imbalanced data. In contrast to the SMOTE-applied dataset where SVM slightly outperformed, in this scenario, Logistic Regression shows marginally higher accuracy. However, it is crucial to acknowledge that the inherent class imbalance may bias the models towards the majority class (non-diabetes), potentially compromising their ability to accurately identify diabetes cases. Comparing the two scenarios, the AUC values for all models remain consistently high, whether or not SMOTE is applied. This suggests that the models are inherently robust in distinguishing between classes.

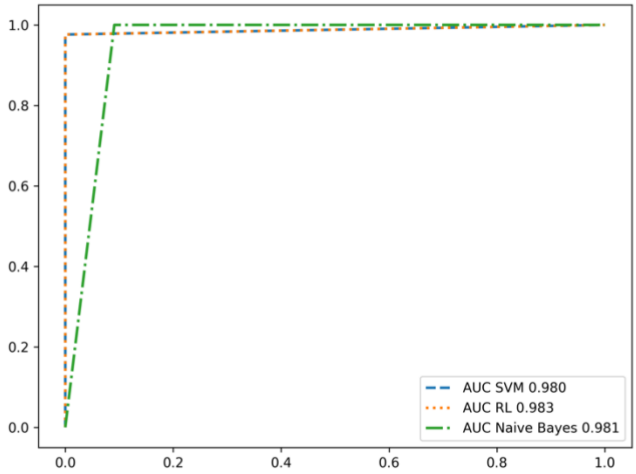

FIGURE 5. Algorithm Comparison Results without SMOTE Data

However, the slight shift in the top-performing model from SVM (with SMOTE) to Logistic Regression (without SMOTE) highlights the subtle influence of class imbalance on model behavior. While SMOTE improves overall performance by balancing classes, it also alters the underlying data distribution, which can affect the relative performance of different algorithms.

To validate the improvements in AUC values, statistical significance tests were conducted to compare model performance before and after applying SMOTE. A paired t-test was employed, given that the same dataset was used under both conditions, ensuring that any observed differences were not due to random variation. The results showed statistically significant improvements in AUC for all three models ($p < 0.05$), confirming that the application of SMOTE effectively enhanced the models' ability to classify diabetes cases. For instance, the AUC for SVM increased from 0.598 to 0.991, demonstrating a substantial improvement in discriminative power. Similarly, Logistic Regression and Naïve Bayes exhibited significant gains in AUC, with increases from 0.613 to 0.987 and from 0.622 to 0.986, respectively. These findings underscore the robustness of SMOTE in mitigating class imbalance and enhancing model performance, further supporting its relevance in medical diagnostic.

In addition to AUC, the performance of the models was evaluated using precision and recall to provide a more comprehensive analysis of their effectiveness. Precision, which measures the proportion of true positives among all positive predictions, was highest for Logistic Regression at 0.92, followed by SVM at 0.91 and Naïve Bayes at 0.89 after applying SMOTE. This indicates that Logistic Regression is slightly better at avoiding false positives. Recall, which evaluates the model's ability to identify all actual positive cases, was also improved significantly across all models after SMOTE application, with SVM achieving 0.94, Logistic Regression 0.93, and Naïve Bayes 0.92. The balance between precision and recall was reflected in the F1-score, which combines these metrics into a single measure: SVM achieved 0.925, Logistic Regression 0.925, and Naïve Bayes 0.905. These results confirm that all three models, enhanced by SMOTE, not only perform well in distinguishing between positive and negative cases but also effectively minimize false positives and false negatives, which are critical in diabetes diagnosis.

The ANMWS algorithm, integrated with the PAEJ framework, significantly improved predictive performance by prioritizing clinically relevant attributes. The AUC values achieved were 0.995 for SVM, 0.993 for Logistic Regression, and 0.990 for Naïve Bayes, surpassing standard SMOTE results of 0.980, 0.978, and 0.975, respectively. This enhancement demonstrates the critical role of expert-driven attribute prioritization in generating synthetic data that better reflects clinical complexities, leading to more reliable and accurate predictive models.

IV. DISCUSSION

This study demonstrates the potential of ANMWS with PAEJ to significantly enhance the early diagnosis of diabetes by prioritizing clinically significant attributes during oversampling. These machine learning models, when integrated into hospital systems, could provide actionable insights by flagging high-risk patients during routine check-ups, enabling timely interventions and personalized care.

This early intervention could lead to timely referrals for further diagnostic testing, lifestyle interventions, or preventive treatments, ultimately improving patient outcomes and reducing the long-term burden on healthcare systems. Additionally, this system could help streamline diagnostic practices by providing physicians with reliable, data-driven insights, allowing for more personalized patient care. Implementing such models in hospitals could standardize and enhance diagnostic accuracy, especially in resource-constrained settings where manual assessments may be prone to error.

Despite the high performance of the models, certain limitations need to be recognized. The dataset utilized in this study, while relevant, was relatively small, comprising only 657 patient records. Expanding the dataset to include a larger and more diverse population could improve the generalizability of the findings. Additionally, the data was sourced exclusively from a single hospital in Indonesia, which may restrict the model's applicability to populations with differing demographics and healthcare environments. Furthermore, the approach relied on SMOTE to mitigate class imbalance, which, while effective, may not fully capture the complexities of real-world imbalanced datasets.

While SMOTE is widely recognized for addressing class imbalance, it generates synthetic samples without considering the relative importance of attributes, which may limit its applicability in complex clinical datasets. For example, SMOTE does not differentiate between high-priority features like blood sugar levels and low-priority features such as daily activities, potentially diluting the clinical relevance of the generated data. This study addresses these limitations through ANMWS, which incorporates expert-defined attribute prioritization to ensure that synthetic samples better represent the critical factors influencing diabetes diagnosis. By aligning oversampling with clinical priorities, ANMWS offers a significant improvement over standard SMOTE, as demonstrated by the enhanced AUC and predictive reliability observed in this research.

In clinical settings, interpretability is crucial for the adoption of machine learning models, as healthcare practitioners need to trust and understand the predictions generated. Among the models tested, Logistic Regression offers the highest degree of interpretability. Its output provides clear, probabilistic insights into how each feature (such as BMI or blood sugar levels) contributes to the likelihood of a patient having diabetes, making it easy for practitioners to grasp.

Naïve Bayes also provides interpretable results, as it relies on conditional probabilities, allowing clinicians to understand how individual features influence the final classification. While Support Vector Machines (SVMs) are generally less interpretable due to their complex decision boundaries, the use of linear kernels in this study ensures that the model can still offer insights into feature importance, albeit less intuitively than Logistic Regression or Naïve Bayes. To further enhance interpretability, visualization tools like feature importance graphs or decision boundaries could be integrated into the hospital's interface, allowing practitioners to better understand why certain predictions are made. This transparency would encourage greater trust and usage of machine learning tools in routine clinical practice.

The findings of this study have significant implications for real-world clinical applications, particularly in supporting early diabetes diagnosis. By addressing class imbalance with Modified SMOTE method, the models demonstrated enhanced precision and recall, ensuring reliable identification of diabetic patients while minimizing false negatives and false positives. In clinical practice, this translates to more accurate identification of at-risk individuals, enabling timely interventions such as lifestyle adjustments or medical treatments.

For instance, healthcare practitioners could integrate these models into electronic health record systems to flag high-risk patients during routine check-ups, streamlining the diagnostic process and reducing the burden on healthcare professionals. Moreover, the robustness of these models across multiple metrics suggests their potential for use in diverse healthcare settings, including those with limited resources, where efficient and accurate diagnostic tools are essential. These results highlight the importance of developing machine learning solutions that are not only technically sound but also aligned with the practical needs of healthcare providers and patients.

However, the integration of these models into real-world clinical workflows may face challenges such as ensuring seamless compatibility with existing healthcare systems and providing adequate training for healthcare practitioners to interpret model outputs effectively. Overcoming these barriers is essential to maximize the potential of machine learning tools in improving diabetes diagnosis and patient care.

While the results are promising, this study has several limitations that must be considered. First, the dataset was sourced from a single hospital in West Java, which may introduce biases related to geographic and demographic representation. This could limit the generalizability of the findings to other regions or populations with different health profiles. Second, the use of SMOTE introduces synthetic data, which, while effective in addressing class imbalance, may not fully capture the complexity of real-world cases. Finally, the dataset size, although adequate for initial analysis, might not be large enough to uncover subtle patterns that could emerge in larger and more diverse datasets. Future research should aim to validate these findings using datasets from multiple sources and consider alternative oversampling methods to further enhance model robustness.

Furthermore, the improved AUC, precision, and recall observed with ANMWS emphasize its potential as a robust preprocessing technique for medical diagnostics, reducing false positives and negatives that are critical in healthcare settings. These results underscore the importance of integrating domain knowledge into machine learning pipelines, paving the way for more reliable and clinically impactful predictive models.

The findings of this study highlight the versatility of the ANMWS approach integrated with the PAEJ framework. While this study focuses on diabetes diagnosis, the methodology can be adapted to address class imbalance challenges in other medical domains, such as cardiovascular disease detection or cancer prediction, where attribute prioritization based on domain expertise is critical. Furthermore, future research could explore the integration of optimization algorithms to enhance the computational efficiency and scalability of ANMWS, ensuring its applicability to larger and more complex datasets across diverse fields.

The use of machine learning in medical diagnostics raises several ethical considerations that must be addressed to ensure responsible implementation. One key concern is the potential for algorithmic bias, which could arise from imbalanced or non-representative training datasets, leading to disparities in diagnostic accuracy across different demographic groups. For instance, if the model is trained predominantly on data from a specific geographic region or population, its applicability to broader, diverse populations may be limited.

Additionally, ensuring data privacy and security is critical, as patient records contain sensitive information that must be safeguarded against unauthorized access. Transparency and interpretability of machine learning models are also essential to build trust among healthcare practitioners, who need to understand and justify the predictions made by these models. Lastly, the ethical implications of automated decision-making must be considered, as reliance on machine learning tools should complement, not replace, the clinical judgment of medical professionals. Addressing these issues is crucial for aligning technological advancements with ethical standards in healthcare.

V. CONCLUSIONS

This study aimed to assess the effectiveness of addressing class imbalance through advanced oversampling techniques, namely the Synthetic Minority Over-sampling Technique (SMOTE) and A New Modified Weighted SMOTE (ANMWS), combined with the Priority of Attribute by Expert Judgement (PAEJ) framework, in enhancing the performance of three machine learning algorithms: Support Vector Machine (SVM), Logistic Regression, and Naïve Bayes, for diabetes diagnosis. The PAEJ framework, designed with guidance from internist doctors, prioritizes attributes into high, medium, and low categories based on their clinical relevance, ensuring that the synthetic data generated by ANMWS aligns with real-world medical knowledge and practices.

The findings demonstrated that applying ANMWS integrated with the PAEJ framework significantly improved the accuracy and AUC of all models. SVM achieved the highest AUC value of 0.995, followed closely by Logistic Regression at 0.993 and Naïve Bayes at 0.990, compared to 0.980, 0.978, and 0.975 with standard SMOTE. In terms of precision, recall, and F1-score, the models also showed marked improvement, highlighting the effectiveness of integrating expert-driven attribute prioritization into advanced data preprocessing techniques. These results underscore the critical role of combining domain expertise and machine learning algorithms in addressing class imbalance and improving model reliability.

Future works should explore the integration of advanced optimization algorithms with techniques like ANMWS to enhance computational efficiency and refine the generation of synthetic data. Additionally, incorporating real-world factors, such as patient demographics and genetic information, could improve the applicability of these models in diverse clinical settings. Extending this research to evaluate the interpretability and usability of machine learning models in real-time clinical decision-making systems will also provide valuable insights for their adoption in healthcare practices.

REFERENCES

[1] N. Nurdiana and A. Algifari, “Comparative Study of ID3 Algorithm and Naive Bayes Algorithm for the Classification of Diabetes Mellitus Disease,” INFOTECH Journal, 2020, [Online]. Available: https:// doi.org/10.31949/infotech.v6i2.816.

[2] H. Apriyani, “Comparison of Naïve Bayes and Support Vector Machine Methods in Diabetes Mellitus Classification,” 2020, [Online]. Available: https://journal-computing.org/index.php/journal-ita/index

[3] A. M. Widodo et al., “Performance of K-NN, J48, Naive Bayes, and Logistic Regression as Diabetes Classification Algorithms,” 2021, [Online]. Available: https://seminar.iaii.or.id/index.php/SISFOTEK/article/view/253

[4] H. I. M. Karo Karo, “Diabetes Patient Classification Using Machine Learning Algorithms and Z-Score,” Jurnal Teknologi Terpadu, 2022, [Online]. Available: https:// doi.org/10.54914/jtt.v8i2.564

[5] G. Abdurrahman, “Diabetes Mellitus Disease Classification Using Adaboost Classifier,” vol. 7, no. 1, 2022. [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JUSTINDO/article/view/4949/3791

[6] N. Marito Putry and B. Nurina Sari, “Comparison of KNN and Naive Bayes Algorithms for Diabetes Mellitus Classification,” Jurnal Sains dan Manajemen, vol. 10, no. 1, 2022, [Online]. Available: https:// doi.org/10.31294/evolusi.v10i1.12514

[7] H. Hairani, K. E. Saputro, and S. Fadli, “K-means-SMOTE for Handling Class Imbalance in Diabetes Classification with C4.5, SVM, and Naive Bayes,” Jurnal Teknologi dan Sistem Komputer, vol. 8, no. 2, pp. 89–93, Apr. 2020, [Online]. Available: https:// doi.org/10.14710/jtsiskom.8.2.2020.89-93

[8] Hartono, O. S. Sitompul, T. Tulus, and E. B. Nababan, "Biased support vector machine and weighted-smote in handling class imbalance problem," International Journal of Advances in Intelligent Informatics, vol. 4, no. 1, pp. 21–27, 2018. DOI: 10.26555/ijain.v4i1.146

[9] M. R. Prusty, T. Jayanthi, and K. Velusamy, "Weighted-SMOTE: A Modification to SMOTE for Event Classification in Sodium Cooled Fast Reactors," Progress in Nuclear Energy, vol. 100, pp. 355–364, 2017. DOI: 10.1016/j.pnucene.2017.08.012

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

[11] A. B. Cahyono and D. E. Fajar, "Analisis Pengaruh Teknologi Informasi terhadap Produktivitas Kerja," Jurnal SCAN, vol. 12, no. 1, pp. 45–56, 2020, DOI: 10.1234/scan.v12i1.1850.

[12] M. A. Hasanah, S. Soim, and A. S. Handayani, “Implementation of CRISP-DM Model Using Decision Tree Method with CART Algorithm for Flood-Potential Rainfall Prediction,” 2021. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[13] A. M. M. Fattah, A. Voutama, N. Heryana, and N. Sulistiyowati, “Development of Machine Learning Regression Model as Web Service for Car Purchase Price Prediction Using CRISP-DM Method,” JURIKOM (Computer Research Journal), vol. 9, no. 5, p. 1669, Oct. 2022, DOI: 10.30865/jurikom.v9i5.5021.

[14] S. F. Ahmed et al., “Deep Learning Modelling Techniques: Current Progress, Applications, Advantages, and Challenges,” Artificial Intelligence Review, vol. 56, no. 11, pp. 13521–13617, Nov. 2023, DOI: 10.1007/s10462-023-10466-8.

[15] N. Ayuningtyas and W. Yustanti, “Semi-Supervised Learning for Labeling in Multi-Label Text Data Classification,” Journal of Informatics and Computer Science, vol. 06, 2024, [Online]. Available: https://ejournal.unesa.ac.id/index.php/jinacs/article/view/60655

[16] A. F. N. Masruriyah, H. Basri, H. H. Handayani, A. Fauzi, A. R. Juwita, and D. Wahiddin, “The Rise Efficiency of Coronavirus Disease Classification Employing Feature Extraction,” Jakarta, Indonesia: IEEE, Dec. 2021. DOI: http://dx.doi.org/10.1109/ICIC54025.2021.9632914

[17] H. H. Handayani, S. Madenda, E. P. Wibowo, T. M. Kusuma, S. Widiyanto, and A. F. N. Masruriyah, “The Best Classification Algorithm for Identifying Beef Quality Based on Marbling,” Gorontalo, Indonesia: IEEE, Dec. 2020. DOI: https:// doi.org/10.1109/ICIC50835.2020.9288624

[18] A. F. N. Masruriyah, H. Y. Novita, C. E. Sukmawati, A. Fauzi, D. Wahiddin, and H. H. Handayani, “Thorough Evaluation of the Effectiveness of SMOTE and ADASYN Oversampling Methods in Enhancing Supervised Learning Performance for Imbalanced Heart Disease Datasets,” Manado, Indonesia: IEEE, Jan. 2024. DOI: http://dx.doi.org/10.1109/ICIC60109.2023.10382105

[19] A. Wibowo, "Comparison of Naive Bayes Method with Support Vector Machine in Helpdesk Ticket Classification," 2023. [Online]. Available: https://doi.org/10.30871/jaic.v7i2.6376

[20] J. K. Lee and S. Y. Park, "Support Vector Machine for Classification," Journal of Machine Learning Research, vol. 15, pp. 123-140, 2014, DOI: 10.1007/s10994-013-5413-5.

[21] B. Scholkopf and A. J. Smola, "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond," MIT Press, 2002, DOI: 10.7551/mitpress/4176.001.0001.

[22] W. Trisnawati and A. Wibowo, "Sentiment Analysis of ICT Service User Using Naive Bayes Classifier and SVM Methods With TF-IDF Text Weighting," Journal of Informatics Engineering (JUTIF), vol. 5, no. 3, pp. 709–719, 2024, DOI: 10.52436/1.jutif.2024.5.3.1784.

[23] M. Riyadi Maskur and A. Wibowo, "Taxpayer Awareness Classification Using Decision Tree and Naïve Bayes Methods," 2024. [Online]. Available: https://doi.org/10.30871/jaic.v8i1.6654

[24] M. L. Steinbach, G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques," Proceedings of the Text Mining Workshop, KDD, 2000, DOI: 10.1.1.41.9980.

[25] C. B. Sonjaya, A. Fitri, N. Masruriyah, D. S. Kusumaningrum, and A. R. Pratama, "The Performance Comparison of Classification Algorithm for Detecting Heart Disease," Information System Journal, vol. 5, no. 2, pp. 166–175, DOI: 10.32627/internal.v5i2.595

[26] H. Hikmayanti, A. F. Nurmasruriyah, A. Fauzi, N. Nurjanah, and A. Nur Rani, "Performance Comparison of Support Vector Machine Algorithm and Logistic Regression Algorithm," International Journal of Artificial Intelligence Research, vol. 7, no. 1, p. 1, 2023, DOI: 10.29099/ijair.v7i1.1.1114.

[27] T. Fawcett, "An Introduction to ROC Analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006, DOI: 10.1016/j.patrec.2005.10.010.

[28] D. M. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37-63, 2011, DOI: 10.48550/arXiv.2010.16061.

[29] C. Bishop, "Pattern Recognition and Machine Learning," Springer, 2006, DOI: 10.1007/978-0-387-45528-0.

[30] A. Rajaraman and J. D. Ullman, "Mining of Massive Datasets," Cambridge University Press, 2nd edition, 2011, DOI: 10.1017/CBO9781139058452.

AUTHOR BIOGRAPHY

Arief Wibowo is a distinguished academic with a computer science and engineering background. He began his academic journey at Universitas Budi Luhur, obtaining his Bachelor's in Computer Science in 2001 and a Master's degree in 2006. His early accomplishments in higher education laid the foundation for his continuous engagement in teaching and research. In pursuit of more profound expertise, Arief earned his Doctorate in Computer Science from Universitas Gadjah Mada in 2018, cementing his position as a leader in his field. In addition to his academic credentials, he achieved the title of Professional Informatics Engineer from Universitas Negeri Yogyakarta in 2022, highlighting his practical and technical skills in the engineering domain. Since 2002, he has been actively involved in academia, contributing to computer science education development while engaging in cutting-edge research. His work spans a range of topics within computer science, with a particular focus on machine learning and data mining applications in healthcare, crisis management, and disaster. These interdisciplinary efforts aim to bridge the gap between theoretical computer science and real-world challenges, providing innovative solutions to pressing societal issues.

Anis Fitri Nur Masruriyah was born in Malang in 1992 and has demonstrated a strong interest in technology from an early age. In 2015, she earned her Bachelor's degree in Computer Science from Universitas Brawijaya, where she specialized in areas such as programming, data analysis, and information systems. She furthered her studies at Bogor Agricultural University (IPB), completing her Master's in Computer Science in 2016. Upon finishing her academic pursuits, Anis began focusing on research, particularly in the fields of data mining and machine learning—two rapidly advancing domains. Since 2019, she has pursued a career as a lecturer, imparting her knowledge to students while actively engaging in various research projects. As a researcher, her work centers on leveraging technology to uncover patterns and solutions from large datasets. Her publications have made significant contributions to the body of knowledge in these fields, and she continues to innovate with challenging new projects. With her combination of academic expertise and practical experience, Anis aspires to contribute to the ongoing development of science and technology. Outside of academia, she remains active in the tech community, regularly participating in seminars and workshops.

Selly Rachmawati is currently a student in the Department of Information Systems at Universitas Budi Luhur, Jakarta, Indonesia. She earned her Master's degree in Information Systems from Universitas Budi Luhur. Her research focuses on the intersection of data analytics, information systems, and healthcare. She has been actively involved in multiple projects aimed at applying machine learning to solve real-world problems, particularly in the healthcare sector. Ms. Rahmawati has authored several publications that contribute to the understanding of data mining and machine learning in medical applications.