Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal Vol. 6, No. 4, October 2024, pp: 343-354; eISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

Manuscript received May 2, 2024; revised May 23, 2024; accepted May 27, 2024; date of publication August 8, 2024
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v6i4.434
Copyright © 2024 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).
How to cite: M. Khairul Rezki, Muhammad Itqan Mazdadi, Fatma Indriani, Muliadi, Triando Hamonangan Saragih and Vijay Annant Athavale, "Application of Smote to Address Class Imbalance in Diabetes Disease Categorization Utilizing C5.0, Random Forest, and Support Vector Machine", Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 6, no. 4, pp. 343-354, October 2024.

Application of Smote to Address Class Imbalance in Diabetes Disease Categorization Utilizing C5.0, Random Forest, and Support Vector Machine

M. Khairul Rezki[1], Muhammad Itqan Mazdadi[1], Fatma Indriani[1], Muliadi[1], Triando Hamonangan Saragih[1], and Vijay Annant Athavale[2]

[1] Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University, Indonesia
[2] Department of Computer Science and Engineering, Walchand Institute of Technology, Solapur, Maharashtra, India

Corresponding author: Muhammad Itqan Mazdadi (email: mazdadi@ulm.ac.id).

This work was supported by Lambung Mangkurat University for providing valuable resources and support.

ABSTRACT The use of categorization techniques for classifying diabetes often leads to poor results because of the complex nature of the dataset and the uneven class distribution. To address the class imbalance, SMOTE is often applied, but this also results in suboptimal outcomes due to the dataset's complexity and the numerous influencing factors. As a result, multiple tests were carried out to assess the accuracy of different classification methods. This study seeks to assess the accuracy of C5.0, Random Forest, and SVM classification models using both standard and SMOTE-based methods. The approach involves selecting appropriate datasets, reviewing classification algorithms like C5.0, Random Forest, and SVM, applying the SMOTE technique, validating through split validation, preprocessing with min-max normalization, and evaluating performance with confusion matrices and AUC analysis. The dataset was sourced from Kaggle to address the class imbalance in a diabetes dataset using the SMOTE technique. The dataset comprises 768 instances, with 268 representing individuals with diabetes and 500 representing those without. Before applying SMOTE, the accuracy of classification using C5.0, Random Forest, and SVM was 0.714, 0.733, and 0.746, respectively. The AUC values for the dataset were 0.745, 0.824, and 0.799. After applying the SMOTE technique, the accuracy values for the same models were 0.603, 0.727, and 0.727, with corresponding AUC values of 0.734, 0.831, and 0.794. This analysis indicates that SMOTE had a minimal impact on the performance of the three classification models. The decrease in performance, including precision and AUC scores, is likely due to the risk of overfitting on the dataset. This happens because the models become overly reliant on the synthetic data generated for the minority classes, which adversely affects their overall effectiveness.

INDEX TERMS SMOTE, C5.0, Random Forest, SVM, Diabetes.

I.INTRODUCTION

Diabetes mellitus (DM) continues to be a major global health issue, with its prevalence steadily increasing. According to the World Health Organization, more than 422 million people worldwide suffer from diabetes, with the majority of cases found in lowand middle-income countries. Additionally, diabetes is responsible for 1.5 million deaths each year. One common issue faced by individuals with diabetes mellitus (DM) is the development of chronic wounds that heal slowly and often progress into diabetic ulcers [1]. Diabetes mellitus is a long-term condition resulting from multiple causes, characterized by high blood sugar levels and impaired carbohydrate, lipid, and protein metabolism due to insufficient insulin secretion or action. This condition necessitates greater

attention to both management and prevention strategies [2]-[4]. Class imbalance occurs when there is a significant disparity in the number of samples among different categories within a dataset. Typically, classification algorithms perform better at identifying the majority class, struggling with the minority class. The Synthetic Minority Over-sampling Technique (SMOTE) addresses this problem by generating synthetic data to enhance the representation of minority classes through oversampling. In the study conducted by [5] , using ANN combined with SMOTE achieved a precision rate of 87.06%, whereas using ANN without SMOTE resulted in a precision rate of 86.35%. The use of the Synthetic Minority Over-sampling Technique (SMOTE) has proven effective in addressing class imbalance, resulting in enhanced classification performance.

C5.0 is a classification algorithm known for its high accuracy, precision, and recall. In a study by [6], C5.0 was compared to the C4.5 algorithm, revealing that the inclusion of a boosting stage in C5.0 enhances its precision. As a result, C5.0 achieved a precision rate of 99.33%, whereas C4.5 attained 87.61%.

Random Forest is a supervised learning algorithm used for classification tasks. Its strength comes from its capability to select features and nodes randomly, which helps reduce errors during processing [7]. The study conducted by [8] determined that the Random Forest algorithm surpasses both Naïve Bayes and Decision Tree in classification tasks. The study reported accuracy rates of 78% for Naïve Bayes, 76% for Decision Tree, and 84% for Random Forest. These results indicate that Random Forest is the most effective and suitable method for classifying this dataset.

The Support Vector Machine (SVM) is a statistical method used in classification tasks. It works by finding a hyperplane that can efficiently separate data points into two different classes [9]. In study [10], a comparison was made between the C4.5 and SVM classification methods for categorizing cardiovascular disease. The results showed that the C4.5 algorithm had lower precision than SVM, achieving an accuracy of 82% compared to SVM's 88%.

Based on the description above, researchers choose to use the C5.0, Random Forest, and SVM classification methods because previous studies have shown that these methods achieve a high level of precision. Therefore, this study will compare these three classification models to determine which one produces the highest AUC and confusion matrix values before and after applying SMOTE to address data imbalance.

The aim of this research is to evaluate the accuracy of different analytical models, including C5.0, Random Forest, and SVM, both with and without the application of SMOTE. By incorporating SMOTE, it is expected that the performance of these models in accurately classifying diabetes will improve. The results of this study are expected to provide valuable insights, such as:

a. Improve understanding of how classification algorithms are applied to diabetes datasets.
b. Aid in enhancing decision-making processes through analytical evaluation.
c. Use the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance.

II.METHODS

The research methodology provides a detailed explanation of the datasets used and describes the underlying principles of the C5.0, Random Forest, and Support Vector Machine algorithms. It also outlines the application of SMOTE and Min-Max Normalization techniques. The study discusses performance evaluation methods, including Confusion Matrix and AUC analysis. The data is split into 80% for training and 20% for testing. The following study procedures are described, with Figure 1 illustrating the study's workflow.

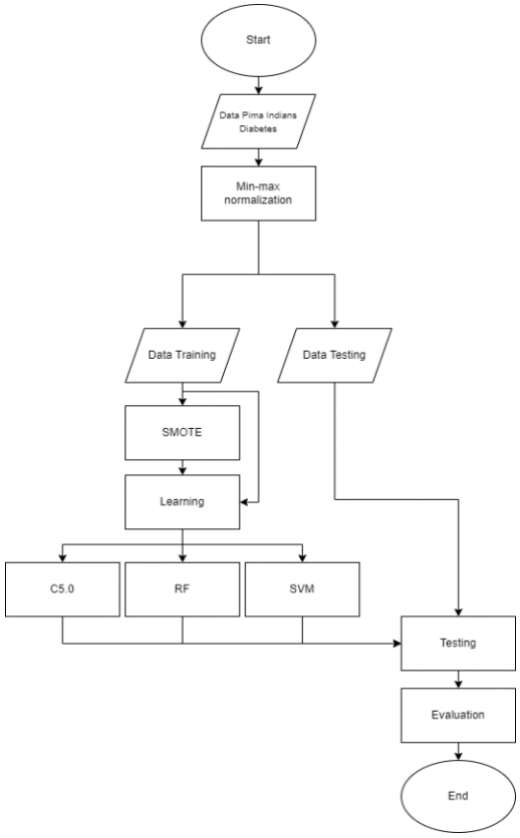

FIGURE 1. Flowchart of the proposed method

A. DATA COLLECTION

This research uses the Pima Indian Diabetes dataset sourced from the Kaggle Datasets platform. The dataset addresses issues related to diabetes and consists of 768 entries with eight predictor variables and a target variable represented by a single label attribute. It provides comprehensive health information for each individual, including data on pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function scores, and age. Additionally, it includes an outcome attribute that indicates the classification associated with each patient's medical condition.

In the dataset, patients are categorized into two groups: those with diabetes (True) and those without diabetes (False). The

data includes 268 instances of patients with diabetes and 500 instances of patients without the condition. [3]. The following information outlines the features and descriptions of the Pima Indian Diabetes dataset, as presented in Table 1.

TABLE 1
Surgery data attribute description

No	Attribute	Description	Category
1	Pregnancies	Number of pregnancies	Numeric
2	Glucose	Plasma glucose concentration after 2 hours during an oral glucose tolerance test	Numeric
3	BloodPressure	Diastolic blood pressure in millimeters of mercury (mm Hg)	Numeric
4	SkinThickness	Triceps skinfold thickness in millimeters (mm)	Numeric
5	Insulin	Serum insulin level after 2 hours in micro-units per milliliter (mu U/ml)	Numeric
6	BMI	Body mass index (BMI), calculated as weight in kilograms divided by height in meters squared (kg/m²)	Numeric
7	DiabetesPedigreeFunction	Diabetes pedigree function score	Numeric
8	Age	Age in years	Numeric
9	Outcome	Class label or outcome variable	Binary

B. C5.0

The C5.0 algorithm, a type of Decision Tree method, functions by examining data and creating a set of rules that guide decision-making processes [11]. This algorithm functions by breaking down data into a series of decisions based on the existing features, thereby enhancing the understanding and interpretation of patterns within the data. The resulting rules enable more informed and effective decision-making in various contexts [12], [13].

This algorithm determines which attributes to process based on the concept of "information gain." When selecting attributes to categorize objects into distinct classes, the objective is to identify the attribute that offers the most informational value. The attribute with the highest "information gain" is selected as the basis for the next node in building the decision tree structure [14], [15]. Eq. (1) is used to calculate the entropy value:

$$Entrophy\ (S) = \sum_{i=0}^{n} - pi * log_2 p_i \quad (1)$$

In this context, "S" represents the dataset or information being analyzed, and "n" denotes the number of subdivisions or subsets created from the dataset "S." The term "pi" indicates the number of instances within the initial subset or portion of the data. In simpler terms, "S" is the observed data, "n" refers to the number of segments derived from this data, and "pi" signifies the count of instances in the first segment. Eq. (2) is used to calculate the gain value:

$$Gain\ (S, A) = Entrophy\ (S) - \sum_{i=1}^{n} - \frac{|Si|}{|S|} * Entrophy(Si) \quad (2)$$

The variable "S" denotes the dataset or collection of cases being examined. "n" indicates the number of distinct segments derived from the attributes within set A. The variable |Si| refers to the number of instances within the i-th segment. The symbol |S| stands for the cardinality of the set S, which is the total number of cases in the dataset. In other words, "S" represents the entire dataset, "n" is the number of segments created from the set of attributes, |Si| is the number of instances in each segment, and |S| is the total number of instances in the dataset S. [14], [15].

C. RANDOM FOREST

Random Forest is an ensemble method composed of multiple decision trees, each constructed from randomly selected samples and different node-splitting criteria. This model utilizes a subset of features for each tree and aims to determine the optimal threshold for dividing the data [16], [17]. As a result, the model generates a collection of trees trained using various techniques, each providing distinct predictions. [18], [19]. Within data classification using the Random Forest algorithm, the Gini Index serves as a criterion for assessing the diversity or impurity of the nodes created at each branch of the decision tree. The Gini Index guides the algorithm in dividing the data into more homogeneous groups to achieve more accurate classification results. The calculation of the Gini Index is done using Eq. (3):

$$Gini = 1 - \sum_{i=1}^{c} (pi)^2 \quad (3)$$

Eq. (4) is used to calculate the entropy value:

$$Enteropy = \sum_{i=1}^{c} - pi * log_2(pi) \quad (4)$$

where variable "pi" measures the relative frequency of a particular class within the dataset, while "c" represents the total number of unique classes. Both of these elements are crucial in statistical analysis for comprehending and interpreting data distributions. [20].

D. SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a classification algorithm that separates data into two distinct classes by creating a hyperplane as a boundary between them. The algorithm also takes into account the margin, which is the distance between the hyperplane and the support vectors, which are the closest data points from each class [21], [22]. SVM aims to maximize the optimal separation between the two classes by utilizing the support vectors and the margin, enhancing its classification capabilities. Support Vector Machines (SVMs) have gained significant recognition in the field of machine learning, particularly for classification and regression tasks. Non-linear

SVM addresses the limitations of linear SVM by employing kernel functions to achieve higher dimensionality [23], [24]. TABLE 2 presents the equations for both linear and non-linear Support Vector Machines (SVM).

TABLE 2
Linear and non-linear SVM equations

SVM properties	Kernel Type	Formula Definition
Linear SVM	Linear	$K(x,y) = x.y$
Non-Linear SVM	Polynomial	$K(x,y) = (x.y+1)^{p}$
	Gaussian RBF	$K(x,y) = e^{-\|\|X-Y\|\|^2/_{2\sigma^2}}$

E. SMOTE (SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE)

Addressing class imbalance in minority classes involves tackling the problem of unequal sample sizes between majority and minority classes in the dataset. Resampling, specifically oversampling, is a common technique used to handle this issue. One widely used oversampling algorithm is the Synthetic Minority Over-sampling Technique (SMOTE) [25], [26]. SMOTE allows us to generate new synthetic samples for minority classes by combining data from existing samples. This technique increases the number of samples in the underrepresented class, thereby reducing the risk of overfitting the dominant class and improving the model's ability to accurately recognize the underrepresented class [27], [28]. SMOTE is a powerful method for handling class imbalance, allowing the model to better understand the minority class and produce more balanced predictions. Incorporating SMOTE can improve the model's performance and reduce biases caused by data imbalances [29], [30]. Creating new data for the underrepresented class by using a specific Eq. (5).

$$Y' = Y^i + (Y^j - Y^i) * Y \quad (5)$$

where Y'enhances the representation of the underrepresented minority group. Y^i denotes the demographics that lack representation. Y^i is a value chosen randomly from the k-nearest neighbors of the underrepresented class on Y^i. Y is a randomly selected value from a vector ranging between 0 and 1 [27].

F. CONFUSION MATRIX

The effectiveness of the developed system can be assessed by evaluating the performance of the classification model. [31] One technique used to evaluate the system's effectiveness and performance is the Confusion Matrix [32]. A Confusion Matrix is a method used to evaluate the performance of a classification algorithm by calculating its accuracy. This accuracy metric reflects the proportion of data that the algorithm has correctly classified. [33]

TABLE 3
Confusion matrix

Class	Predictions	
	favorable	adverse
favorable	TP (True favorable)	FN (False adverse)
adverse	FP (False favorable)	TN (True adverse)

Referring to Table 3 above, the following terms are defined:

1. True Positive: Data that is positive and has been correctly classified as positive.
2. False Positive: Data that is negative but has been incorrectly classified as positive.
3. False Negative: Data that is positive but has been incorrectly classified as negative.
4. True Negative: Data that is negative and has been correctly classified as negative.

G. AREA UNDER THE ROC (RECEIVER OPERATING CHARACTERISTIC CURVE)

The Area Under the ROC (Receiver Operating Characteristic) Curve is a numerical metric used to assess the performance of a model. It reflects how effectively the model distinguishes between positive and negative observations and indicates the model's success in providing accurate classifications [34], [35]. ROC curves and AUC values are crucial for classification and model evaluation. The AUC metric ranges from 0 to 1, with higher values indicating better performance. An AUC value close to 1 indicates that the model is highly effective at distinguishing between positive and negative classes [36] The following are the classification groups based on the AUC values, as outlined in Table 4:

TABLE 4
categorization CATEGORY According to AUC VALUE

AUC value	Classifier categories
0.90 1.00	Excellent
0.80 0.90	Good
0.70 0.80	Fair
0.60 0.70	Poor
0.50 0.60	Fail

AUC is a useful metric for comparing multiple classification models to identify the best fit. Because it is independent of classification thresholds, AUC offers a more generalized assessment of model quality. [13], [37]. To get the AUC value, Eq. (6) is utilized:

$$AUC = \frac{1+TPR-FPR}{2} \quad (6)$$

1) DATA GATHERING

The data used in this study was obtained from a third-party source, specifically the Pima Indian Diabetes dataset available on Kaggle Datasets. This dataset consists of two categories: individuals with diabetes (True) and those without (False), comprising a total of 768 entries and one label.

2) PREPROCESSING

At this stage, preprocessing is performed with the goal of preparing the data to meet the requirements of the classification algorithm, thereby enhancing its performance. One of the preprocessing methods used is min-max normalization, which aims to improve the system's understanding and processing of the data. [38], [39]. Min-max normalization is done by resizing the data so that it scopes among 0 and 1. [40].

3) DATA SHARING

The data partitioning process involves splitting the dataset into two separate subsets: one for training the model and the other for evaluating its classification performance. Split validation is used to evaluate how well the data is divided into training and testing sets, with an 80:20 ratio used for allocation. The preprocessing procedure consists of multiple stages, which are detailed below.

The first phase of preprocessing involves identifying and addressing problematic data, such as empty fields and inaccuracies. Upon reviewing the collected data, it was confirmed that there were no missing values or duplicates, eliminating the need for further corrective measures. The next phase involves converting the data to match the data types required by the C5.0, Random Forest, SVM, and SMOTE algorithms. This includes applying min-max normalization to scale the data between 0 and 1. Min-max normalization helps prepare the data for analysis and modeling, enhancing the interpretability, stability, and efficiency of the algorithms.

Subsequently, the dataset is divided into two categories: individuals with diabetes (True) and those without diabetes (False). The true class contains 268 data samples, while the false class has 500 data samples. Data partitioning is essential for the subsequent training and evaluation of the classification model's performance.

After completing the preprocessing phase, the Pima Indian Diabetes dataset is collected, verified for accuracy, transformed into suitable data formats, and categorized as needed. This dataset can then be analyzed using the C5.0, Random Forest, SVM, and SMOTE algorithms to train a classification model and assess its ability to differentiate between individuals with and without diabetes.

4) RESAMPLING DATA

Resampling is a technique used to create additional samples from existing samples or populations within a dataset. This study applies resampling to the training dataset, specifically using the Synthetic Minority Over-sampling Technique (SMOTE) to enhance the data. Resampling involves generating new iterations of existing data for statistical analysis, machine learning, or model validation purposes. The primary goal is to address the class imbalance in the dataset. More specifically, resampling is performed on the training data using the SMOTE technique. SMOTE generates synthetic instances for the underrepresented class, increasing the number of samples in that class. By employing SMOTE, class imbalance can be reduced, resulting in a more balanced classification model that can accurately predict the minority class.

5) MODEL MAKING

At this stage, after data partitioning and resampling, the data modeling process begins using the C5.0, Random Forest, and SVM classification algorithms. RapidMiner is the software used for developing, training, and evaluating models in machine learning and data analysis, specifically focusing on C5.0, Random Forest, SVM, and SMOTE techniques. For this study, an 80:20 split ratio is applied to the data, with all other parameters kept at their default settings..

6) EVALUATION OF OUTCOMES

Evaluation is a crucial process aimed at determining the effectiveness of the modeling efforts undertaken. During this stage, the precision of the models will be assessed using a confusion matrix and the Area Under the Curve (AUC) as metrics to compare model performance. This study aims to enhance the efficiency and accuracy of the C5.0, Random Forest, and SVM classification models by addressing class imbalance with the SMOTE resampling technique. SMOTE was employed to correct class imbalance in the dataset while maintaining an 80:20 ratio of training to testing data.

After applying the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance, the C5.0, Random Forest, and Support Vector Machine (SVM) algorithms were evaluated using a confusion matrix. The effectiveness of the classification models is measured with the AUC metric, assisted by SMOTE. The final results compare the accuracy levels of classification models with and without SMOTE to determine if its implementation improves the accuracy of diabetes classification.

III.RESULT

A.C5.0 METHOD RESEARCH RESULTS

The study procured experimental outcomes by employing the C5.0 method. To uphold model integrity, an 80:20 data split and min-max normalization were incorporated during the evaluation phase. Following validation, the model's execution was assessed using a confusion matrix and AUC-ROC analysis (FIGURE 3 (a)). The outcomes are detailed in Table 5.

TABLE 5
precision AND AUC C5.0

Model	precision	AUC
C5.0	0.714	0.745

TABLE 6
CONFUSION MATRIX C5.0

categorization	Predicted Class	
	Postive	adverse
Actual: favorable	17	7
Actual: adverse	37	93

The C5.0 model underwent evaluation using an 80:20 data split and min-max normalization. The precision of a given measurement can be determined by analyzing the outcomes presented by the confusion matrix Table 6. The assessment resulted in a precision of 0.714 and an Area The model's receiver operating characteristic (ROC) curve, as depicted in Figure 2, resulted in an area under the curve (AUC) score of 0.745. This suggests a moderate ability to differentiate between positive and negative classes. An AUC value in the range of 0.70 to 0.80 is generally considered "Fair" by commonly accepted standards. Although the AUC of 0.745 implies that the model performs reasonably well, it also highlights potential areas for improvement. A higher AUC indicates better performance in terms of classification and discrimination ability.

B. RANDOM FOREST RESEARCH RESULTS

The present study produced experimental findings by employing the Random Forest approach. To ensure model consistency, a combination of 80:20 data and min-max normalization was utilized during the evaluation phase. Following validation, execution assessment of the model was conducted by means of confusion matrix and AUC-ROC measures (FIGURE 3 (b)). The outcomes are detailed in TABLE 7:

TABLE 7
precision AND AUC RANDOM FOREST

Model	precision	AUC
Random Forest	0.733	0.824

TABLE 8
CONFUSION MATRIX C5.0

categorization	Predicted Class	
	Postive	adverse
Actual: favorable	22	9
Actual: adverse	32	91

The Random Forest model was assessed using an 80:20 data split and min-max normalization. The precision of a given measurement can be determined by analyzing the outcomes presented by the confusion matrix Table 8. There is no text provided. The evaluation resulted in an accuracy of 0.733 and an Area Under the Curve (AUC) score of 0.824. The model's receiver operating characteristic (ROC) curve, shown in Figure 3, yielded an area under the curve (AUC) value of 0.824. This indicates a decent capacity to distinguish between the positive and negative classes. The AUC result is within the range of "Good categorization," which is commonly considered as 0.80 to 0.90. The AUC value of 0.824 suggests that the model has decent performance, but it also indicates that there is potential for improvement. A higher area under the curve (AUC) indicates superior model performance in terms of its ability to classify and discriminate.

C. SVM METHOD RESEARCH RESULTS

Experimental outcomes in this study were derived using the SVM method. To maintain model integrity, an 80:20 data split and min-max normalization were employed during the evaluation phase. Once validated, the effectiveness of the model was assessed using a confusion matrix and AUC-ROC analysis (FIGURE 3 (c)), with the results shown in TABLE 9.

TABLE 9
precision AND AUC SVM

Model	Precision	AUC
SVM	0.746	0.799

TABLE 10
CONFUSION MATRIX SVM

categorization	Predicted Class	
	Postive	adverse
Actual: favorable	29	14
Actual: adverse	25	86

The efficacy of the support vector machine (SVM) model was evaluated by dividing the data into an 80:20 ratio and applying min-max normalization. The precision of a given measurement can be determined by analyzing the outcomes presented by the confusion matrix Table 10. The model achieved a precision of 0.746 and an Area Under the Curve (AUC) score of 0.799. The receiver operating characteristic (ROC) curve for the model, as shown in Figure 4, yielded an area under the curve (AUC) of 0.799, indicating a decent ability to distinguish between positive and negative classifications. This AUC value falls within the "Fair categorization" range, which is typically defined as 0.70 to 0.80. Although the AUC of 0.799 is indicative of a model with reasonable performance, it highlights that there is room for improvement. A greater Area Under the Curve (AUC) shows that the model has better performance in terms of its ability to classify and discriminate across different categories.

D. C5.0 WITH SMOTE RESEARCH RESULTS

In this study, the C5.0 + SMOTE approach was utilized to obtain experimental outcomes. To maintain model integrity, a combination of an 80:20 data split and min-max normalization was employed during the evaluation process. Following the validation phase, model execution was assessed using a confusion matrix and AUC-ROC analysis (FIGURE 3 (d)). The outcomes are detailed in Table 11.

TABLE 11
precision AND AUC C5.0 + SMOTE

Model	precision	AUC
C5.0 + SMOTE	0.603	0.734

TABLE 12
CONFUSION MATRIX C5.0 + SMOTE

categorization	Predicted Class	
	Postive	adverse
Actual: favorable	50	57
Actual: adverse	4	43

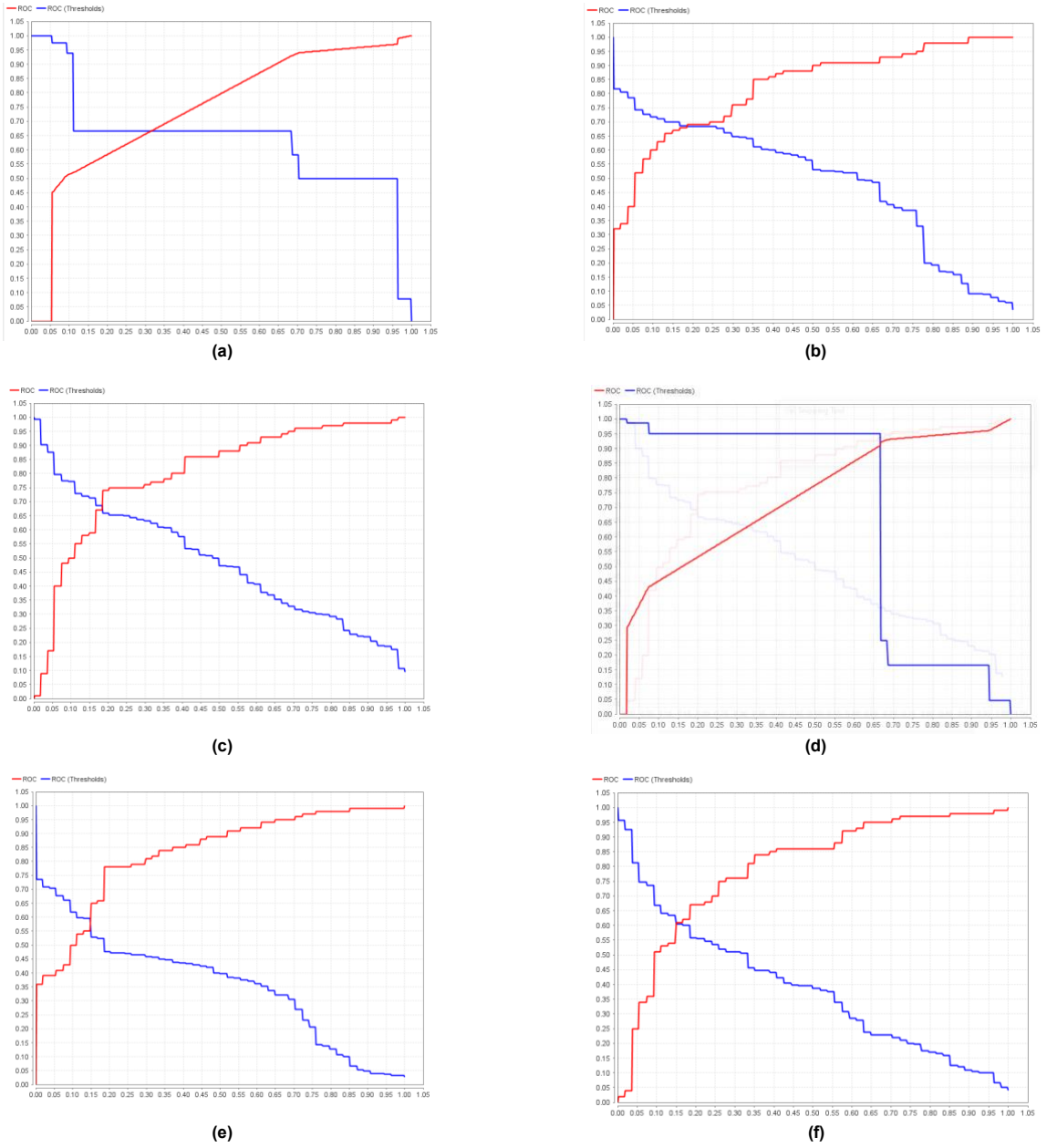

FIGURE 3. (a) ROC curve of C5.0, (b) ROC curve of Random Forest, (c) ROC curve of SVM, (d) ROC curve of C5.0 + SMOTE, (e) ROC curve of Random Forest + SMOTE, (f) ROC curve of SVM + SMOTE

The C5.0+ SMOTE model was assessed using an 80:20 data split and min-max normalization. The precision of a given measurement can be determined by analyzing the outcomes presented by the confusion matrix Table 12. The evaluation resulted in a precision of 0.603 and an Area Under the Curve (AUC) score of 0.734. The model's receiver operating characteristic (ROC) curve, shown in Figure 2, yielded an area under the curve (AUC) value of 0.734. This indicates a decent ability to distinguish between the positive and negative classes. The AUC value is categorized as "Fair" according to the commonly accepted range of 0.70 to 0.80. The AUC value of 0.734 suggests that the model performs reasonably well, but it also indicates that there is potential for improvement. A higher AUC significantly stronger model performance in terms of its ability to classify and discriminate.

E. RANDOM FOREST WITH SMOTE RESEARCH RESULTS

In this study, experimental outcomes were derived by employing the Random Forest + SMOTE method. In order to maintain the integrity of the model, a data split of 80:20 and min-max normalization were utilized during the evaluation phase. Following the validation process, the performance of the model was assessed by utilizing a confusion matrix and conducting AUC-ROC analysis (FIGURE 3 (f)). The findings are consistently displayed in Table 13.

TABLE 13
precision AND AUC C5.0 + SMOTE

Model	precision	AUC
Random Forest + SMOTE	0.727	0.831

TABLE 14
CONFUSION MATRIX RANDOM FOREST + SMOTE

categorization	Predicted Class	
	Postive	adverse
Actual: favorable	44	32
Actual: adverse	10	68

The Random Forest + SMOTE model was assessed using an 80:20 data split and min-max normalization. The precision of a given measurement can be determined by analyzing the outcomes presented by the confusion matrix Table 14. The evaluation yielded an accuracy of 0.727 and an Area Under the Curve (AUC) score of 0.831. The receiver operating characteristic (ROC) curve for the model, depicted in Figure 3, also produced an area under the curve (AUC) value of 0.831, suggesting a satisfactory capability to differentiate between the positive and negative classes. This AUC value falls within the "Good categorization" range, which is typically defined as 0.80 to 0.90. Although the AUC of 0.831 is indicative of a model with reasonable performance, it highlights that there is room for improvement. A higher AUC would suggest better model performance in terms of classification and discrimination capabilities.

F. SVM WITH SMOTE RESEARCH RESULTS

In this study, experimental outcomes were reached by utilizing the SVM + SMOTE method. To maintain model integrity, an 80:20 data split and min-max normalization were employed during the evaluation process. After validation, the model's execution was evaluated using a confusion matrix and AUC-ROC analysis (FIGURE 3 (a)). The results are specified in Table 15.

TABLE 15
precision AND AUC SVM + SMOTE

Model	precision	AUC
SVM + SMOTE	0.727	0.794

TABLE 16
CONFUSION MATRIX SVM + SMOTE

categorization	Predicted Class	
	Postive	adverse
Actual: favorable	36	24
Actual: adverse	18	76

The SVM + SMOTE model's performance was assessed by employing an 80:20 data division and min-max normalization. The precision of a given measurement can be determined by analyzing the outcomes presented by the confusion matrix Table 16. The evaluation yielded an accuracy of 0.727 and an Area Under the Curve (AUC) score of 0.794. The ROC curve of the model, depicted in Figure 3, resulted in an AUC value of 0.794. This suggests a commendable proficiency in discerning between the good and negative categories. The AUC value is categorized as "Fair" according to the range of 0.80 to 0.90. The AUC value of 0.794 suggests that the model has a reasonably good performance, but it also indicates that there is potential for further improvement. A higher AUC signifies a more exceptional model performance in terms of its classification and discrimination capabilities.

IV.DISCUSSION

This study uses the Indian Pima Diabetes dataset, which contains 768 records with 8 attributes and 1 output label. The validation process includes using min-max normalization to scale the dataset values between 0 and 1 to improve model performance. The data is split into an 80:20 ratio for training and testing. The classification methods applied are C5.0, Random Forest, and SVM. Additionally, the SMOTE technique is used to address class imbalance in the dataset.

Using the C5.0 model, the study analyzed diabetes classification in the Indian Pima Diabetes dataset, achieving an accuracy of 0.714 and an AUC of 0.745 according to the confusion matrix results. The AUC benchmark categorizes this result as "Fair classification," indicating that the C5.0 model has a balanced ability to differentiate between diabetic and non-diabetic cases in the dataset.

The implementation of the Random Forest algorithm results in increased effectiveness when classifying diabetes data, with an accuracy of 0.733 and an AUC score of 0.824. According to AUC benchmark standards, this result falls within the "Good classification" range, suggesting that the Random Forest model has a strong ability to distinguish between diabetes and non-diabetes cases.

The application of the SVM model has improved evaluation outcomes for diabetes classification, showing an increase in accuracy to 0.746. However, there was a slight decrease in AUC to 0.799. According to the AUC reference, this result is considered "Fair classification," indicating that the SVM model does not significantly improve in distinguishing between individuals with and without diabetes.

Utilizing the SMOTE and C5.0 amalgamation, the evaluation outcomes obtained by scrutinizing the confusion matrix and AUC for diabetes data categorization exhibit a diminution in efficacy when contrasted by the pre-SMOTE equilibrium phase, presenting an precision of 0.603 and an AUC value of 0.734. According to the AUC benchmark, this outcome drops into the "Fair categorization" category. It suggests that the SMOTE + C5.0 model does not produce substantial outcomes in distinguishing among diabetic and non-diabetic cases.

Employing the combination of SMOTE and Random Forest, the evaluation outcomes derived by the confusion matrix and AUC for diabetes data categorization demonstrate a decrease in execution compared to the pre-SMOTE balancing phase, by an precision value of 0.727. Nevertheless, there has been a rise in the area under the curve (AUC) by a precise amount of 0.831. Based on the AUC benchmark, this output falls inside the "Good categorization" category. The SMOTE + Random Forest model demonstrates a significant capability

to distinguish between persons with diabetes and those without.

By applying the SMOTE-SVM approach, the evaluation of diabetes data categorization through confusion matrix and AUC analysis indicates a decline in execution when compared to the pre-SMOTE balancing phase. The precision value stands at 0.727 while the AUC value is 0.794. As per the AUC benchmark, this outcome drops under "Fair categorization" category. It suggests that the SMOTE + SVM model does not exhibit significant outcomes in distinguishing among diabetic and non-diabetic instances. The evaluation outcomes are outlined in Table 17.

TABLE 17
AUC categorization OF DIABETES DATA

Model	precision	AUC
C5.0	0.714	0.745
Random Forest	0.733	0.824
SVM	0.746	0.799
C5.0 + SMOTE	0.603	0.734
Random Forest + SMOTE	0.727	0.831
SVM + SMOTE	0.727	0.794

In this study, a decline in categorization outcomes was observed subsequent to the utilization of SMOTE for addressing data imbalance. Evaluation findings indicate that the Random Forest model, which has been balanced via SMOTE, exhibits superior execution by respect to AUC values when compared to alternative models. Conversely, the SVM model demonstrates better precision value execution relative to other models. Comparing the outcomes of testing the three models by and byout the utilization of SMOTE, as illustrated in FIGURE 4 below, suggests that there's no notable alteportionn following the implementation of SMOTE to address data imbalance in the analyzed models using the Pima Indian Diabetes dataset.

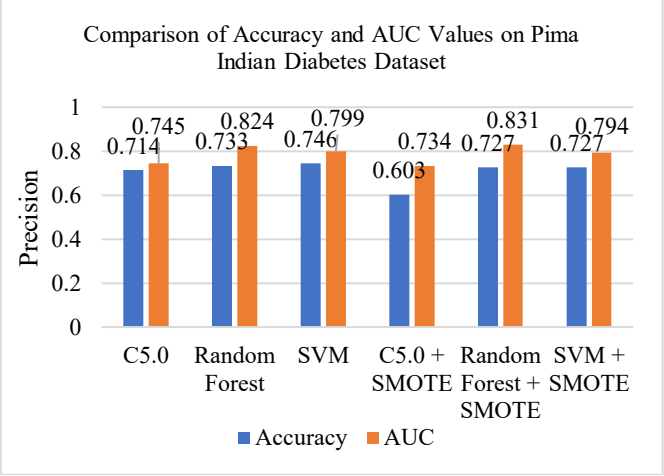

FIGURE 4. **Precision and AUC Value Comparison Chart**

Based on the analysis of the study findings, it is clear that incorporating SMOTE into the C5.0, Random Forest, and SVM algorithms for diabetes classification results in similar accuracy levels before and after SMOTE integration. This suggests that applying the Synthetic Minority Over-sampling Technique (SMOTE) does not improve the classification accuracy when using the Pima Indian Diabetes dataset. Some classification algorithms might tend to overfit after applying SMOTE.

Additionally, the study considered the AUC (Area Under Curve) value. In the Random Forest trial without SMOTE, the AUC was 0.824. However, with SMOTE included, the AUC increased to 0.831, indicating effective classification. Overall, the results suggest that incorporating SMOTE into the C5.0, Random Forest, and SVM techniques for classifying diabetes using the Pima Indian dataset did not result in a significant improvement in classification accuracy. This validation confirms that SMOTE does not significantly enhance the model's ability to detect diabetes. It can be concluded that SMOTE is not a crucial resampling technique for addressing class imbalance to improve the performance and effectiveness of the C5.0, Random Forest, and SVM classification models.

Upon further examination, juxtaposing the findings of this study by prior study reveals that the integration of C5.0, Random Forest, and SVM techniques by SMOTE does not yield superior results in categorizing the Pima Indian Diabetes dataset. This discrepancy arises from the fact that the amalgamated approach generates accuracy and AUC metrics that exhibit minimal variation. Unlike previous research that used other categorization methods, this study utilized the Synthetic Minority Over-sampling Technique (SMOTE). The results of the studies conducted by Sutoyo, Ayu Wahyuning, Kusumarini, and Hasanah [5], [6], [8], [10], while using different categorization techniques and using SMOTE, could not attain accuracy and AUC values that were equivalent to the ones found in the current study. The findings indicate that the utilization of C5.0, Random Forest, and SVM with SMOTE does not significantly improve the accuracy of categorizing the Pima Indian Diabetes dataset. This analysis also facilitates understanding of the efficacy of the tactics and algorithms employed in this study in producing superior outcomes in comparison to prior methodologies.

However, this study is limited by the use of a small dataset, both in terms of the number of patients and the available features. Additionally, there is a significant class imbalance, with the majority of patients belonging to the non-diabetic group. These factors can impact the results. Therefore, future research should use a larger and more diverse dataset that includes a greater number of patients and relevant diabetes-related attributes. This approach will lead to more accurate and comprehensive results. Despite these limitations, the study successfully demonstrated that applying SMOTE to the C5.0, Random Forest, and SVM techniques did not significantly improve the accuracy of diabetes classification.

V.CONCLUSIONS

Based on the results from the study using the Indian Pima Diabetes dataset, the C5.0 classification model achieved an accuracy of 0.714 and an AUC of 0.745, placing it in the "Fair categorization" range. The Random Forest model achieved an accuracy of 0.733 and an AUC of 0.824, classified as "Good

categorization." The SVM model attained an accuracy of 0.746 and an AUC of 0.799, also considered "Fair categorization."

When SMOTE was applied, the C5.0 model had an accuracy of 0.603 and an AUC of 0.734, still within the "Fair categorization" range. The Random Forest model with SMOTE achieved an accuracy of 0.727 and an AUC of 0.831, which is classified as "Good categorization." The SVM model with SMOTE resulted in an accuracy of 0.727 and an AUC of 0.794, which falls under "Fair categorization."

The study concludes that using SMOTE did not significantly improve the performance of the three classification models on the Indian Pima Diabetes data. This is likely due to overfitting, where the model becomes too complex and overly tailored to the training data, making it too reliant on the synthetic data generated by SMOTE for the minority class. As a result, there is a decrease in model performance, leading to lower accuracy and AUC scores.

The results of this study indicate that using the C5.0, Random Forest, and SVM classification algorithms along with the SMOTE technique to address class imbalance does not significantly improve accuracy when categorizing Indian Pima Diabetes data. However, to enhance the effectiveness of these algorithms and techniques, future research should focus on several key areas. One important aspect to address in future studies is the use of larger and more diverse datasets. Additionally, exploring alternative methods to tackle class imbalance could be a promising area for further research, as these efforts may reveal techniques that significantly improve the accuracy of diabetes classification. Finally, future studies should prioritize a comprehensive and varied evaluation process to ensure that any improvements in classification accuracy are genuinely due to the appropriate methods. By focusing on these factors, future research is expected to produce more accurate and thorough results in diabetes data classification.

REFERENCES

[1] E. Subandi and K. Adam, "Modern Dressing Terhadap Penyembuhan Luka Diabetes Melitus Tipe 2 Proses," *J. Kesehat.*, vol. 10, no. 1, pp. 1273–1283, 2019.

[2] M. E. Fitriyanti, H. Febriawati, and L. Yanti, "Pengalaman Penderita Diabetes Mellitus Dalam Pencegahan Ulkus Diabetik," *J. Keperawatan Muhammadiyah Bengkulu*, vol. 07, pp. 597–603, 2019.

[3] M. Abedini, A. Bijari, and T. Banirostam, "categorization of Pima Indian Diabetes Dataset using Ensemble of Decision Tree, Logistic Regression and Neural Network," *Ijarcce*, vol. 9, no. 7, pp. 1–4, 2020, doi: 10.17148/ijarcce.2020.9701.

[4] H. Pangestika, D. Ekawati, and N. S. Murni, "Faktor-Faktor Yang Berhubungan Dengan Kejadian Diabetes Mellitus Tipe 2," *J. 'Aisyiyah Med.*, vol. 7, no. 1, 2022, doi: 10.36729/jam.v7i1.779.

[5] E. Sutoyo and M. A. Fadlurrahman, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network," *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 3, p. 379, 2020, doi: 10.26418/jp.v6i3.42896.

[6] D. Ayu Wahyuning Dewi, I. Cholissodin, and Sutrisno, "Klasifikasi Penyimpangan Tumbuh Kembang Anak Menggunakan Algoritme C5.0," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 10, pp. 10258–10265, 2019, [Online]. Available: http://j-ptiik.ub.ac.id

[7] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data," *J. Biomed. Inform.*, vol. 107, no. May 2019, p. 103465, 2020, doi: 10.1016/j.jbi.2020.103465.

[8] A. I. Kusumarini, P. A. Hogantara, M. Fadhlurohman, and S. K. . M. K. Nurul Chamidah, Perbandingan Algoritma Random Forest, Naive Bayes, Dan Decision Tree Dengan Oversampling Untuk Klasifikasi Bakteri E.Coli, vol. 2, no. 1. 2021.

[9] N. G. Ramadhan, "Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus," *Sci. J. Informatics*, vol. 8, no. 2, pp. 276–282, 2021, doi: 10.15294/sji.v8i2.32484.

[10] H. Hasanah and Nurmalitasari, "Perbandingan Tingkat Akurasi Algoritma Support Vector Machines (SVM) dan C45 dalam Prediksi Penyakit Jantung," *Pros. Semin. Nas. Teknol. dan Sains*, vol. 2, pp. 13–18, 2023.

[11] E. Purwanti, R. U. N. U. Nor, and S. Soelistyono, "Web Design for Stroke Early Detection Using Decision Tree C5.0," *Komputasi J. Ilm. Ilmu Komput. dan Mat.*, vol. 20, no. 2, pp. 135–147, 2023, doi: 10.33751/komputasi.v20i2.8265.

[12] R. N. Amalda, N. Millah, and I. Fitria, "Implementasi Algoritma C5.0 Dalam Menganalisa Kelayakan Penerima Keringanan Ukt Mahasiswa Itk," *Teorema Teor. dan Ris. Mat.*, vol. 7, no. 1, p. 101, 2022, doi: 10.25157/teorema.v7i1.6692.

[13] J. Zhang and L. Chen, "Clustering-based undersampling by random over sampling examples and support vector machine for imbalanced categorization of breast cancer diagnosis," *Comput. Assist. Surg.*, vol. 24, no. sup2, pp. 62–72, 2019, doi: 10.1080/24699322.2019.1649074.

[14] A. C. Wijaya, N. A. Hasibuan, and P. Ramadhani, "Implementasi Algoritma C5.0 Dalam Klasifikasi Pendapatan Masyarakat (Studi Kasus: Kelurahan Mesjid Kecamatan Medan Kota)," *Maj. Ilm. INTI*, vol. 5, 2018.

[15] D. P. Utomo, P. Sirait, and R. Yunis, "Reduksi Atribut Pada Dataset Penyakit Jantung dan Klasifikasi Menggunakan Algoritma C5.0," *J. Media Inform. Budidarma*, vol. 4, no. 4, pp. 994–1006, 2020, doi: 10.30865/mib.v4i4.2355.

[16] M. R. Ansyari, M. I. Mazdadi, F. Indriani, D. Kartini, and T. H. Saragih, "Implementation of Random Forest and Extreme Gradient Boosting in the categorization of Heart Disease using Particle Swarm Optimization Feature Selection," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 5, no. 4, pp. 250–260, 2023, doi: 10.35882/jeeemi.v5i4.322.

[17] H. Tyralis, G. Papacharalampous, and A. Langousis, "A brief review of random forests for water scientists and practitioners and their recent history in water resources," *Water (Switzerland)*, vol. 11, no. 5, 2019, doi: 10.3390/w11050910.

[18] D. H. Depari, Y. Widiastiwi, and M. M. Santoni, "Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung," *Inform. J. Ilmu Komput.*, vol. 18, no. 3, p. 239, 2022, doi: 10.52958/iftk.v18i3.4694.

[19] X. Tan *et al.*, "Wireless sensor networks intrusion detection based on SMOTE and the random forest algorithm," *Sensors (Switzerland)*, vol. 19, no. 1, 2019, doi: 10.3390/s19010203.

[20] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, "Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 3, pp. 677–690, 2022, doi: 10.30812/matrik.v21i3.1726.

[21] Muhamad Fawwaz Akbar, Muhammad Itqan Mazdadi, Muliadi, Triando Hamonangan Saragih, and Friska Abadi, "Implementation of Information Gain Ratio and Particle Swarm Optimization in the Sentiment Analysis categorization of Covid-19 Vaccine Using Support Vector Machine," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 5, no. 4, pp. 261–270, 2023, doi: 10.35882/jeeemi.v5i4.328.

[22] Y. Ferdinand and W. F. Al Maki, "Broccoli leaf diseases categorization using support vector machine by particle swarm optimization based on feature selection," *Int. J. Adv. Intell. Informatics*, vol. 8, no. 3, pp. 337–348, 2022, doi: 10.26555/ijain.v8i3.951.

[23] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection," *IEEE Access*,

vol. 6, pp. 33789–33795, 2018, doi: 10.1109/ACCESS.2018.2841987.
[24] A. Bhavani and B. Santhosh Kumar, "A Review of State Art of Text categorization Algorithms," *Proc. 5th Int. Conf. Comput. Methodol. Commun. ICCMC 2021*, no. April 2021, pp. 1484–1490, 2021, doi: 10.1109/ICCMC51019.2021.9418262.
[25] C. Tantithamthavorn, A. E. Hassan, and K. Matsumoto, "The Impact of Class Rebalancing Techniques on the Performance and Interpretation of Defect Prediction Models," *IEEE Trans. Softw. Eng.*, vol. 46, no. 11, pp. 1200–1219, 2020, doi: 10.1109/TSE.2018.2876537.
[26] A. S. Hussein, T. Li, C. W. Yohannese, and K. Bashir, "A-SMOTE: A new preprocessing approach for highly imbalanced datasets by improving SMOTE," *Int. J. Comput. Intell. Syst.*, vol. 12, no. 2, pp. 1412–1422, 2019, doi: 10.2991/ijcis.d.191114.002.
[27] H. Hairani, A. Anggrawan, and D. Priyanto, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data categorization Using Smote-Tomek Link," *Int. J. Informatics Vis.*, vol. 7, no. 1, pp. 258–264, 2023, doi: 10.30630/joiv.7.1.1069.
[28] H. Al Majzoub and I. Elgedawy, "AB-SMOTE: An Affinitive Borderline SMOTE Approach for Imbalanced Data Binary categorization," *Int. J. Mach. Learn. Comput.*, vol. 10, no. 1, pp. 31–37, 2020, doi: 10.18178/ijmlc.2020.10.1.894.
[29] M. Sulistiyono, Y. Pristyanto, S. Adi, and G. Gumelar, "Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi," *Sistemasi*, vol. 10, no. 2, p. 445, 2021, doi: 10.32520/stmsi.v10i2.1303.
[30] M. F. Ijaz, G. Alfian, M. Syafrudin, and J. Rhee, "Hybrid Prediction Model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, Synthetic Minority Over Sampling Technique (SMOTE), and random forest," *Appl. Sci.*, vol. 8, no. 8, 2018, doi: 10.3390/app8081325.
[31] J. H. J. C. Ortega, "Analysis of Performance of categorization Algorithms in Mushroom Poisonous Detection using Confusion Matrix Analysis," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1.3, pp. 451–456, 2020, doi: 10.30534/ijatcse/2020/7191.32020.
[32] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," *Comput. Oper. Res.*, vol. 152, no. April 2022, p. 106131, 2023, doi: 10.1016/j.cor.2022.106131.
[33] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in categorization performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.
[34] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," *Eur. J. Oper. Res.*, vol. 267, no. 2, pp. 687–699, 2018, doi: 10.1016/j.ejor.2017.12.001.
[35] Shalehah, Muhammad Itqan Mazdadi, Andi Farmadi, Dwi Kartini, and Muliadi, "Implementation of Particle Swarm Optimization Feature Selection on Naïve Bayes for Thoracic Surgery categorization," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 5, no. 3, pp. 150–158, 2023, doi: 10.35882/jeemi.v5i3.305.
[36] V. Sari, F. Firdausi, and Y. Azhar, "Perbandingan Prediksi Kualitas Kopi Arabika dengan Menggunakan Algoritma SGD, Random Forest dan Naive Bayes," *Edumatic J. Pendidik. Inform.*, vol. 4, no. 2, pp. 1–9, 2020, doi: 10.29408/edumatic.v4i2.2202.
[37] D. Pramadhana, "Klasifikasi Penyakit Diabetes Menggunakan Metode CFS dan ROS dengan Algoritma J48 Berbasis Adaboost," *Edumatic J. Pendidik. Inform.*, vol. 5, no. 1, pp. 89–98, 2021, doi: 10.29408/edumatic.v5i1.3336.
[38] S. Sinsomboonthong, "Performance Comparison of New Adjusted Min-Max by Decimal Scaling and Statistical Column Normalization Methods for Artificial Neural Network categorization," *Int. J. Math. Math. Sci.*, vol. 2022, 2022, doi: 10.1155/2022/3584406.
[39] S. A. D. Prasetyowati, M. Ismail, E. N. Budisusila, D. R. I. M. Setiadi, and M. H. Purnomo, "Dataset Feasibility Analysis Method based on Enhanced Adaptive LMS method by Min-max Normalization and Fuzzy Intuitive Sets," *Int. J. Electr. Eng. Informatics*, vol. 14, no. 1, pp. 55–75, 2022, doi: 10.15676/ijeei.2022.14.1.4.
[40] A. Ambarwari, Q. J. Adrian, and Y. Herdiyeni, "Analisis Pengaruh Data Scaling Terhadap Performa Algoritme Machine Learning untuk Identifikasi Tanaman," *J. RESTI(Rekayasa Sist. dan Teknol. Inf.)*, vol. 1, no. 3, pp. 117–122, 2017.

AUTHORS BIOGRAPHY

M. Khairul Rezki was born in Barabai, South Kalimantan. Since 2018, he has pursued his academic endeavors as a student Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Lambung Mangkurat. His current area of study lies by the realm of Data Science. Additionally, his final project entailed conducting study that centered around the categorization of Diabetes. His dedication to exploring the field of Data Science has driven him to achieve a deep understanding of complex algorithms and machine learning techniques. With a keen focus on innovative solutions, he aims to contribute significantly to the advancement of healthcare through data-driven approaches. His commitment to academic excellence and research in the area of Diabetes categorization showcases his passion for utilizing technology to make a positive impact on society.

Muhammad Itqan Mazdadi is a lecturer in the Department of Computer Science, Lambung Mangkurat University. His study interest is centered on Data Science and Computer Networking. Before becoming a lecturer, he completed his undergraduate program in the Computer Science Department at Lambung Mangkurat University In 2013. He then completed his master's degree from Department of Informatics at Islamic Indonesia University, Yogyakarta. Currently, he serves as the Secretary of the Computer Science Department at Lambung Mangkurat University. He is dedicated to fostering a collaborative and innovative learning environment that encourages students to explore and excel in the field of computer science.

Fatma Indriani is a lecturer in the Department of Computer Science, Lambung Mangkurat University. Her study interest is focused on Data Science. Before becoming a lecturer, she completed her undergraduate program in the Informatics Department at Bandung Institute of Technology. In 2008, she started working as a lecturer in the Computer Science department at Lambung Mangkurat University. She then completed her master's degree at Monash University, Australia in 2012. And her latest education is a doctorate degree in Bioinformatics at Kanazawa University, Japan, which was completed in 2022. The study fields she focuses on are Data Science and Bioinformatics. She is dedicated to fostering a collaborative and innovative learning environment that encourages students to explore and excel in the field of computer science.

Muliadi is a lecturer in the Department of Computer Science at Lambung Mangkurat University, where he specializes in Artificial Intelligence, Decision Support Systems, and Data Science. His academic journey began by a bachelor's degree in Informatics Engineeringfrom STMIK Akakomin 2004, followed by the attainment of a master's degree in Computer Science from Gadjah Mada University in 2009. by

expertise in Data Science, he also brings valuable skills in Start-up Business Development, Digital Entrepreneurship, and Data Management Staff. His expertise in the field of Data Science has allowed him to contribute significantly to Start-up Business Development, Digital Entrepreneurship, and Data Management Staff. His background and experience make him a valuable asset to the academic community and the tech field. Through his dedication to research and teaching, he continues to inspire and educate future generations of technologists and innovators.

Triando Hamonangan Saragih is a lecturer in Department of Computer Science, Lambung Mangkurat University. His study interest is focused on Data Science. He completed his bachelor's degree in Informatics at Brawijaya University, Malang in 2016. After that, he pursued a master's degree in Computer Science Brawijaya University, Malang in 2018. The study field he is involved in is Data Science. He has contributed to various research projects related to data analysis and machine learning. He actively mentors students and junior researchers, sharing his knowledge and passion for data science. His dedication to advancing the field of data science through education and research makes him a valuable asset to the academic community.

Professor Dr. Vijay Anant Athavale is a distinguished academic and professional with extensive experience in computer science and engineering. He holds a Ph.D. in Computer Science from Barkatullah University, Bhopal, and has served in various prestigious roles, including Dean of Engineering and Professor at Panipat Institute of Engineering & Technology, Haryana. Dr. Athavale has been a key figure in numerous institutions, contributing significantly to their academic and administrative advancements. He is a life member of several professional bodies, such as the Computer Society of India and ISTE. His research interests include machine learning, IoT, and data management, with numerous publications and patents to his name. Dr. Athavale has also chaired and organized several international conferences, reflecting his commitment to advancing technology and education.