Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics e-ISSN: 2656-8624
Homepage: https://ijeeemi.org/; Vol. 8, No. 2, pp. 202-217, May 2026

RESEARCH PAPER

OPEN ACCES

A Two-Stage Hybrid Oversampling and Ensemble Learning Framework for Improved Type 2 Diabetes Mellitus Classification

Siti Fatimah Nurdiah Permatasari[1] and Ermatita[2]

Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia

Abstract

Type 2 Diabetes Mellitus (T2DM) screening using clinical tabular data commonly suffers from class imbalance, where non-diabetic records dominate diabetic cases, causing models to bias toward the majority class and yield poor detection of the positive (diabetic) class. This study aims to improve T2DM classification on an imbalanced dataset by increasing minority-class detection while maintaining acceptable overall performance. The main contribution is a leakage-safe framework that integrates two-stage hybrid oversampling (RandomOverSampler followed by Borderline-SMOTE) and soft-voting ensemble learning to obtain more balanced predictions. Experiments were conducted on the Diabetes Bangladesh (DiaBD) dataset, containing 5,288 clinical records with a binary target, diabetic (Yes/No, mapped to 1/0). The data were stratified into train_full/test splits (80/20) and further into train/validation splits (80/20 of train_full). Features were normalized using MinMaxScaler fitted only on the training set and applied to validation and test sets to prevent data leakage. Class imbalance handling was applied only on the training set using the proposed two-stage oversampling (ROS Borderline-SMOTE; borderline-1, k_neighbors=3). Classification models included SVM (RBF), Random Forest, and Gradient Boosting, as well as soft-voting ensembles of two and three models. Results show that the baseline setting (No OS) can achieve high accuracy but low minority detection; for instance, SVM (No OS) reached an accuracy of 0.9374 with a Recall_pos of 0.0909 and an F1_pos of 0.1587. After oversampling, SVM (OS) improved minority recall to 0.7273 with F1_pos 0.4188, although accuracy decreased to 0.8688 due to increased false positives. The best-balanced performance was achieved by the SVM + RandomForest soft-voting ensemble (OS) with accuracy 0.9125, Recall_pos 0.6545, and the highest F1_pos 0.4932. Overall, the proposed two-stage hybrid oversampling combined with soft-voting ensembles improves T2DM detection on imbalanced tabular data, and the findings highlight that model selection should prioritize Recall_pos and F1_pos rather than accuracy alone.

Paper history

Received Dec. 9, 2025
Revised Feb. 6, 2026
Accepted March 28, 2026
Published May 3, 2026

Keywords

Diabetes Mellitus;
Imbalance Data;
Oversampling;
SMOTE;
Ensemble Learning;
Classification.

Author Email

diaharsyad2024@gmail.com
ermatita@unsri.ac.id

I. introduction

Diabetes mellitus is a chronic disease whose prevalence continues to increase globally [1]. The International Diabetes Federation (IDF, 2021) reports that 537 million people worldwide have diabetes and projects that this number will reach 783 million by 2045 [2]. Although these statistics represent all types of diabetes, epidemiological studies indicate that approximately 90%–95% of diabetes cases are classified as Type 2 Diabetes Mellitus (T2DM) [3]. Therefore, T2DM constitutes the dominant form of diabetes and remains a primary focus in clinical screening and predictive modeling studies. In Indonesia, the number of sufferers reached 19.5 million in 2021, and projections show it will increase to 28.6 million by 2045 [4], [5]. The high number of undiagnosed cases further worsens the situation, leading to complications being identified only after severe symptoms emerge [6]. Early detection of T2DM is an important step toward reducing the risk of complications and improving patients' quality of life [7], [8], [9]. However, early detection of T2DM remains hampered by limited medical personnel, including high patient burden, uneven distribution of medical personnel, and limited time and examination resources [10]. To overcome these limitations, a computer-based system capable of classifying patient risk quickly, accurately, and consistently using machine learning methods on medical data is needed [11], [12], [13]. In recent years, machine learning-based classification methods have been widely applied to identify patients at risk of diabetes and have demonstrated competitive performance [14].

Classification of medical data is the process of grouping patients into risk categories based on patterns learned from clinical data, enabling rapid, accurate automated risk determination [15], [16]. Several studies have classified diabetes using algorithms such as Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine (SVM), achieving accuracies of 76%–98% [17], [18], [19]. However, these studies have

Corresponding author: Ermatita, ermatita@unsri.ac.id, Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.308

not addressed the problem of data imbalance, in which the number of non-diabetic cases is significantly greater than that of diabetic cases. It leads to the model being biased toward the majority class and reducing its ability to detect patients at risk of diabetes [20]. Therefore, data preprocessing requires special techniques, one of which is oversampling.

Oversampling is a technique for increasing the number of samples in the minority class to equalize its proportion to that of the majority class, so that the model does not learn only from the patterns of the dominant class [21]. Several studies have applied oversampling to diabetes classification and achieved accuracies of 92-95% [22], [23], [24]. Some of these studies demonstrated good performance, but still relied on simple duplication-based oversampling methods such as RandomOverSampler. This approach has the potential to introduce data redundancy and cannot optimally represent patient characteristics in the minority class [25]. To address these limitations, one widely used method is the Synthetic Minority Oversampling Technique (SMOTE).

SMOTE is an oversampling technique that addresses data imbalance by generating synthetic samples from the minority class [26]. Several studies have applied SMOTE to diabetes classification and achieved accuracies ranging from 75% to 98% [27], [28], [29]. Several studies have demonstrated satisfactory performance. However, SMOTE still has several limitations. The generated synthetic samples do not account for areas prone to misclassification, particularly the boundaries between the positive and negative classes [26]. It can lead to model bias in borderline regions. Consequently, risk determination is not entirely optimal. One method designed to address this issue is the Borderline-SMOTE method [30].

Borderline-SMOTE is an extension of SMOTE that focuses on generating synthetic samples in borderline regions where minority-class samples are near the majority class [31]. Borderline-SMOTE has the advantage of more selective oversampling, thus strengthening data representation in areas most prone to misclassification [31]. Borderline-SMOTE has the advantage of more selective oversampling, thus strengthening data representation in areas most prone to misclassification [32], [33]. Although Borderline-SMOTE improves minority-class representation, researchers can further enhance classification results on complex medical data by using more robust modeling methods [34]. One widely used approach to improving model generalization is ensemble learning.

Ensemble learning is a method that combines predictions from multiple base learners to produce a more stable and accurate final decision than a single model [35]. Ensemble learning has the advantage of reducing variance, minimizing the risk of overfitting, and capturing more complex patterns in data, including imbalanced data [36]. Several studies have applied ensemble learning to medical classification and achieved accuracies and F1 Scores of 92%–98%, demonstrating its superiority over a single model [37], [38], [39].

Based on these issues, this study proposes a Two-Stage Hybrid Oversampling approach combined with ensemble learning to improve the classification performance of Type 2 Diabetes Mellitus. The Two-Stage Hybrid Oversampling method improves the representation of minority classes, while ensemble learning strengthens the model's ability to capture complex patterns in medical data. The performance evaluation in this study uses accuracy, recall, precision, and F1-score. The main contributions of this research include: (1) proposing a novel integration of a Two-Stage Hybrid Oversampling workflow, combining RandomOverSampler and Borderline-SMOTE, specifically designed to address severe class imbalance in Type 2 Diabetes Mellitus prediction; (2) integrating this oversampling method with an ensemble-based classification strategy to enhance robustness; and (3) demonstrating improved predictive performance that supports earlier and more accurate identification of patients at risk for Type 2 Diabetes Mellitus. This research aims to provide a more comprehensive understanding of the system's ability to identify patients at risk of diabetes earlier and with greater accuracy.

This study is structured as follows: Section II describes the dataset used, the proposed methods, and the training and testing schemes. Section III presents the experimental results. Section IV discusses the interpretation of the results, compares them with findings from previous studies, and outlines the limitations of the proposed approach. Finally, Section V concludes the study by restating the objectives, summarizing the key findings, and providing directions for future research.

II. Materials and Method

This study proposes a classification framework for Type 2 Diabetes Mellitus (T2DM) on imbalanced tabular data through the following stages: (1) description of the DiaBD dataset, (2) data preprocessing including data cleaning, categorical feature encoding, feature-label separation, stratified data partitioning, and numeric feature normalization, (3) handling class imbalance in training data using two-stage hybrid oversampling (RandomOverSampler followed by Borderline-SMOTE), (4) training of classification models (SVM, Random Forest, Gradient Boosting) and formation of soft-voting ensemble, and (5) performance evaluation using confusion matrix and relevant metrics for imbalanced data, especially recall and f1-score for the positive class (diabetes). The research flow is shown in Fig. 1.

A. Dataset

The dataset used in this study is the Diabetes Bangladesh Dataset (DiaBD), collected by CMED Health Ltd. and the Palli Karma-Sahayak Foundation (PKSF), Bangladesh, and published through the UCI Machine Learning Repository and Data in Brief (Elsevier) [40]. This dataset contains clinical patient data used to analyze and classify Type 2 Diabetes Mellitus (T2DM). Each entry in the dataset represents a patient with numeric and categorical attributes, such as age, gender, body mass index (BMI), blood pressure, glucose level, insulin level, and other

Corresponding author: Ermatita, ermatita@unsri.ac.id, Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.308

relevant clinical features. The target column is diabetic with the class No as a non-diabetic patient and Yes as a patient diagnosed with diabetes (T2DM). To provide a clear picture of the data structure and characteristics, Fig. 1 presents a list of attributes in the DiaBD dataset, along with their data types and brief descriptions. These attributes include clinically relevant predictor variables and one target attribute used in the classification process.

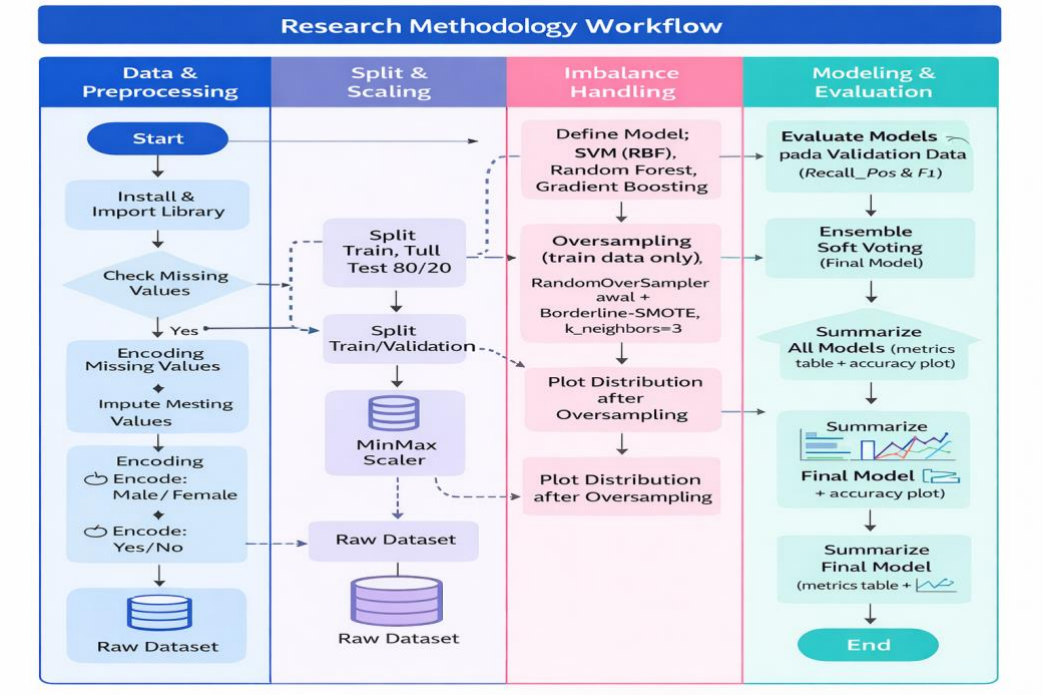

Fig. 1. Research Methodology Workflow.

Table 1. Description of the DiaBD Dataset Attributes

No	Attribute	Data Type	Description
1	Age	Numeric	Patient's age (years)
2	Gender	Categorical	Patient's gender (Male/Female)
3	Pulse Rate	Numeric	Heart rate per minute
4	Systolic_BP	Numeric	Systolic blood pressure (mmHg)
5	Diastolic_BP	Numeric	Diastolic blood pressure (mmHg)
6	Glucose	Numeric	Blood glucose level (mg/dL)
7	Height	Numeric	Patient's height (cm)
8	Weight	Numeric	Patient's weight (kg)
9	BMI	Numeric	Body Mass Index (kg/m²), calculated from weight/height
10	Family_Diabetes	Categorical	Family history of diabetes (1 = Yes / 0 = No)
11	Hypertensive	Categorical	Hypertension status (1 = Yes / 0 = No)
12	Family_Hypertension	Categorical	Family history of hypertension (1 = Yes / 0 = No)
13	Cardiovascular_Disease	Categorical	History of cardiovascular disease (1 = Yes / 0 = No)
14	Stroke	Categorical	History of stroke (1 = Yes / 0 = No)
15	Diabetic (Target)	Categorical	Target label: 0 = Non-diabetic, 1 = Diabetic

Based on Table 1, the DiaBD dataset is dominated by numeric attributes representing the patient's physiological condition, while categorical attributes are binary variables describing disease history and risk factors. The combination of these two types of attributes provides comprehensive clinical information to support the classification of type 2 diabetes mellitus. In addition to attribute characteristics, understanding the distribution of target classes is an important aspect in dataset analysis, especially for medical classification problems. The raw DiaBD dataset consists of 5,288 patient records containing both numeric and categorical clinical attributes. The dataset includes 14 predictive features related to demographic and medical examination variables, along with one target variable indicating diabetes status. Prior to preprocessing, the dataset reflects real-world clinical conditions, including class imbalance between diabetic and non-diabetic cases. A detailed description of each attribute is presented in Table 1. Table 2 presents the distribution of sample counts and percentages for each target class in the DiaBD dataset. Table 2 shows a significant class imbalance, with the non-diabetic class significantly outnumbering the diabetic class. This condition reflects the distribution of real-world clinical data and poses a major challenge in developing classification models, as it can potentially bias the model toward the

Corresponding author: Ermatita, ermatita@unsri.ac.id, Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.308

majority class. Therefore, the characteristics of this dataset are highly relevant for evaluating classification approaches designed to address class imbalance.

Table 2. Class Distribution of the Dataset

Label	Samples	Percentage
0 (Non-diabetic)	4,946	93.50%
1 (Diabetic)	342	6.50%
Total	5,288	100%

B. Data Preprocessing

The preprocessing stage is carried out to ensure data quality before entering the model training process. The main steps are as follows.

1. Data Cleaning

The dataset was checked to ensure there were no missing values or duplicate data. This is important because the presence of blank values can disrupt the model training process, while duplicate data can potentially introduce bias [41].

2. Encoding Categorical Variables

Several attributes in the dataset are categorical and therefore need to be converted to numerical representation. The gender feature (Male/Female) is encoded into a numeric format for processing by the machine learning algorithm. Furthermore, the target diabetic label, originally Yes/No, is converted to a binary label, with No = 0 and Yes = 1. Meanwhile, attributes such as family_diabetes, hypertensive, family_hypertension, cardiovascular_disease, and stroke are already binary (0/1), so they do not require additional encoding [42]. This binary encoding approach was selected because all categorical variables in the dataset are dichotomous. Binary transformation preserves feature simplicity, avoids unnecessary dimensional expansion associated with One-Hot Encoding, and is suitable for clinical tabular datasets with limited categorical levels.

3. Dataset Splitting

The dataset was divided using a stratified split strategy to ensure that class proportions were maintained in each data subset. The splitting process was carried out in two stages: (1) initial splitting of the dataset into 80% full training data (train_full) and 20% test data (test), and (2) further splitting of the full training data into 80% training data (training) and 20% validation data (validation). With this scheme, the final proportion of the dataset consisted of 64% training data, 16% validation data, and 20% test data. The application of stratification at each splitting stage ensured that the minority class, namely diabetes patients, remained proportionally represented, thereby reducing the potential for bias in the model training and evaluation process [43].

4. Data Normalization

All numeric variables, including age, pulse_rate, systolic_bp, diastolic_bp, glucose, height, weight, and BMI, were normalized using Min–Max scaling to the range [0,1] [44]. This normalization aims to standardize feature scales and prevent variables with larger numerical ranges from dominating the model training process [45]. This normalization aims to standardize feature scales and prevent variables with larger numerical ranges from dominating the model training process.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

In Eq. (1) [46], x denotes the original feature value, x' represents the normalized value, x_{min} and x_{max} are the minimum and maximum values of the corresponding feature. o avoid data leakage, the parameter x_{min} and x_{max} are calculated using only the training (fit) data. These parameters are then used to normalize the validation and test data using the same transformation, so that information from the validation and test data does not affect the model training process.

C. Two-Stage Hybrid Oversampling (RandomOverSampler and Borderline-SMOTE)

One of the primary challenges associated with the dataset used in this study is class imbalance, where the number of non-diabetic samples substantially exceeds that of diabetic samples. Such an imbalance may introduce bias in classification models, as learning algorithms tend to optimize performance toward the majority class while insufficiently capturing the characteristics of the minority class. To address this issue, this study adopts a two-stage hybrid oversampling strategy, consisting of RandomOverSampler (ROS) followed by Borderline-SMOTE, both applied sequentially to the training data.

1. RandomOverSampler (ROS)

In the first stage, RandomOverSampler (ROS) is employed to increase the proportion of the minority class through random duplication of minority samples until a more balanced class distribution is achieved. This step aims to ensure an adequate number of minority instances, thereby enabling more stable and reliable generation of synthetic samples in the subsequent stage, particularly under severe class imbalance conditions.

2. Borderline-SMOTE

In the second stage, Borderline-SMOTE is utilized to generate synthetic samples in borderline regions, i.e., areas near the decision boundary between classes that are typically difficult to classify. Unlike standard SMOTE, which generates synthetic samples across the entire minority feature space, Borderline-SMOTE focuses on minority samples located close to the majority class. This targeted oversampling strategy is expected to enhance the model's discriminative capability in critical boundary regions. In general, the synthetic sample generation process within the SMOTE is shown in Eq. (2) [47].

$$x_{new} = x_i + \delta\,(x_{nn} - x_i) \quad (2)$$

where x_i denotes a minority-class sample, x_n represents one of its k-nearest neighbors, and δ is a random value drawn from the interval [0,1]. Equation (2) produces a synthetic sample x_{new} located along the linear interpolation between x_i and its neighbor. In this study, Borderline-SMOTE is applied using the configuration kind = borderline-1 and k_neighbors = 3, following the initial reinforcement of the minority-class distribution through

Corresponding author: Ermatita, ermatita@unsri.ac.id, Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.308

ROS. The selection of Borderline-SMOTE with the borderline-1 configuration and k_neighbors = 3 was determined based on both methodological considerations and prior studies. The borderline-1 variant focuses on generating synthetic samples primarily in regions where minority samples are surrounded by majority samples, which helps improve class boundary learning while reducing noise generation [48][49]. The k_neighbors parameter was set to 3 to maintain sensitivity to local neighborhoods and avoid excessive smoothing in clinical tabular data, where minority samples are relatively sparse. Smaller k values are commonly recommended in highly imbalanced medical datasets to preserve local data characteristics and reduce the risk of generating synthetic samples in irrelevant regions [50].

The DiaBD dataset exhibits severe class imbalance, where minority samples are sparsely distributed in the feature space. Distance-based oversampling techniques such as Borderline-SMOTE rely on neighborhood structures to generate synthetic samples. When minority samples are extremely limited, applying Borderline-SMOTE directly may produce unstable synthetic samples due to insufficient minority neighbors, which can increase noise and class overlapping. To address this limitation, RandomOverSampler (ROS) is first applied to increase the representation of the minority class and stabilize sample density. By duplicating existing minority samples, ROS ensures sufficient neighborhood availability for the subsequent generation of synthetic samples. After balancing the dataset using ROS, Borderline-SMOTE is then applied to generate synthetic samples near decision boundaries, where misclassification is more likely to occur. This targeted oversampling strategy improves the representation of ambiguous minority samples and enhances boundary learning. Therefore, the proposed two-stage oversampling sequence improves data distribution stability while preserving boundary sensitivity.

3. Application to Training Data Only

The proposed two-stage hybrid oversampling approach (ROS → Borderline-SMOTE) is applied exclusively to the training dataset after feature normalization. The validation and test datasets remain unaltered and preserve their original class [51]distributions to ensure an objective and realistic evaluation of model performance. Applying oversampling techniques to the validation or test sets may lead to biased performance estimates, as the inclusion of synthetic samples can inadvertently influence evaluation outcomes. Therefore, to strictly adhere to data leakage prevention principles, oversampling is confined solely to the training phase.

D. Ensemble Training Stage

The next stage of the proposed framework involves the construction of classification models using an ensemble learning approach. The selection of Random Forest, Support Vector Machine, and Gradient Boosting was motivated by their complementary learning characteristics and their proven effectiveness in modeling nonlinear relationships and complex patterns in clinical tabular datasets. Ensemble learning refers to a methodology that combines multiple base learners to produce predictions that are generally more accurate and robust than those generated by a single model [52]. In this study, three complementary classifiers are employed as base learners: Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting (GB).

1. Random Forest (RF)

Random Forest is a bagging-based ensemble method that constructs multiple decision trees by introducing randomness in both sample selection through bootstrap sampling and feature selection at each split. The final prediction is obtained by majority voting for classification tasks or by averaging the outputs for regression tasks [53]. Mathematically, given M decision trees $h_m(x)$, Random Forest classification prediction is defined as in Eq. (3) [54].

$$\hat{y} = mode\{h_1(x), h_2(x), \ldots, h_M(x)\} \quad (3)$$

where $\hat{y}$ denotes the final predicted class label, $h_m(x)$ represents the prediction of the m-th decision tree for input x, and M is the total number of trees in the Random Forest ensemble

2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm designed to construct an optimal [55] decision boundary by maximizing the margin between classes, making it effective for binary classification tasks. Mathematically, based on Eq. (4) [56][57], the Support Vector Machine (SVM) classification function is defined as.

$$\hat{y} = sign(\sum_{i=1}^{N} \propto_i y_i K(x_i, x) + b) \quad (4)$$

where x denotes the input feature vector, y_i is the class label of the i-th training sample, $K(\cdot)$ represents the kernel function, $\propto_i$ are the Lagrange multipliers corresponding to the support vectors, and b is the bias.

3. Gradient Boosting (GB)

Gradient Boosting is a boosting method that builds decision trees incrementally, with each new tree trained to correct errors from the previous model. This process is performed by optimizing the loss function using a gradient descent approach [58], [59]. Gradient Boosting optimizes the loss function $L(y, F(x))$ in an iterative manner, where the model at iteration m is updated according to Eq (5) [60].

$$F_m(x) = F_{m-1}(x) + v \cdot h_m(x) \quad (5)$$

where $F_m(x)$ denotes the model at the m-th iteration, $h_m(x)$ is a newly trained decision tree based on the error gradient, and v is the learning rate controlling the contribution of each tree.

4. Soft Voting Ensemble

To improve classification performance, this study combines Support Vector Machine (SVM), Random Forest, and Gradient Boosting using a soft-voting ensemble approach. In this method, the class probabilities predicted by each base model are aggregated with predefined weights, and the class with the highest combined probability is selected as the final prediction [61], [62]. In practice, soft voting combines the predicted class probabilities generated by each base classifier and computes the average probability for each class. The class label corresponding to the highest

Corresponding author: Ermatita, ermatita@unsri.ac.id, Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.308

averaged probability is then assigned as the final prediction. The soft-voting formulation is given in Eq (6) [63].

$$\hat{y} = arg\ \max_{k} \sum_{m-1}^{M} w_m . P_m(y = k \mid x) \quad (6)$$

where $P_m(y = k \mid x)$ denotes the predicted probability of class k from the m-th model and w_m represents the corresponding model weight. In this study, all models were assigned equal weights, following the default setting of the *Voting Classifier* in scikit-learn, so that eac5 classifier contributed equally to the final decision. Soft voting was selected over hard voting because it leverages the predicted class probabilities produced by each base classifier rather than relying solely on discrete class labels. This probabilistic aggregation allows the ensemble to account for model confidence and provides smoother decision boundaries, which is particularly beneficial for imbalanced medical datasets. In clinical screening tasks, soft voting enables more flexible trade-offs between recall and precision, thereby improving minority-class detection while reducing abrupt decision shifts that may occur in hard-voting schemes.

5. Training and Validation Process

The models were trained using the training set balanced with SMOTE. Validation was conducted on the validation set (15%) to select the best-performing model, while the final evaluation was performed on the testing set (15%) to assess generalization capability. This strategy aims to leverage the strengths of Random Forest in terms of stability and handling non-linear data, as well as the high accuracy of Gradient Boosting through its boosting mechanism, enabling the ensemble model to achieve superior performance compared with single algorithms [64].

E. Evaluation Model

The best-performing model was evaluated using the testing set, which was not involved in either the training or validation stages, ensuring an unbiased assessment of the model's generalization capability. Model performance was measured using commonly adopted classification metrics for imbalanced data, including accuracy, precision, recall, specificity, F1-score, ROC–AUC, and PR–AUC. These metrics are derived from the confusion matrix, which consists of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), and provides the basis for a comprehensive evaluation of classification performance. Accuracy measures the proportion of correctly classified samples among all samples and is defined in Eq. (7) [65] as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

Accuracy alone is insufficient for imbalanced datasets, as a model may achieve high accuracy by predominantly predicting the majority class [66]. Precision quantifies the proportion of correct positive predictions and is defined in Eq. (8) [65] as:

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

This metric is important for controlling the number of false positives [67]. Recall measures the model's ability to correctly identify positive samples and is expressed in Eq. (9) [65] as:

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

Recall is particularly critical in medical applications, where failing to detect diabetic patients (false negatives) is more detrimental than misclassifying healthy individuals [68]. Specificity evaluates the model's ability to correctly identify negative samples and is defined in Eq. (10) [65] as.

$$Specificity = \frac{TN}{TN + FP} \quad (10)$$

This metric assesses whether the model incorrectly classifies healthy patients as diabetic [28]. The F1-score represents the harmonic mean of precision and recall, providing a balanced assessment of both metrics, as shown in Eq. (11) [65].

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (11)$$

The F1-score is well-suited for imbalanced datasets, as it accounts for both false positives and false negatives [69].

III. Results

A. Data Description

The dataset used in this study is the DiaBD_A – Diabetes Dataset for Enhanced Risk Analysis and Research in Bangladesh. Based on the initial exploratory analysis, the dataset consists of 5,288 records with 15 features, including 4,946 (93.50%) non-diabetic and 342 (6.50%) diabetic samples, indicating a significant class imbalance, comprising multiple predictor features and one target variable. The relatively large number of samples provides a sufficient basis for training and evaluating machine learning models. Each record represents an individual, while the available attributes describe various characteristics relevant to the risk of Type 2 Diabetes Mellitus. The detailed description of the dataset attributes has been presented in Table 1 in the Methodology section, while a sample of the dataset records is provided in Table 3. Based on Table 3, the DiaBD_A dataset consists of diverse demographic and clinical attributes, including age, gender, pulse rate, blood pressure, blood glucose levels, and anthropometric measurements. The sample records demonstrate considerable variability across individuals in terms of age, body mass index, and physiological indicators. In addition to numerical features, the dataset also contains several categorical variables, such as gender, and binary health-status attributes related to family history and comorbid conditions. The presence of both numerical and categorical data highlights the need for appropriate preprocessing before applying machine learning models for Type 2 Diabetes Mellitus classification.

Corresponding author: Ermatita, ermatita@unsri.ac.id, Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.308

Table 3. Sample Records from the DiaBD_A Dataset

No	Age	Gen	PR	SBP	DBP	Glu	H (m)	W (kg)	BMI	FD	HTN	FHTN	CVD	Str	D
1	42	F	66	110	73	5.88	1.65	70.2	25.75	No	No	No	No	No	No
2	35	F	60	125	68	5.71	1.47	42.5	19.58	No	No	No	No	No	No
3	62	F	57	127	74	6.85	1.52	47.0	20.24	No	No	No	No	No	No
4	73	M	55	193	112	6.28	1.63	57.4	21.72	No	No	No	No	No	No
5	68	F	71	150	81	5.71	1.42	36.0	17.79	No	No	No	No	No	No

Notes: *Gen* denotes gender (M = male, F = female); *PR* represents pulse rate; *SBP* and *DBP* indicate systolic and diastolic blood pressure, respectively; *Glu* refers to fasting blood glucose; *H* and *W* denote height and weight; *FD* indicates family history of diabetes; *HTN* denotes hypertensive status; *FHTN* represents family history of hypertension; *CVD* indicates cardiovascular disease; *Str* denotes stroke; and *D* represents the diabetic status used as the class label in this study.

B. Data Preprocessing

1. Data Cleaning

The first preprocessing step involved assessing the quality of the dataset by checking for missing values and duplicated instances. The evaluation results showed that all columns, including age, gender, pulse_rate, systolic_bp, diastolic_bp, glucose, height, weight, bmi, family_diabetes, hypertensive, family_hypertension, cardiovascular_disease, stroke, and diabetic, contained 0 missing values. Therefore, no imputation was required. In addition, duplicate checking confirmed that the dataset contained 0 duplicated records, indicating that the data is clean and free from redundancy. Consequently, the dataset could be directly processed in subsequent steps without requiring data correction.

2. Encoding Categorical Variables

Several categorical attributes were included in the dataset, such as gender, family_diabetes, hypertensive, family_hypertension, and stroke, which originally used string-based labels. Since machine learning algorithms require numerical representation, categorical encoding was applied. A binary conversion method was used to map values such as Male/male/M to 0 and Female/female/F to 1, and likewise, No/no was mapped to 0 and Yes/yes to 1. After this conversion, the dataset was re-examined to identify any remaining object-type features, which were then encoded using one-hot encoding when more than two categories existed. After all encoding operations were completed, the dataset consisted of 14 features, as reflected by the final data shape: (5288, 14).

3. Dataset Splitting

Following the encoding process, the dataset was separated into feature variables (X) and the target variable (y), with the diabetic attribute designated as the prediction target. The dataset consisted of 5288 instances and exhibited class imbalance, with 4946 samples in class 0 (non-diabetic) and 342 in class 1 (diabetic). To preserve the proportional representation of both classes and prevent model bias during the learning process, a stratified splitting strategy was implemented. The dataset was first divided into a training set (80%) and a test set (20%). Subsequently, the training portion was partitioned again into training and validation sets using an additional 80:20 split, yielding 3384 samples for training, 846 for validation, and 1058 for testing. This structured partitioning ensures that model training, hyperparameter tuning, and final performance evaluation are conducted on mutually exclusive subsets, thereby maintaining objectivity and reducing the risk of overfitting.

4. Data Normalization

The final step in the preprocessing phase was feature scaling to ensure uniform value ranges and enhance algorithm learning efficiency. In this study, Min-Max Scaling was employed to normalize all numerical attributes into a [0, 1] range. To prevent data leakage, the scaler was fitted only on the training set (X_train), and the learned parameters were then used to transform the validation set (X_val) and the test set (X_test). This ensures that the model does not obtain prior statistical knowledge from unseen data, thereby maintaining the validity of the experimental evaluation.

C. Two-Stage Hybrid Oversampling

This study aims to reveal whether there is a To address the substantial class imbalance in the dataset, the preprocessing stage employed a Two-Stage Hybrid Oversampling technique. This method consists of two sequential steps designed to balance the minority class without increasing the risk of overfitting commonly associated with single random oversampling. First, Random Oversampling was applied to replicate minority class samples and improve the initial representation of the diabetic class. This was followed by BorderlineSMOTE, a SMOTE variant that synthesizes new samples near the decision boundary between classes to generate more informative and representative minority data, particularly in regions prone to misclassification. This combined strategy not only increases the number of minority samples but also enhances their diversity and relevance, allowing the model to better learn decision boundaries and reducing bias toward the majority class. A visual comparison of the class distribution before and after oversampling is provided in Fig. 2. Fig. 2 clearly shows the effectiveness of this approach. Before oversampling, the dataset was highly imbalanced with 3165 samples in class 0 (non-diabetic) and only 219 samples in class 1

Corresponding author: Ermatita, ermatita@unsri.ac.id, Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.308

(diabetic). After applying the combined Random Oversampling and Borderline-SMOTE method, the distribution became fully balanced with 3165 samples in each class. The left chart in Fig. 2 corresponds to the Before Oversampling condition with severe underrepresentation of the minority class, while the right chart represents the After Combined Oversampling condition with a symmetric class distribution. These results confirm that the two-stage oversampling strategy successfully resolved the class imbalance issue, enabling a fairer and less biased model training process.

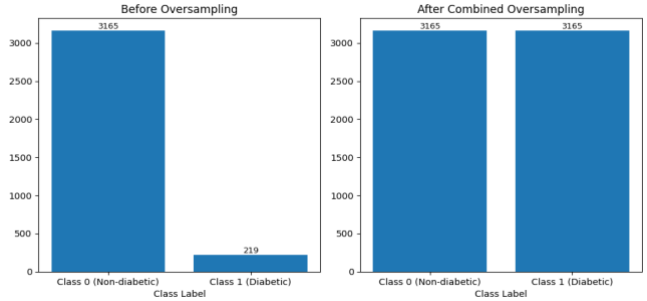

Fig. 2. Class Distribution Before and After Combined Oversampling.

D. Training Result

1. Single Classifier Without Oversampling

The first set of experiments used single classifiers: Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting (GB) trained on the original imbalanced dataset without any oversampling. SVM separates classes by finding an optimal hyperplane, RF constructs an ensemble of decision trees using bagging, and GB sequentially builds weak learners to minimize prediction errors. These baseline models rely entirely on the natural class distribution, which is heavily skewed toward the majority class.

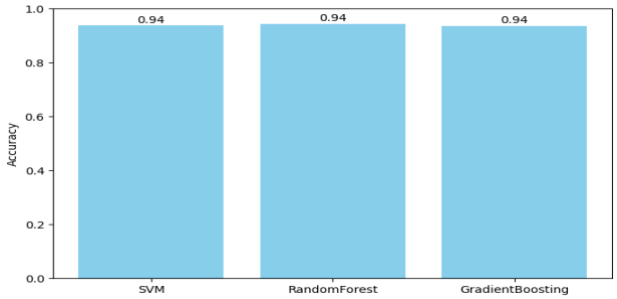

Fig. 3. Baseline Model Performance (Without Oversampling)

As shown in Fig. 3, Random Forest achieved the highest accuracy of 94.44%, followed by SVM (93.74%) and Gradient Boosting (93.62%). However, the recall for the diabetic class was very low (SVM: 0.0909, RF: 0.1818, GB: 0.2000), indicating strong bias toward the majority class. These results highlight that high accuracy alone is insufficient for evaluating model performance on imbalanced datasets, as the models failed to detect minority cases effectively, with recall values below 0.20.

2. Single Classifiers With Oversampling

To mitigate class imbalance, oversampling techniques were applied to the training data, generating additional instances of the minority class. The same single classifiers (SVM, RF, GB) were then retrained on this augmented dataset. Oversampling enables models to better learn the characteristics of the diabetic class, aiming to improve detection performance while maintaining overall accuracy.

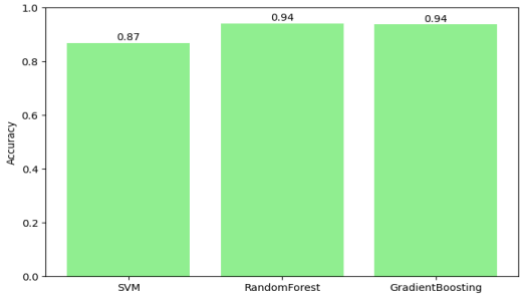

Fig. 4. Single Classifiers with Oversampling

Fig. 4 shows that the SVM's overall accuracy decreased to 0.8688, but its recall for the diabetic class substantially increased to 0.7273 (from 0.0909 in the baseline), demonstrating improved minority-class detection. Gradient Boosting's recall increased to 0.3273, while Random Forest maintained relatively high accuracy (0.9409) with a recall of 0.2364. These findings illustrate the trade-off between overall accuracy and minority class recognition when applying oversampling.

3. Ensemble Models with Oversampling

Ensemble learning combines multiple classifiers to leverage their complementary strengths, improving robustness and generalization. Various combinations of SVM, RF, and GB were trained on the oversampled dataset, including SVM+RF, SVM+GB, RF+GB, and a three-model ensemble SVM+RF+GB. Ensemble methods aim to achieve more balanced performance for both majority and minority classes.

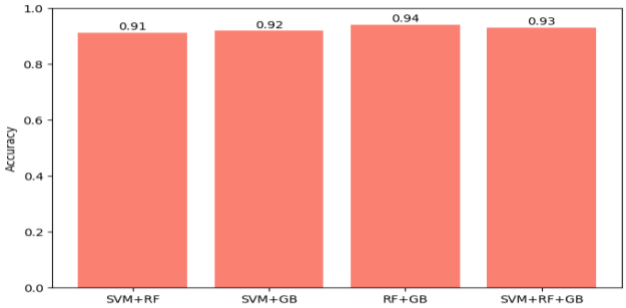

Fig. 5. Ensemble Models with Oversampling

As illustrated in Fig. 5, the RF–GB ensemble achieved the highest overall accuracy of 94.21% and a recall of 30.91% for the diabetic class, outperforming both individual classifiers and other ensemble combinations. Ensemble models exhibited more balanced performance, with improved recall (up to 0.3091) while maintaining high accuracy (above 94%), demonstrating their effectiveness in detecting minority-class instances.

E. Best Model Performance on Testing

Corresponding author: Ermatita, ermatita@unsri.ac.id, Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.308

The best-performing model based on validation results was the Random Forest–Gradient Boosting (RF+GB) ensemble with oversampling. This model combines the strengths of both classifiers to enhance minority class detection while maintaining high overall accuracy. It was further evaluated on the independent testing set to assess generalization performance. Key evaluation metrics included accuracy, precision, recall, and F1-score for each class, as well as the confusion matrix to illustrate the distribution of correct and incorrect predictions.

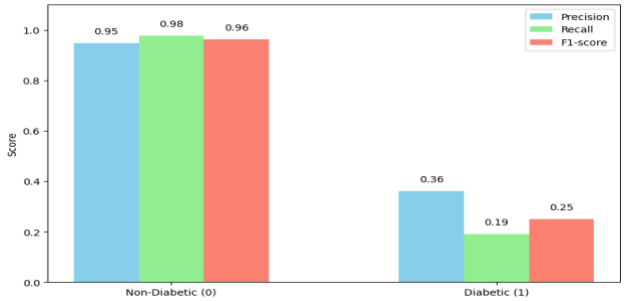

Fig. 6. Test Set Performance Metrics for RF+GB Ensemble Oversampling

Fig. 6 presents the precision, recall, and F1-score for the RF+GB ensemble on the test set. The model achieved an overall accuracy of 92.63%. While performance on the majority class (Non-Diabetic) was very high (precision: 0.9462, recall: 0.9768, F1-score: 0.9612), the recall for the minority class (Diabetic) was limited to 0.1912, resulting in a lower F1-score of 0.2500. These results indicate that although the ensemble model can accurately predict most non-diabetic cases, detecting diabetic cases remains challenging due to class imbalance.

Table 4. Validation Results of All Models

Model	Accuracy	Precision_pos	Recall_pos	F1_pos
RF(No OS)	0.944	0.833	0.182	0.299
SVM (No OS)	0.937	0.625	0.091	0.159
GB (No OS)	0.936	0.524	0.200	0.289
RF (OS)	0.941	0.619	0.236	0.342
GB (OS)	0.937	0.529	0.327	0.404
SVM (OS)	0.869	0.294	0.727	0.419
RF + GB (Ensemble OS)	0.942	0.607	0.309	0.410
SVM + GB (Ensemble OS)	0.920	0.386	0.400	0.393
SVM + RF + GB (Ensemble OS)	0.930	0.458	0.400	0.427
SVM + RF (Ensemble OS)	0.913	0.396	0.655	0.493

Among all evaluated models, the SVM + Random Forest ensemble with oversampling achieved the best performance, with an accuracy of 0.9125, a recall of 0.6545, and an F1-score of 0.4932. These results indicate that combining two-stage hybrid oversampling with ensemble learning significantly improves minority class detection in imbalanced medical datasets.

IV. Discussion

A. Baseline (No Oversampling): High Accuracy but Poor Minority Detection

In imbalanced medical classification problems, models can achieve high accuracy by favoring the majority class, as shown by Random Forest, which achieved 0.944 accuracy but only 0.1818 recall for the diabetic class, while failing to correctly identify the minority (positive) class [70]. This behavior is clearly observed in our baseline results. SVM (No OS) reached 0.9374 accuracy, yet the positive-class performance was very low (Recall_pos = 0.0909, F1_pos = 0.1587), indicating that most diabetes cases were not detected, which is problematic in clinical screening scenarios.

Table 5. Baseline Models Without Oversampling

Model	Accuracy	Precision_pos	Recall_pos	F1_pos
RF	0.944	0.833	0.182	0.299
SVM	0.937	0.625	0.091	0.159
GB	0.936	0.524	0.200	0.289

In clinical screening contexts, false negatives (diabetic patients predicted as non-diabetic) are typically more critical than false positives because missed diagnoses can delay intervention and increase complication risk [11]. Similarly, RandomForest (No OS) achieved the highest accuracy among the baseline models (0.9444) but still had weak minority detection (Recall_pos = 0.1818, F1_pos = 0.2985), reinforcing that accuracy alone is insufficient for imbalanced datasets [71].

B. Impact of Two-Stage Hybrid Oversampling (ROS Borderline-SMOTE)

This study applies a two-stage hybrid oversampling strategy on the training set (RandomOverSampler followed by Borderline-SMOTE). The rationale is to (1) increase minority representation quickly (ROS) and (2) generate synthetic samples near difficult decision regions (borderline) to better shape the class boundary. The effect is evident from the SVM comparison: after oversampling, SVM (OS) substantially improved diabetes detection, achieving Recall_pos = 0.7273 and F1_pos = 0.4188, compared to the baseline SVM (No OS) with Recall_pos = 0.0909 and F1_pos = 0.1587. However, this improvement in minority recall was accompanied by a decrease in precision, where Precision_pos dropped from 0.6250 (No OS) to 0.2941 (OS), indicating an increase in false positive predictions. As a result, the overall accuracy also decreased to 0.8688. This quantitative comparison highlights the inherent trade-off between minority recall and false positives in imbalanced classification. Therefore, model selection in clinical screening tasks should prioritize minority-sensitive metrics such as Recall_pos and F1_pos rather than relying solely on accuracy.

It should be noted that this study focuses on validation-based performance comparisons rather than formal statistical significance testing. No hypothesis-based statistical tests were conducted to quantify the significance of the observed performance differences. However, the improvements in Recall_pos and F1_pos

Corresponding author: Ermatita, ermatita@unsri.ac.id, Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.308

were consistently observed across baseline, oversampling, and ensemble configurations, indicating a stable and reproducible performance trend. In the context of clinical screening, such consistent improvements in minority-class detection are practically meaningful, as they directly relate to reducing missed diagnoses. Future work may incorporate repeated cross-validation and statistical significance testing to further assess the robustness of these performance differences.

C. Role of Soft-Voting Ensemble: Balancing Recall–Precision Trade-offs

Oversampling can make certain classifiers overly sensitive to the minority class (higher recall but lower precision due to increased false positives). In such conditions, soft-voting ensembles can help stabilize predictions by combining probabilistic outputs from complementary models. Based on your validation results, SVM + RandomForest (Ensemble OS) produced the highest F1_pos = 0.4932 with Recall_pos = 0.6545 and Accuracy = 0.9125. This suggests a more balanced minority-class performance than SVM(OS) alone, which achieved the highest recall but much lower precision. Thus, if the goal is to improve diabetes detection while keeping false alarms at a manageable level, the SVM + RandomForest (Ensemble OS) configuration is the most suitable among the tested models, as it maximizes F1_pos (a balance between precision and recall) while maintaining reasonable accuracy.

In this study, a “manageable level” of false alarms refers to a classification outcome in which the increase in false positive predictions remains acceptable when balanced against a substantial improvement in minority-class detection. Specifically, the SVM + Random Forest (Ensemble OS) configuration achieves a higher Recall_pos (0.6545) while maintaining a moderate Precision_pos (0.3956), resulting in the highest F1_pos among the evaluated models. In clinical screening contexts, false positives are generally more tolerable than false negatives, as additional follow-up examinations can be conducted to confirm diagnosis. Therefore, the observed precision–recall balance of the proposed ensemble is considered manageable and suitable for early diabetes risk screening.

The differences observed among ensemble configurations can be attributed to the complementary characteristics of the base classifiers. The SVM + Random Forest ensemble achieved a more balanced performance because SVM effectively captures complex decision boundaries, while Random Forest provides stability through ensemble averaging and robustness to noise. In contrast, ensembles involving Gradient Boosting tended to exhibit either higher variance or sensitivity to oversampled data, as boosting-based methods iteratively emphasize misclassified samples, which may amplify noise introduced by synthetic instances. Similarly, combining all three models did not consistently improve performance, as including multiple highly correlated decision tree–based learners may reduce ensemble diversity. These findings highlight that effective ensemble performance depends not only on the number of combined models but also on their complementary learning behaviors.

Although oversampling is known to increase false positive predictions by amplifying minority-class sensitivity, the proposed ensemble framework helps mitigate this limitation through probabilistic aggregation. By combining predictions from classifiers with complementary behaviors, the soft-voting ensemble reduces overconfidence from individual models that may become overly sensitive after oversampling. In particular, while SVM (OS) exhibits very high recall with reduced precision, the inclusion of Random Forest in the ensemble contributes stabilizing effects by leveraging decision tree averaging and robustness to noise. This interaction results in a more balanced precision–recall trade-off, indicating that the ensemble effectively suppresses excessive false positives while preserving improved minority-class detection.

D. Comparison with Related Studies

Any prior studies on diabetes prediction investigate (i) resampling methods such as SMOTE variants, and/or (ii) ensemble learning (including voting) to improve classification performance. As summarized in Table 4, these approaches generally report improvements in overall accuracy and AUC compared to baseline models. However, cross-paper numerical comparisons must be interpreted carefully because datasets (e.g., PIMA vs. CDC/BRFSS vs. local clinical cohorts), class distributions, preprocessing steps, and evaluation protocols often differ, which can substantially affect the reported performance metric. The studies in Table 4 show that the most recent diabetes prediction research focuses on addressing class imbalance using SMOTE-based oversampling and ensemble learning techniques, which consistently improve model accuracy and AUC. Approaches such as Random Forest, boosting, and hybrid resampling (e.g., SMOTE-Tomek) demonstrate strong performance in detecting diabetes, although minority class metrics like Recall and F1-score are still often underreported or remain suboptimal. Compared with prior work, the proposed method follows the same trend and achieves competitive overall accuracy while improving sensitivity to positive diabetes cases, indicating progress in reducing false-negative predictions, an essential aspect for reliable early diagnosis.

Compared with prior studies summarized in Table 4, the proposed method differs in several key methodological aspects. First, while most existing works employ single-stage oversampling techniques such as standard SMOTE and its variants, this study adopts a two-stage hybrid strategy combining RandomOverSampler and Borderline-SMOTE to address both severe class imbalance and decision boundary refinement. Second, regarding ensemble strategies, many related studies rely on single classifiers or homogeneous ensembles, whereas the proposed approach integrates heterogeneous classifiers (SVM, Random Forest, and Gradient Boosting) using a soft-voting scheme to leverage complementary learning behaviors. Finally, in contrast to studies that primarily report overall accuracy

Corresponding author: Ermatita, ermatita@unsri.ac.id, Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.308

or AUC, this work emphasizes minority-sensitive metrics such as Recall_pos and F1_pos, which are more relevant for clinical screening scenarios. These distinctions highlight the proposed framework's contribution to improving minority-class detection while maintaining balanced predictive performance. Several studies summarized in Table 4 employ benchmark datasets with relatively smaller sample sizes, such as the PIMA Indian Diabetes Dataset. Compared with such datasets, the DiaBD dataset used in this study contains 5,288 clinical records, offering greater sample diversity and more realistic class distributions. Prior research has shown that larger and more diverse clinical datasets tend to support more stable model training and reduce performance variance caused by data sparsity. Therefore, differences in reported performance across studies should be interpreted in light of dataset scale and complexity, in addition to methodological variations.

Table 6. Summary of Related Studies on Handling Class Imbalance in Diabetes Prediction

Author	Method	Dataset	Reported Performance
Siti et al [22]	SMOTE + Random Forest	BRFSS Survey Data (Kaggle) — diabetes risk factors	Accuracy improved from 86% → 92%; for minority class: Precision 96%, Recall 88%, F1-score 92%
Pradeepa et al [27]	SMOTE + Ensemble learning (AdaBoost / XGBoost)	Clinical diabetes dataset (hospital-based)	AUC = 0.968 ± 0.015
Younseo et al [80]	SMOTE + RUS + Soft Voting Ensemble	Scikit-learn Diabetes Dataset	Accuracy = 0.8764; AUC = 0.9227 (PMC)
Ganie et al [81]	Boosting ensemble + data preprocessing + upsampling	Pima Indian Diabetes Dataset (UCI)	AUC = 0.950; Specificity = 0.934; Accuracy = 88.84%; Precision = 84.32%; Sensitivity = 78.0% (PMC)
Sukamto et al [82]	SMOTE-Tomek Links + Random Forest	Local healthcare diabetes data (2,075 records)	Accuracy improved from 97% → 99.64% after applying SMOTE-Tomek
García-Ordás et al. [83] vgarcia	Deep Learning (VAE + SAE + CNN) + Oversampling	PIMA Indian Diabetes Dataset (UCI)	Accuracy 92.31%
Proposed Method	ROS → Borderline-SMOTE + Soft-Voting Ensemble	DiaBD Dataset	Accuracy = 0.9125; Recall (positive) = 0.6545; F1-score (positive) = 0.4932

E. Key Advantages of This Study (Strengths / Novelty)

1. Explicit two-stage oversampling design (ROS-Borderline-SMOTE)
   Instead of using a single oversampling technique, the proposed pipeline applies a structured two-step strategy that improves minority representation and targets hard boundary regions [72].
2. Leakage-safe training protocol
   Scaling is performed with a MinMaxScaler fitted only to the training set, and oversampling is also applied only to the training set, preventing contamination of validation/test distributions and preserving evaluation validity [73].
3. Minority-focused evaluation aligned with clinical screening needs
   Rather than optimizing accuracy alone, the study emphasizes Recall_pos and F1_pos, which are more appropriate for imbalanced clinical prediction tasks and safer in screening-oriented settings [74].
4. Ensemble as a practical solution to oversampling trade-offs
   The SVM + RandomForest soft-voting ensemble achieved the highest F1_pos while maintaining strong recall and acceptable accuracy, indicating a balanced operating point suitable for real screening pipelines [75].
5. Use of the DiaBD dataset
   Using DiaBD strengthens the relevance of the study beyond overused small benchmarks and supports a more realistic evaluation of larger tabular clinical data [76].

F. Limitations and Future Work

Although oversampling improves minority recall, it may increase false positives, especially for highly sensitive models. Future work can explore decision-threshold tuning, probability calibration, or cost-sensitive learning to reduce false alarms while maintaining recall [77]. In addition, employing stratified k-fold cross-validation can yield more robust performance estimates compared to a single hold-out split [78]. Finally, adding interpretability analysis (e.g., feature importance/SHAP) would support clinical trust and transparency [79].

V. Conclusion

This study proposes a Type 2 Diabetes Mellitus (T2DM) classification framework on the DiaBD dataset (5,288 samples; binary target `diabetic`) by integrating MinMaxScaler normalization (fitted only on the training set) and class-imbalance handling via a two-stage hybrid

Corresponding author: Ermatita, ermatita@unsri.ac.id, Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.308

oversampling strategy (RandomOverSampler → Borderline-SMOTE) applied exclusively to the training data to prevent data leakage. Under the baseline setting without oversampling (No OS), the models tended to achieve high overall accuracy but performed poorly at detecting diabetic cases; for example, SVM (No OS) reached an accuracy of 0.9374 while achieving only 0.0909 positive-class recall (Recall_pos) and 0.1587 positive-class F1 (F1_pos). After applying oversampling (OS), minority detection improved substantially; SVM (OS) achieved the highest Recall_pos of 0.7273 (F1_pos = 0.4188), although this came with a trade-off in overall accuracy (0.8688) due to increased false positives. To obtain a more balanced performance on the minority class, soft-voting ensembles provided a more stable operating point; the best validation result was achieved by the SVM + RandomForest (Ensemble OS) configuration, with the highest F1_pos of 0.4932 and Recall_pos of 0.6545 at an accuracy of 0.9125. Overall, these results indicate that combining two-stage hybrid oversampling with ensemble learning is effective for improving T2DM classification on imbalanced tabular data, and that model selection should prioritize positive-class metrics (Recall_pos and F1_pos) rather than accuracy alone. Future work will focus on strengthening the evaluation using stratified k-fold cross-validation and reporting final results on the test set, while further reducing false positives through decision-threshold tuning, probability calibration, or cost-sensitive learning. In addition, interpretability analysis (e.g., feature importance/SHAP) is recommended to improve transparency and support clinical decision-making adoption.

References

[1] O. A. Ojo, H. S. Ibrahim, D. E. Rotimi, A. D. Ogunlakin, and A. B. Ojo, “Diabetes mellitus: From molecular mechanism to pathophysiology and pharmacology,” *Med. Nov. Technol. Devices*, vol. 19, p. 100247, Sep. 2023, doi: 10.1016/j.medntd.2023.100247.

[2] A. Kumar, R. Gangwar, A. Ahmad Zargar, R. Kumar, and A. Sharma, “Prevalence of Diabetes in India: A Review of IDF Diabetes Atlas 10th Edition,” *Curr. Diabetes Rev.*, vol. 20, no. 1, Jan. 2024, doi: 10.2174/1573399819666230413094200.

[3] M. Ortiz-Martínez, M. González-González, A. J. Martagón, V. Hlavinka, R. C. Willson, and M. Rito-Palomares, “Recent Developments in Biomarkers for Diagnosis and Screening of Type 2 Diabetes Mellitus,” *Curr. Diab. Rep.*, vol. 22, no. 3, pp. 95–115, Mar. 2022, doi: 10.1007/s11892-022-01453-4.

[4] B. B. Duncan, C. Stein, and A. Basit, “Edinburgh Research Explorer IDF Diabetes Atlas,” *Glob. Reg. country-level diabetes Preval. Estim. 2021 Proj. 2045*, 2021.

[5] International Diabetes Federation, “Indonesia – IDF Diabetes Atlas (10th Edition),” Brussels, Belgium.

[6] J. Zhang, Z. Zhang, K. Zhang, X. Ge, R. Sun, and X. Zhai, “Early detection of type 2 diabetes risk: limitations of current diagnostic criteria,” *Front. Endocrinol. (Lausanne).*, vol. 14, Nov. 2023, doi: 10.3389/fendo.2023.1260623.

[7] N. S. Skolnik and A. J. Style, “Importance of Early Screening and Diagnosis of Chronic Kidney Disease in Patients with Type 2 Diabetes,” *Diabetes Ther.*, vol. 12, no. 6, pp. 1613–1630, Jun. 2021, doi: 10.1007/s13300-021-01050-w.

[8] S. Shah *et al.*, “Diabetic retinopathy in newly diagnosed Type 2 diabetes mellitus: Prevalence and predictors of progression; a national primary network study,” *Diabetes Res. Clin. Pract.*, vol. 175, p. 108776, May 2021, doi: 10.1016/j.diabres.2021.108776.

[9] R. Manuela, “Positive Effekte einer frühen Diabetestherapie greifen lebenslang,” *Diabetes aktuell*, vol. 22, no. 06, pp. 234–234, Oct. 2024, doi: 10.1055/a-2420-2434.

[10] M. Masdiana, R. Hidayat, and T. Febrianti, “Efektifitas Intervensi Berbasis Komunitas Terhadap Penderita Diabetes Mellitus Tipe 2 : A Systematic Review,” *J. Ners*, vol. 9, no. 2, pp. 3115–3124, Apr. 2025, doi: 10.31004/jn.v9i2.44463.

[11] J. J. Boutilier, T. C. Y. Chan, M. Ranjan, and S. Deo, “Risk Stratification for Early Detection of Diabetes and Hypertension in Resource-Limited Settings: Machine Learning Analysis,” *J. Med. Internet Res.*, vol. 23, no. 1, p. e20123, Jan. 2021, doi: 10.2196/20123.

[12] A. M. Rahmani *et al.*, “Machine Learning (ML) in Medicine: Review, Applications, and Challenges,” *Mathematics*, vol. 9, no. 22, p. 2970, Nov. 2021, doi: 10.3390/math9222970.

[13] S. M. D. A. C. Jayatilake and G. U. Ganegoda, “Involvement of Machine Learning Tools in Healthcare Decision Making,” *J. Healthc. Eng.*, vol. 2021, pp. 1–20, Jan. 2021, doi: 10.1155/2021/6679512.

[14] H. Habehh and S. Gohel, “Machine Learning in Healthcare,” *Curr. Genomics*, vol. 22, no. 4, pp. 291–300, Dec. 2021, doi: 10.2174/1389202922666210705124359.

[15] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, “Machine learning towards intelligent systems: applications, challenges, and opportunities,” *Artif. Intell. Rev.*, vol. 54, no. 5, pp. 3299–3348, Jun. 2021, doi: 10.1007/s10462-020-09948-w.

[16] E. Barbierato and A. Gatti, “The Challenges of Machine Learning: A Critical Review,” *Electronics*, vol. 13, no. 2, p. 416, Jan. 2024, doi: 10.3390/electronics13020416.

[17] R. Venkatesh *et al.*, “Evaluation of Systemic Risk Factors in Patients with Diabetes Mellitus for Detecting Diabetic Retinopathy with Random Forest Classification Model,” *Diagnostics*, vol. 14, no. 16, p. 1765, Aug. 2024, doi: 10.3390/diagnostics14161765.

Corresponding author: Ermatita, ermatita@unsri.ac.id, Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.308

[18] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, "Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM)," *Diagnostics*, vol. 11, no. 9, p. 1714, Sep. 2021, doi: 10.3390/diagnostics11091714.

[19] C. Dewi, J. Zendrato, and H. J. Christanto, "Improvement of support vector machine for predicting diabetes mellitus with machine learning approach," *J. Auton. Intell.*, vol. 7, no. 2, Dec. 2023, doi: 10.32629/jai.v7i2.888.

[20] I. Abousaber, H. F. Abdallah, and H. El-Ghaish, "Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets," *Front. Artif. Intell.*, vol. 7, Jan. 2025, doi: 10.3389/frai.2024.1499530.

[21] Y. Chen, J. Zou, L. Liu, and C. Hu, "Improved Oversampling Algorithm for Imbalanced Data Based on K-Nearest Neighbor and Interpolation Process Optimization," *Symmetry (Basel).*, vol. 16, no. 3, p. 273, Feb. 2024, doi: 10.3390/sym16030273.

[22] S. K. Nisa, M. A. Barata, and P. E. Yuwita, "Optimization of Random Forest Algorithm with SMOTE Method to Improve the Accuracy of Early Diabetes Prediction," *Sci. J. Informatics*, vol. 12, no. 3, pp. 387–396, Aug. 2025, doi: 10.15294/sji.v12i3.22986.

[23] S. Z. R. Krishandhie and A. Purwinarko, "Random Forest Algorithm Optimization using K-Nearest Neighbor and SMOTE on Diabetes Disease," *Recursive J. Informatics*, vol. 3, no. 1, pp. 43–50, Mar. 2025, doi: 10.15294/rji.v3i1.1576.

[24] S. Rahmawati, A. Wibowo, and A. F. N. Masruriyah, "Improving Diabetes Prediction Accuracy in Indonesia: A Comparative Analysis of SVM, Logistic Regression, and Naive Bayes with SMOTE and ADASYN," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 8, no. 5, pp. 607–614, Oct. 2024, doi: 10.29207/resti.v8i5.5980.

[25] R. S. Abdulsadig and E. Rodriguez-Villegas, "A comparative study in class imbalance mitigation when working with physiological signals," *Front. Digit. Heal.*, vol. 6, Mar. 2024, doi: 10.3389/fdgth.2024.1377165.

[26] R. Mohammed and E. M. Karim, "FCM-CSMOTE: Fuzzy C-Means Center-SMOTE," *Expert Syst. Appl.*, vol. 248, p. 123406, Aug. 2024, doi: 10.1016/j.eswa.2024.123406.

[27] P. Sampath *et al.*, "Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique," *Sci. Rep.*, vol. 14, no. 1, p. 28984, Nov. 2024, doi: 10.1038/s41598-024-78519-8.

[28] P. Talari *et al.*, "Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2," *PLoS One*, vol. 19, no. 1, p. e0292100, Jan. 2024, doi: 10.1371/journal.pone.0292100.

[29] M. Kivrak, U. Avci, H. Uzun, and C. Ardic, "The Impact of the SMOTE Method on Machine Learning and Ensemble Learning Performance Results in Addressing Class Imbalance in Data Used for Predicting Total Testosterone Deficiency in Type 2 Diabetes Patients," *Diagnostics*, vol. 14, no. 23, p. 2634, Nov. 2024, doi: 10.3390/diagnostics14232634.

[30] I. M. Bermudez Vera, J. Mosquera Restrepo, and D. F. Manotas-Duque, "Data Mining for the Adjustment of Credit Scoring Models in Solidarity Economy Entities: A Methodology for Addressing Class Imbalances," *Risks*, vol. 13, no. 2, p. 20, Jan. 2025, doi: 10.3390/risks13020020.

[31] H. Sun, J. Li, and X. Zhu, "A Novel Expandable Borderline Smote Over-Sampling Method for Class Imbalance Problem," *IEEE Trans. Knowl. Data Eng.*, vol. 37, no. 5, pp. 2183–2199, May 2025, doi: 10.1109/TKDE.2025.3544284.

[32] Y. Chachoui, N. Azizi, R. Hotte, and T. Bensebaa, "Enhancing algorithmic assessment in education: Equi-fused-data-based SMOTE for balanced learning," *Comput. Educ. Artif. Intell.*, vol. 6, p. 100222, Jun. 2024, doi: 10.1016/j.caeai.2024.100222.

[33] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, p. 87, Jun. 2024, doi: 10.1186/s40537-024-00943-4.

[34] S. Gholampour, "Impact of Nature of Medical Data on Machine and Deep Learning for Imbalanced Datasets: Clinical Validity of SMOTE Is Questionable," *Mach. Learn. Knowl. Extr.*, vol. 6, no. 2, pp. 827–841, Apr. 2024, doi: 10.3390/make6020039.

[35] S. M. Ganie, P. K. D. Pramanik, and Z. Zhao, "Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets," *Sci. Rep.*, vol. 15, no. 1, p. 13912, Apr. 2025, doi: 10.1038/s41598-025-97547-6.

[36] M. E. Ahsen, "Harnessing Unsupervised Ensemble Learning for Biomedical Applications: A Review of Methods and Advances," *Mathematics*, vol. 13, no. 3, p. 420, Jan. 2025, doi: 10.3390/math13030420.

[37] B. O. Olorunfemi *et al.*, "Efficient diagnosis of diabetes mellitus using an improved ensemble method," *Sci. Rep.*, vol. 15, no. 1, p. 3235, Jan. 2025, doi: 10.1038/s41598-025-87767-1.

[38] H. Ulutas, R. B. Günay, and M. E. Sahin, "Detecting diabetes in an ensemble model using a unique PSO-GWO hybrid approach to hyperparameter optimization," *Neural Comput. Appl.*, vol. 36, no. 29, pp. 18313–18341, Oct. 2024, doi: 10.1007/s00521-024-10160-y.

[39] M. S. Reza, R. Amin, R. Yasmin, W. Kulsum, and S. Ruhi, "Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data," *Heliyon*, vol. 10,

Corresponding author: Ermatita, ermatita@unsri.ac.id, Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.308

no. 2, p. e24536, Jan. 2024, doi: 10.1016/j.heliyon.2024.e24536.

[40] T. T. Prama, M. J. Rahman, M. Zaman, F. Sarker, and K. A. Mamun, "DiaBD: A diabetes dataset for enhanced risk analysis and research in Bangladesh," *Data Br.*, vol. 61, p. 111746, Aug. 2025, doi: 10.1016/j.dib.2025.111746.

[41] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *J. Big Data*, vol. 8, no. 1, p. 140, Oct. 2021, doi: 10.1186/s40537-021-00516-9.

[42] J. Meng and R. Xing, "Inside the 'black box': Embedding clinical knowledge in data-driven machine learning for heart disease diagnosis," *Cardiovasc. Digit. Heal. J.*, vol. 3, no. 6, pp. 276–288, Dec. 2022, doi: 10.1016/j.cvdhj.2022.10.005.

[43] S. Badawi, A. M. Saeed, S. A. Ahmed, P. A. Abdalla, and D. A. Hassan, "Kurdish News Dataset Headlines (KNDH) through multiclass classification," *Data Br.*, vol. 48, p. 109120, Jun. 2023, doi: 10.1016/j.dib.2023.109120.

[44] A. Ambarwari, Q. Jafar Adrian, and Y. Herdiyeni, "Analysis of the Effect of Data Scaling on the Performance of the Machine Learning Algorithm for Plant Identification," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 1, pp. 117–122, Feb. 2020, doi: 10.29207/resti.v4i1.1517.

[45] V. V. Starovoitov and Y. I. Golub, "Data normalization in machine learning," *Informatics*, vol. 18, no. 3, pp. 83–96, Sep. 2021, doi: 10.37661/1816-0301-2021-18-3-83-96.

[46] Z. Xue, J. Yang, R. Chen, Q. He, Q. Li, and X. Mei, "AR-Assisted Guidance for Assembly and Maintenance of Avionics Equipment," *Appl. Sci.*, vol. 14, no. 3, p. 1137, Jan. 2024, doi: 10.3390/app14031137.

[47] A. Wantoro, A. F. Yuliana, D. Y. A. Andini, I. Awaliyani, and W. Caesarendra, "Optimizing Type 2 Diabetes Classification with Feature Selection and Class Balancing in Machine Learning," *J. Tek. Inform.*, vol. 6, no. 4, pp. 2625–2637, Aug. 2025, doi: 10.52436/1.jutif.2025.6.4.5166.

[48] H. Hairani, T. Widiyaningtyas, and D. Dwi Prasetya, "Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies," *JOIV Int. J. Informatics Vis.*, vol. 8, no. 3, p. 1310, Sep. 2024, doi: 10.62527/joiv.8.3.2283.

[49] Y. Yang, H. A. Khorshidi, and U. Aickelin, "A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems," *Front. Digit. Heal.*, vol. 6, Jul. 2024, doi: 10.3389/fdgth.2024.1430245.

[50] I. Riadi, A. Yudhana, and G. C. Kurniawan, "Evaluating Synthetic Minority Oversampling Technique Strategies for Diabetes Mellitus Classification using K-Nearest Neighbors Algorithm," *J. Tek. Inform.*, vol. 6, no. 5, pp. 3958–3970, Oct. 2025, doi: 10.52436/1.jutif.2025.6.5.5189.

[51] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[52] H. B. Kibria, M. Nahiduzzaman, M. O. F. Goni, M. Ahsan, and J. Haider, "An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI," *Sensors*, vol. 22, no. 19, p. 7268, Sep. 2022, doi: 10.3390/s22197268.

[53] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.

[54] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[55] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[56] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. in Springer Series in Statistics. New York, NY: Springer New York, 2009. doi: 10.1007/978-0-387-84858-7.

[57] J. E. Black, J. K. Kueper, and T. S. Williamson, "An introduction to machine learning for classification and prediction," *Fam. Pract.*, vol. 40, no. 1, pp. 200–204, Feb. 2023, doi: 10.1093/fampra/cmac104.

[58] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

[59] D. Boldini, F. Grisoni, D. Kuhn, L. Friedrich, and S. A. Sieber, "Practical guidelines for the use of gradient boosting for molecular property prediction," *J. Cheminform.*, vol. 15, no. 1, p. 73, Aug. 2023, doi: 10.1186/s13321-023-00743-7.

[60] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *Ann. Stat.*, vol. 29, no. 5, Oct. 2001, doi: 10.1214/aos/1013203451.

[61] S. W. A. Sherazi, J.-W. Bae, and J. Y. Lee, "A soft voting ensemble classifier for early prediction and diagnosis of occurrences of major adverse cardiovascular events for STEMI and NSTEMI during 2-year follow-up in patients with acute coronary syndrome," *PLoS One*, vol. 16, no. 6, p. e0249338, Jun. 2021, doi: 10.1371/journal.pone.0249338.

[62] S. Ali *et al.*, "A Soft Voting Ensemble-Based Model for the Early Prediction of Idiopathic Pulmonary Fibrosis (IPF) Disease Severity in Lungs Disease Patients," *Life*, vol. 11, no. 10, p. 1092, Oct. 2021, doi: 10.3390/life11101092.

Corresponding author: Ermatita, ermatita@unsri.ac.id, Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.308

[63] *Pattern Recognition and Machine Learning*. Springer New York, 2006. doi: 10.1007/978-0-387-45528-0.

[64] D. Li, Z. Liu, D. J. Armaghani, P. Xiao, and J. Zhou, "Novel ensemble intelligence methodologies for rockburst assessment in complex and variable environments," *Sci. Rep.*, vol. 12, no. 1, p. 1844, Feb. 2022, doi: 10.1038/s41598-022-05594-0.

[65] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008. doi: 10.1017/CBO9780511809071.

[66] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Dec. 2020, doi: 10.1186/s12864-019-6413-7.

[67] I. Eegdeman, I. Cornelisz, M. Meeter, and C. van Klaveren, "Identifying false positives when targeting students at risk of dropping out," *Educ. Econ.*, vol. 31, no. 3, pp. 313–325, May 2023, doi: 10.1080/09645292.2022.2067131.

[68] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informatics*, vol. 17, no. 1, pp. 168–192, Jan. 2021, doi: 10.1016/j.aci.2018.08.003.

[69] G. Velarde *et al.*, "Tree boosting methods for balanced and imbalanced classification and their robustness over time in risk assessment," *Intell. Syst. with Appl.*, vol. 22, p. 200354, Jun. 2024, doi: 10.1016/j.iswa.2024.200354.

[70] X. Wang *et al.*, "Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, p. 105, Dec. 2021, doi: 10.1186/s12911-021-01471-4.

[71] T. Usuzaki, K. Takahashi, and R. Inamori, "Be Careful About Metrics When Imbalanced Data Is Used for a Deep Learning Model," *Chest*, vol. 165, no. 3, pp. e87–e89, Mar. 2024, doi: 10.1016/j.chest.2023.10.039.

[72] T. Zhu, X. Liu, and E. Zhu, "Oversampling with Reliably Expanding Minority Class Regions for Imbalanced Data Learning," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2022, doi: 10.1109/TKDE.2022.3171706.

[73] A. Demircioğlu, "Applying oversampling before cross-validation will lead to high bias in radiomics," *Sci. Rep.*, vol. 14, no. 1, p. 11563, May 2024, doi: 10.1038/s41598-024-62585-z.

[74] M. Owusu-Adjei, J. Ben Hayfron-Acquah, T. Frimpong, and G. Abdul-Salaam, "Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems," *PLOS Digit. Heal.*, vol. 2, no. 11, p. e0000290, Nov. 2023, doi: 10.1371/journal.pdig.0000290.

[75] J. Adeoye, L.-W. Zheng, P. Thomson, S.-W. Choi, and Y.-X. Su, "Explainable ensemble learning model improves identification of candidates for oral cancer screening," *Oral Oncol.*, vol. 136, p. 106278, Jan. 2023, doi: 10.1016/j.oraloncology.2022.106278.

[76] S. Friedrich and T. Friede, "On the role of benchmarking data sets and simulations in method comparison studies," *Biometrical J.*, vol. 66, no. 1, Jan. 2024, doi: 10.1002/bimj.202200212.

[77] T. Prexawanprasut and T. Banditwattanawong, "Improving Minority Class Recall through a Novel Cluster-Based Oversampling Technique," *Informatics*, vol. 11, no. 2, p. 35, May 2024, doi: 10.3390/informatics11020035.

[78] V. Singh, M. Pencina, A. J. Einstein, J. X. Liang, D. S. Berman, and P. Slomka, "Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging," *Sci. Rep.*, vol. 11, no. 1, p. 14490, Jul. 2021, doi: 10.1038/s41598-021-93651-5.

[79] M. Hasan, W. Wu, and X. Zhao, "SHAP-Driven Feature Analysis Approach for Epileptic Seizure Prediction," *J. Med. Syst.*, vol. 49, no. 1, p. 77, Jun. 2025, doi: 10.1007/s10916-025-02211-1.

[80] Y. Jang, "Feature-based ensemble modeling for addressing diabetes data imbalance using the SMOTE, RUS, and random forest methods: a prediction study," *Ewha Med. J.*, vol. 48, no. 2, p. e32, Apr. 2025, doi: 10.12771/emj.2025.00353.

[81] S. M. Ganie, P. K. D. Pramanik, M. Bashir Malik, S. Mallik, and H. Qin, "An ensemble learning approach for diabetes prediction using boosting techniques," *Front. Genet.*, vol. 14, Oct. 2023, doi: 10.3389/fgene.2023.1252159.

[82] T. F. Sukamto, C. L. Prameswary, D. Royadi, and D. Sofia, "Diabetes Disease Prediction on Unbalanced Data Using SMOTE-Tomek Links and Random Forest Algorithm," *G-Tech J. Teknol. Terap.*, vol. 9, no. 3, pp. 1194–1203, Jul. 2025, doi: 10.70609/g-tech.v9i3.7164.

[83] M. T. García-Ordás, C. Benavides, J. A. Benítez-Andrades, H. Alaiz-Moretón, and I. García-Rodríguez, "Diabetes detection using deep learning techniques with oversampling and feature augmentation," Feb. 2024, doi: 10.1016/j.cmpb.2021.105968.

AUTHOR BIOGRAPHY

Siti Fatimah Nurdiah Permatasari is a Master's student in Computer Science at the Faculty of Computer Science, Universitas Sriwijaya, Indonesia. She received her bachelor's degree from STMIK GI Multi Data Palembang (now Universitas MDP), Indonesia. She is also a civil servant at the Ministry of Health of the Republic of Indonesia and works at RS Mohammad Hoesin. Her current research focuses on machine learning for healthcare, particularly on the classification of Type 2 Diabetes Mellitus from imbalanced clinical tabular data

Corresponding author: Ermatita, ermatita@unsri.ac.id, Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.308

using a two-stage hybrid oversampling approach (RandomOverSampler–Borderline-SMOTE) and soft-voting ensemble learning. Her research interests include medical data analytics, imbalanced learning, oversampling methods, ensemble learning, and clinical decision support systems.

Ermatita received a bachelor's degree in mathematics from the University of Wollongong in 2015. Ermatita received a master's degree in computer science from Universitas Indonesia and a doctoral degree in Computer Science from Universitas Gadjah Mada. She is currently working in the Department of Computer Science at the Faculty of Computer Science, Sriwijaya University, Indonesia. Her research includes artificial intelligence, data mining, machine learning, and information systems. Her most-cited research articles relate to electric methods for group decision support systems in bioinformatics, particularly for gene mutation detection simulation. She can be contacted at the email: ermatita@unsri.ac.id

Corresponding author: Ermatita, ermatita@unsri.ac.id, Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v8i2.308