Advancing Early-Stage Diabetes Prediction: A Comparative Study of Data Oversampling and Classification Methods

Michelle Melody d'Viola[1] and Raymond Sunardi Oetama[2]

[1,2,3]Universitas Multimedia Nusantara, Tangerang, Indonesia

raymond@umn.ac.id

Abstract. Diabetes is a common and increasing chronic disease worldwide with a large number of undiagnosed cases. An optimal early-stage prediction model is needed to support early prevention. This study has examined the combination of oversampling techniques with classifiers to determine the optimal models for early-stage diabetes prediction. Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) methods are used to overcome class imbalance. Random Forest (RF), Extreme Gradient Boosting (XGB), Multilayer Perceptron (MLP), and hybrid models combining two of these classifiers were evaluated to determine the most effective model. The combinations of the Synthetic Minority Over-sampling Technique with Random Forest (SMOTE + RF), Adaptive Synthetic Sampling with Random Forest (ADASYN + RF), as well as Adaptive Synthetic Sampling with Extreme Gradient Boosting and Multilayer Perceptron (ADASYN + XGB-MLP) demonstrated the highest accuracy at 99.04% with an F1-score of 0.99. Combining oversampling techniques with these classifiers enhances the accuracy of prediction while addressing class imbalance, and thus a useful tool for early diabetes detection.

Keywords: Classification, Early Stage Diabetes, Oversampling

1 Introduction

Diabetes is a global non-communicable chronic disease defined by high glucose levels resulting from inadequate insulin production or resistance. It could cause life-threatening complications, including cardiovascular disease, nerve damage, kidney damage, and alzheimer [1]. More than 500 million individuals worldwide are affected by diabetes, including all age groups and genders. In the next three decades, the number are projected to double to 1.3 billion. Diabetes accounts for one of the top ten leading causes of death and disability worldwide, with an estimated global prevalence of 6.1%. One of the main issues is the high proportion of undiagnosed cases. About 44.7% of diabetes patients are unaware of

their condition [2]. It can increase the risk of complications as symptoms go unnoticed and untreated.

One effective solution to improve early diabetes detection is the application of classification methods, which analyze patient history data to identify symptom patterns that indicate diabetes. However, class imbalance in diabetes data can result in a skewed model, as the majority class dominates predictions and reduces sensitivity to the minority class. Previous research has applied the Synthetic Minority Over-sampling Technique to adjust the class distribution of a dataset containing heart disease symptoms with an imbalance ratio of 60:40 [3]. The Synthetic Minority Over-sampling Technique has also been applied on the diabetes data and it showed an improvement in C5.0, Random Forest, and Support Vector Machine classifiers performance [4]. The performance of several classifiers (Support Vector Machine, Logistic Regression, and Naive Bayes) is evaluated through a comparative approach of the Synthetic Minority Over-sampling Technique and Adaptive Synthetic Sampling to show the superiority of the over-sampling techniques in relation to the class imbalance problem. The highest accuracy of 95.8% was achieved by the combination of Synthetic Minority Over-sampling Technique and Support Vector Machine [5].

Several studies have analyzed different classification methods through various algorithms to find the optimal classifier for predicting diabetes. A study compared Logistic Regression, Support Vector Machine (linear and non-linear kernels), Decision Tree, Adaptive Boosting Classifier, K-Nearest Neighbor, Random Forest, and Naive Bayes, with Random Forest attained the highest accuracy of 98% [6]. Another study found Extreme Gradient Boosting outperformed Support Vector Machine, Random Forest, and k-nearest Neighbor, with an accuracy rate of 89.09% [7]. A further comparative study found that Multilayer Perceptron outperformed Support Vector Machine, achieving 77.47% accuracy [8]. This research will utilize the most optimal classifiers based on these three studies.

Beyond single classifiers approaches, some studies utilize a hybrid approach that combines multiple classifiers, to enhance accuracy, handle complex data, and offset individual weaknesses. This approach has been used and shown to perform optimally in various cases, including improving water quality index prediction [9], handling imbalanced medical data [10], predicting internal corrosion levels in multiphase pipelines [11], estimating flood vulnerability [12], lung cancer prediction [13], object tracking in videos [14], medium-term forecasting of crude oil pipeline electricity consumption [15], and heart disease prediction [16]. Research on early-stage diabetes prediction has also been conducted by combining multiple classifiers, such as Multiple Linear Regression, Random Forest, and Extreme Gradient Boosting [17], Common Scrambling Algorithm and Feedforward

Neural Network [18], stacking methods with k-Nearest Neighbors, Logistic Regression, and Random Forest [19], as well as Genetic Algorithm and Stacking with Decision Tree, Convolutional Neural Network, and Support Vector Machine [20].

This study examines the combination of oversampling techniques with classifiers in predicting early-stage diabetes, which has not been used in previous studies examining the same data. The oversampling techniques used are Synthetic Minority Over-sampling Technique and Adaptive Synthetic Sampling. Due to their strong performance in previous studies, the classification methods to be evaluated include Random Forest, Extreme Gradient Boosting, and Multilayer Perceptron. The hybrid approach will also be explored by combining these single classifiers. The optimal oversampling classification approaches will be determined based on accuracy, precision, recall, and F1-score. The results of this study can be used as a reference for the use of oversampling techniques in handling imbalanced early-stage diabetes data.

2 Research Method

This section explains the dataset used, the preprocessing applied, and the methods used in the experiment. The process begins with selecting and determining the dataset to be used. The data is then preprocessed to prepare it for the main procedure. Two dataset scenarios are created using different oversampling techniques. Each oversampled data is used for modeling with various classifiers. Performance evaluation is conducted to determine the most optimal model. This experimental flow is illustrated in Fig. 1.

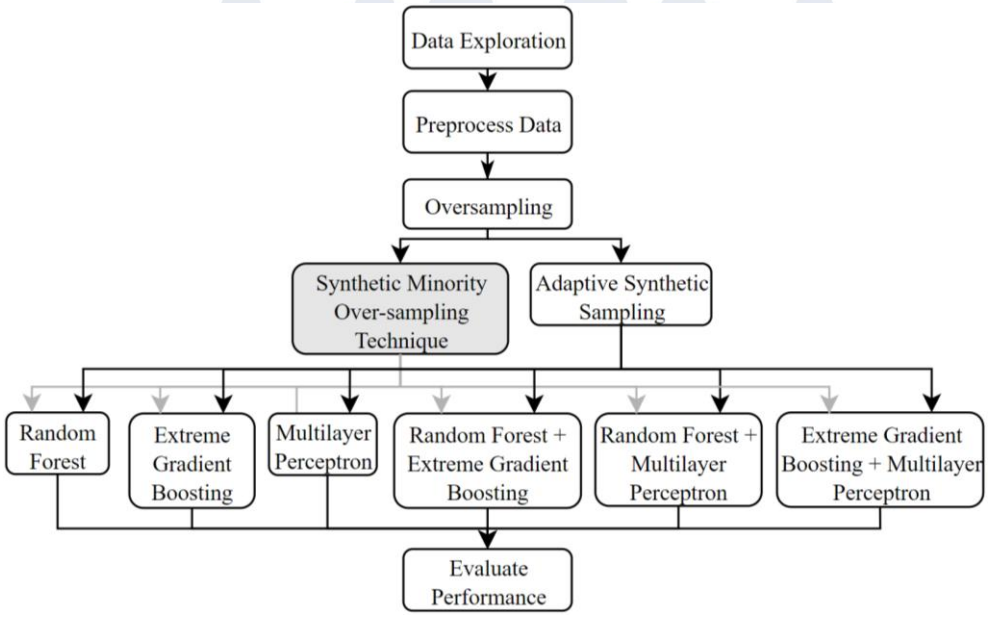

Fig. 1. *Experimental Flow*

1.1 Dataset

This research uses the Early Stage Diabetes Risk Prediction dataset, sourced from the UCI Machine Learning Repository [21]. A questionnaire was used to gather data from patients at Sylhet Diabetes Hospital located in Bangladesh, with approval from the doctors. The dataset is composed of 520 rows and 17 columns. The attributes represent early-stage diabetes symptoms, with one as the class label, including 320 positive and 200 negative cases. The sample of one row of data used is as shown in Table I. A description of the dataset attributes is provided in Table II.

Table 1. *Dataset Sample - First Row (Wrapped Across 3 lines for Readability)*

Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness
40	Male	No	Yes	No	Yes

Polyphagia	Genital thrush	visual blurring	Itching	Irritability
No	No	No	Yes	No

delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
Yes	No	Yes	Yes	Yes	Positive

Table 2. *Dataset Description*

No	Attributes Name	Description	Value
1	Age	-	20 – 65
2	Gender	-	Male/Female
3	Polyuria	Excessive urination	Yes/No
4	Polydipsia	Excessive thirst	Yes/No
5	sudden weight loss	-	Yes/No
6	weakness	-	Yes/No
7	Polyphagia	Excessive hunger	Yes/No
8	Genital thrush	-	Yes/No
9	visual blurring	-	Yes/No
10	Itching	-	Yes/No
11	Irritability	-	Yes/No
12	delayed healing	-	Yes/No
13	partial paresis	muscle weakness	Yes/No
14	muscle stiffness	-	Yes/No
15	Alopecia	hair loss or thinning	Yes/No
16	Obesity	Excess body fat	Yes/No
17	class	-	Positive/Negative

1.2 Data Exploration

The exploration stage is carried out to obtain an initial picture of the data characteristics. The exploratory data functions are applied to analyzes data type, quantity, and missing values. Class distribution analysis revealed class imbalance including 320 people with diabetes and 200 non-diabetic individuals, with the visualization shown in Fig. 2. The distribution of numerical data is shown in Fig. 3 which displays a normal distribution, where the majority of individuals are in the age range of 35 to 60 years, with the peak distribution being around the age of 40 to 50 years. This distribution tends to resemble a normal distribution.

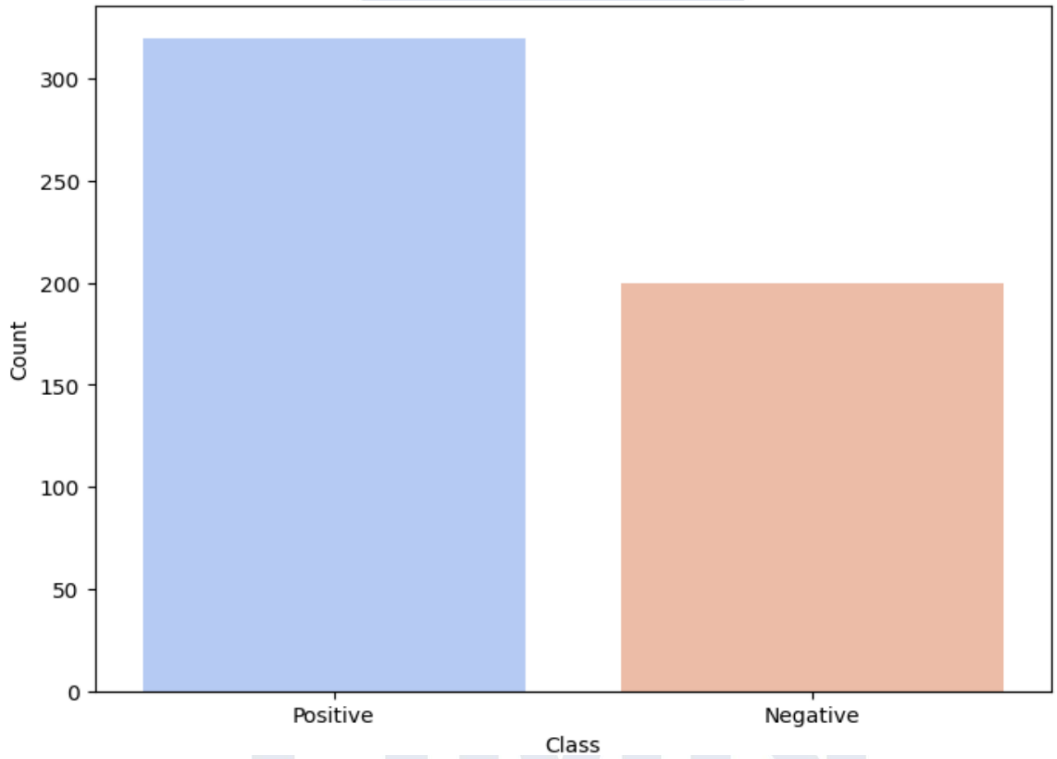

Fig. 2. *Class Distribution*

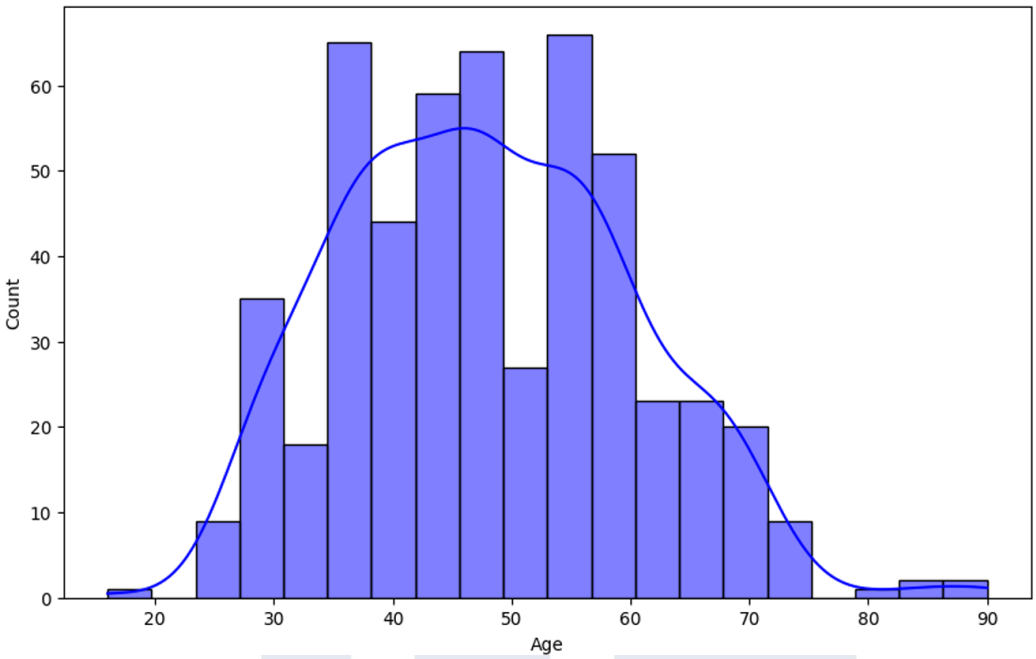

Fig. 3. *Age Distribution*

1.3 Preprocessing

Before performing the main processing on the data, preprocessing is carried out to prepare the it for modeling. Encoding is applied to convert categorical data into numerical form so the classifiers can process it. Numerical data is normalized using MinMaxScaler provided by scikit-learn's preprocessing module, to ensure a consistent range for all values (0-1) to improve the stability and performance of the model. The final step in preprocessing for this experiment involves splitting the data into training and testing subsets with the train_test_split function provided by the scikit-learn library, with an 80:20 ratio, where each partition is employed to train and evaluate the model, respectively. After going through the oversampling process using the Synthetic Minority Over-sampling Technique, the number of each class becomes 256 rows, which makes the total training data 512 rows and 616 if added with the test data. While the number of data using the Adaptive Synthetic Sampling method is 257, 514, and 618 respectively.

1.4 Methods

This experiment performed data oversampling using Synthetic Minority Over-sampling Technique and Adaptive Synthetic Sampling to tackle data imbalance. The oversampled data was then paired with the Random Forest, Extreme Gradient Boosting, and Multilayer Perceptron classifiers, individually and in hybrid combinations, resulting in 12 models. Performance evaluation was conducted to determine the optimal model.

1.4.1 Synthetic Minority Over-sampling Technique

This technique is an oversampling method that uses interpolation to generate new samples between existing minority class data points. Minority class samples are randomly selected, and their nearest neighbors identified using k-Nearest Neighbors. Synthetic samples are generated by linear interpolation between each sample and a neighbor. This process results in a more balanced dataset, minimizing bias toward the dominant class and enhancing model learning [22].

1.4.2 Adaptive Synthetic Sampling

Adaptive Synthetic Sampling is a technique for generating synthetic samples that build upon Synthetic Minority Over-sampling Technique. This method targets hard-to-classify samples, generating synthetic data around minority instances in low-density areas based on the data distribution [23].

1.4.3 Random Forest

Random Forest is an algorithm for classification that leverages an ensemble of decision trees from data subsets, each producing an individual prediction known as a weak classifier. A decision tree constructs a binary tree structure for classification, and Random Forest determines the final predictions based on the majority vote to improve accuracy and stability. The Gini Index metric in a decision tree measures the purity of nodes within the tree, facilitating the grouping of data into more similar subsets aligned with the target class [24].

1.4.4 Extreme Gradient Boosting

This algorithm is a classification algorithm that optimizes the Gradient Boosting Decision. This classifier operates by constructing decision trees gradually and iteratively to correct the prediction errors of the prior iteration. The final prediction is updated by combining the predictions from all previous trees with the latest tree [7].

1.4.5 Multilayer Perceptron

Multilayer Perceptron is a deep learning algorithm derived from the Feedforward Neural Network, an artificial neural network architecture inspired by the human brain and consists of several hidden layers in between input and output layers. The input layer transmits signals hidden layers, where weighted inputs are processed through activation functions to extract features. The final output is produced by the output layer [25].

1.4.6 Hybrid Model

A hybrid algorithm employs an approach that combines two or more learning methods and utilizes the strengths of each to produce the model's prediction [26]. This paper will evaluate hybrid models created by combining two classifiers from Random Forest, Extreme Gradient Boosting, and Multilayer Perceptron. Table III displays the specific role of every classifier in the hybrid setups.

Table 3. *Classifier Contribution in Hybrid Models*

No	Hybrid Model	Classifier	Task
1	Random Forest + Multilayer Perceptron	Random Forest	Feature Selection
		Multilayer Perceptron	Classifier
2	Random Forest + Extreme Gradient Boosting	Random Forest	Feature Selection
		Extreme Gradient Boosting	Classifier
3	Extreme Gradient Boosting + Multilayer Perceptron	Extreme Gradient Boosting	Feature Selection
		Multilayer Perceptron	Classifier

1.4.7 Hyperparameter Tuning and Feature Selection

This experiment does not aim to find the best hyperparameters, but all models underwent tuning via GridSearchCV with 5-fold cross-validation for fair comparison, using commonly recommended values from past studies. Table IV summarizes the parameters consistently applied to single and hybrid classifiers to prevent bias in model performances. Feature selections were treated equally, using a threshold between 0.01 and 0.1.

Table 4. *Hyperparameter Tuning*

Hyperparameters	Random Forest	Extreme Gradient Boosting	Multilayer Perceptron
n_estimators	50, 100, 200	50, 100, 200	N/A
max_depth	10, 20, None	3, 6, 9	N/A
alpha	N/A	N/A	0.0001, 0.001, 0.01
hidden_layer_sizes	N/A	N/A	(50,), (100,), (50, 50)

1.5 Evaluation

This study employs four evaluation metrics calculated from the counts of: correctly identified positive cases (True Positive, TP); negative cases incorrectly

identified as positive (False Positive, FP); positive cases incorrectly identified as negative (False Negative, FN); and correctly identified negative cases (True Negative, TN) [27].

1.5.1 Accuracy

This metric measures the model's accuracy by dividing correct predictions by the total predictions. The formula applied to compute accuracy is given by:

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

1.5.2 Precision

High precision means the model's positive predictions are usually correct. This is important in medical data to avoid wrongly labeling healthy individuals as diseased [27]. The formula applied to compute precision is given by:

$$Presisi = \frac{TP}{TP+FP} \quad (2)$$

1.5.3 Recall

High recall in medical data means the model correctly identifies most diseased individuals. In medical data, balancing precision and recall helps avoid misclassifying both healthy and diseased individuals [28]. The formula applied to compute recall is given by:

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

1.5.4 F1-Score

This metric indicates whether the model performs well in both aspects. The formula applied to compute F1-score is given by:

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

3 Result and Analysis

This section analyzes the experimental results obtained by combining oversampling techniques with classifiers to predict diabetes based on early symptoms. The analysis aims to identify the most optimal model combination.

The comparison between the initial data distribution and the oversampled data using Synthetic Minority Over-sampling Technique and Adaptive Synthetic Sampling is shown in Fig. 4. The plot compares the numerical age attribute within the dataset. The dataset enhanced through the application of the Synthetic Minority

Over-sampling Technique exhibits a distribution that closely follows the original data indicated by the curve shape resembling the original pattern, as this technique generates synthetic data through interpolation between values of the minority class. The distribution of data oversampled using Adaptive Synthetic Sampling shows more fluctuations and deviations as it produces synthetic samples by considering the distribution density, placing greater emphasis on underrepresented instances that are more difficult for the model to capture. The peak at 0.2 to 0.3 indicates that this technique generates more synthetic samples for ages 20–30 to compensate for fewer younger individuals, ensuring sufficient data for each age group to ensure balanced data and predictive accuracy.

The distribution of target class values in the original data and after oversampling using both oversampling methods is shown in Fig. 5. The original data is imbalanced, with individuals diagnosed with early-stage diabetes making up 62% of the dataset and negative cases account for 38%. Both oversampling methods add synthetic samples to balance the dataset, which ensures that each class is represented equally at 50%. This approach helps the model avoid overfitting to the majority group and minimizes bias toward the more frequent class.

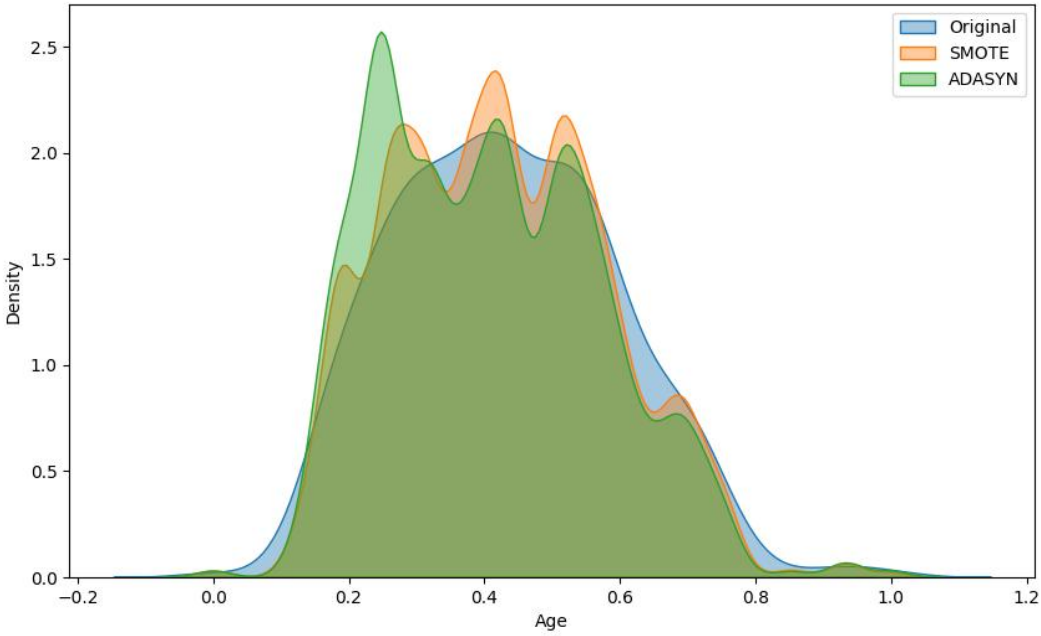

Fig. 4. *Data Distribution for Numerical Attribute*

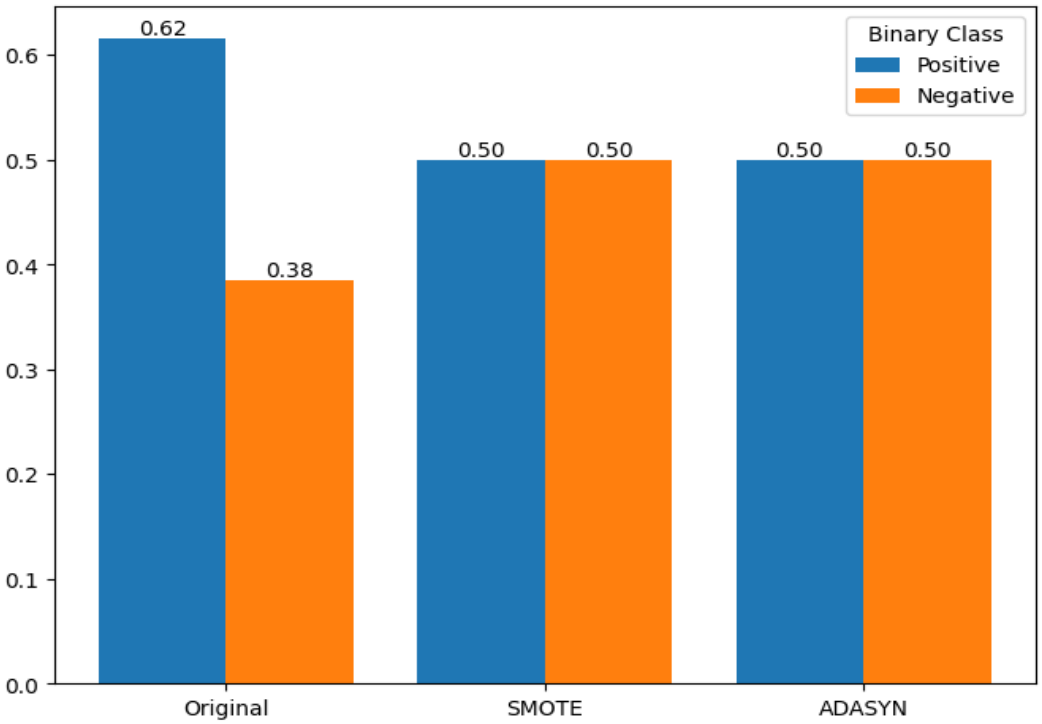

Fig. 5. *Data Distribution for Binary Class*

Table 5. *Performance Evaluation of Classifiers with Different Oversampling Techniques*

Model		**Precision**			**Recall**			**F1 – Score**			**Accuracy**
Oversampling	***Classifier***	***Neg***	***Pos***	***Avg***	***Neg***	***Pos***	***Avg***	***Neg***	***Pos***	***Avg***	
SMOTE	**Random Forest**	**0.98**	**1.00**	**0.99**	**1.00**	**0.98**	**0.99**	**0.99**	**0.99**	**0.99**	**99.04%**
	Extreme Gradient Boosting	0.95	1.00	0.98	1.00	0.97	0.98	0.98	0.98	0.98	98.08%
	Multilayer Perceptron	0.95	0.97	0.96	0.95	0.97	0.96	0.95	0.97	0.96	96.15%
	Random Forest + Extreme Gradient Boosting	0.93	0.97	0.95	0.95	0.95	0.95	0.94	0.96	0.95	95.19%
	Random Forest + Multilayer Perceptron	0.97	0.98	0.98	0.97	0.98	0.98	0.97	0.98	0.98	98.08%
	Extreme Gradient Boosting + Multilayer Perceptron	0.93	0.98	0.96	0.97	0.95	0.96	0.95	0.97	0.96	96.15%
ADASYN	**Random Forest**	**0.98**	**1.00**	**0.99**	**1.00**	**0.98**	**0.99**	**0.99**	**0.99**	**0.99**	**99.04%**
	Extreme Gradient Boosting	0.95	1.00	0.98	1.00	0.97	0.98	0.98	0.98	0.98	98.08%
	Multilayer Perceptron	0.91	0.98	0.95	0.97	0.94	0.95	0.94	0.96	0.95	95.19%
	Random Forest + Extreme Gradient Boosting	0.95	0.98	0.97	0.97	0.97	0.97	0.96	0.98	0.97	97.12%
	Random Forest + Multilayer Perceptron	0.97	0.98	0.98	0.97	0.98	0.98	0.97	0.98	0.98	98.08%
	Extreme Gradient Boosting + Multilayer Perceptron	**0.98**	**1.00**	**0.99**	**1.00**	**0.98**	**0.99**	**0.99**	**0.99**	**0.99**	**99.04%**

Table 6. *Hyperparameter Settings and Selected Features*

Model 1: SMOTE + Random Forest			
Hyperparameters	***Random Forest***	***Extreme Gradient Boosting***	***Multilayer Perceptron***
n_estimators	50	N/A	N/A
max_depth	10	N/A	N/A
alpha	N/A	N/A	N/A
hidden_layer_sizes	N/A	N/A	N/A
Model 2: ADASYN + Random Forest			
Hyperparameters	***Random Forest***	***Extreme Gradient Boosting***	***Multilayer Perceptron***
n_estimators	50	N/A	N/A
max_depth	10	N/A	N/A
alpha	N/A	N/A	N/A
hidden_layer_sizes	N/A	N/A	N/A
Model 3: ADASYN + Extreme Gradient Boosting + Multilayer Perceptron			
Selected features	Age, Gender, Polyuria, Polydipsia, sudden weight loss, Genital thrush, visual blurring, Irritability, muscle stiffness, Alopecia, Obesity		
Hyperparameters	***Random Forest***	***Extreme Gradient Boosting***	***Multilayer Perceptron***
n_estimators	N/A	50	N/A
max_depth	N/A	9	N/A
alpha	N/A	N/A	0.0001
hidden_layer_sizes	N/A	N/A	(50,)

Table 7. *Optimal Models with and Without Oversampling*

Model	Accuracy Without Oversampling	Accuracy With Oversampling
Model 1	97.12%	99.04%
Model 2	97.12%	99.04%
Model 3	97.12%	99.04%

Table V presents how each tested model performed in terms of classification accuracy. According to the findings, the optimal models include combinations of the Synthetic Minority Over-sampling Technique with Random Forest, Adaptive Synthetic Sampling with Random Forest, as well as Adaptive Synthetic Sampling

with Extreme Gradient Boosting and Multilayer Perceptron, each achieving an accuracy of 99.04%. The high F1-score demonstrates that the model successfully balances precision against recall, which is crucial to ensure the model does not misclassify individuals with and without diabetes.

Table VI presents the optimal hyperparameters for each proposed model. In the hybrid model (model 3), 11 of 16 features were selected using a 0.1 threshold. The selected features, including Age [29], Gender [29], Polyuria [30], Polydipsia [30], sudden weight loss [30], Genital thrush [31], visual blurring [30], Irritability [28], muscle stiffness [32], Alopecia [33], and Obesity [30], have been validated in previous studies as risk factors or linked to high blood sugar.

Table VII presents how the performance accuracy differs between the best-performing models that apply oversampling methods and those that do not. All three models achieved higher accuracy when the majority and minority classes were balanced. It indicates that the oversampling technique plays a role in enhancing accuracy and creating a more optimal model by balancing the class distribution in the dataset. By adding minority class samples through data augmentation, the model gains a better representation of the minority class, enhancing generalization and reducing prediction bias [3].

Table 8. *Comparison of Previous Studies and Proposed Models on the Same Dataset*

Model	Accuracy
Previous Studies	
K-Nearest Neighbor + Logistic Regression + Random Forest [20]	99.60%
k-Nearest Neighbor + Gradient Boosting Machine + Light Gradient Boosting Machine [35]	99.23%
Multilayer Regression + Random Forest + Extreme Gradient Boosting [18]	99.20%
Crow Search Algorithm + Feed-Forward Neural Network [19]	99.04%
Random Forest [7]	98%
Random Forest [36]	97%
Decision Tree, Convolutional Neural Network, and Support Vector Machine [21]	85.88%
This Study	
SMOTE and Random Forest	**99.04%**
ADASYN and Random Forest	**99.04%**
ADASYN, Extreme Gradient Boosting, and Multilayer Perceptron	**99.04%**

Furthermore, table VIII compares the accuracy of previous studies that used the same dataset but implemented different models. Some earlier studies achieved accuracy that was either lower or similar to this study. However, three studies

produced models with higher accuracy, reaching up to 99.60%. Several algorithms were not used in this study but were applied in those three studies, including K-Nearest Neighbor, Logistic Regression, variations of Gradient Boosting Machine, and Multilayer Regression. Although those studies achieved higher accuracy, there is no evidence that oversampling techniques were used to address data imbalance. Therefore, this study should be further developed by incorporating those algorithms.

4 Discussion

The use of oversampling techniques in classification is currently still a debate involving both critics and supporters. Some critics argue that synthetic data may occasionally deviate from the original minority class distribution and fail to accurately represent the true data pattern. A study concluded that this method may create synthetic data that is not accurately representative, which may result in model overfitting and reduced reliability of outcomes [34].

Despite this criticism, many studies show oversampling benefits classification of medical data. A study on classification of diabetes showed that oversampling methods provide more information to the classifier so that the model can learn both classes equally, enhancing accuracy [35]. In general, studies have shown that oversampling effectively handles imbalanced medical data by creating synthetic samples similar to existing ones, which helps diversify minority classes and improve model performance [36].

5 Conclusion and Future Work

This study successfully examines combinations of oversampling and classifiers to predict early-stage diabetes symptoms. SMOTE with Random Forest, ADASYN with Random Forest, and ADASYN with Extreme Gradient Boosting and Multilayer Perceptron performed optimally, with all three achieving accuracy, precision, recall, and F1-score of 99.04%, 0.99, 0.99, and 0.99.

This study strongly recommends the use of oversampling techniques when data imbalance is present. While this study focuses on two oversampling methods (SMOTE and ADASYN) as well as three classification algorithms (Random Forest, Extreme Gradient Boosting, and Multilayer Perceptron), further research can be done by incorporating K-Nearest Neighbor, Logistic Regression, Gradient Boosting Machine variations, Multilayer Regression, and oversampling techniques such as SMOTE, ADASYN, and several others for further development.

Acknowledgements

The authors would like to express their gratitude to the Institution of Research and Community Services at Universitas Multimedia Nusantara for the support and assistance provided throughout the completion of this work.

References

[1] "Diabetes: Pengertian, Gejala, Penyebab, dan Pengobatan." Accessed: May 21, 2025. [Online]. Available: https://hellosehat.com/diabetes/diabetes-melitus/

[2] K. Ogurtsova *et al.*, "IDF diabetes Atlas: Global estimates of undiagnosed diabetes in adults for 2021," *Diabetes Res Clin Pract*, vol. 183, Jan. 2022, doi: 10.1016/J.DIABRES.2021.109118,.

[3] M. C. Untoro, L. Rizta, A. Perdana, N. A. Wijaya, and N. Ferdiyanto, "Penerapan K-Means Clustering pada Imbalance Dataset Gejala Penyakit Jantung," *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, vol. 5, no. 1, pp. 1–7, Apr. 2023, doi: 10.28926/ILKOMNIKA.V5I1.455.

[4] M. K. Rezki, M. I. Mazdadi, F. Indriani, Muliadi, T. H. Saragih, and V. A. Athavale, "Application Of SMOTE To Address Class Imbalance In Diabetes Disease Classification Utilizing C5.0, Random Forest, And SVM," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 4, pp. 343–354, Aug. 2024, doi: 10.35882/JEEEMI.V6I4.434.

[5] J. Resti *et al.*, "Improving Diabetes Prediction Accuracy in Indonesia: A Comparative Analysis of SVM, Logistic Regression, and Naive Bayes with SMOTE and ADASYN," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 5, pp. 607–614, Oct. 2024, doi: 10.29207/RESTI.V8I5.5980.

[6] M. Rady, K. Moussa, M. Mostafa, A. Elbasry, Z. Ezzat, and W. Medhat, "Diabetes Prediction Using Machine Learning: A Comparative Study," *NILES 2021 - 3rd Novel Intelligent and Leading Emerging Sciences Conference, Proceedings*, pp. 279–282, 2021, doi: 10.1109/NILES53778.2021.9600091.

[7] L. Wang, X. Wang, A. Chen, X. Jin, and H. Che, "Prediction of type 2 diabetes risk and its effect evaluation based on the xgboost model," *Healthcare (Switzerland)*, vol. 8, no. 3, 2020, doi: 10.3390/HEALTHCARE8030247,.

[8] K. Thaiyalnayaki, "Classification of diabetes using deep learning and svm techniques," *Int J Curr Res Rev*, vol. 13, no. 1, pp. 146–149, Jan. 2021, doi: 10.31782/IJCRR.2021.13127.

[9] D. T. Bui, K. Khosravi, J. Tiefenbacher, H. Nguyen, and N. Kazakis, "Improving prediction of water quality indices using novel hybrid machine-learning algorithms," *Science of the Total Environment*, vol. 721, Jun. 2020, doi: 10.1016/j.scitotenv.2020.137612.

[10] 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, p. 103465, Jul. 2020, doi: 10.1016/J.JBI.2020.103465.

[11] S. Peng, Z. Zhang, E. Liu, W. Liu, and W. Qiao, "A new hybrid algorithm model for prediction of internal corrosion rate of multiphase pipeline," *J Nat Gas Sci Eng*, vol. 85, Jan. 2021, doi: 10.1016/J.JNGSE.2020.103716.

[12] E. Dodangeh *et al.*, "Novel hybrid intelligence models for flood-susceptibility prediction: Meta optimization of the GMDH and SVR models with the genetic algorithm and harmony search," *J Hydrol (Amst)*, vol. 590, p. 125423, Nov. 2020, doi: 10.1016/J.JHYDROL.2020.125423.

[13] P. Nanglia, S. Kumar, A. N. Mahajan, P. Singh, and D. Rathee, "A hybrid algorithm for lung cancer classification using SVM and Neural Networks," *ICT Express*, vol. 7, no. 3, pp. 335–341, Sep. 2021, doi: 10.1016/J.ICTE.2020.06.007.

[14] S. S. Moghaddasi and N. Faraji, "A hybrid algorithm based on particle filter and genetic algorithm for target tracking," *Expert Syst Appl*, vol. 147, Jun. 2020, doi: 10.1016/J.ESWA.2020.113188.

[15] L. Xu *et al.*, "Mid-term prediction of electrical energy consumption for crude oil pipelines using a hybrid algorithm of support vector machine and genetic algorithm," *Energy*, vol. 222, p. 119955, May 2021, doi: 10.1016/J.ENERGY.2021.119955.

[16] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," *Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021*, pp. 1329–1333, Jan. 2021, doi: 10.1109/ICICT50816.2021.9358597.

[17] S. Gündoğdu, "Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique," *Multimed Tools Appl*, vol. 82, no. 22, pp. 34163–34181, Sep. 2023, doi: 10.1007/S11042-023-15165-8/METRICS.

[18] A. Yasar, "Data Classification of Early-Stage Diabetes Risk Prediction Datasets and Analysis of Algorithm Performance Using Feature Extraction Methods and Machine Learning Techniques," *International Journal of*

Intelligent Systems and Applications in Engineering, vol. 9, no. 4, pp. 273–281, Dec. 2021, doi: 10.18201/ijisae.2021473767.

[19] I. Cinar, Y. S. Taspinar, and M. Koklu, "Development of Early Stage Diabetes Prediction Model Based on Stacking Approach," *Tehnicki Glasnik*, vol. 17, no. 2, pp. 153–159, 2023, doi: 10.31803/TG-20211119133806.

[20] Y. Tan, H. Chen, J. Zhang, R. Tang, and P. Liu, "Early Risk Prediction of Diabetes Based on GA-Stacking," *Applied Sciences 2022, Vol. 12, Page 632*, vol. 12, no. 2, p. 632, Jan. 2022, doi: 10.3390/APP12020632.

[21] "Early Stage Diabetes Risk Prediction - UCI Machine Learning Repository." Accessed: May 21, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/529/early%2Bstage%2Bdiabetes%2Brisk%2Bprediction%2Bdataset

[22] W. Rahayu *et al.*, "Synthetic Minority Oversampling Technique (SMOTE) for Boosting the Accuracy of C4.5 Algorithm Model," *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, vol. 3, no. 3, pp. 624–630, Jun. 2024, doi: 10.59934/JAIEA.V3I3.469.

[23] G. Ahmed *et al.*, "DAD-Net: Classification of Alzheimer's Disease Using ADASYN Oversampling Technique and Optimized Neural Network," *Molecules 2022, Vol. 27, Page 7085*, vol. 27, no. 20, p. 7085, Oct. 2022, doi: 10.3390/MOLECULES27207085.

[24] A. S. Saud, S. Shakya, and B. Neupane, "Analysis of Depth of Entropy and GINI Index Based Decision Trees for Predicting Diabetes," *Indian Journal of Computer Science*, vol. 6, no. 6, p. 19, Jan. 2021, doi: 10.17010/IJCS/2021/V6/I6/167641.

[25] S. Abirami and P. Chitra, "Energy-efficient edge based real-time healthcare support system," *Advances in Computers*, vol. 117, no. 1, pp. 339–368, Jan. 2020, doi: 10.1016/BS.ADCOM.2019.09.007.

[26] M. N. Ab Wahab, A. Nazir, A. T. Z. Ren, M. H. M. Noor, M. F. Akbar, and A. S. A. Mohamed, "Efficientnet-Lite and Hybrid CNN-KNN Implementation for Facial Expression Recognition on Raspberry Pi," *IEEE Access*, vol. 9, pp. 134065–134080, 2021, doi: 10.1109/ACCESS.2021.3113337.

[27] Y. Qian, G. Zeng, Y. Pan, Y. Liu, L. Zhang, and K. Li, "A Prediction Model for High Risk of Positive RT-PCR Test Results in COVID-19 Patients Discharged From Wuhan Leishenshan Hospital, China," *Front Public Health*, vol. 9, Nov. 2021, doi: 10.3389/FPUBH.2021.778539,.

[28] A. S. Wahyuningsih, R. A. Agastya, and S. Sutrisno, "Analisis Hubungan Mood Swings terhadap Kadar Gula Darah Sewaktu pada Penderita Diabetes

Mellitus," *Jurnal Keperawatan Jiwa*, vol. 11, no. 2, p. 429, May 2023, doi: 10.26714/JKJ.11.2.2023.429-436.

[29] "Risk Factors for Type 2 Diabetes - NIDDK." Accessed: May 21, 2025. [Online]. Available: https://www.niddk.nih.gov/health-information/diabetes/overview/risk-factors-type-2-diabetes

[30] "Diabetes." Accessed: May 21, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[31] J. Talapko, T. Meštrović, and I. Škrlec, "Growing importance of urogenital candidiasis in individuals with diabetes: A narrative review," *World J Diabetes*, vol. 13, no. 10, p. 809, Oct. 2022, doi: 10.4239/WJD.V13.I10.809.

[32] J. H. Choi, H. R. Kim, and K. H. Song, "Musculoskeletal complications in patients with diabetes mellitus," *Korean Journal of Internal Medicine*, vol. 37, no. 6, pp. 1099–1110, Nov. 2022, doi: 10.3904/KJIM.2022.168,.

[33] S. Ali, M. Collins, S. C. Taylor, K. Kelley, E. Stratton, and M. Senna, "Type 2 diabetes mellitus and central centrifugal cicatricial alopecia severity," *J Am Acad Dermatol*, vol. 87, no. 6, pp. 1418–1419, Dec. 2022, doi: 10.1016/j.jaad.2022.08.031.

[34] I. M. Alkhawaldeh, I. Albalkhi, and A. J. Naswhan, "Challenges and limitations of synthetic minority oversampling techniques in machine learning.," *World J Methodol*, vol. 13, no. 5, pp. 373–378, Dec. 2023, doi: 10.5662/wjm.v13.i5.373.

[35] G. R. Ashisha, X. A. Mary, E. G. M. Kanaga, J. Andrew, and R. J. Eunice, "Random Oversampling-Based Diabetes Classification via Machine Learning Algorithms," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, pp. 1–17, Dec. 2024, doi: 10.1007/S44196-024-00678-3/TABLES/9.

[36] P. Sampath *et al.*, "Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique," *Sci Rep*, vol. 14, no. 1, pp. 1–15, Dec. 2024, doi: 10.1038/S41598-024-78519-8;SUBJMETA=114,1537,631,692,699,700;KWRD=COMPUTATIONAL+BIOLOGY+AND+BIOINFORMATICS,DISEASES,HEALTH+CARE,HEALTH+OCCUPATIONS.