Scientific Journal of Informatics
Vol. 12, No. 3, August 2025
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Optimization of Random Forest Algorithm with SMOTE Method to Improve the Accuracy of Early Diabetes Prediction

Siti Khoirun Nisa[1*], Mula Agung Barata[2], Pelangi Eka Yuwita[3]
[1, 2]Department of Informatics Engineering, Nahdlatul Ulama Sunan Giri University, Indonesia
[3]Department of Mechanical Engineering, Nahdlatul Ulama Sunan Giri University, Indonesia

Abstract.
Purpose: This research aims to examine the performance of the random forest algorithm in diabetes risk classification with data balancing using the Synthetic Minority Oversampling Technique (SMOTE) method to improve the representation of minority classes and increase the prediction accuracy value.
Methods: The study used the Behavioral Risk Factor Surveillance System (BRFSS) dataset, obtained from Kaggle, which contains health-related survey data used to identify individuals at risk of diabetes. The Random Forest algorithm was applied to classify diabetes. To balance the data, the SMOTE method was used. The model's performance was evaluated using 10-fold cross-validation by comparing result before and after SMOTE.
Result: The results showed that the application of the SMOTE method improved the performance of the Random Forest classification model, especially in minority classes. Model performance in minority classes without SMOTE had poor evaluation metrics with precision of 49%, recall of 18%, and F1-score of 26%. After applying SMOTE, these values increased to precision of 96%, recall of 88%, and F1-score of 92%. Representing improvements of 47 percentage points in precision, 70 points in recall, and 66 points F1-score. The overall accuracy of the Random Forest model also increased from 86% to 92%, showing a 6 percentage point improvement.
Novelty: This study use integrating the Random Forest algorithm with the SMOTE technique and validating the results using 10-fold cross-validation. The combination significantly improves minority class prediction performance in early diabetes detection, addressing the common limitations of previous studies in handling imbalanced datasets effectively.

Keywords: Diabetes, Random forest, SMOTE, Class imbalance, Cross-validation
Received April 2025 / **Revised** June 2025 / **Accepted** June 2025

INTRODUCTION

Diabetes is now one of the major health problems in the world caused by increasing obesity and lack of physical activity, as well as genetic and environmental factors. It is estimated that rapid urbanization will increase the difficulty in managing diabetes [1]. As reported in the IDF Diabetes Atlas 2021, an estimated 536.6 million adults suffer from diabetes, representing 10.5% of the global population, which is likely to increase in the coming years [2]. The most common type of diabetes, Type II Diabetes, reaches 90% to 95% and is characterized by insufficient insulin sensitivity, resulting in increased blood sugar levels. Uncontrolled diabetes poses many health risks.

Predicting the likelihood of diabetes cases at an early stage with the help of data science methods such as machine learning has shown much promise. Detecting diabetes at an early stage requires the classification of certain criteria including but not limited to glucose levels, blood pressure, age, lifestyle, etc. Classification is a part of data mining, the purpose of which is to sort data into different classes determined by patterns obtained from previous data [3]. To achieve better accuracy while also avoiding overfitting, the Random Forest algorithm which is an ensemble method of machine learning, constructs multiple decision trees during the training phase and uses the predictions from these trees in a voting system. Each tree is trained on randomly selected subsets of the data, along with random subsets of features to ensure diversity and robustness within the model [4]. As a commonly implemented technique for classification problems in machine learning, random forests are praised for their utility, performance, scalability, and user-friendliness

*Corresponding author.
Email addresses: stkhnnisa10@gmail.com (Nisa)*, mula.ab26@gmail.com (Barata), pelangi.ardata@gmail.com (Yuwita)
DOI: 10.15294/sji.v12i3.22986

[5]. They are also able to manage vast amounts of intricate data reliably while maintaining a high level of accuracy [6].

In diabetes detection poses several challenges, one of them being class imbalance: the ratio of a diabetes patient to a patient without diabetes is heavily skewed. This imbalance often leads to the model being biased towards the positive class which diminishes accuracy [7]. To deal with this issue, samples of the minority class are artificially generated using the synthetic minority over-sampling technique (SMOTE) to create a more balanced data distribution [8]. The expectation is that using SMOTE with random forest will improve the model's accuracy for predicting diabetes risk while achieving a more balanced outcome [9].

The Random Forest algorithm for diabetes prediction was initially investigated by Shengyu Wan [10]. In his study, Wan compared the performance of the Random Forest algorithm with that of logistic regression. The results demonstrated that Random Forest significantly outperformed logistic regression, achieving an accuracy of 83.6% compared to 74.8%. This finding highlights the potential of Random Forest in improving early detection and treatment planning in healthcare.

In another study that concerned heart disease prediction using Random Forest, she applied the technique of Synthetic Minority Over-sampling Technique (SMOTE) for treating class imbalance [7]. SMOTE reduced overfitting of the model and improved performance on all metrics using data from the UCI Machine Learning Repository. This proves that it successfully solves class imbalance, thus improving heart disease prediction accuracy.

Another study presented the application of the Random Forest algorithm for the classification of imbalanced diabetes data using the Smote-Tomek Link method [11]. The result indicated that the combination improved accuracy, sensitivity, precision, and F1 score compared to the utilization of Random Forest without the application of any data balancing methods. This confirms that the combination method is appropriate for improving the model's performance on datasets with class continuity.

Arifiyanti and Wahyuni also carried out another study on credit card fraud detection [12]. The authors in research showcased that the imbalance issue addressed by SMOTE increased the performance of most classification algorithms, especially with respect to Logistic Regression, KNN, and Naive Bayes. With the exception of Decision Tree, the results remained unchanged. While precision and recall metrics improved, the overall balanced accuracy slightly decreased, emphasizing the importance of data balancing in enhancing classification for the minority class.

Pertiwi performed diabetes diagnosis using the KNN algorithm and SMOTE [13]. The findings from the study showed that applying SMOTE improved model accuracy by 8.25% when compared to the model without SMOTE, proving that SMOTE effectively mitigates the imbalanced data issue.

Studies such as Akdeas demonstrated that SMOTE significantly improves classification performance across multiple metrics, particularly in multiclass imbalanced datasets [14]. Similarly, Michael highlighted that combining Random Forest with oversampling methods like SMOTE or Random Oversampling yields robust performance improvements, especially in highly imbalanced scenarios [15]. These findings reinforce the importance of incorporating SMOTE into classification pipelines for health-related or risk-prediction data where minority class detection is critical.

The research analyzed the dataset of Pima Indian Diabetes concerning the application of SMOTE and Random Oversampling [16]. The results showed that while random over-sampling was used, the performance of the random forest model was SMOTE enhanced to a greater degree than any other methods used.

Several previous studies have combined Random Forest with SMOTE for diabetes prediction, but they are generally limited to datasets with few features. This study explores datasets with more features. In addition, comprehensive evaluation of model stability in terms of F1 score and sensitivity to positive cases has not been widely studied. Therefore, this study aims to optimize the Random Forest algorithm with the SMOTE technique on a dataset with richer features, and evaluate its effect on the stability of model performance in detecting diabetes cases early. The purpose of this study is to produce a prediction model that is not only accurate, but also reliable and consistent in handling imbalanced data.

METHODS

This research is based on the Diabetes Health Indicators data set that is publicly available on Kaggle. Data is freely accessible to the public. The dataset includes records from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) study conducted by the Centers for Disease Control and Prevention (CDC). Table 1 depicts the dataset which contains 253.680 data entries and 21 health indicators, along with their complete descriptions in Table 2.

Table 1. Dataset diabetes

No	HighBP	HighChol	….	Diabetes_binary
1	1	1	….	0
2	0	0	….	0
….	….	….	….	….
253.680	1	1	….	1

Table 2. Attribute

No.	Atribut	Description
1	HighBP	High blood pressure (1 = yes, 0 = no).
2	HighChol	High cholesterol (1 = yes, 0 = no).
3	CholCheck	Checking cholesterol in the last five years (1 = yes, 0 = no).
4	BMI	Body Mass Index (BMI).
5	Smoker	Smoking (1 = smoker, 0 = non-smoker).
6	Stroke	History of Stroke (1 = yes, 0 = no).
7	HeartDiseaseorAttack	Heart disease (1 = yes, 0 = no).
8	PhysActivity	Physical activity (1 = active, 0 = inactive).
9	Fruits	Fruit consumption (1 = often, 0 = rarely).
10	Veggies	Vegetable consumption (1 = often, 0 = rarely).
11	HvyAlcoholConsump	Alcohol consumption (1 = often, 0 = rarely).
12	AnyHealthcare	Access to health services (1 = yes, 0 = no).
13	NoDocbcCost	No doctor's fees (1 = yes, 0 = no).
14	GenHlth	General health (scale 1-5).
15	MentHlth	Mental health (scale 1-30).
16	PhysHlth	Physical health (scale 1-30).
17	DiffWalk	Difficulty walking (1 = yes, 0 = no).
18	Sex	Gender (1 = male, 0 = female).
19	Age	Respondent's age.
20	Education	Education (scale 1-6).
21	Income	Respondent's income (scale 1-8).

Model performance was evaluated using cross-validation after class assignment and handling with SMOTE, data processing, alignment, random forest model training, and evaluation. The research flow is shown in Figure 1.

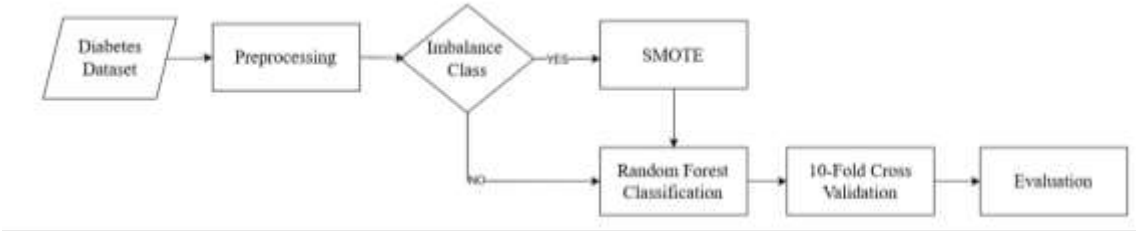

Figure 1. Research flow

The research begins with the Diabetes Health Indicators dataset leveraged from Behavioral Risk Factor Surveillance System (BRFSS) and publicly available on the Kaggle platform. Preprocessing which includes dealing with missing values and overwriting duplicates is the first initial stage. Approximately 24.206 duplicate records were recorded and overwritten in order to enhance data quality. An assessment of class imbalance is conducted post splitting into 'train' and 'test' datasets. The identified imbalance is addressed through the application of SMOTE on the training dataset [17]. Class imbalance can result in the model being biased towards the majority class, which can negatively impact its predictive performance for the minority class [7], [18] .

SMOTE

Oversampling and undersampling are two main techniques in data-driven approaches to handling class imbalance [19]. Oversampling works by increasing the number of samples from the minority class through

replication or synthetic data generation, while undersampling reduces imbalance by removing some samples from the majority class [20]. One of the most widely used oversampling methods is SMOTE (Synthetic Minority Oversampling Technique) [21], is widely acknowledged method for addressing class imbalance by generating synthetic samples rather than duplicating existing ones [22]. This approach enhances model generalization and mitigates the risk of overfitting often caused by random oversampling [23]. Additionally, the application of SMOTE enhances the sensitivity of the classification model [24].

Synthetic samples using the SMOTE technique are created by blending existing samples of the under-represented class [25]. In the SMOTE technique, K-Nearest Neighbors (KNN) plays a key role in generating synthetic data for the minority class. After selecting a random sample from this class, the algorithm identifies its *k* closest neighboring samples measured using Euclidean distance [20]. This process introduces new, plausible samples that enhance the representation of the minority class without merely duplicating existing data [26], [27]. After determining the k nearest neighbors for a selected minority class instance, SMOTE proceeds to generate synthetic data by interpolating between the selected point and one of its neighbors [21]. This is achieved using the formula 1.

$$X_{new} = X_{min} + \lambda \times (X_{neighbor} - X_{min}) \quad (1)$$

where λ is a random number between 0 and 1, X_{min} is the feature vector of a minority instance, and $X_{neighbor}$ is one of its k nearest neighbors [20],[28],[29]. This technique results in more adequate representation for class distribution of the minority class while ensuring that there are no copies of imbalanced data [22]. The following stage involves setting the model with Random Forest.

Cross validation

Cross-validation is a widely used model evaluation approach for objectively and comprehensively measuring the performance of machine learning algorithms [30]. The basic principle of this method is the separation of training data and testing data, so that the model is trained using one subset of data (training set) and then tested on another subset (validation/test set) that the model has never seen before. This allows for testing the model's performance under conditions that are more representative of real data [31].

The framework is rigorously validated using 10-Fold Cross Validation [32]. Which increases the model's accuracy assurance. It is independent of specific data partitions and thus will work in all scenarios and configurations of the data [33]. In every 10-Fold Cross Validation, 9 folds are used for training and 1 is set aside for testing. Each fold has to be tested once. This technique helps reduce the risk of overfitting and ensures that the model performance evaluation does not depend on a specific data split [34].

Random forest

Random Forest, as described by Breiman in 2001, is an ensemble method that combines many decision trees formed by the bootstrap aggregating (bagging) technique [7] [35][36]. This method, as discussed in The Elements of Statistical Learning by Hastie, Tibshirani and Friedman, is not sensitive to noise and missing values [37], [38].

Random Forest outperforms a logistic regression or a single decision tree in predicting outcomes because it employs a voting system based on the predictions of multiple trees [39], [40]. The algorithm first calculates the Gini index of all the nodes, and during the decision tree construction, selects the node with the smallest Gini index. In this case, C represents the classes and Gini coefficient G is defined for every value of i as which can be calculated using the following formula 2.

$$Gini(S) = 1 - \sum_{i=1}^{C} p_i^2 \quad (2)$$

Where C values denote the count of classes and pi is the sample ratio of class I. Hence, the Gini index would be lowest when a sample node is dominated by a single stronger class which makes it purer In the case of optimal feature selection, post dataset division using a feature, reduction of impurity can be gauged using Gini gain [41]. Following formula, Gini Gain can be calculated using formula 3.

$$Gini\ Gain\ (A) = Gini\ (S) - \sum_{t \epsilon T} \frac{|S_t|}{|S|} \cdot Gini(S_t) \quad (3)$$

Explanation:
- S is the initial dataset
- T represents all partitions created by attribute A
- |St| represents the number of cases in partition t
- |S| refers to the total number of cases in the dataset S
- Gini (St) indicates the Gini Index value for partition t.

RESULTS AND DISCUSSIONS

This is the result section of diabetes prediction experiments using random forests classifiers with and without oversampling model. The experiments were carried out in Google Colab using 10-Fold Cross Validation, and a model without oversampling is compared with a model that implements SMOTE.

The original dataset suffered from class imbalance, as non-diabetic samples dominated the population. The model is bound to perform poorly due to its inherent bias towards a singular classification which is shown in the Figure 2.

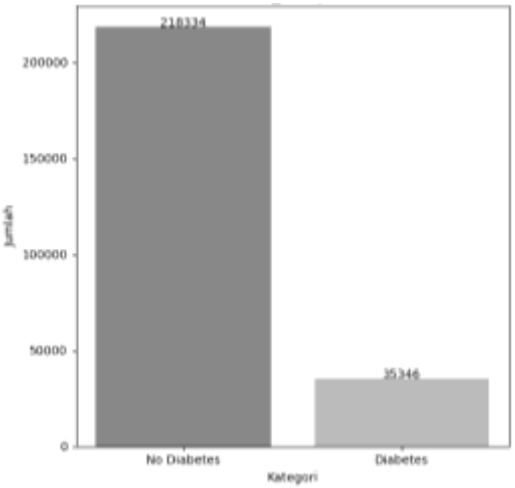

Figure 2. Class distribution before SMOTE

The application of SMOTE ensures class imbalance by augmenting the underrepresented class non-diabetic population. Balance is achieved which will enhance the model's ability to accurately detect both classes as demonstrated in the Figure 3.

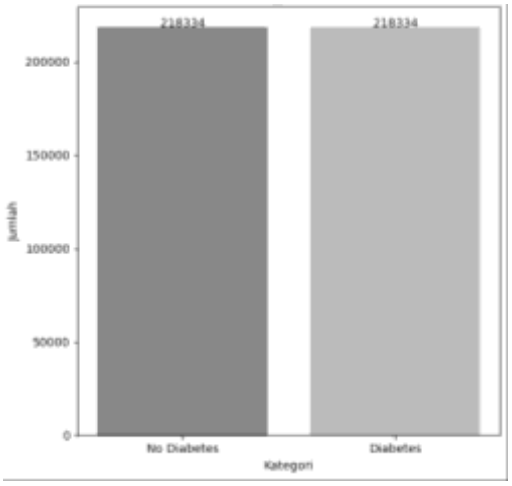

Figure 3. Distribution of classes after SMOTE

A performance comparison of the Random Forest model was done after application of SMOTE and before it. The results from experiments indicate the model tends toward being biased to the majority class without SMOTE. The diabetes class does achieve a high level of accuracy but a lower level of recall. A number of positive diabetes cases can be frequently difficult for the model in order to identify.

SMOTE application balances the data distribution to a greater extent, improving recall without substantial accuracy sacrifice. The model shows capabilities with SMOTE for diabetes case detection, which is reflected in a more stable F1-score. We did perform a 10-Fold Cross-Validation in order to assess the effectiveness of the Random Forest model so as to ensure a thorough assessment of its performance across different subsets of the data, as can be seen in Table 3.

Table 3. Result of 10-fold cross-validation

K-Fold	1	2	3	4	5	6	7	8	9	10
RF	86%	86%	85%	86%	85%	85%	85%	85%	86%	85%
RF + SMOTE	92%	91%	91%	91%	91%	92%	92%	92%	91%	92%

The Random Forest model's performance is assessed through use of a confusion matrix. It displays, for each class, the accurate as well as inaccurate prediction counts. This matrix offers assistance toward comprehension of how the model classifies data, especially when it comes to positive diabetes case detection [35].

The model tends to classify samples as the majority class (non-diabetes) [22] because of the class imbalance when SMOTE is not used [32]. As a result of it, in spite of it exactly attaining, the model might have such constrained capacity for recognizing the positive diabetes cases. The confusion matrix that is show in Figure 4 below is for the model without SMOTE.

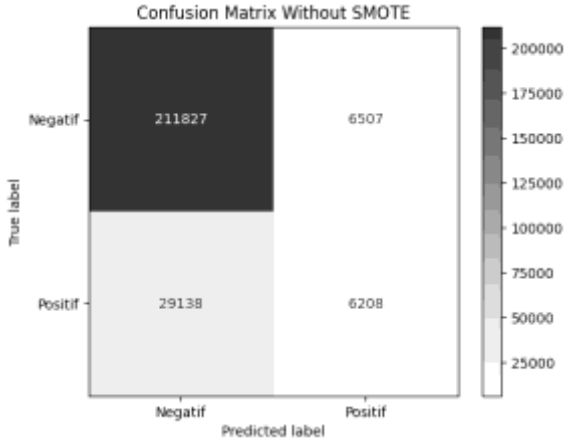

Figure 4. Confusion matrix without SMOTE

After the SMOTE (Synthetic Minority Oversampling Technique) application, the Random Forest model's performance was greatly better than the model without SMOTE [42]. The confusion matrix provides one overview of data classification by the model, after it has oversampled the minority class (diabetes).

The model effectively improves its ability for detecting positive diabetes cases as well as reduces bias toward the majority class (non-diabetes) through SMOTE. This higher recall score exemplifies some improvement, and indicates this model identifies more actual diabetes cases, quite successfully. Additionally, such F1-score is more balanced in nature. It shows some optimization in the balance between both precision and recall. Presented below is that confusion matrix following when SMOTE was applied for highlighting of data balancing's impact upon classification performance:

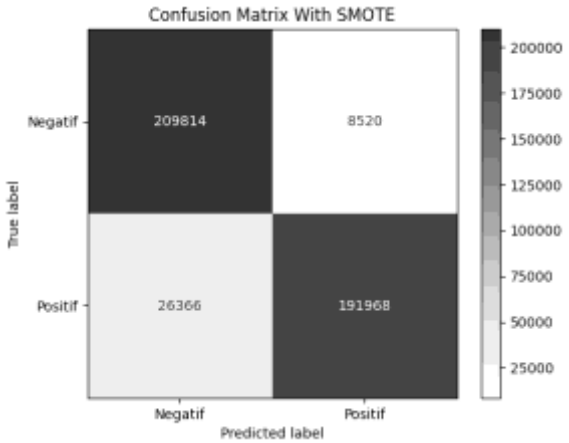

Figure 5. Confusion matrix with SMOTE

Additionally, the model performance F1-score metrics, precision, and also recall are included in the evaluation in order to have a more thorough analysis of the classification results. Accuracy for Random Forest increases up to 92% through applying SMOTE from Table 4 versus Random Forest at 86%. This kind of improvement indicates SMOTE balances data to help out the model recognize more of the positive diabetes cases and it improves the overall performance without sacrificing accuracy for the majority class. Table 4 is presented below, displaying the Random Forest classification report that shows a metrics evaluation comparison between models, with and without SMOTE:

Table 4. Comparison of random forest algorithm results without and with SMOTE in accuracy, recall, precision, and F1 score

Measurement	Without SMOTE		With SMOTE	
	0	1	0	1
Precision	88%	49%	89%	96%
Recall	97%	18%	96%	88%
F1 – Score	92%	26%	92%	92%
Accuracy	86%		92%	

The model evaluation results do show that the SMOTE use within the Random Forest algorithm improves on the functionality of the model. It especially identifies the smaller group, helping diabetes's early discovery [43]. This technique makes the model to be more effective by adjustment of class distribution in the handling of imbalanced datasets. According to the evaluation results, it is indicated that SMOTE helps to improve recall and the F1-score, and thus the model becomes more capable of detecting positive cases [9]. Prior to the application of SMOTE, classification imbalance hurt the model because many positive diabetes cases were missed on account of a recall of 18% for class 1 coupled with an F1-score of 26%. A certain important improvement was observed following SMOTE's implementation.

Table 5. Related research

No	Author	Topic	Result	Relevance
1	S. Wang	Diabetes prediction using Random Forest	Random Forest outperform logistic regression dengan akurasi 83.6% vs 74.8%.	Demonstrating the potential of Random Forest in early prediction of diabetes as a baseline.
2	Erlin, Y. Desnelita, N. Nasution, L. Suryati, F. Zoromi	Heart disease prediction with Random Forest and SMOTE	SMOTE reduces overfitting, improving model performance on imbalanced data.	Supports the use of SMOTE to address class imbalance in datasets.
3	H. Hairani, A. Anggrawan, and D. Priyanto	Diabetes classification with Random Forest and SMOTE-Tomek Link	The combination of methods improves accuracy, sensitivity, precision, and F1-score compared to Random Forest alone.	Validation that data balancing techniques can improve Random Forest performance.
4	A. A. Arifiyanti and E. D. Wahyuni	Detect credit card fraud with SMOTE	SMOTE improves the performance of most algorithms except Decision Tree; improves precision & recall.	Demonstrates the importance of data balancing to improve minority class detection.
5	A. G. Pertiwi, N. Bachtiar, R. Kusumaningrum, I. Waspada, and A. Wibowo	Diabetes diagnosis using KNN and SMOTE	SMOTE improves model accuracy by 8.25% compared to without SMOTE.	Evidence of the effectiveness of SMOTE in dealing with imbalanced data for classification.

CONCLUSION

Based on the results of the study, it can be concluded that the Random Forest algorithm shows good performance in classifying diabetes risk. The application of the Synthetic Minority Oversampling Technique (SMOTE) method has proven to be effective in addressing data imbalance issues, particularly in improving the representation of minority classes. Imbalanced data was addressed by way of a pre-processing stage that the dataset underwent, and F1-score, accuracy, recall, and precision were used in order to evaluate the model's total performance. Experimental results show Random Forest with SMOTE performs better. Classes without SMOTE had poor evaluation metrics with precision of 49%, recall of 18%, and F1-score of 26%. After applying SMOTE, these values increased to precision of 96%, recall of 88%, and F1-score of 92%. These certain findings confirm SMOTE resamples effectively, toward addressing class imbalance toward improving model performance. Future research needs to select features in order to reduce data dimensionality and to improve computational efficiency. Recursive feature elimination, also principal component analysis, and also feature importance exemplify how it is that these techniques analyze data right from Random Forest for the purpose of identifying the most relevant features in order to predict diabetes. Exploring additional models such as XGBoost or LightGBM may provide useful comparisons. The goal is for achieving of the best-performing predictive model.

REFERENCES

[1] Fitriani Nasution, Andilala Siregar, and Ambali Azwar, “Faktor Risiko Kejadian Diabetes Melitus,” vol. 9, no. 2, p. 6, 2021.

[2] 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.

[3] E. F. Laili, Z. Alawi, R. Rohmah, and M. A. Barata, “KOMPARASI ALGORITMA DECISION TREE DAN SUPPORT VECTOR MACHINE (SVM) DALAM KLASIFIKASI SERANGAN JANTUNG,” vol. 8, no. 1, pp. 67–76, 2025.

[4] D. Wulandari, S. Aziz, S. Adrianto, F. Pratiwi, and R. M. Sari, *Teori Dan Implementasi Machine Learning Menggunakan Python*. Serasi Media Teknologi, 2025.

[5] M. A. Barata, Edi Noersasongko, Purwanto, and Moch Arief Soeleman, “Improving the Accuracy of C4.5 Algorithm with Chi-Square Method on Pure Tea Classification Using Electronic Nose,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 2, pp. 226–235, 2023, doi: 10.29207/resti.v7i2.4687.

[6] M. D. Purbolaksono, M. Irvan Tantowi, A. Imam Hidayat, and A. Adiwijaya, “Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 2, pp. 393–399, 2021, doi: 10.29207/resti.v5i2.3008.

[7] Erlin, "Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 3, pp. 677–690, 2022, doi: 10.30812/matrik.v21i3.1726.

[8] 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 Categorization Utilizing C5.0, Random Forest, and Support Vector Machine," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 4, pp. 343–354, 2024, doi: 10.35882/jeeemi.v6i4.434.

[9] S. Pokhrel, "Implementasi Algoritma Random Forest dan Teknik Smote Untuk Klasifikasi Penyakit Diabetes," vol. 15, no. 1, pp. 37–48, 2024.

[10] S. Wang, "Diabetes Prediction Using Random Forest in Healthcare," *Highlights Sci. Eng. Technol.*, vol. 92, pp. 210–217, 2024, doi: 10.54097/5ndh9a05.

[11] H. Hairani, A. Anggrawan, and D. Priyanto, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link," *Int. J. Informatics Vis.*, vol. 7, no. 1, pp. 258–264, 2023, doi: 10.30630/joiv.7.1.1069.

[12] A. A. Arifiyanti and E. D. Wahyuni, "Smote: Metode Penyeimbang Kelas Pada Klasifikasi Data Mining," *SCAN - J. Teknol. Inf. dan Komun.*, vol. 15, no. 1, pp. 34–39, 2020, doi: 10.33005/scan.v15i1.1850.

[13] A. G. Pertiwi, N. Bachtiar, R. Kusumaningrum, I. Waspada, and A. Wibowo, "Comparison of performance of k-nearest neighbor algorithm using smote and k-nearest neighbor algorithm without smote in diagnosis of diabetes disease in balanced data," *J. Phys. Conf. Ser.*, vol. 1524, no. 1, 2020, doi: 10.1088/1742-6596/1524/1/012048.

[14] A. O. Widodo, B. Setiawan, and R. Indraswari, "Machine Learning-Based Intrusion Detection on Multi-Class Imbalanced Dataset Using SMOTE," *Procedia Comput. Sci.*, vol. 234, pp. 578–583, 2024, doi: 10.1016/j.procs.2024.03.042.

[15] M. G. Wijaya, M. F. Pinaringgi, A. Y. Zakiyyah, and Meiliana, "Comparative Analysis of Machine Learning Algorithms and Data Balancing Techniques for Credit Card Fraud Detection," *Procedia Comput. Sci.*, vol. 245, no. C, pp. 677–688, 2024, doi: 10.1016/j.procs.2024.10.294.

[16] H. Hasbi and T. B. Sasongko, "Optimasi Performa Random Forest dengan Random Oversampling dan SMOTE pada Dataset Diabetes," *J. Media Inform. Budidarma*, vol. 8, no. 3, p. 1756, 2024, doi: 10.30865/mib.v8i3.7855.

[17] S. K. M. K. Dony Novaliendry, *Deep Learning Untuk Pemula Jilid 1*. Penerbit CV. SARNU UNTUNG.

[18] Y. Wang, M. M. Rosli, N. Musa, and L. Wang, "Improving clustering-based and adaptive position-aware interpolation oversampling for imbalanced data classification," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 10, p. 102253, 2024, doi: 10.1016/j.jksuci.2024.102253.

[19] 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.

[20] J. Li, Q. Zhu, Q. Wu, and Z. Fan, "A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors," *Inf. Sci. (Ny).*, vol. 565, pp. 438–455, 2021, doi: 10.1016/j.ins.2021.03.041.

[21] R. M. Pereira, Y. M. G. Costa, and C. N. Silla, "Toward hierarchical classification of imbalanced data using random resampling algorithms," *Inf. Sci. (Ny).*, vol. 578, pp. 344–363, 2021, doi: 10.1016/j.ins.2021.07.033.

[22] D. Dash, M. Kumar, A. Kumar, A. Ganguly, and A. Technology-, "ScienceDirect Healthcare Fraud Detection Using an Integrated ML Approach with Healthcare Fraud Detection Using an Integrated ML Approach with Healthcare Fraud Detection Using an Integrated ML Approach with Healthcare Fraud Detection ML Approach with a Usi," *Procedia Comput. Sci.*, vol. 258, pp. 800–810, 2025, doi: 10.1016/j.procs.2025.04.312.

[23] S. B. Belhaouari, A. Islam, K. Kassoul, A. Al-Fuqaha, and A. Bouzerdoum, "Oversampling techniques for imbalanced data in regression," *Expert Syst. Appl.*, vol. 252, no. PB, p. 124118, 2024, doi: 10.1016/j.eswa.2024.124118.

[24] M. Kivrak, U. Avci, and H. Uzun, "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 and Type 2 [2]. Type 2 diabetes constitutes around 90 ease is fundamentally characterized by increasing in creasing insulin secretion over time , triggered by 16 life posed individuals [3]. Hypogonadism is a from syndrome that is a clinical a tosterone de fi studies have Type 2 diabetes have hypogonadism , of is the clinical

features of clude erectile dysfunction , libido , Many studies have shown a relationship betwee ides (TG) [8], as well as hypertension (HT) [9 , 10]. strongly associated with Type 2 diabetes (T D), with s approximately one-third of men with T 2 D [11]. Low tes have been reported to be linked with insulin resistanc mechanism underlying T 2 D . Long-term testosterone," 2024.

[25] F. S. Pratiwi *et al.*, "IMPLEMENTASI METODE SMOTE DAN RANDOM OVER- SAMPLING PADA ALGORITMA MACHINE LEARNING UNTUK," vol. 8, no. 1, pp. 87–98, 2025.

[26] M. Shamshuzzoha, T. T. B. Audry, M. J. Alam, Z. A. Bhuiyan, M. Motaharul Islam, and M. M. Hassan, "A novel framework for seasonal affective disorder detection: Comprehensive machine learning analysis using multimodal social media data and SMOTE," *Acta Psychol. (Amst).*, vol. 256, no. April, p. 105005, 2025, doi: 10.1016/j.actpsy.2025.105005.

[27] A. Wibowo, A. F. N. Masruriyah, and S. Rahmawati, "Refining Diabetes Diagnosis Models: The Impact of SMOTE on SVM, Logistic Regression, and Naïve Bayes for Imbalanced Datasets," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 1, pp. 197–207, 2025, doi: 10.35882/jeeemi.v7i1.596.

[28] V. V. N. Raju *et al.*, "Enhancing emotion prediction using deep learning and distributed federated systems with SMOTE oversampling technique," *Alexandria Eng. J.*, vol. 108, no. July, pp. 498–508, 2024, doi: 10.1016/j.aej.2024.07.081.

[29] J. Fonseca and F. Bacao, "Geometric SMOTE for imbalanced datasets with nominal and continuous features," *Expert Syst. Appl.*, vol. 234, no. July, p. 121053, 2023, doi: 10.1016/j.eswa.2023.121053.

[30] M. Rafało, "Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis," *ICT Express*, vol. 8, no. 2, pp. 183–188, 2022, doi: 10.1016/j.icte.2021.05.001.

[31] F. Marchetti, "A fast surrogate cross validation algorithm for meshfree RBF collocation approaches," *Appl. Math. Comput.*, vol. 481, no. July, 2024, doi: 10.1016/j.amc.2024.128943.

[32] M. Shuja, S. Mittal, and M. Zaman, "Effective Prediction of Type II Diabetes Mellitus Using Data Mining Classifiers and SMOTE," in *Advances in Computing and Intelligent Systems*, H. Sharma, K. Govindan, R. C. Poonia, S. Kumar, and W. M. El-Medany, Eds., Singapore: Springer Singapore, 2020, pp. 195–211.

[33] A. H. Pratama, *BELAJAR MUDAH DAN SINGKAT MACHINE LEARNING: Panduan Praktis dengan Studi Kasus, Kode Program, dan Dataset*. Penerbit Andi, 2024.

[34] T. Leinonen, D. Wong, A. Wahab, R. Nadarajah, M. Kaisti, and A. Airola, "Empirical investigation of multi-source cross-validation in clinical machine learning," *Comput. Biol. Med.*, vol. 183, no. October, p. 109271, 2024, doi: 10.1016/j.compbiomed.2024.109271.

[35] Dikan Ismafillah, Tatang Rohana, and Yana Cahyana, "Analisis algoritma pohon keputusan untuk memprediksi penyakit diabetes menggunakan oversampling smote," *INFOTECH J. Inform. Teknol.*, vol. 4, no. 1, pp. 27–36, 2023, doi: 10.37373/infotech.v4i1.452.

[36] D. Truong, *Data Science and Machine Learning for Non-Programmers: Using SAS Enterprise Miner*. in Chapman \& Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press, 2024.

[37] N. Sharfina and N. G. Ramadhan, "Analisis SMOTE Pada Klasifikasi Hepatitis C Berbasis Random Forest dan Naïve Bayes," *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 8, no. 1, p. 33, 2023, doi: 10.31328/jointecs.v8i1.4456.

[38] H. Ohanyan *et al.*, "Associations between the urban exposome and type 2 diabetes: Results from penalised regression by least absolute shrinkage and selection operator and random forest models," *Environ. Int.*, vol. 170, no. October, p. 107592, 2022, doi: 10.1016/j.envint.2022.107592.

[39] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. in Springer Series in Statistics. Springer New York, 2013.

[40] W. Xu, J. Zhang, Q. Zhang, and X. Wei, "Risk prediction of type II diabetes based on random forest model," *Proc. 3rd IEEE Int. Conf. Adv. Electr. Electron. Information, Commun. Bio-Informatics, AEEICB 2017*, pp. 382–386, 2017, doi: 10.1109/AEEICB.2017.7972337.

[41] A. F. Fadhlullah and T. Widiyaningtyas, "Comparative Analysis of Decision Tree and Random Forest Algorithms for Diabetes Prediction," vol. 8, no. 4, pp. 1121–1132, 2024.

[42] D. V. Ramadhanti, R. Santoso, and T. Widiharih, "Perbandingan Smote Dan Adasyn Pada Data Imbalance Untuk Klasifikasi Rumah Tangga Miskin Di Kabupaten Temanggung Dengan Algoritma K-Nearest Neighbor," *J. Gaussian*, vol. 11, no. 4, pp. 499–505, 2023, doi: 10.14710/j.gauss.11.4.499-505.

[43] A. Nugroho and D. Harini, "Teknik Random Forest untuk Meningkatan Akurasi Data Tidak Seimbang," *JSITIK J. Sist. Inf. dan Teknol. Inf. Komput.*, vol. 2, no. 2, pp. 128–140, 2024, doi: 10.53624/jsitik.v2i2.379.