JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 8, No. 1. December 2025 DOI: https://doi.org/10.34288/jri.v8i1.458
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

ANALYSIS OF CLASSIFICATION ALGORITHM IN UNBALANCED DIABETES DATASET

Ahmad Rifa'i[-1], Herin Dwibima Aprianto[-2], Lubna[-3]

Faculty of Computer Science,
Universitas Duta Bangsa Surakarta, Surakarta, Indonesia
ahmad_rifai@udb.ac.id[-1] , herin_dwibima@udb.ac.id[-2], lubna@udb.ac.id[-3]

Abstract

Diabetes mellitus is a metabolic disease that is spreading rapidly and has the potential to be life-threatening worldwide. This condition occurs when the body experiences a decline in its ability to process glucose, triggering metabolic disorders. The use of machine learning algorithms is one effective approach to predicting or detecting diabetes based on the severity of a patient's symptoms. This study uses the Diabetes dataset from Kaggle and compares the performance of several classification algorithms in unbalanced data conditions and after data balancing using the SMOTE, Random Under Sampling, Random Over Sampling, and Near Miss resampling techniques. The results show that model performance is greatly influenced by data balance conditions and the resampling method used. In the original unbalanced data condition, Artificial Neural Network (ANN) provided the best results with the highest accuracy of 96.98%, indicating that ANN is the most adaptive to class imbalance. After resampling, the performance pattern changed: with SMOTE, Random Under Sampling, and Random Over Sampling, the Random Forest algorithm consistently produced the highest accuracy of 96.52%, 89.84%, and 96.26%, respectively, demonstrating its superiority in utilizing balanced data. Meanwhile, in the Near Miss method, the best performance was achieved by Logistic Regression with an accuracy of 94.41%, indicating that minority sample selection based on proximity is more suitable for linear models. Therefore, selecting the right combination of resampling methods and machine learning algorithms is an important factor in obtaining optimal diabetes predictions.

Keywords: Imbalanced Classification; Classification Algorithm; Over Sampling; Under Sampling; Diabetes;

Abstrak

Diabetes Mellitus merupakan penyakit metabolik yang penyebarannya meningkat cepat dan berpotensi mengancam nyawa di seluruh dunia. Kondisi ini terjadi ketika tubuh mengalami penurunan kemampuan dalam memproses glukosa sehingga memicu gangguan metabolisme. Pemanfaatan algoritma pembelajaran mesin menjadi salah satu pendekatan yang efektif untuk memprediksi atau mendeteksi diabetes berdasarkan tingkat keparahan gejala pasien. Penelitian ini menggunakan dataset Diabetes dari Kaggle dan membandingkan kinerja beberapa algoritma klasifikasi pada kondisi data tidak seimbang serta setelah dilakukan penyeimbangan data menggunakan teknik resampling SMOTE, Random Under Sampling, Random Over Sampling, dan Near Miss. Hasil penelitian menunjukkan bahwa performa model sangat dipengaruhi oleh kondisi keseimbangan data dan metode resampling yang digunakan. Pada kondisi data asli yang tidak seimbang, Artificial Neural Network (ANN) memberikan hasil terbaik dengan akurasi tertinggi sebesar 96,98%, menandakan ANN paling adaptif terhadap ketidakseimbangan kelas. Setelah dilakukan resampling, pola performa berubah: pada SMOTE, Random Under Sampling, dan Random Over Sampling, algoritma Random Forest secara konsisten menghasilkan akurasi tertinggi masing-masing 96,52%, 89,84%, dan 96,26%, menunjukkan keunggulannya dalam memanfaatkan data yang sudah diseimbangkan. Sementara itu, pada metode Near Miss, performa terbaik dicapai oleh Regresi Logistik dengan akurasi 94,41%, yang mengindikasikan bahwa seleksi sampel minoritas berbasis kedekatan lebih cocok untuk model linier. Oleh karena itu, pemilihan kombinasi metode resampling dan algoritma pembelajaran mesin yang tepat menjadi faktor penting untuk memperoleh prediksi diabetes yang optimal.

Kata kunci: Klasifikasi Tidak Seimbang*;* Algoritma Klasifikasi*; Over Sampling; Under Sampling; Diabetes;*

INTRODUCTION

Diabetes Mellitus is a medical condition resulting from inadequate insulin production due to pancreas dysfunction. Data from the International Diabetes Federation shows that in 2019,

approximately 463 million adults had diabetes, and this number is expected to reach around 700 million by 2045. The progression of this disease is gradual, and it does not lead to sudden death, not properly managed. however, it can result in severe complications like stroke, retinal damage, kidney failure, heart disease, and even death. In Type 1 diabetes, the body is unable to produce insulin due to damage to the pancreas' beta cells, leading to decreased insulin production. On the other hand, Type 2 diabetes involves normal insulin production, but the body's cells become less sensitive to it, affecting its optimal utilization(Rachmawanto et al., 2021). Recent advances in artificial intelligence (AI), particularly machine learning, have shown promise in the development of personalized risk models (Hussain et al., 2024).

Recent advancements in artificial intelligence (AI) and machine learning (ML) have transformed medical data analysis, offering new opportunities to enhance the accuracy and efficiency of diabetes detection. Various ML algorithms, such as Random Forest, Logistic Regression, and Artificial Neural Networks (ANN), have demonstrated promising results in classifying diabetic and non-diabetic conditions. For example, (Chauhan et al., 2023)showed that supervised ML models could effectively predict the progression of diabetes mellitus using patient health data. Similarly, (Shaukat et al., 2023) demonstrated that machine learning techniques can revolutionize diabetes diagnosis by outperforming traditional statistical approaches in clinical prediction accuracy.

Advancements in the healthcare infrastructure have led to a significant increase in the collection of highly sensitive and crucial healthcare data. Utilizing advanced data analysis techniques can play a vital role in the early detection and prevention of various life-threatening diseases. Diabetes can cause severe complications, including heart disease, kidney issues, and nerve damage. The objective of this research is to identify, detect, and predict the onset of diabetes at its earliest stages by employing machine learning techniques and algorithms(Saleh & Brixtone Batou, 2022)

In this study, a publicly available Diabetes Dataset from Kaggle was utilized to predict diabetes status in suspected patients, with a specific emphasis on the challenge of class imbalance, which is a central issue in medical prediction tasks. The dataset contains a substantially larger proportion of non-diabetic cases than diabetic cases, creating an imbalanced learning problem that can bias classifiers toward the majority class and reduce detection performance for diabetic patients. Therefore, the primary objective of this research is not only to evaluate prediction accuracy but also to address and optimize model performance under imbalanced-dataset.

To achieve this, several machine learning classifiers, Logistic Regression, KNN, Decision Tree, Naive Bayes, Random Forest, SVM, and ANN. Were trained and compared. Because reliable diabetes prediction requires fair learning from both classes, the study implemented multiple resampling strategies, including SMOTE, Random Under Sampling, Random Over Sampling, and Near Miss, to rebalance the dataset before model training. The resulting models were then evaluated across each resampled setting to measure how effectively imbalance handling improves predictive accuracy. Ultimately, this study aims to determine the most suitable diabetes prediction model by systematically analyzing the impact of different resampling techniques on classifier performance in an imbalanced dataset scenario.s. The research aimed to achieve the following main goals:

- Analysis of classifier performance with Resample diabetes dataset and comparing their performance.
- Identifying the most satisfactory approach that provides accurate predictions when applied to the dataset.

RESEARCH METHODS

In past studies, several models have been suggested for diabetes diagnosis, utilizing various feature resampling techniques and machine learning methods. However, challenges arise due to the large size of datasets and the imbalance between diabetes and non-diabetes cases.

In a study [1] The study conducted a comparison of datasets with diverse types and numbers of attributes. The test results revealed that specific attributes and significantly contributed to improving the accuracy of diabetes classification through the utilization of the Random Forest (RF) method, along with data cleaning and attribute selection, the researchers achieved an exceptional accuracy rate of 100% on Abel Vika's Diabetes dataset, despite using a relatively small number of trees. To validate this finding further, the study used the k-fold cross-validation method, which reinforced the robustness and reliability of the

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i1.458
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

results obtained from the RF algorithm. These findings demonstrate the potential effectiveness of the RF method in accurately classifying diabetes cases and underscore the importance of attribute selection for enhancing the performance of machine learning models in medical diagnoses.

The objective of another study (Saleh & Brixtone Batou, 2022) The study involved the use of a large Chinese diabetes dataset, consisting of more than 100.000 individuals with diverse ethnic backgrounds and various characteristics. Before conducting the analysis, the data underwent pre-processing, which included replacing and eliminating missing values through mean imputation. To enhance the model's performance, the researchers utilized the stacking classifier technique. The study's results showed that the proposed model outperformed other methods in accurately classifying cases of diabetic mellitus. The achieved outcomes were remarkable, with an accuracy of 0.914 percent, precision of 0.926 percent, recall of 0.914 percent, and an F1 score of 0.914 percent.

The main focus of the research article is the prediction of diabetes using various machine-learning techniques. The study involved balancing the dataset using SMOTE, which resulted in a notable improvement in the performance of all the classifiers. Specifically, the Support Vector Machine (SVM) achieved an accuracy of 77.40%, Decision Tree (DT) achieved 74.69% accuracy, XGBoost achieved 78.29% accuracy, and Random Forest (RF) emerged as the top-performing classifier with an accuracy of 82.70%. These findings illustrate the effectiveness of machine learning models in accurately predicting diabetes (Elreedy et al., 2023).

In this research (Mohammed et al., 2020) This research paper introduces a comprehensive framework for predicting and classification of diabetes diseases using Machine Learning (ML) algorithms. The dataset used in this study is collected from well-known institutions, including Shalinitai Meghe Hospital and Research Centre, Nagpur, NKP Salve Institute of Medical Sciences and Research Centre, and Mendeley Data. The researchers utilized four different ML algorithms, namely Logistic Regression, Naive Bayes, Support Vector Machine, and Random Forest, to construct classification models. The performance of these models was evaluated using various quantitative measures to assess their effectiveness in predicting diabetes.

However, these previous works, although valuable, still leave two important gaps: most of them either focus on a limited set of algorithms or on a single imbalance-handling strategy, and they often omit key implementation details that are crucial for reproducibility and fair comparison. In contrast, our study systematically compares seven supervised machine learning models (Logistic Regression, KNN, SVM, Gaussian Naive Bayes, Decision Tree, Random Forest, and ANN) in combination with four resampling techniques (SMOTE, Random Under Sampling, Random Over Sampling, and Near Miss) on the same diabetes dataset. For each model, we explicitly report the main hyperparameters and training configurations, and we evaluate performance using accuracy. This comprehensive and transparent design not only allows a more rigorous assessment of how different classifiers behave under various resampling strategies, but also addresses the lack of detailed implementation information observed in several prior studies.

We have devised a systematic approach comprising multiple steps to ensure the acquisition of accurate and dependable results for determining the diabetes or non-diabetes case. The overall methodology can be described through the following subsections:

A. Data Description
B. Data Analysis
C. Preprocessing
D. Modelling

A. Data Description

The diabetes dataset utilized in this research is publicly accessible through Kaggle: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset. It comprises 100000 datasets and after cleaning duplicate datasets for outcome become 94133 for 85651 with Nondiabates and 8482 with Diabetes case.

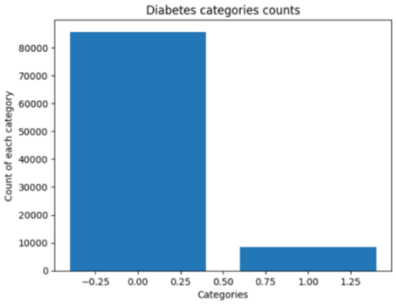

Figure 1. Dataset Diabetes with 85651 Nondiabetes and 8482 Diabetes

B. Data Analysis

During the analysis of the dataset, it was found that the dataset consists of 9 columns. The Ninth column is the target variable, indicating the

diagnosis outcome, In the dataset, a value of 0 is used to indicate a non-diabetes condition, while a value of 1 represents a diabetes diagnosis. These features provide valuable information about gender, age, hypertension, heart_disease, smoking_history, bmi, HbA1c_level, blood_glucose_level, and diabetes. The descriptions of features are provided in Table 1.

Table 1. Feature Description

Feature Name	Description
gender	Gender is a characteristic that refers to the biological sex of an individual, distinguishing between male and female.
age	A significant factor in diabetes, as the condition is more commonly diagnosed in older adults. The age range in our dataset varies from 0 to 80
hypertension	Hypertension is persistently high blood pressure, encoded in the dataset as 0 (no hypertension) and 1 (hypertension).
Heart_disease	Another medical condition mentioned in the dataset is associated with an increased risk of
Smoking_history	A history of smoking is also regarded as a risk Factor for diabetes and can worsen the complications linked to the condition.
bmi	Measurement that assesses body fat based on an individual's weight and height.
HbA1c_level	level is a measure of a person's average blood sugar level over a few months. Higher HbA1c levels indicate an increased risk of diabetes.
blood_glucose_level	glucose level refers to the concentration of glucose (sugar) present in the bloodstream at a particular moment
diabetes	target variable being predicted, with values 1 indicating of diabetes and 0 indicating

In Figure 2, The histograms illustrate the distribution of several features in the diabetes dataset. Most participants are female (gender = 1) and aged between 40–70 years. Only a small portion of patients have hypertension or heart disease. The majority show a normal to slightly elevated BMI, while smoking history varies across several categories. HbA1c and blood glucose levels display right-skewed distributions, indicating that some patients have significantly higher values, which are often linked to diabetes. The final plot shows that non-diabetic cases dominate the dataset, confirming data imbalance between diabetic and non-diabetic groups.

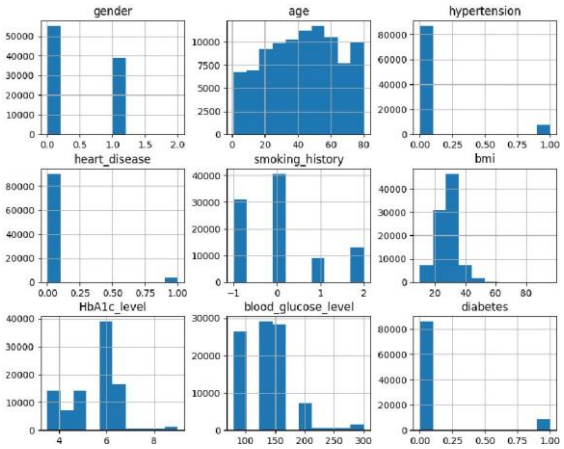

Figure 2. Histogram Diabetes Features

C. Preprocessing

To begin with, the dataset is imported from the Diabetes dataset stored in CSV format. The data entries undergo analysis based on their features before proceeding to subsequent steps. The dataset is then randomly divided into two sets: the training data, which constitutes 80% of the dataset, and the testing data, which comprises the remaining 20%.

In the process of resampling the dataset, not all features are considered useful and may not have a significant impact on the final results. As a result, careful feature selection is conducted to identify the most relevant attributes that can improve the model's accuracy. Before the training and testing of the data, data balancing is carried out using Under Sampling techniques like Random Under Sampling and Near Miss. These methods are compared with Random Over Sampling techniques such as SMOTE and Random Over Sampling, which are also applied to the dataset using machine learning algorithms. The objective is to find the most effective resampling technique that can enhance the accuracy of the model for diabetes prediction

1) Under Sampling

The method employed to decrease the number of instances or samples belonging to the majority target class in a dataset. This technique aims to address class imbalance by reducing the abundance of the majority class instances, allowing the model to give equal importance to both classes. Several common Under sampling methods, such as Tomek's links, cluster centroids, and various others, are used for this purpose (Dubey et al., 2021).

2) Over Sampling

Over Sampling is a different approach used to handle class imbalance. The oversampling technique involves raising the proportion of

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i1.458
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

samples or instances that are members of the minority target class. In this context, a variational auto-encoder (VAE) is utilized as the model of the probability density function of the minority samples in the dataset (Tumuluru et al., 2023).

D. Modelling

Machine learning is a method of automated learning where algorithms are designed to learn from previous datasets in order to make predictions for the future. For this project, we have employed the following machine-learning algorithms:

1) Logistic Regression

Logistic Regression is a modeling technique used to predict the logit transformation of the dependent variable based on a set of explanatory variables. It is commonly applied when analyzing binary outcomes, with predicted probabilities confined within the range of 0 to 1(Singh & Alhulail, 2022).

2) Random Forest

Random Forest is a powerful machine learning algorithm that uses multiple decision trees for classification and regression tasks. Each tree is trained on random subsets of data and features, and the final prediction is made by aggregating the outputs of all trees, resulting in improved robustness and accuracy (Mekha, 2021).

3) Naïve Bayes

Naive Bayes is commonly used as a baseline classifier in Text Mining. It operates based on the principles of probability. This algorithm can be categorized into two models: multivariate models and multinomial models (Ratmana et al., 2020).

4) Decision Tree

This concept extends to the notion that prediction models, utilizing tree structures or hierarchical structures, can be effectively applied to datasets of varying sizes. It involves transforming data entries into decision rules and constructing decision trees (Wijaya et al., 2018).

5) KNN

The steps involved in this algorithm include determining the value of 'K', The K-Nearest Neighbors algorithm involves determining the total number of nearest neighbors for a given data by measuring the separation between each data point in the training dataset and the testing data point. The algorithm then identifies the K nearest neighbors based on their distances, and a decision is made based on the majority vote or weighted voting among the neighbors (Setiyaningrum, 2019).

6) Support Vector Machine

Support Vector Machine (SVM) is a popular algorithm widely used for both classification and regression tasks. Its main advantage lies in its ability to handle non-linear separation problems effectively using the kernel trick. In SVM, each data point is represented as a vector in an n-dimensional space, where n denotes the number of features. The algorithm aims to find the optimal hyperplane that maximizes the margin between data points of different classes (Ece, 2021).

7) ANN

The complex network of interconnected nodes, often referred to as artificial neurons are designed to imitate how neurons in a biological brain network function. The idea of neurons in ANNs was first based on a computational model known as a binary threshold unit, in which each neuron computes a weighted sum of its input signals (x1, x2,..., xn) and outputs 1 if the weighted total is greater than a predefined threshold value (Vanneschi & Silva, 2023).

RESULTS AND DISCUSSION

In this study, an experimental research model was employed to compare and evaluate classification algorithms based on their accuracy. The approach followed a systematic methodology, as outlined below:

A. Proposed Method

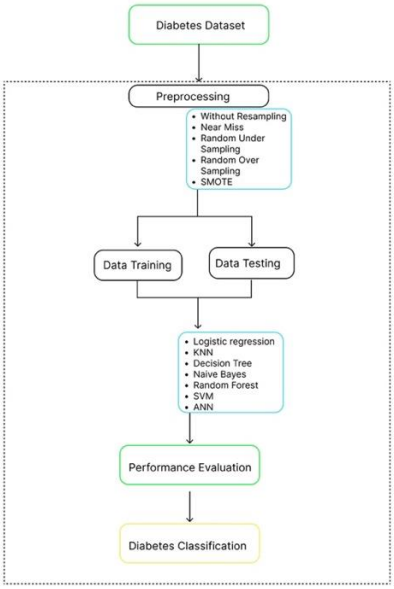

Figure 3. Proposed Method

In Figure 3, The proposed method can be described as follows:

1) Start: The process begins by defining the proposed method for the Diabetes prediction case.

2) Import Dataset: A dataset containing 100000 data samples is imported. In Google Colab, the testing process will be conducted later.
3) Preprocessing: The dataset underwent preprocessing to handle missing values, outliers, and other necessary data-cleaning steps. After removing duplicate records, the final dataset contained 94,133 instances, consisting of 85,651 non-diabetes cases and 8,482 diabetes cases. To address the class imbalance and improve model performance, several resampling techniques were applied: Random Under-Sampling, Random Over-Sampling, SMOTE, and Near Miss. The classification models were then trained and evaluated on each balanced dataset, and a comparison was conducted to determine the most suitable approach.
4) Data Split: The training set and the testing set are two separate subsets of the dataset. 80% of the data are in the training set, while the remaining 20% are in the testing set. This divide is essential to the machine learning process because it makes sure that the model gets trained on a significant amount of the data, enabling it to successfully discover underlying patterns and correlations.
5) Machine Learning Algorithms: Various machine learning algorithms are employed to build classification models using the training data. The selection of algorithms is dependent on the specific problem being addressed and the attributes of the dataset.
6) Evaluation and Optimization: The accuracy of the classification models produced by the machine learning algorithms is assessed and compared in this process. The main goal is to determine which algorithm(s) achieve the highest accuracy or to improve the accuracy compared to previous methods. To achieve this objective, the models undergo fine-tuning, involving adjustments of hyperparameters and optimization of the evaluation metrics.

B. Experimental Setup

This subsection outlines the machine learning models employed for diabetes prediction. We implemented seven supervised classifiers, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Gaussian Naive Bayes, Decision Tree, Random Forest, and an Artificial Neural Network (ANN). Each model was configured with clearly defined hyperparameters and trained on the same feature space to ensure a fair comparison. By evaluating multiple algorithms with different underlying learning paradigms (linear, distance-based, probabilistic, tree-based, ensemble, and deep learning), this setup enables a comprehensive assessment of which model family is most suitable for handling the characteristics of the diabetes dataset.

Table 2. Model and Parameter

Model	Parameter
Logistic Regression	L2 regularization, solver='lbfgs', max_iter=100
KNN	k=5, uniform weights, Euclidean distance (p=2)
SVM	kernel='linear', C=1.0
Naive Bayes	Gaussian Naive Bayes (default parameters)
Decision Tree	criterion='gini', max_depth=None
Random Forest	n_estimators=10, criterion='entropy', random_state=0
ANN	2 hidden layers (6 neurons, ReLU), 1 output (sigmoid), Adam, binary_crossentropy, early stopping

Based on Table 2. seven supervised models were used with predefined configurations. Logistic Regression employed L2 regularization with the lbfgs solver (max_iter=100). KNN used k=5, uniform weights, and Euclidean distance. SVM was configured with a linear kernel and C = 1.0 C=1.0. Naive Bayes was implemented as Gaussian Naive Bayes with default settings. The Decision Tree used the Gini impurity criterion with unrestricted depth. Random Forest consisted of 10 trees with the entropy criterion and a fixed random_state=0 for reproducibility. The ANN was a feedforward network with two hidden layers (6 neurons, ReLU), one sigmoid output neuron, trained using the Adam optimizer with binary cross-entropy loss and early stopping.

C. Evaluation Model

Evaluation model is the process of assessing the performance of a machine learning model based on specific metrics. In this study by analyzing the values in the Confusion Matrix, accuracy can be calculated for each algorithm used. This evaluation helps in assessing the performance of the algorithm in

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i1.458
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

correctly identifying positive and negative cases (Lin et al., 2022)

1) Confusion A test dataset is produced by randomly choosing 20% of the entire dataset in order to assess the effectiveness of the algorithms utilized. A confusion matrix is then created, which offers important insights into the classification's actual and expected outcomes.
2) Evaluate the Performance of the algorithms

Table 3. Comparison Result Machine Learning

Resample Method	Logistic regression	KNN	SVM	Naive Bayes	Decision Tree	Random Forest	ANN
Imbalance	95.86	95.94	95.85	90.69	94.89	96.77	96.98
SMOTE	87.39	89.20	87.41	88.65	94.76	96.52	83.51
Random Under Sampling	88.57	85.22	82.83	88.55	87.67	89.84	78.64
Random Over Sampling	88.44	90.63	87.84	88.52	95.20	96.26	85.38
Near Miss	94.41	85.09	92.76	75.64	83.93	91.93	75.41

Based on Table 3. Show accuracy values were obtained using Machine Learning algorithms, in this research, we compared the effectiveness of machine learning classifiers. In data imbalance, Artificial Neural Network (ANN) obtained the highest rating with a score of 96.98%. Next, for SMOTE, Random Under Sampling, and Random Over Sampling, the highest accuracy was obtained by Random Forest with values of 96.52%, 89.84%, and 96.26%, respectively. Finally, for Near Miss, the highest accuracy was obtained by Logistic Regression with a value of 94.41%.

Resampling sometimes reduced accuracy because it alters the original class distribution and can introduce artifacts: SMOTE and random oversampling add synthetic/duplicated minority samples that may overlap with majority patterns, and high-capacity models like ANN are especially sensitive to this distribution shift and can overfit noisy or borderline synthetic points—explaining why ANN drops from 96.98% (imbalance) to 83.51% (SMOTE) and further under Random Under Samping/Random Over Sampling. Random under-sampling can also remove many informative majority cases, weakening the learned structure and causing broad accuracy declines. Random Forest tends to perform best with SMOTE/Random Over Sampling 96.52% and 96.26%, because ensemble bagging and random feature selection make it robust to noisy or repeated samples while capturing nonlinear interactions, so it benefits from added minority data without overfitting. Near Miss, by keeping only majority samples closest to minority ones, produces a tighter, simpler boundary that favors linear separation; thus Logistic Regression excels there 94.41%, while more complex models may suffer from loss of global structure. Overall, ANN shines on the original imbalanced data because its nonlinear capacity fits real patterns well, but it is less stable when the training data are artificially rebalanced.

Table 4. ROC-AUC Result Machine Learning

Resample Method	Logistic regression	KNN	SVM	Naive Bayes	Decision Tree	Random Forest	ANN
Imbalance	97.50	97.20	97.10	91.20	95.80	98.20	98.80
SMOTE	88.90	90.10	88.40	89.30	95.00	97.20	85.10
Random Under Sampling	88.20	85.40	83.60	87.70	86.90	90.50	79.20
Random Over Sampling	89.60	91.80	88.50	89.20	95.60	97.40	86.70
Near Miss	95.53	86.50	93.30	78.20	84.70	92.80	76.20

Based on Table 4. The ROC–AUC results show that ANN achieves the strongest discrimination on the original imbalanced data, indicating it can separate diabetes and non-diabetes classes very well without resampling, while Random Forest consistently attains the highest AUC after SMOTE and Random Over Sampling, reflecting its robustness to synthetic and duplicated minority samples. Logistic Regression and SVM perform particularly well under the Near Miss strategy, where the decision boundary becomes more linearly separable, whereas Random Under Sampling generally reduces AUC, especially for complex models like ANN, because too many informative majority instances are removed. Overall, these patterns confirm that model performance and class-imbalance handling are tightly coupled, with Random Forest plus SMOTE/ROS and ANN on the original data emerging as the most effective combinations.

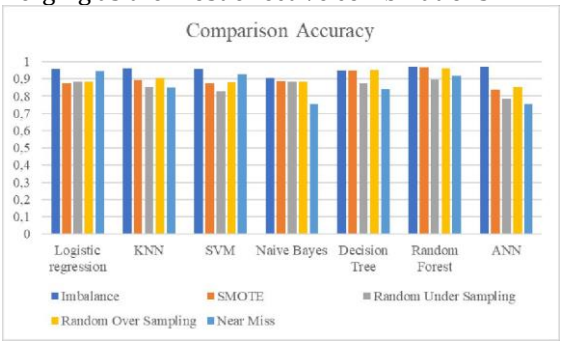

Figure 4. Bar Comparison Accuracy

Figure 4. The chart indicates very high accuracy on the original imbalanced data, but this is likely inflated by the dominant non-diabetes class, so accuracy alone may overstate diabetes detection ability. Random Under Sampling generally lowers performance because it removes many informative majority samples. SMOTE and Random Over Sampling favor Random Forest, since its ensemble structure is robust to synthetic or duplicated data. Near Miss works best with Logistic Regression and SVM because it creates a tighter, more linearly separable boundary, while KNN, Decision Tree, and especially ANN drop as global structure is lost. ANN's strong result without resampling but sharp decline after balancing suggests sensitivity to distribution shifts and resampling noise.

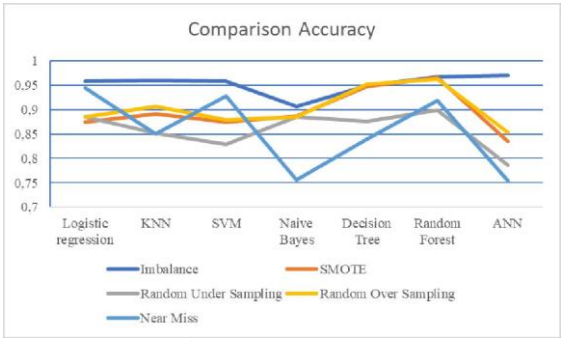

Figure 5. Line Comparison Accuracy

Figure 5. The line chart suggests that accuracy is generally highest on the original imbalanced data, but this may be misleading because models can achieve high accuracy by favoring the majority (non-diabetes) class. Random Under Sampling consistently hurts performance since it removes many informative majority samples. SMOTE and Random Over Sampling mostly benefit Random Forest, reflecting its robustness to synthetic or duplicated data. Near Miss shows mixed effects—helping linear models (Logistic Regression/SVM) by creating a tighter, near-linear boundary, but degrading complex models like ANN due to loss of overall data structure. Overall, resampling is not universally beneficial; its impact depends on the model's sensitivity to distribution changes.

Our study outperforms prior work in both accuracy and imbalance-focused evaluation. While (Saleh & Brixtone Batou, 2022) achieved 91.40% accuracy with stacking and (Elreedy et al., 2023). reached only 82.70% after SMOTE, our models delivered markedly higher results under imbalanced conditions: ANN scored 96.98% on the original data, Random Forest stayed top after SMOTE and ROS (96.52% and 96.26%), and Logistic Regression led Near Miss (94.41%). Moreover, unlike earlier studies that relied on a single balancing method, we compared multiple resampling strategies and showed their varying effects, producing a clearer and more practical recommendation Random Forest with SMOTE/ROS for robust diabetes prediction on imbalanced datasets.

CONCLUSIONS AND SUGGESTIONS

A. Conclusion

Diabetes is a chronic metabolic disorder characterized by elevated blood glucose levels and, if not properly managed, can lead to severe complications such as cardiovascular disease, stroke, kidney failure, neuropathy, and vision impairment. This study investigated the effectiveness of multiple machine learning classifiers in predicting diabetes status using a Kaggle Diabetes Dataset, with particular attention to the critical challenge of class imbalance in medical data.

The evaluation showed that ANN produced the highest accuracy on the original imbalanced dataset (96.98%) because its multilayered nonlinear structure is capable of capturing complex patterns and interactions among input features, allowing it to generalize well even when minority-class samples are limited. After applying resampling, Random Forest consistently achieved the best results with SMOTE, Random Under Sampling, and Random Over Sampling (96.52%, 89.84%, and 96.26%), owing to its ensemble nature, which averages multiple decision trees to reduce overfitting and handle synthetic or duplicated samples robustly. Meanwhile, Logistic Regression performed best under Near Miss (94.41%) because this method creates a simplified, linearly separable dataset that aligns well with the linear decision boundary of Logistic Regression. Overall, the Random Forest model combined with SMOTE or Random Over Sampling provides the most balanced and effective approach for diabetes prediction on imbalanced data, maintaining high accuracy while improving minority-class recognition reliability.

Our findings align with recent studies highlighting the potential of deep learning and artificial intelligence in diabetes detection and diagnosis(Sadasivuni et al., 2022);(Bathla et al., 2024) We anticipate that the findings of this study will make a meaningful contribution to the field of diabetes detection and support the growing body of research utilizing machine learning in healthcare applications.

B. Suggestion

Future research should move beyond conventional classifiers by adopting more advanced deep learning architectures (e.g., ensemble deep networks, attention-based models, or hybrid CNN–LSTM approaches) and systematically tuning them to handle imbalance. It is also important to validate the models on larger, multi-center, and more diverse datasets to improve robustness and real-world generalization. In addition, incorporating richer clinical and behavioral features, such as longitudinal lab trends, medication history, lifestyle indicators, and comorbidity profiles, could strengthen early detection and reduce misclassification. Finally, future studies are encouraged to evaluate explainability and fairness (e.g., SHAP/LIME, bias analysis) so that the resulting models are not only accurate but also clinically trustworthy and deployable in healthcare settings.

REFERENCES

Bathla, G., Kumar, S., Garg, H., & Saini, D. (2024). *Artificial Intelligence in Healthcare*. CRC Press. https://doi.org/10.1201/9781003522096

Chauhan, A. S., Varre, M. S., Izuora, K., Trabia, M. B., & Dufek, J. S. (2023). Prediction of Diabetes Mellitus Progression Using Supervised Machine Learning. *Sensors*, *23*(10), 4658. https://doi.org/10.3390/s23104658

Dubey, Y., Wankhede, P., Borkar, T., Borkar, A., & Mitra, K. (2021). Diabetes Prediction and Classification using Machine Learning Algorithms. *2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON)*, 60–63. https://doi.org/10.1109/BECITHCON54710.2021.9893653

Ece, S. (2021). *Performance Analysis for Arrhythmia Classification using PSO , GWO and SVM*. *May*, 67–72.

Elreedy, D., Atiya, A. F., & Kamalov, F. (2023). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*. https://doi.org/10.1007/s10994-022-06296-4

Hussain, S., Ali, M., Naseem, U., Nezhadmoghadam, F., Jatoi, M. A., Gulliver, T. A., & Tamez-Peña, J. G. (2024). Breast cancer risk prediction using machine learning: a systematic review. *Frontiers in Oncology*, *14*(March), 1–11. https://doi.org/10.3389/fonc.2024.1343627

Lin, P., Soto-Ferrari, M., & Chams-Anturi, O. (2022). A Logistic Regression Assessment to Measure Radiotherapy Clinical Pathway Concordance for Early Stages Breast Cancer Patients. *Procedia Computer Science*, *203*, 559–564. https://doi.org/10.1016/j.procs.2022.07.080

Mekha, P. (2021). *Image Classification o f Rice Leaf Diseases Using Random Forest Algorithm*. 165–169.

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *2020 11th International Conference on Information and Communication Systems, ICICS 2020*, 243–248. https://doi.org/10.1109/ICICS49469.2020.239556

Rachmawanto, E. H., Ignatius Moses Setiadi, D. R., Rijati, N., Susanto, A., Wahyu Mulyono, I. U., & Rahmalan, H. (2021). Attribute Selection Analysis for the Random Forest Classification in Unbalanced Diabetes Dataset. *Proceedings - 2021 International Seminar on Application for Technology of Information and Communication: IT Opportunities and Creativities for Digital Innovation and Communication within Global Pandemic, ISemantic 2021*, 82–86. https://doi.org/10.1109/iSemantic52711.2021.9573181

Ratmana, D. O., Shidik, G. F., Fanani, A. Z., & Pramunendar, R. A. (2020). *Evaluation of Feature Selections on Movie Reviews Sentiment*. 567–571.

Sadasivuni, K. K., Cabibihan, J.-J., A M Al-Ali, A. K., & Malik, R. A. (Eds.). (2022). *Advanced Bioscience and Biosystems for Detection and Management of Diabetes* (Vol. 13). Springer International Publishing. https://doi.org/10.1007/978-3-030-99728-1

Saleh, A. Y., & Brixtone Batou, B. (2022). Diabetes Mellitus Classification Using Hybrid Machine Learning With Stacking Technique. *2022 2nd International Conference on Emerging Smart Technologies and Applications (ESmarTA)*, 1–7. https://doi.org/10.1109/eSmarTA56775.2022.9935383

Setiyaningrum, Y. D. (2019). Classification of Twitter Contents using Chi-Square and K-Nearest Neighbour Algorithm. *2019 International Seminar on Application for Technology of Information and Communication (ISemantic)*, 1–4.

https://doi.org/10.1109/ISEMANTIC.2019.8884290

Shaukat, Z., Zafar, W., Ahmad, W., Haq, I. U., Husnain, G., Al-Adhaileh, M. H., Ghadi, Y. Y., & Algarni, A. (2023). Revolutionizing Diabetes Diagnosis: Machine Learning Techniques Unleashed. *Healthcare*, *11*(21), 2864. https://doi.org/10.3390/healthcare11212864

Singh, H. P., & Alhulail, H. N. (2022). Predicting Student-Teachers Dropout Risk and Early Identification : A Four-Step Logistic Regression Approach. *IEEE Access*, *10*, 6470–6482. https://doi.org/10.1109/ACCESS.2022.3141992

Tumuluru, P., Daniel, R., Mahesh, G., Lakshmi, K. D., Mahidhar, P., & Kumar, M. V. (2023). Class Imbalance of Bio-Medical Data by Using PCA-Near Miss for Classification. *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*, 1832–1839. https://doi.org/10.1109/ICIRCA57980.2023.10220757

Vanneschi, L., & Silva, S. (2023). Artificial Neural Networks. *Natural Computing Series*, 161–204. https://doi.org/10.1007/978-3-031-17922-8_7

Wijaya, S. H., Pamungkas, G. T., & Sulthan, M. B. (2018). *Improving Classifier Performance Using Particle Swarm Optimization on Heart Disease Detection*. 603–608.