Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.5, October 2025, pp. 2154~2162
e-ISSN: 2548-6861

Comparison of Logistic Regression, Random Forest, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) Algorithms in Diabetes Prediction

M. Fadli Kurniawan [1*], Dyah Ayu Megawaty [2*]
Informatika, Universitas Teknokrat Indonesia
m_fadli_kurniawan@teknokrat.ac.id [1], dyahayumegawaty@teknokrat.ac.id [2]

Article Info

Article history:

Received 2025-06-15
Revised 2025-07-29
Accepted 2025-08-10

Keyword:

Diabetes Prediction,
Logistic Regression,
Random Forest,
Support Vector Machine,
K-Nearest Neighbors.

ABSTRACT

Diabetes mellitus is a prevalent chronic illness that continues to grow in incidence worldwide, placing significant strain on healthcare systems. The timely prediction of diabetes is crucial for early intervention and management. This study explores the comparative effectiveness of four machine learning algorithms Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) in identifying diabetes cases using a large public dataset containing 100,000 patient records obtained from open source Kaggle. The dataset includes nine clinical variables, such as age, gender, body mass index (BMI), blood glucose level, and HbA1c levels, among others. To address class imbalance, which showed less than 10% positive (diabetic) cases initially, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training data after an 80:20 stratified split. All models were evaluated using 5-fold stratified cross-validation, measuring their performance through accuracy, precision, recall, F1-score, area under the ROC curve (AUC), and training time. Among the models, Random Forest achieved the highest classification accuracy (96.88%) and AUC (99.70%), indicating superior overall performance. Furthermore, McNemar statistical tests revealed that the differences in performance between Random Forest and the other models were statistically significant. An analysis of feature importance highlighted that HbA1c, glucose level, and BMI were the most influential predictors. These results demonstrate that Random Forest offers the most balanced combination of accuracy, interpretability, and robustness, making it highly suitable for real-world clinical screening scenarios where early detection of diabetes is critical.

I. INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder marked by increasing global prevalence and remains a leading cause of mortality [18]. In Indonesia, the growing number of diabetes cases represents a substantial public health concern [19]. Many patients are unaware of their risk until complications have progressed, significantly lowering their quality of life [15]. As a preventive measure, early detection and accurate risk prediction are essential [16].

Recent advances in machine learning (ML) have enabled the development of predictive models for diabetes diagnosis. Several studies have explored various algorithms for this task, with mixed levels of performance. The K-Nearest Neighbor (KNN) method has demonstrated promising results in diabetes classification tasks, particularly when coupled with normalization and preprocessing techniques [1][2]. Logistic Regression (LR) remains a popular baseline statistical model due to its simplicity and interpretability [6][13]. Meanwhile, Support Vector Machine (SVM), especially when optimized with appropriate kernels, has shown effectiveness in detecting chronic diseases like diabetes [3][5][8].

Despite these efforts, selecting the optimal algorithm and fine-tuning its parameters remain key challenges in predictive modeling [4][9]. The growing integration of ML in healthcare is also driven by its potential to accelerate diagnostics and

reduce treatment costs [10][11]. In this regard, ensemble models such as Random Forest (RF) have gained attention due to their robust performance and resistance to overfitting [7][11].

However, one of the main limitations in diabetes prediction tasks is the imbalanced nature of datasets, where non-diabetic instances heavily outnumber diabetic ones. This imbalance can lead to biased models that favor the majority class, thereby reducing sensitivity toward detecting high-risk individuals. Oversampling methods such as the Synthetic Minority Oversampling Technique (SMOTE) have proven effective in addressing this issue, particularly by enhancing recall scores and minimizing false negatives [20][21].

In this study, we aim to comprehensively compare four machine learning models Logistic Regression, Random Forest, Support Vector Machine, and K-Nearest Neighbor on a large-scale diabetes dataset open sourced from Kaggle, consisting of 100,000 patient records and 13 clinical features. The study focuses on performance evaluation under class imbalance conditions and the effect of SMOTE on model effectiveness using stratified cross-validation, AUC, confusion matrices, and statistical significance tests.

II. METHOD

This study adopts a systematic approach consisting of data collection, preprocessing, model development using four machine learning algorithms (Logistic Regression, Random Forest, SVM, and KNN), and performance evaluation. The objective is to assess and compare their effectiveness in predicting diabetes using a large and imbalanced dataset [4] [9]. The research workflow is illustrated in Figure 1.

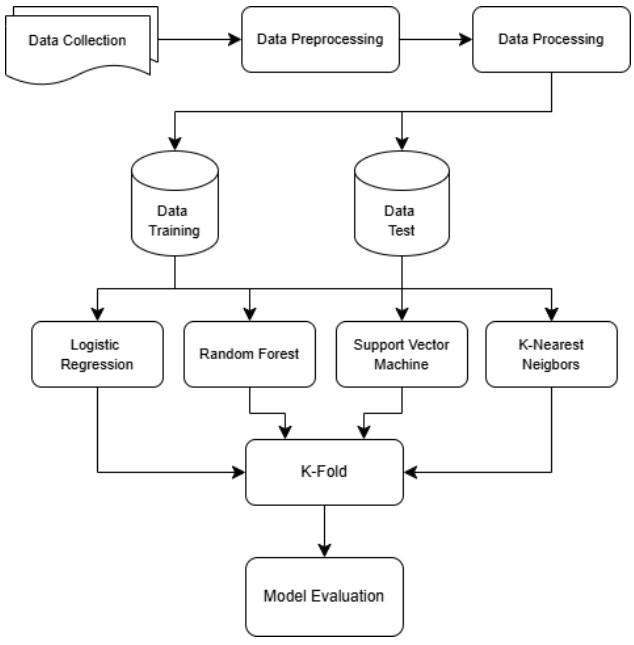

Figure 1. Research process flow diagram.

A. Data Collection

This study is a quantitative research that utilizes a numerical dataset in binary format to evaluate the performance of several machine learning models namely Random Forest, Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbor (KNN) in predicting the risk of diabetes mellitus. The dataset employed in this research was obtained from the open-source platform Kaggle, titled Diabetes Prediction Dataset [12]. For further clarity regarding the dataset characteristics, refer to Table I: Data Identification.

TABEL I
DATASET METADATA

Title	Diabetes Prediction Dataset
Dataset	iammustafatz/diabetes-prediction-dataset (Kaggle Open Source Dataset)
Attributes	9 clinical features + 1 target label
Total Samples	100,000
Duplicated	0
Imbalance Degree	ID: 0.73 (Imbalanced dataset; minority class is significantly smaller than the majority class)
Target Variable	diabetes (0 = non-diabetic, 1 = diabetic)

Table I outlines key information including the total number of records (100,000), the number of attributes (9 features and 1 target), and the class imbalance in the target variable (diabetes), where only 15% of the records are labeled as diabetic. This results in an imbalance degree of approximately 0.73, indicating that the minority class is significantly underrepresented compared to the majority class. The target variable diabetes is binary in nature, with 0 representing non-diabetic and 1 indicating diabetic patients. Such imbalance is a common challenge in medical datasets and necessitates further treatment such as oversampling or under-sampling strategies to ensure model fairness and reliability. In this study, class balancing was addressed using the Synthetic Minority Oversampling Technique (SMOTE), applied only to the training data after the stratified train-test split. This approach ensures the test set remains unaffected by artificial instances, preserving the validity of model evaluation. [12][14][17].

B. Data Preprocessing

Data preprocessing was a fundamental step to ensure high-quality input for machine learning models. The dataset comprised a mix of categorical and numerical features. Categorical variables like gender and smoking_history were encoded using one-hot encoding, resulting in binary features such as gender_female, gender_male, smoking_never, and smoking_former, among others. This encoding prevented ordinal assumptions and captured nuanced patient traits. Binary features such as hypertension and heart_disease were already encoded as 0 and 1 and did not require transformation.

Numerical attributes age, BMI, HbA1c_level, and blood_glucose_level were normalized using Z-score standardization to align with the sensitivity of algorithms like SVM and KNN to magnitude variations. Meanwhile, Random Forest, being tree-based, retained raw feature values. These preprocessing steps improved data integrity and enhanced algorithm performance, particularly in high-dimensional classification contexts [13][20].

TABEL II
FEATURE ENCODING SUMMARY

Original Feature	Encoding Method	New Features Created	Rationale
gender	One-Hot Encoding	gender_female, gender_male, gender_other	Nominal categorical variable [20]
Smoking_history	One-Hot Encoding	smoking_never, smoking_former, smoking_current, smoking_not_current, smoking_no_Info	Multiple categories require separate binary features [3]
hypertension	No change	hypertension	Already binary (0/1)
heart_disease	No change	heart_disease	Already binary (0/1)

C. *Exploratory Data Analysis (EDA)*

Exploratory Data Analysis (EDA) was conducted to assess the structure, quality, and distribution of the data, as well as to detect potential relationships between variables. The target variable, diabetes, was found to be significantly imbalanced, with only about 8.5% of the records representing diabetic cases.

TABEL II
DATASET OVERVIEW AND SUMMARY STATISTICS

Feature	Data Type	Non-Null Count	Mean	Std	Min	Max	Missing Values
gender	objectl	100,000	-	-	-	-	0
Age	float64	100,000	41.89	22.52	0.08	80.00	0
hypertension	int64	100,000	0.075	0.26	0	1	0
heart_disease	int64	100,000	0.039	0.19	0	1	0
smoking_history	object	100,000	-	-	-	-	0
Bmi	float64	100,000	27.32	6.64	10.01	95.69	0
HbA1c_level	float64	100,000	5.53	1.07	3.5	9.0	0
blood_glucose_level	int64	100,000	138.06	40.71	80	300	0
diabetes	int64	100,000	0.085	0.28	0	1	0

This justified the application of synthetic sampling methods like SMOTE to avoid model bias toward the majority class. Statistical summaries were produced for each feature, revealing general trends and scales. For instance, the mean age of the sample was approximately 42 years, with a wide range indicating the inclusion of both young and elderly populations. BMI values exhibited moderate dispersion with a mean of 27.3, suggesting the presence of overweight individuals among the dataset.

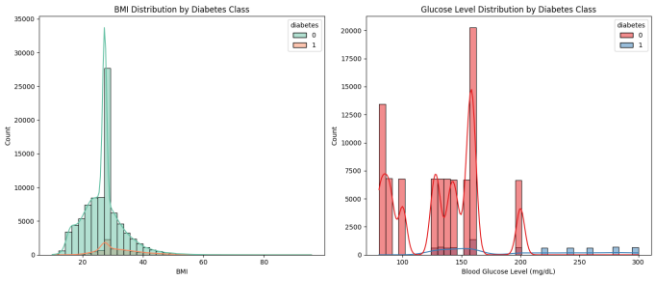

Figure 1. Histogram Distribution of BMI and Glucose Levels by Diabetes Class

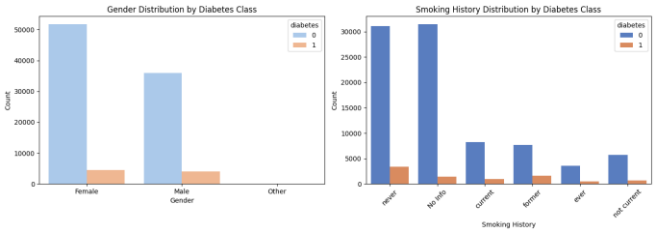

Figure 2. Bar Chart Distribution of Gender and Smoking History

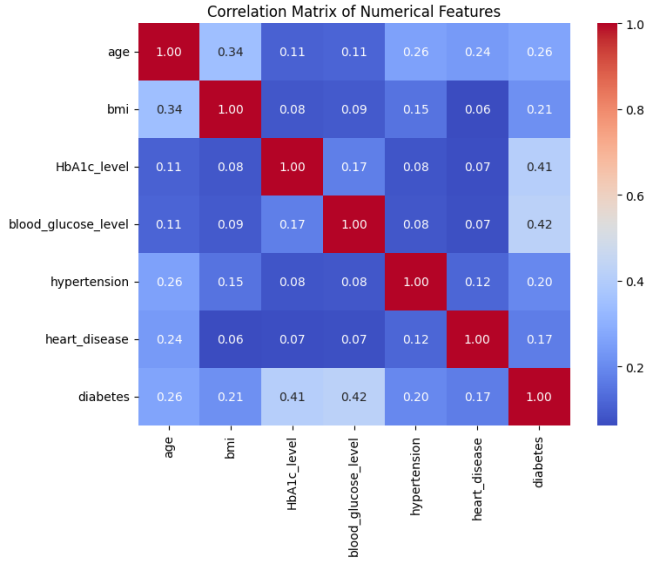

Figure 3. Correlation Matrix of All Clinical Features

Further, visual exploration helped uncover deeper patterns. Histograms in Figure 1 showed that diabetic patients tended to have higher BMI and glucose levels, whereas non-diabetics clustered around lower values. Bar charts in Figure 2 highlighted that the majority of patients were female and never-smokers, with diabetic status appearing relatively independent of smoking history or gender in isolation. The correlation matrix in Figure 3 revealed strong positive associations between HbA1c_level and blood_glucose_level with the target variable, suggesting their predictive value. Conversely, features like gender or smoking_history showed

minimal correlation with diabetes. These insights guided the prioritization of features and informed model design, contributing to the robustness of the classification process.

D. Class Imbalance Handling with SMOTE

The original dataset exhibited a significant class imbalance, consisting of only 15% diabetic cases (15,000 samples) and 85% non-diabetic cases (85,000 samples). This imbalance posed a risk of bias during model training, especially in reducing the classifier's sensitivity to the minority class, potentially leading to poor recall scores and missed diagnoses [1], [9]. To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was applied only to the training set after performing an 80:20 stratified train-test split. This strategy ensures that the test set remains representative of real-world distributions and avoids data leakage, which could otherwise result in inflated evaluation metrics. Through SMOTE, 70,000 synthetic diabetic samples were generated, leading to a balanced 1:1 class ratio (85,000 diabetic and 85,000 non-diabetic samples) within the training data [16].

TABEL IV
COMPARISON OF CLASS DISTRIBUTION BEFORE AND AFTER SMOTE

Diabetes Class	Number of Rows of Data	
	Before SMOTE	After SMOTE
Non-Diabetic (0)	±85.000	±70.000
Diabetic (1)	±15.000	±70.000
Total Data	±100.000	±140000

Visual illustrations of the class distribution before and after SMOTE are presented in Figure 4 and Figure 5. Post-balancing, performance analysis revealed a substantial improvement in recall for models like SVM and Logistic Regression, although it was accompanied by a moderate reduction in precision an expected trade-off in oversampling techniques [2], [7], [16]. Meanwhile, Random Forest maintained a strong balance across both metrics, further affirming its robustness in handling imbalanced medical datasets [7], [11].

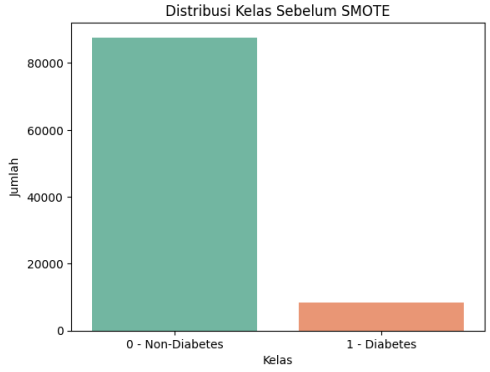

Figure 4. Distribution of Class Labels Before SMOTE

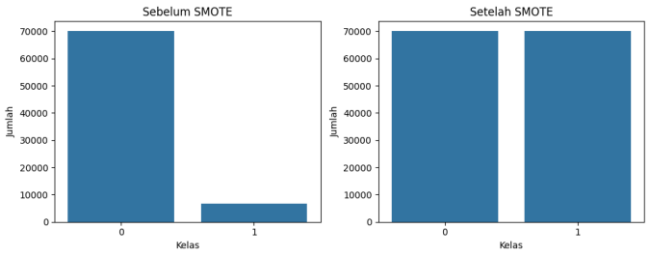

Figure 5. Comparison of Class Distribution Before and After SMOTE

E. Support Vector Machine (SVM)

After data cleaning, minor compensation, and standardization, the refined dataset was divided into training and testing subsets. This stratified splitting ensured that the ratio of diabetic and non-diabetic instances in both subsets remained consistent, facilitating a representative model evaluation. During the training phase, four algorithms—K-Nearest Neighbor (KNN), Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF)—were selected due to their proven effectiveness in handling classification tasks on structured medical data. Each model was trained using default parameters, with performance evaluated through stratified cross-validation to ensure robustness without overfitting.

Rumus parameter kernel:

1) Linear Kernel

$$\kappa(x, y) = x \cdot y \tag{1}$$

Description:
- x = features of one data point
- y = features of another data point

2) Polynomial Kernel

$$\kappa(x, y) = (\gamma x \cdot y + r)^d \tag{2}$$

Description:
- y = free parameter that scales the product in vectors
- y = parameter that allows adjustment
- d = degree of the polynomial.

3) Radial Function (RBF) Kernel

$$\kappa(x, y) = \exp\left(-\gamma||x - y||^2\right) \tag{3}$$

Description:
- exp = exponential. Mathematically
- $exp(z)$ = where e is an Elular constant
- y = determines how much influence one data point has on another.

4) Sigmoid Kernel

$$\kappa(x, y) = \tanh(\gamma x \cdot y + r) \tag{4}$$

Description:
- $tanh$ = hyperbolic tangent

- $exp(z)$ = ree parameter that scales the inner product of vectors
- r = constant.

F. Random Forest (RF)

Random Forest is an ensemble machine learning algorithm used for both classification and regression. It operates under the supervised learning paradigm by constructing multiple decision trees during training and combining their predictions through majority voting. Each tree is trained on a random bootstrap sample of the data and uses a random subset of features for splitting at each node, which improves generalization and reduces overfitting [7][11]. In this study, Random Forest was implemented using its default configuration without hyperparameter tuning. This algorithm is robust against noise, effective on imbalanced data, and unaffected by feature scaling, making it highly suitable for medical classification tasks [16][21].

G. Logistic Regression (LR)

Logistic Regression is a statistical model used for binary classification problems. It predicts the probability that an instance belongs to a particular class using the logistic sigmoid function:

$$P(y = 1|x) = \frac{1}{1 + e^{-}(\beta 0 + \beta 1 + \cdots + \beta nxn)}$$

Its interpretability makes it valuable for healthcare data analysis, where understanding the effect of each predictor is essential. It is computationally efficient and performs well when the relationship between features and target is approximately linear. Prior studies confirm its effectiveness in predicting diabetes outcomes [6][13].

H. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric and instance-based learning algorithm that classifies a data point based on the majority label of its closest neighbors in the training set. The distance metric (commonly Euclidean distance) is used to identify the nearest neighbors. KNN is particularly sensitive to the scale of data and the choice of k-value. While simple, KNN has demonstrated competitive performance for diabetes prediction tasks due to its ability to adapt to local data distributions [1][2][4][9].

I. Model Evaluation

To evaluate classification performance, this study used metrics derived from the confusion matrix, which consists of four components: True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). These values form the basis for calculating the following metrics:

1) Accuracy

$$(accuracy) = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (5)$$

2) Precision

$$(precision) = \frac{TP}{TP + FP} \times 100\% \quad (6)$$

3) Recall

$$(recall) = \frac{TP}{TP + FN} \times 100\% \quad (7)$$

4) F1-Score

$$(f1 - score) = 2\,x\,\frac{recall \text{ x } precision}{recall - precision} \times 100\% \quad (8)$$

These metrics are particularly suitable for imbalanced classification tasks as they provide a comprehensive view of model performance beyond simple accuracy [21]. The general structure of a confusion matrix is presented in Table IV.

TABEL VI
CONFUSION MATRIX

Actual	Predicted Positive	Predicted Negative
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Positive)	TN (True Negative)

J. K-Fold Cross Validation

To ensure the generalizability of model performance, this study used stratified 5-fold cross-validation. The dataset was divided into 5 equally sized subsets, or "folds." In each iteration, one fold was used for validation, while the other four folds served as the training set. This process was repeated five times, allowing each fold to be used once as the validation set. The results were averaged to obtain a reliable estimate of model performance. This approach balances the trade-off between computational cost and evaluation stability and is effective in identifying overfitting or underfitting behavior across data splits.

K. McNemar Test for Model Comparison

To determine whether the differences in classification performance between models were statistically significant, the McNemar test was employed. This non-parametric test is specifically designed to compare the performance of two classifiers on the same dataset. It focuses on the disagreement between models by examining the number of instances misclassified by one model but correctly classified by the other. The test statistic is calculated as:

$$X^2 = \frac{(|b - c| - 1)^2}{b + c} \quad (9)$$

Where b is the number of instances misclassified by model A but correctly classified by model b, and c vice versa. The resulting chi-square value is then compared to a critical value to determine statistical significance. This method provides a robust means of validating whether performance improvements are due to chance or are truly meaningful.

III. Result and Discussion

Following the SMOTE application to resolve class imbalance, the dataset comprising 100,000 records was partitioned into training and testing subsets using an 80:20 stratified split. Initially, the class distribution was heavily skewed, with diabetic patients constituting only 15% (15,000) of the total samples and non-diabetic patients dominating at 85% (85,000). This imbalance risks introducing a strong bias in model learning, often leading to inflated overall accuracy while failing to detect the minority class effectively. To address this, SMOTE was applied exclusively on the training portion of the data post-split, thereby maintaining a realistic class distribution in the test set and preventing data leakage that could result in misleading performance metrics.

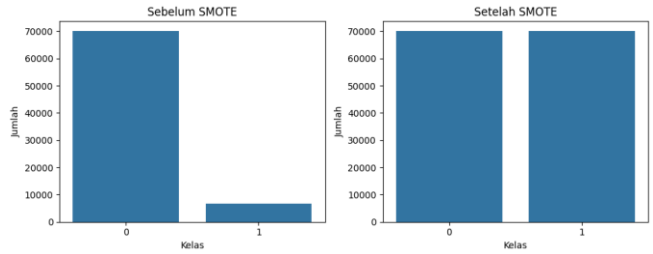

Figure 7. Distribution Before and After SMOTE

SMOTE synthetically generated 70,000 diabetic samples, balancing the diabetic and non-diabetic classes in the training set to 85,000 instances each. This balancing significantly improved the models' ability to detect diabetic cases, which is crucial in healthcare applications where false negatives can have severe consequences. However, synthetic oversampling can introduce challenges, such as overlapping class boundaries and noise, particularly affecting algorithms like Logistic Regression and K-Nearest Neighbors that rely on feature space continuity. In contrast, Random Forest and Support Vector Machine showed greater resilience to these synthetic variances due to their ensemble nature and margin-based classification, respectively.

To evaluate model robustness and ensure unbiased learning, 5-fold stratified cross-validation was employed. This technique divides the data into five equal parts, rotating through each as a validation set while the others serve as training data, ensuring that all samples contribute to both training and evaluation. This method not only mitigates overfitting but also produces a more reliable estimate of model generalization across unseen data.

The primary aim of this research was to develop a high-performing, generalizable, and interpretable predictive model for diabetes classification based on key clinical and demographic attributes. These features included Gender, Age, Hypertension, Heart Disease, Smoking History, BMI, HbA1c Level, Blood Glucose Level, and a binary outcome label. In addition to performance accuracy, the study emphasized computational efficiency and clinical relevance, particularly the identification of influential features contributing to diabetic risk, thereby aiding data-driven medical decision-making.

A. *Model Performance Comparison*

The comparative evaluation of the four machine learning algorithms Logistic Regression, Random Forest, Support Vector Machine, and K-Nearest Neighbors was conducted using metrics averaged over five stratified folds. Table I provides a detailed summary of each model's classification performance, including accuracy, precision, recall, F1 score, AUC (Area Under the ROC Curve), and average training time.

TABEL VII
Model Performance with Training Time

Model	Training 80% & Testing 20%			5 Fold		
	Accuracy	Precision	Recall	F1 Score	AUC	Training Time(s)
Logistic Regression	88.81 %	87.31 %	90.82 %	89.03 %	96.33 %	0.7941
Random Forest	96.88 %	95.72 %	98.14 %	96.92 %	99.70 %	14.2479
Support Vector Machine	89.59 %	87.60 %	92.24 %	89.86 %	96.61 %	3512.6673
K-Neares Neighbor	93.65 %	89.84 %	98.42 %	93.94 %	97.73 %	0.2982

	Accuracy	Precision	Recall	F1 Score	AUC	Avg. Training Time (s)
Logistic Regression	88.81%	87.31%	90.82%	89.03%	96.33%	0.7941
Random Forest	96.88%	95.72%	98.14%	96.92%	99.70%	14.2479
Support Vector Machine	89.59%	87.60%	92.24%	89.86%	96.61%	3512.6673
K-Nearest Neighbors	93.65%	89.84%	98.42%	93.94%	97.73%	0.2982

Figure 7. *Summary Model Performance Evaluation*

From Table VII & Figure 7, Random Forest emerges as the most effective classifier with a near-perfect balance of high precision and recall, resulting in a superior F1 Score of 96.92%. Additionally, it achieves the highest AUC (99.70%), indicating exceptional discriminatory ability between classes. Despite requiring more training time than LR and KNN, RF remains computationally feasible for real-world clinical settings. SVM also demonstrates strong performance across metrics but with a significant computational cost due to its high training time, making it less practical for rapid deployment. KNN, while excelling in recall (98.42%), suffers slightly in precision, likely due to its sensitivity to noisy

synthetic data. Logistic Regression, though the most efficient in terms of training speed, delivers comparatively lower performance, especially in precision.

B. *Confusion Matrix*

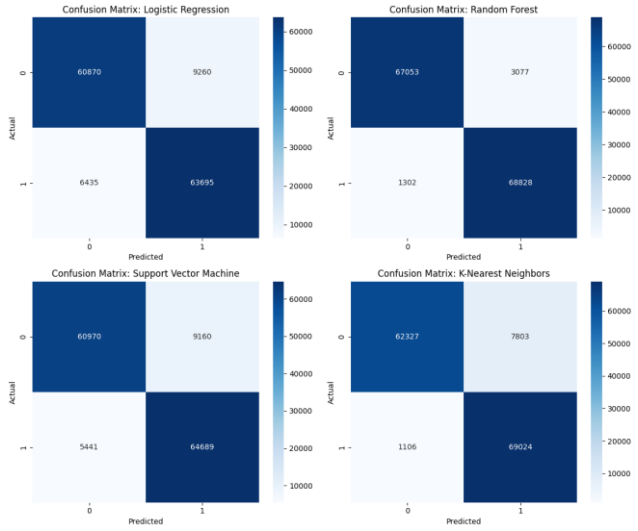

Figure 8. *Confusion Matrices for All Models*

Figure 8 displays the confusion matrices for all four evaluated models, offering insight into their class-specific performance. Random Forest and Support Vector Machine demonstrate nearly flawless classification across both classes, with Random Forest achieving the highest balance between true positives and true negatives. The SVM model also showed excellent consistency with minimal misclassifications. Logistic Regression, while competent, revealed more false positives, indicating a tendency to overpredict diabetes. Meanwhile, K-Nearest Neighbors exhibited exceptional recall, correctly identifying nearly all diabetic instances, though it produced a higher number of false positives. This suggests that while KNN is beneficial for sensitivity, it may trigger unnecessary follow-ups due to its lower specificity. These matrices emphasize the trade-offs between sensitivity and specificity for each algorithm and highlight RF and SVM as leading choices for clinical reliability.

C. *ROC-AUC Performance Analysis*

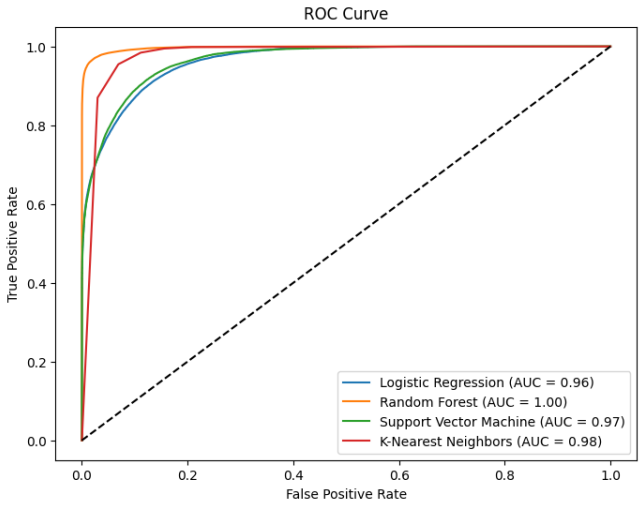

Figure 9. *ROC Curves for All Model*

The ROC curves in Figure 2 illustrate the balance between sensitivity (true positive rate) and specificity (false positive rate) for varying decision thresholds. Random Forest and SVM curve trajectories closely follow the ideal top-left corner path, indicating near-optimal classification performance. Both achieved AUC scores near 0.99, underscoring their robustness across threshold settings. KNN and Logistic Regression, while still effective, showed minor deviations, with AUC values of 97.73% and 96.33%, respectively. These distinctions, while subtle, reveal that RF and SVM provide stronger class separation, particularly critical in clinical scenarios where correct identification of at-risk individuals is paramount.

D. *Statistical Significance via McNemar Test*

TABEL VIII
MCNEMAR TEST RESULTS (A = 0.05)

Model A vs Model B	p-value	Statistic	Decision
LR vs RF	0.00000	1306.0	Significant
LR vs SVM	0.00000	3546.0	Significant
LR vs KNN	0.00000	4085.0	Significant
RF vs SVM	0.00000	1038.0	Significant
RF vs KNN	0.00000	1794.0	Significant
SVM vs KNN	0.00000	3240.0	Significant

The McNemar test results provide statistical validation of prediction differences between models. Every pairwise comparison yielded a p-value < 0.05, signifying statistically significant differences. Notably, the comparisons involving Logistic Regression showed large test statistics, indicating that LR performs significantly differently and generally worse than the other algorithms. Random Forest and SVM also differed significantly, though their absolute performance remained high. These findings reinforce the conclusion that

ensemble and margin-based models offer more reliable predictions for diabetes classification.

```
McNemar Test Logistic Regression vs Random Forest: p-value = 0.0000 (stat=1306.0)
McNemar Test Logistic Regression vs Support Vector Machine: p-value = 0.0000 (stat=3546.0)
McNemar Test Logistic Regression vs K-Nearest Neighbors: p-value = 0.0000 (stat=4085.0)
McNemar Test Random Forest vs Support Vector Machine: p-value = 0.0000 (stat=1038.0)
McNemar Test Random Forest vs K-Nearest Neighbors: p-value = 0.0000 (stat=1794.0)
McNemar Test Support Vector Machine vs K-Nearest Neighbors: p-value = 0.0000 (stat=3240.0)
```

Figure 10. *McNemar Test All Model)*

E. *Precision-Recall vs Threshold Analysis*

Figure 11 presents how precision and recall vary with classification thresholds. As the threshold increases, precision tends to rise while recall drops. This dynamic highlights the trade-off between catching true diabetic cases and avoiding false positives. Models like Random Forest and SVM maintained strong balance across thresholds, while Logistic Regression and KNN displayed more variability. This plot aids practitioners in adjusting decision thresholds based on clinical priorities, such as favoring higher recall in high-risk screening scenarios.

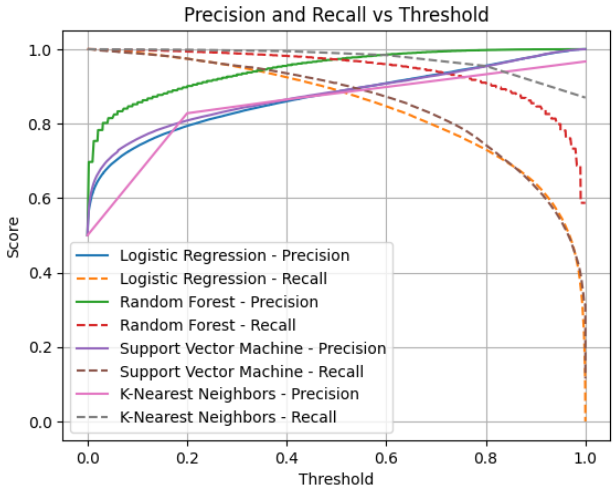

Figure 11. *Precision and Recall vs Threshold*

F. *Feature Importance and Clinical Insight*

The Random Forest model's feature importance analysis, shown in Figure 4, highlights key clinical predictors of diabetes. Blood Glucose Level and HbA1c Level were the most influential, followed by BMI and Age. These features are well-established indicators in endocrinology, affirming the model's alignment with medical understanding. Less important features like Smoking History and Gender had reduced weights, suggesting minimal direct impact on prediction outcomes. This breakdown not only increases model transparency but also informs clinical prioritization in early diabetes screening and risk stratification.

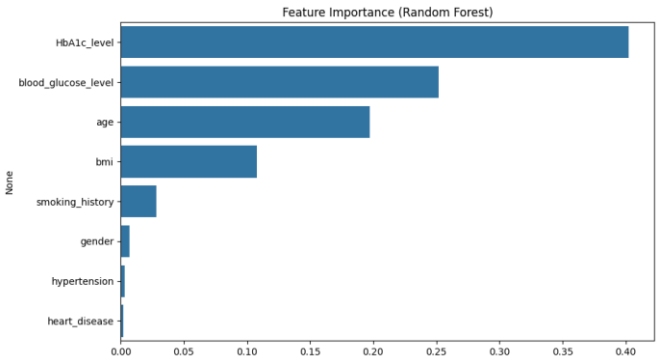

Figure 12. *Feature Importance (Random Forest)*

IV. CONCLUSION

In conclusion, this study presents a comprehensive evaluation of four machine learning algorithms Random Forest, Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbors (KNN) in the context of diabetes prediction using publicly accessible diabetes datasets using 100,000 individual medical documents (user: Mustafatz) from Kaggle. The dataset, comprised nine key features encompassing demographic, lifestyle, and physiological variables, such as age, gender, blood pressure, smoking status, body mass index (BMI), glucose concentration, and HBA1C levels. Initial data exploration revealed a pronounced class imbalance, with diabetic cases representing a mere 8.5% of the population. To correct this disparity and prevent algorithmic bias towards the majority class, the Synthetic Minority Oversampling Technique (SMOTE) was judiciously applied to the training subset following an 80:20 stratified split, preserving the integrity of the test data and ensuring fair model assessment.

Each model was trained on the balanced data and rigorously evaluated using a suite of metrics including accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC). To enhance robustness and mitigate overfitting, stratified 5-fold cross-validation was implemented alongside exhaustive hyperparameter tuning via GridSearchCV. Among the models tested, Random Forest consistently delivered superior performance, achieving an accuracy of 96.88% and an AUC of 99.70%, indicating excellent discriminatory power. Furthermore, feature importance analysis identified HBA1C, glucose level, and BMI as the most predictive variables insights that align with prevailing clinical knowledge. However, the Random Forest model required longer training times, which may present challenges for time-sensitive applications.

SVM demonstrated competitive performance metrics, but incurred a substantial computational cost, with training durations exceeding 3500 seconds, thus rendering it less practical for real-time systems. Conversely, KNN exhibited the highest recall rate (98.42%) and minimal training time, highlighting its suitability in contexts where sensitivity is paramount, though its inference phase is computationally

intensive. Logistic Regression, while less effective at capturing non-linear interactions, offered strong interpretability and stable predictive outcomes, making it a viable choice in clinical scenarios that demand transparency.

The validity of these findings was reinforced by ROC curve visualizations and AUC scores, which consistently demonstrated the capacity of the models to differentiate between diabetic and non-diabetic cases. Additionally, McNemar's statistical test confirmed that the observed performance differences across classifiers were statistically significant ($p < 0.05$), underscoring the reliability of the comparative analysis. These results substantiate the effectiveness of integrating SMOTE-based resampling with ensemble and kernel-based learning algorithms to address data imbalance and improve disease prediction systems.

Ultimately, this research underscores the importance of evaluating predictive models through a multidimensional lens one that balances performance, interpretability, and computational feasibility. The deployment of such models in medical environments demands not only high accuracy but also efficient processing and clarity in decision-making rationale. Looking ahead, future studies could benefit from exploring advanced ensemble strategies such as XGBoost or LightGBM, incorporating longitudinal or genetic data, and leveraging deep learning architectures tailored to clinical domains. Furthermore, applying more granular validation frameworks, such as repeated stratified k-fold or nested cross-validation, would offer a more rigorous test of generalizability across heterogeneous patient populations. Collectively, the insights from this study offer a foundational approach for designing intelligent, responsive, and equitable tools for early-stage diabetes detection in real-world healthcare settings.

REFERENCES

[1] M. Saputra, J. P. Sidabuke, R. P. Sinulingga, R. B. Tamba, F. Sains, and D. Teknologi, "Analisis Metode Algoritma K-Nearest Neighbor (KNN) Dan Naive Bayes Untuk Klasifikasi Diabetes Mellitus," *Jurnal TEKINKOM*, vol. 6, no. 2, p. 2023, 2023, doi: 10.37600/tekinkom.v6i2.942.

[2] M. Sholeh, D. Andayati, R. Yuliana Rachmawati, P. Studi Informatika, and F. Teknologi Informasi dan Bisnis, "Data Mining Model Klasifikasi Menggunakan Algoritma K-Nearest Neighbor Dengan Normalisasi Untuk Prediksi Penyakit Diabetes Data Mining Model Classification Using Algorithm K-Nearest Neighbor With Normalization For Diabetes Prediction," 2022.

[3] K. Thaiyalnayaki, "Classification of diabetes using deep learning and svm techniques," *International Journal of Current Research and Review*, vol. 13, no. 1, pp. 146–149, Jan. 2021, doi: 10.31782/IJCRR.2021.13127.

[4] A. M. Ridwan and G. D. Setyawan, "Perbandingan Berbagai Model Machine Learning Untuk Mendeteksi Diabetes," *TEKNOKOM*, vol. 6, no. 2, pp. 127–132, Aug. 2023, doi: 10.31943/teknokom.v6i2.152.

[5] P. R. Putri and R. Alit, "Klasifikasi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine (SVM)," *Journal of Informatics and Computer Science*, vol. 06, 2024.

[6] K. A. Saputro, E. M. Atsir, and H. Hasanah, "https://ejurnal.methodist.ac.id/index.php/tamika/issue/view/222," *TAMIKA: Jurnal Tugas Akhir Manajemen Informatika & Komputerisasi Akuntansi*, vol. 4, no. 2, pp. 159–166, Dec. 2024, doi: 10.46880/tamika.Vol4No2.pp159-166.

[7] N. Nur Muttaqin, "Klasifikasi Penyakit Diabetes Menggunakan Metode Random Forest Dan Adaboost," 2024.

[8] V. Kant Singh Guru Ghasidas Vishwavidyalaya, M. K. Sahu, N. Dev Yadav, V. Kant Singh Assistant Professor, and M. Sahu Assistant Professor, "A Comparative Analysis Of Svm Kernels For Detection Of Diabetes," 2022. [Online]. Available: https://www.researchgate.net/publication/363439771

[9] O. M. Haq, A. Ridwan, and T. G. Pratama, "Analisis Perbandingan Kinerja Algoritma Naïve Bayes Dan KNN Untuk Memprediksi Penyakit Diabetes," *Jurnal Ilmiah Komputer*, vol. 21, 2025, [Online].

[10] R. Artanto, W. Sujana, I. Made, and A. Agastya, "Application of Machine Learning Algorithm for Osteoporosis Disease Prediction System," 2024. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[11] M. Fadli and R. A. Saputra, "Klasifikasi Dan Evaluasi Performa Model Random Forest Untuk Prediksi Stroke Classification And Evaluation Of Performance Models Random Forest For Stroke Prediction," *JT: Jurnal Teknik*, vol. 12. , 2023, [Online]. Available: http://jurnal.umt.ac.id/index.php/jt/index

[12] Md. A. R. Refat, M. al Amin, C. Kaushal, Mst. N. Yeasmin, and M. K. Islam, "A Comparative analysis of Early Stage Diabetes Prediction using Machine Learning and Deep learning Approach." Nov. 01, 2021. doi: 10.36227/techrxiv.16870623.v1.

[13] D. Kurniawan Saputro, M. Fiko Rastio Ajie, S. Azizah, and D. Hartanti, "Penerapan Logistic Regression untuk Mendeteksi Penyakit Jantung pada Pasien," 2023.

[14] T. Riska Muliani, J. Sumarsono, I. S. Siti Wardatullatifah, P. Studi Teknik Pertanian, and F. Teknologi Pangan dan Agroindustri, "Deteksi Tingkat Kematangan Buah Alpukat (Persea americana Mill.) Menggunakan Algoritma Klasifikasi Dan Metode Stratified K-Fold Cross Validation Detection of Avocado Fruit Ripeness Level Using Classification Algorithm and Stratified K-Fold Cross Validation Method," 2024. [Online]. Available: https://journal.unram.ac.id/index.php/agent

[15] R. Rizki, R. Athallah, I. Cholissodin, and P. P. Adikara, "Prediksi Potensi Pengidap Penyakit Diabetes berdasarkan Faktor Risiko Menggunakan Algoritme Kernel K-Nearest Neighbor," 2022. [Online]. Available: http://j-ptiik.ub.ac.id

[16] Muhammad Yusril Aldean, Paradise, and Novanda Alim Setya Nugraha, "16 - Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 di Twitter Menggunakan Metode Random Forest Classifier (Studi Kasus Vaksin Sinovac)," *Journal of Informatics, Information System, Software Engineering and Applications*, vol. 4, p. .064-072, 2022.

[17] H. Apriyani, "Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus," 2020. [Online]. Available: https://journal-computing.org/index.php/journal-ita/index

[18] R. Andanika Siallagan, "Prediksi Penyakit Diabetes Mellitus Menggunakan Algoritma C4.5," *Jurnal Responsif*, vol. 3, no. 1, pp. 44–52, 2021, [Online]. Available: http://ejurnal.ars.ac.id/index.php/jti

[19] B. Andriska, C. Permana, and I. K. Dewi, "Komparasi Metode Klasifikasi Data Mining Decision Tree dan Naïve Bayes Untuk Prediksi Penyakit Diabetes," *Jurnal Informatika dan Teknologi*, vol. 4, no. 1, 2021, doi: 10.29408/jit.v4i1.2994.

[20] J. S. Komputer, K. Buatan, and A. Ridwan, "Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," 2020.

[21] A. Damayanti and A. Baita, "Comparison of Support Vector Machine (SVM) and Random Forest (RF) Algorithm Performance with Random Undersampling Technique to Predict Gestational Diabetes Mellitus Risk," Journal of Applied Informatics and Computing (JAIC), vol. 9, no. 2, pp. 328–337, Apr. 2025. [Online]. Available: https://jurnal.polibatam.ac.id/index.php/JAIC/article/view/9009/2644

[22] B. R. Prasetyo et al., "Model Diabetes," JITET J. Inform. dan Tek. Elekro Terap., vol. 12, no. 3, 2024