Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16133
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Multi-Metric Evaluation of Machine Learning Algorithms for Diabetes Prediction Using Feature Importance and ROC Analysis

Fendi Setiawan[1], Tri Sugihartono[2]
Informatics Engineering, Faculty of Information Technology, ISB Atma luhur
2211500078@mahasiswa.atmaluhur.ac.id , trisugihartono@atmaluhur.ac.id

Submitted :April 22, 2026 | **Accepted** : May 14, 2026 | **Published** : July 5, 2026

Abstract: Diabetes mellitus has become a major global health threat, and many undiagnosed cases remain undetected due to some limitations of the conventional diagnostic methods. Despite the promising results of machine learning (ML) for early diabetes diagnosis, the majority of the current research assessing algorithms either uses insufficient metrics or does not follow a consistent assessment approach. This paper addresses that gap by utilising an integrated evaluation framework. The framework includes feature importance analysis, Pearson correlation assessment, confusion matrix decomposition, and ROC-AUC comparison. It applies this framework to the Pima Indians Diabetes Dataset (mde) and four popular ML classification algorithms: Naive Bayes, Decision Tree, Random Forest, and Logistic Regression. The most significant predictors, according to our feature analysis, were glucose (27.6%), body mass index (16.0%), age (12.7%), and diabetes pedigree function (12.7%). Among the classifiers, Random Forest exhibited the greatest accuracy (76.0%) and precision (68.1%), Naive Bayes the best recall (64.8%), and Logistic Regression the highest AUC-ROC (82.3%). For patients at high risk, the models' virtual projections across all three risk profiles were in agreement. Model selection should be determined by the unique clinical screening aim, since these findings suggest that there is no one better universal method. Random Forest and Logistic Regression are the most promising for assisting in preliminary diabetes prediction, although further validation on diversity datasets is needed prior to clinical deployment.

Keywords: Diabetes Prediction; Machine Learning; Pima Indians Diabetes Dataset; Random Forest; Logistic Regression; AUC-ROC; Feature Importance.

INTRODUCTION

With an increase in incidence from 108 million in 1980 to 422 million by 2014 and around 1.4 million fatalities related to DM in 2019 alone, diabetes mellitus (DM) is a major noncommunicable illness globally (WHO). If not addressed, it may have serious consequences such as heart disease, renal failure, neuropathy, and retinopathy. Conventional diagnostics are time-consuming, labor-intensive, and costly, even if early detection is vital. Machine learning's (ML) capacity to extract intricate patterns from clinical datasets has made it a very attractive tool for enhancing medical diagnosis. Disease forecasting tasks have made extensive use of ML classification models, with performance assessed using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. However, studies usually only use one measure to evaluate algorithms, which makes it hard to compare how well they work in practice.

Smith et al. made the Pima Indians Diabetes Data Set freely available via the UCI Machine Learning Repository, and it is among the most utilised datasets for diabetes prediction. Glucose level, body mass index (BMI), age, blood pressure, insulin, skin thickness, number of pregnancies, and the diabetes pedigree function are eight clinical features that indicate the likelihood of a diabetes diagnosis. Previous studies have indicated that glucose and BMI always appear as the strongest predictors in different classification models. Overall, evidence supporting ML-based diabetes is prevalent with numerous gaps. Historically, many studies investigate accuracy only and ignore clinically relevant metrics like recall, FNR or AUC-ROC. In addition, there are few studies that combine feature importance analysis, correlation assessment, confusion matrix decomposition and ROC curve comparison within

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16133
e-ISSN : 2541-2019
p-ISSN : 2541-044X

a common experimental framework. Furthermore, high false negative rates are especially harmful in medical screening but this dimension is rarely addressed directly.

Consequently, the purpose of this research is to examine the Pima Indians Diabetes Dataset using four popular ML classification algorithms: Naive Bayes, Decision Tree, Random Forest, and Logistic Regression. The main advance of this research is the consolidation of several evaluation methods into a single, cohesive framework. These methods include multi-metric evaluation, feature importance analysis, correlation assessment, confusion matrix decomposition, AUC-ROC comparison, and multi-scenario prediction simulation. Compared to single-metric comparisons, this method allows for a more comprehensive and therapeutically useful evaluation. What follows is an outline of the rest of the paper: Section II gives a summary of the relevant literature, Section III lays out the steps used to conduct the study, Section IV offers the data from the experiments, Section V analyses the results, and Section VI draws all the necessary conclusions.

LITERATURE REVIEW

There has been a lot of research on using ML for illness prediction, and the results always show that classification-based techniques may help with clinical diagnosis. This section synthesizes the relevant literature according to four thematic threads: (1) studies supporting the dominance of ensemble methods such as Random Forest; (2) studies highlighting the advantages of Logistic Regression and AUC-based evaluation; (3) studies emphasizing the clinical importance of recall and false negative minimization; and (4) studies using feature importance or feature selection methodologies relevant to diabetes prediction.

Several studies have shown that ensemble approaches, and Random Forest in particular, are the best in predicting diabetes and other diseases. On the Pima dataset, Ghosh et al. (2021) assessed RF in conjunction with the MRMR feature selection method. They confirmed that glucose and body mass index are the most important predictors, and they reported an accuracy of 99.35% with near-perfect sensitivity. Like Random Forest, Zhao (2025) showed that on the same dataset, Logistic Regression, Decision Tree, SVM, Gradient Boosting, and KNN all performed worse in terms of accuracy. Ghosh et al. (2021) extended this finding to liver disease prediction, where RF achieved 83.76% accuracy with strong recall and AUC scores. Across domains, Pai et al. (2021) further confirmed that RF consistently outperformed other classifiers in multi-category classification tasks. The common thread across these studies is that ensemble tree methods reduce overfitting through variance averaging, which explains their generalisation advantage over single-tree models.

Several studies demonstrate that AUC-ROC rather than accuracy alone provides a more clinically meaningful evaluation criterion. Kangra and Singh (2023) compared six ML algorithms on the Pima dataset and found that while SVM achieved the highest accuracy (74.3%), Logistic Regression obtained a superior ROC area of 0.83, highlighting that accuracy alone is insufficient for medical classification. Sheth et al. (2022) evaluated four classifiers across five datasets and found that no single algorithm universally dominated, reinforcing the importance of multi-metric evaluation. These findings directly motivate the present study's use of both accuracy and AUC-ROC as co-primary evaluation criteria.

In medical screening contexts, recall and false negative rate are clinically more critical than overall accuracy, because missing a true positive (i.e., a diabetic patient) carries greater consequences than a false alarm. Ghosh et al. (2021) observed a precision-recall trade-off in liver disease prediction, where Decision Tree produced high recall but poor overall accuracy. This pattern is equally relevant to diabetes prediction, where classifiers with lower accuracy but higher sensitivity may be preferable for preliminary screening. The present study explicitly addresses this clinical trade-off by reporting and comparing false negative counts across all classifiers.

Feature importance and correlation analysis have been widely used to identify the most predictive clinical variables. Ghosh et al. (2021) confirmed Glucose and BMI as dominant predictors through MRMR feature correlation, a finding replicated across multiple Pima-based studies. Kangra and Singh (2023) demonstrated that different features contribute differently across datasets, underscoring the value of conducting feature analysis specific to the dataset and task at hand.

Based on this synthesis, several research gaps remain unaddressed. First, most studies evaluate algorithms using a single metric or without a unified evaluation framework that incorporates feature importance, correlation analysis, confusion matrix decomposition, and ROC curve comparison simultaneously. Second, few studies include multi-scenario prediction simulation to demonstrate model behaviour across varying risk profiles. Third, interpretable algorithms such as Naive Bayes and Logistic Regression are often excluded from comparisons dominated by advanced ensemble methods, leaving gaps in understanding how simpler models perform under equivalent experimental conditions. The present study addresses these gaps through a comprehensive, multi-metric evaluation framework designed to provide clinically actionable insights for diabetes screening.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16133
e-ISSN : 2541-2019
p-ISSN : 2541-044X

METHOD

The purpose of this work is to evaluate four different machine learning classification algorithms for diabetes prediction using a quantitative experimental method. Preparing the dataset, preparing the data, doing exploratory data analysis, training and testing the models, and finally, evaluating their performance are all sequential steps in the research technique. Figure 1 shows the whole research procedure.

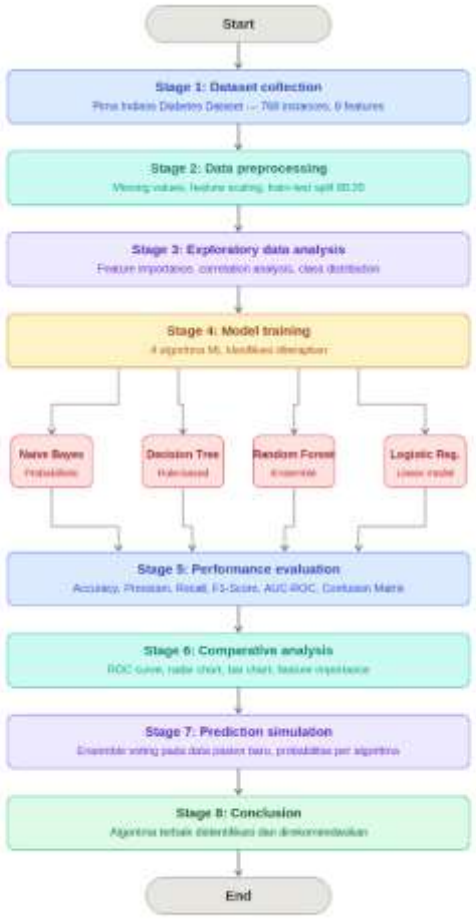

Figure 1. Research Workflow of Machine Learning Classification for Diabetes Prediction

Dataset

The Pima Indians Diabetes Dataset was retrieved from the UCI Machine Learning Repository, which is the original source. With 768 instances and 8 clinical input parameters, the dataset only includes one binary output variable that shows whether a patient has diabetes or not. Age, gestational age, blood sugar, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin level, BMI, function of the diabetic pedigree, and plasma glucose concentration are the eight elements that need to be taken into account. Because of its clinical relevance and accessibility, the dataset has been extensively used as a standard in research on diabetes prediction.

Data Preprocessing

The dataset is subjected to many stages of preprocessing before the model is trained. Since 0 cannot possibly be a biologically viable value for physiologically implausible characteristics such as glucose, blood pressure, skin thickness, insulin, and body mass index (BMI), these values were considered missing in the Pima Indians Diabetes Dataset. An industry-standard method for dealing with missing data in structured medical datasets, the feature-specific mean was used to replace these zero values. Although median imputation might be a better option for parameters like Insulin and Skin Thickness that have very skewed distributions, mean imputation was used consistently in this research for consistency's sake. For distance-sensitive and probabilistic algorithms like Naive Bayes, it is crucial to normalise the range of all input variables before applying feature scaling via standardisation (z-score normalisation). To maintain the initial class distribution across both sets, the dataset was subdivided into

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

614 instances for training and 154 instances for testing using a stratified 80:20 split ratio. The dataset had a significant class imbalance, with around 65% of the participants not having diabetes and 35% having the disease, hence stratified splitting was necessary.

Exploratory Data Analysis

In order to derive the dataset's underlying structure and connections, Exploratory Data Analysis (EDA) was performed. In order to determine how important each feature was in predicting the result, we used the Random Forest method. It was possible to measure the degree of linear association between each attribute and the diabetes outcome variable by using Pearson correlation coefficients. The most significant predictors, according to the results, were glucose level (27.6% of the total) and body mass index (16.0% of the whole), with age and diabetic pedigree function following at 12.7% each.

Classification Algorithms

Due to their extensive use in medical prediction literature, four machine learning classification methods were chosen and applied to the preprocessed dataset [3][4]. A probabilistic classifier that relies on Bayes' theorem and presupposes conditional independence among features is Naive Bayes (NB). Although it is computationally efficient and has a simple design, it competes well on medical datasets. An interpretable model structure is produced using the rule-based technique known as Decision Tree (DT), which splits the feature space via a sequence of hierarchical binary choices. Random Forest (RF) is an ensemble learning method that constructs many decision trees during training and utilises the majority vote as the final prediction. This helps to enhance generalizability and prevent overfitting. A linear classification approach, Logistic Regression (LR) models the likelihood of a binary outcome using the sigmoid function. It performs well on linearly separable data and has excellent interpretability.

Performance Evaluation

A full suite of performance measures were extracted from the confusion matrix, which provides information on the amounts of TP, TN, FP, and FN, and used to assess each approach. Here are the definitions of the metrics: “The total percentage of right estimates is what we call accuracy; The accuracy of a forecast is defined as the percentage of correct predictions; The percentage of true positives that are accurately recognized is called recall (sensitivity); Among these metrics, we find specificity, which is defined as the percentage of true negative cases correctly identified, FNR, which measures the percentage of true positive cases missed by the model, and AUC-ROC, which measures the model's capacity to distinguish between positive and negative classes across all classification thresholds, and finally, F1-Score, which is the harmonic mean of recall and precision.” Because a missed actual diabetes case (false negative) has more health implications than a false alarm, Recall and False Negative Rate are given special clinical importance within the framework of diabetes screening. So, if the false negative count is significant, even the most accurate model may not be the best option for medical screening.

Prediction Simulation

To demonstrate the practical applicability of the proposed approach, an illustrative prediction scenario was conducted using three representative patient profiles reflecting different risk levels. Profile 1 (Low Risk): Glucose = 85, BMI = 22.0, Age = 25; Profile 2 (Moderate Risk): Glucose = 120, BMI = 28.5, Age = 38; Profile 3 (High Risk): Glucose = 148, BMI = 33.6, Age = 50. Each of the four algorithms independently predicted the diabetes probability for each profile, and an ensemble voting mechanism was applied to produce a combined decision. This multi-scenario simulation enables a more meaningful assessment of model behaviour across varying clinical conditions, rather than relying on a single patient example that may not be representative of the broader range of clinical presentations.

RESULT

This section presents the experimental results obtained from applying four machine learning classification algorithms to the Pima Indians Diabetes Dataset. The findings are organized according to the evaluation framework described in the methodology: confusion matrix analysis, feature importance and correlation, ROC curve comparison, performance metric visualization, and multi-scenario prediction simulation.

Confusion Matrix – All Algorithms

Figure 2 displays the findings of the confusion matrices for the four classification methods that were used to analyse the Pima Indians Diabetes Dataset. From the test set, which consists of 154 samples (20% of 768 occurrences), each matrix shows the number of True Negatives (TN), False Positives (FP), and True Negatives (FN).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16133
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Naive Bayes achieved an accuracy of 70.8%, with TN=74, FP=26, FN=19, and TP=35. Decision Tree improved accuracy to 72.1% with TN=85, FP=15, FN=28, and TP=26. Random Forest obtained the highest accuracy of 76.0% with TN=85, FP=15, FN=22, and TP=32. Logistic Regression achieved 71.4% accuracy with TN=82, FP=18, FN=26, and TP=28.

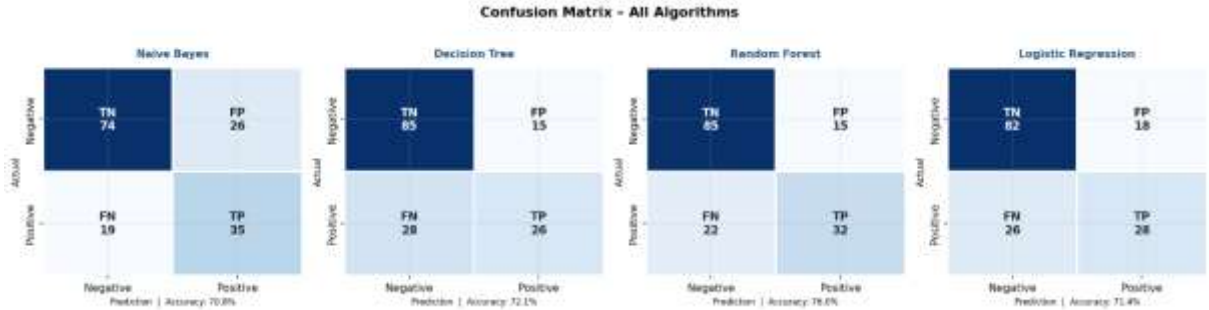

Figure 2. Confusion Matrix for All Classification Algorithms (Naive Bayes, Decision Tree, Random Forest, Logistic Regression)

Feature Importance and Feature Correlation

Figure 3 shows the feature significance scores and their link with the diabetes result produced from the Random Forest model. Glucose (27.6% relevance score), body mass index (16.0% score), age (12.7% score), and diabetes pedigree function (12.7% score) are the top four predictors. Diabetes (7.2%), Skin Thickness (6.6%), Blood Pressure (8.6%), and Pregnancies (8.4%).

The results of the correlation study back up these claims: The result is positively correlated with glucose at ~0.47, followed by body mass index at ~0.29, age at ~0.24, and pregnancies at ~0.22. Having a higher score for any of these eight characteristics is typically related with an increased risk of developing diabetes, since they all show positive associations with the diabetes result.

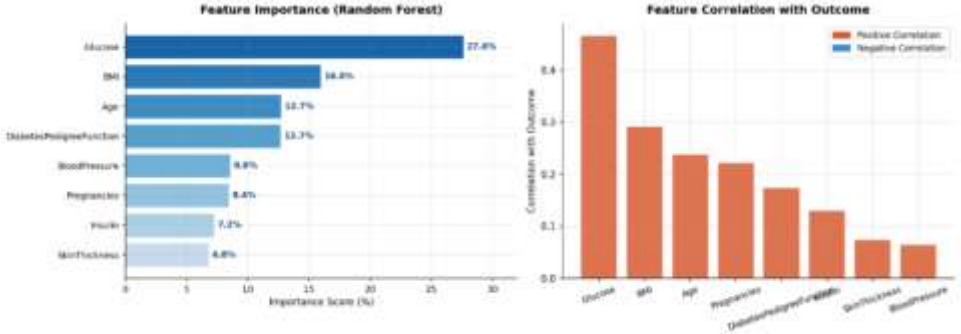

Figure 3. Feature Importance (Random Forest) and Feature Correlation with Diabetes Outcome

ROC Curve – Algorithm Comparison

Figure 4 for a comparison of the four algorithms' Receiver Operating Characteristic (ROC) curves with a random classifier baseline (AUC = 0.50). With an Area Under the Curve (AUC) of 82.30%, Logistic Regression had the greatest discriminative potential. Random Forest followed closely with AUC = 81.47%, while Naive Bayes reached AUC = 77.28%. Decision Tree performed the lowest among the four with AUC = 66.57%.

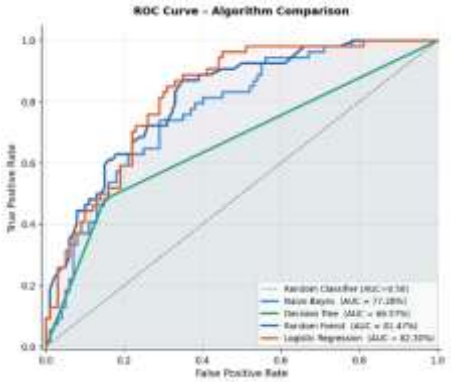

Figure 4. ROC Curve Comparison Across All ML Algorithms

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16133
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparative Performance Visualization

Figure 5 presents a comprehensive comparison of evaluation metrics across all algorithms using a grouped bar chart and a radar chart. The metrics evaluated include Accuracy, Precision, Recall, F1-Score, and AUC-ROC. Random Forest leads in Accuracy (76.0%) and Precision (68.1%), while Naive Bayes achieves the highest Recall (64.8%) and F1-Score (60.9%). Logistic Regression records the highest AUC-ROC (82.3%), and the radar chart confirms that no single algorithm dominates across all metrics, indicating different trade-offs. Decision Tree consistently underperforms in most metrics, particularly in Recall (48.1%) and AUC-ROC (66.6%).

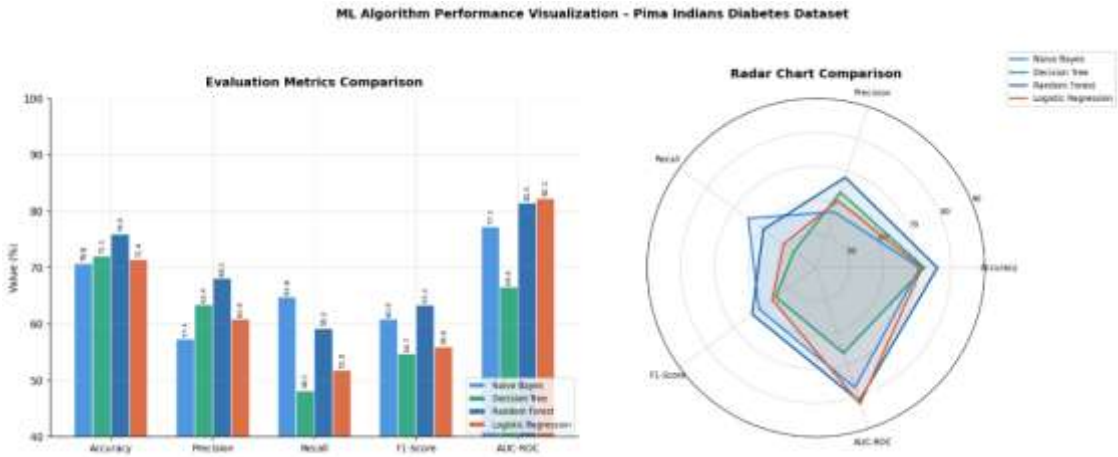

Figure 5. Evaluation Metric Comparison (Bar Chart and Radar Chart) – Pima Indians Diabetes Dataset

Prediction Simulation

An illustrative prediction scenario was conducted using three patient profiles representing low, moderate, and high diabetes risk levels. The results for each algorithm and the ensemble voting outcome for all three profiles are presented in Figure 6. For the low-risk profile (Glucose = 85, BMI = 22.0, Age = 25), all four algorithms predicted a negative outcome with ensemble consensus of 0 votes positive. For the moderate-risk profile (Glucose = 120, BMI = 28.5, Age = 38), model predictions diverged, with Naive Bayes and Logistic Regression predicting positive at 53.2% and 51.4% probability respectively, while Random Forest predicted negative at 38.0% and Decision Tree predicted negative at 0.0%, resulting in a split ensemble vote of 2 positive versus 2 negative. For the high-risk profile (Glucose = 148, BMI = 33.6, Age = 50), all four algorithms predicted a positive diabetes diagnosis with unanimous consensus: Naive Bayes 56.9%, Decision Tree 100.0%, Random Forest 73.0%, and Logistic Regression 66.0%, yielding 4 votes positive and 0 negative. These results demonstrate that model agreement is strongest at the extremes of the risk spectrum, while divergence in moderate-risk cases highlights the importance of multi-model ensemble voting for borderline presentations.

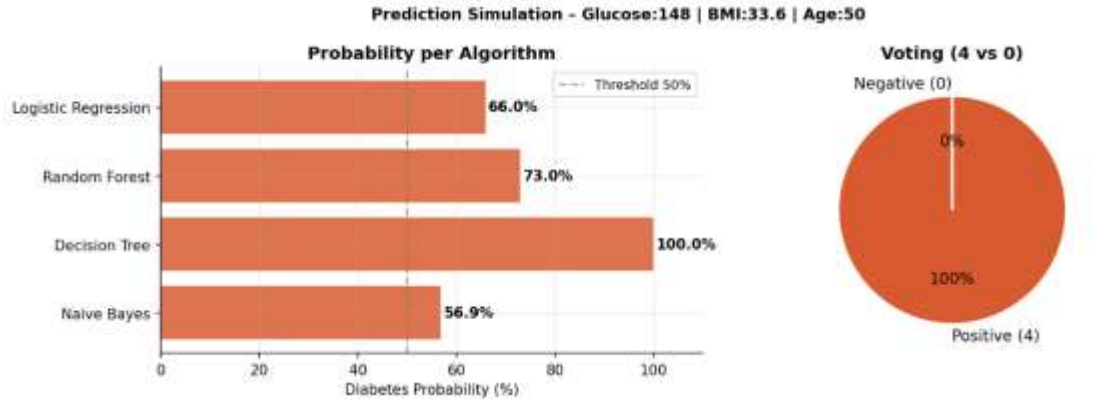

Figure 6. Illustrative Prediction Scenario Results for Three Patient Risk Profiles (Low, Moderate, High Risk) – Probability per Algorithm and Ensemble Voting

DISCUSSIONS

This section discusses the experimental results obtained from applying four machine learning classification algorithms Naive Bayes, Decision Tree, Random Forest, and Logistic Regression to the Pima Indians Diabetes

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16133
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Dataset. The analysis covers feature relevance, model performance, discriminative capability, and prediction behavior.

Feature Importance and Correlation Analysis

The Random Forest feature importance analysis identifies Glucose as the most influential predictor with a score of 27.6%, more than double that of the second-ranked feature, BMI (16.0%). Age and Diabetes Pedigree Function both contribute 12.7%, while the remaining features (Blood Pressure, Pregnancies, Insulin, Skin Thickness) each contribute below 9%. This ranking is strongly corroborated by Pearson correlation analysis, where Glucose also holds the highest positive correlation (~0.47) with the diabetic outcome. The consistency between both analyses confirms glucose metabolism as the primary clinical indicator, aligning with established medical diagnostic criteria.

Table 1. Feature Importance Score and Correlation with Diabetic Outcome

Feature	Importance Score (%)	Correlation with Outcome	Direction
Glucose	27.6	0.47	Positive
BMI	16.0	0.29	Positive
Age	12.7	0.24	Positive
DiabetesPedigreeFunction	12.7	0.17	Positive
BloodPressure	8.6	0.07	Positive
Pregnancies	8.4	0.22	Positive
Insulin	7.2	0.13	Positive
SkinThickness	6.8	0.07	Positive

Comparative Algorithm Performance

Table 2 summarizes the performance of all four classifiers across five evaluation metrics. Random Forest achieves the highest accuracy (76.0%) and precision (68.1%), indicating the strongest ability to minimize false positives. This advantage stems from its ensemble architecture: by aggregating predictions from multiple decision trees, Random Forest reduces overfitting through variance averaging a characteristic weakness of single-tree models like Decision Tree on small datasets. Naive Bayes records the highest recall (64.8%), as its probabilistic generative approach tends to produce softer decision boundaries that favour sensitivity toward the positive class, making it more likely to flag borderline cases as positive. This is clinically valuable in screening contexts where missing a true diabetic case carries greater consequences than a false alarm. Decision Tree, despite competitive accuracy (72.1%), yields the lowest recall (48.1%) and F1-Score (54.7%), reflecting a tendency to overfit the training data and produce rigid boundaries that fail to generalize well to positive cases. For AUC-ROC, Logistic Regression leads at 82.3%, which can be explained by the relatively linear relationship between the primary features (Glucose, BMI, Age) and the diabetic outcome a pattern that logistic regression exploits effectively through its sigmoid activation. Its threshold-independent discriminative power makes it particularly suitable for ranking patients by diabetes risk across a spectrum of clinical thresholds.

These results confirm that no single algorithm universally dominates across all metrics. From a clinical trade-off perspective, the appropriate model depends on the screening objective: Random Forest is preferable in contexts where false positive minimization is critical (e.g., to avoid unnecessary follow-up costs); Naive Bayes is preferable when minimizing false negatives is the priority (e.g., in population-level early screening where missing a case is more dangerous); and Logistic Regression is best suited for risk stratification tasks requiring reliable probability scores across all risk thresholds. Decision Tree is not recommended as a standalone classifier in this dataset due to its high false negative rate (FN = 28), which represents the worst miss rate among all models and is clinically unacceptable in a screening setting.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16133
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 2. Performance Metrics Comparison Across All Algorithms (denotes best in metric)*

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
Naive Bayes	70.8	57.4	64.8	60.9	77.3
Decision Tree	72.1	63.4	48.1	54.7	66.6
Random Forest	**76.0***	**68.1***	**59.3**	**63.4***	**81.5**
Logistic Regression	71.4	60.9	51.9	56.0	**82.3***

Confusion Matrix Analysis

The confusion matrix results in Table 3 reveal class-level prediction patterns that aggregate metrics alone cannot capture. Decision Tree and Random Forest both achieve the highest True Negative count (TN = 85) and lowest False Positive count (FP = 15), demonstrating superior specificity. However, Decision Tree records the highest False Negative count (FN = 28), meaning it misses the most diabetic patients a significant clinical concern. Random Forest reduces this to FN = 22 while maintaining TP = 32, offering the best overall trade-off. Naive Bayes, with TP = 35 and FN = 19, provides the most sensitive detection of diabetic cases at the cost of more false positives (FP = 26). In medical screening contexts where missing a true positive carries greater risk than a false alarm, Naive Bayes's profile may be considered more acceptable.

Table 3. Confusion Matrix Summary for All Algorithms (denotes best accuracy)*

Algorithm	TN	FP	FN	TP	Accuracy (%)
Naive Bayes	74	26	19	35	70.8
Decision Tree	85	15	28	26	72.1
Random Forest	**85**	**15**	**22**	**32**	**76.0***
Logistic Regression	82	18	26	28	71.4

ROC Curve and Prediction Simulation

The ROC curve analysis confirms that Logistic Regression (AUC = 82.30%) and Random Forest (AUC = 81.47%) are the two strongest classifiers in terms of threshold-independent discrimination. Decision Tree's low AUC (66.57%) confirms that its deterministic splits do not generalise well across thresholds, consistent with its susceptibility to overfitting on small datasets. The illustrative prediction scenarios across three risk profiles demonstrated that model consensus was strongest at the extremes: all four models agreed on a negative outcome for the low-risk profile and a positive outcome for the high-risk profile (Glucose = 148, BMI = 33.6, Age = 50). For the moderate-risk profile, however, model predictions diverged, with a split ensemble vote of 2 positive versus 2 negative. This divergence highlights the value of ensemble voting in borderline cases, where individual model limitations are more pronounced. It is important to note that these scenarios are illustrative only and do not constitute clinical validation; the models have not been tested against external datasets or real-world clinical populations.

Overall, the results suggest that an ensemble approach combining these classifiers could leverage their complementary strengths Random Forest's precision, Naive Bayes's sensitivity, and Logistic Regression's discriminative capability to support more balanced preliminary diabetes prediction. Future work should explore advanced ensemble techniques such as stacking, class imbalance correction via SMOTE, external validation on diverse datasets, and the inclusion of additional clinical biomarkers to strengthen the applicability and generalisability of the findings.

CONCLUSION

This study compared four interpretable machine learning algorithms Naive Bayes, Decision Tree, Random Forest, and Logistic Regression for diabetes prediction using the Pima Indians Diabetes Dataset, employing an integrated evaluation framework combining feature importance, correlation analysis, confusion matrix decomposition, ROC-AUC comparison, and multi-scenario prediction simulation.

Glucose concentration was found to be the most dominant predictive feature, contributing 27.6% of Random Forest feature importance and exhibiting the highest Pearson correlation (~0.47) with the diabetic outcome. This

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

finding is consistent with established clinical diagnostic criteria, reinforcing that glucose metabolism remains the primary biological marker for diabetes detection. BMI, Age, and Diabetes Pedigree Function were also identified as meaningful contributors, collectively accounting for over 40% of predictive power.

In terms of overall classification performance, Random Forest achieved the best accuracy (76.0%) and precision (68.1%), making it the most suitable model when minimizing false positives is the priority. Naive Bayes demonstrated the highest recall (64.8%), which is particularly valuable in medical screening contexts where missing a true diabetic case carries greater clinical risk than generating a false alarm. Logistic Regression led in AUC-ROC (82.3%), indicating the strongest threshold-independent discriminative capability across all classification boundaries. Decision Tree, while offering reasonable accuracy (72.1%), exhibited the highest false negative rate and the lowest AUC, suggesting it is prone to overfitting and poor generalization to positive cases.

The illustrative prediction scenarios confirmed that all four models converged on unanimous positive predictions for the high-risk profile, while model divergence at moderate risk levels reinforced the value of ensemble voting strategies for borderline clinical presentations. These findings suggest that Random Forest, Logistic Regression, and Naive Bayes may support preliminary diabetes screening under different clinical priorities, though further external validation is required before any deployment in real-world settings.

Overall, the results demonstrate that no single classifier universally outperforms the others across all metrics. The selection of an appropriate model must therefore be driven by the clinical objective: precision-focused tasks favor Random Forest, sensitivity-focused screening favors Naive Bayes, and probabilistic ranking tasks favor Logistic Regression.

Limitations

This study acknowledges several limitations that constrain the scope and generalizability of its findings. One limitation is that the Pima Indians Diabetes Dataset only includes female patients who are 21 years old and have Pima Indian ancestry. This means that the findings may not apply to a more varied or larger clinical group. Second, even after preprocessing, characteristics like Glucose, Insulin, Blood Pressure, BMI, and Skin Thickness still have physiologically implausible zero values. Mean imputation was used, but residual noise could still be present. Third, the class imbalance between non-diabetic and diabetic cases was not explicitly corrected through techniques such as SMOTE or class weighting, which may have contributed to the lower recall observed in some classifiers. Fourth, no external validation dataset was used to test the generalizability of the trained models beyond the Pima dataset. Fifth, hyper parameter tuning was not performed; all models were run using default configurations, which may not represent their optimal performance. Sixth, cross-validation was not applied, meaning the reported results are based on a single train-test split and may be sensitive to data partitioning. Finally, this study did not include comparisons with more advanced ensemble models such as XGBoost or LightGBM, which may offer superior performance for this dataset.

Recommendations for Future Work

Future research should address the limitations of the current study in a prioritized and systematic manner. First, k-fold cross-validation (e.g., 10-fold) should replace the single train-test split to produce more robust and statistically reliable performance estimates. Second, hyper parameter tuning via grid search or Bayesian optimization should be applied to each classifier to ensure comparisons reflect optimized rather than default configurations. Third, class imbalance should be explicitly addressed through SMOTE or class-weighting strategies to improve recall for the positive (diabetic) class. Fourth, external validation on demographically diverse datasets beyond the Pima cohort is essential to assess the generalizability of the trained models to broader populations. Fifth, advanced ensemble models such as XGBoost, LightGBM, and stacking ensembles should be included to benchmark against the interpretable classifiers evaluated in this study. Finally, incorporating additional clinical biomarkers such as HbA1c, fasting insulin, and lifestyle indicators may strengthen predictive accuracy and clinical relevance, and the integration of explainable AI techniques such as SHAP or LIME would improve model transparency for clinical decision support.

ACKNOWLEDGMENT

In particular, the authors are grateful to the NIDDK for making the Pima Indians Diabetes Dataset freely accessible to the public and to the UCI Machine Learning Repository for storing and disseminating the dataset for scholarly study. The academic supervisors and teachers who provided invaluable input and direction throughout the research process also deserve our gratitude. No particular grant money was allocated to this study by any governmental, private, or nonprofit organisation.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

REFERENCES

Abdelsattar, M., AbdelMoety, A., & Emad-Eldeen, A. (2025). Comparative analysis of machine learning techniques for temperature and humidity prediction in photovoltaic environments. *Scientific Reports*, *15*(1), 1–22. https://doi.org/10.1038/s41598-025-98607-7

Ajala, A. A., Adeoye, O. L., Salami, O. M., & Jimoh, A. Y. (2025). An examination of daily CO2 emissions prediction through a comparative analysis of machine learning, deep learning, and statistical models. *Environmental Science and Pollution Research*, *32*(5), 2510–2535. https://doi.org/10.1007/s11356-024-35764-8

Alsabhan, W., & Alfadhly, A. (2025). Effectiveness of machine learning models in diagnosis of heart disease: a comparative study. *Scientific Reports*, *15*(1), 1–19. https://doi.org/10.1038/s41598-025-09423-y

Ayhan, B., Ayan, E., Karadağ, G., & Bayraktar, Y. (2025). Evaluation of Caries Detection on Bitewing Radiographs: A Comparative Analysis of the Improved Deep Learning Model and Dentist Performance. *Journal of Esthetic and Restorative Dentistry*, *37*(7), 1949–1961. https://doi.org/10.1111/jerd.13470

Azam, Z., Islam, M. M., & Huda, M. N. (2023). Comparative Analysis of Intrusion Detection Systems and Machine Learning-Based Model Analysis Through Decision Tree. *IEEE Access*, *11*, 80348–80391. https://doi.org/10.1109/ACCESS.2023.3296444

Ennab, M., & Mcheick, H. (2025). Advancing AI Interpretability in Medical Imaging: A Comparative Analysis of Pixel-Level Interpretability and Grad-CAM Models. *Machine Learning and Knowledge Extraction*, *7*(1). https://doi.org/10.3390/make7010012

Fattah, F. H., Salih, A. M., Salih, A. M., Asaad, S. K., Ghafour, A. K., Bapir, R., Abdalla, B. A., Othman, S., Ahmed, S. M., Hasan, S. J., Mahmood, Y. M., & Kakamad, F. H. (2025). Comparative analysis of ChatGPT and Gemini (Bard) in medical inquiry: a scoping review. *Frontiers in Digital Health*, *7*(February), 1–7. https://doi.org/10.3389/fdgth.2025.1482712

Ghosh, M., Raihan, M. M. S., Raihan, M., Akter, L., Bairagi, A. K., Alshamrani, S. S., & Masud, M. (2021). A comparative analysis of machine learning algorithms to predict liver disease. *Intelligent Automation and Soft Computing*, *30*(3), 917–928. https://doi.org/10.32604/iasc.2021.017989

Ghosh, P., Azam, S., Karim, A., Hassan, M., Roy, K., & Jonkman, M. (2021). A comparative study of different machine learning tools in detecting diabetes. *Procedia Computer Science*, *192*, 467–477. https://doi.org/10.1016/j.procs.2021.08.048

Houssein, E. H., Gamal, A. M., Younis, E. M. G., & Mohamed, E. (2025). Explainable artificial intelligence for medical imaging systems using deep learning: a comprehensive review. In *Cluster Computing* (Vol. 28, Number 7). Springer US. https://doi.org/10.1007/s10586-025-05281-5

Hussain, I., Ching, K. B., Uttraphan, C., Tay, K. G., & Noor, A. (2025). Evaluating machine learning algorithms for energy consumption prediction in electric vehicles: A comparative study. *Scientific Reports*, *15*(1), 1–20. https://doi.org/10.1038/s41598-025-94946-7

Kangra, K., & Singh, J. (2023). Comparative analysis of predictive machine learning algorithms for diabetes mellitus. *Bulletin of Electrical Engineering and Informatics*, *12*(3), 1728–1737. https://doi.org/10.11591/eei.v12i3.4412

Kumar, A., Saini, R., & Kumar, R. (2024). A Comparative Analysis of Machine Learning Algorithms for Breast Cancer Detection and Identification of Key Predictive Features. *Traitement Du Signal*, *41*(1). https://doi.org/10.18280/ts.410110

Martinović, M., Dokic, K., & Pudić, D. (2025). Comparative Analysis of Machine Learning Models for Predicting Innovation Outcomes: An Applied AI Approach. *Applied Sciences (Switzerland)*, *15*(7), 1–44. https://doi.org/10.3390/app15073636

Mohammadagha, M. (2025). Hyperparameter Optimization Strategies for Tree-Based Machine Learning Models Prediction: A Comparative Study of AdaBoost, Decision Trees, and Random Forest. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5226457

Nishat, M. H., Khan, M. H. R. B., Ahmed, T., Hossain, S. N., Ahsan, A., El-Sergany, M. M., Shafiquzzaman, M., Imteaz, M. A., & Alresheedi, M. T. (2025). Comparative analysis of machine learning models for predicting water quality index in Dhaka's rivers of Bangladesh. *Environmental Sciences Europe*, *37*(1). https://doi.org/10.1186/s12302-025-01078-w

Pai, V., Devidas, & Adesh, N. D. (2021). Comparative analysis of Machine Learning algorithms for Intrusion Detection. *IOP Conference Series: Materials Science and Engineering*, *1013*(1). https://doi.org/10.1088/1757-899X/1013/1/012038

Rabee, A., Anwar, Z., AbdelMoety, A., Abdelsallam, A., & Ali, M. (2025). Comparative analysis of automated foul detection in football using deep learning architectures. *Scientific Reports*, *15*(1), 1–20. https://doi.org/10.1038/s41598-025-96945-0

Sheth, V., Tripathi, U., & Sharma, A. (2022). A Comparative Analysis of Machine Learning Algorithms for

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Classification Purpose. *Procedia Computer Science*, *215*, 422–431. https://doi.org/10.1016/j.procs.2022.12.044

Srivastava, S., Divekar, A. V., Anilkumar, C., Naik, I., Kulkarni, V., & Pattabiraman, V. (2021). Comparative analysis of deep learning image detection algorithms. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00434-w

Zhao, Y. (2025). Comparative Analysis of Diabetes Prediction Models Using the Pima Indian Diabetes Database. *ITM Web of Conferences*, *70*, 02021. https://doi.org/10.1051/itmconf/20257002021

Zhong, J., & Wang, Y. (2025). *Enhancing Thyroid Disease Prediction Using Machine Learning: A Comparative Study of Ensemble Models and Class Balancing Techniques*. 0–20.

*name of corresponding author