NPRC Journal of Multidisciplinary Research
Volume 2, Issue 14, December, 2025 Pages: 1-15
ISSN: 3059-9148 (Online)
DOI: 10.3126/nprcjmr.v2i14.86923

OPEN ACCESS

Comparative Study of Diabetes Prediction using Machine Learning Approaches

Ramesh Prasad Bhatta
Assistant Professor
Central Department of CSIT
Far Western University, Mahendranagar, Nepal
rpb.mcs@gmail.com
https://orcid.org/0009-0005-0554-9072

Received: October 21, 2025 Revised & Accepted: December 27, 2025

Abstract

Background: Diabetes is a rapidly growing global health problem. Diabetes mellitus is a chronic disease that occurs when one's pancreas no longer able to produce enough insulin. The long-term hyperglycemia during diabetes causes chronic damage and dysfunction of various tissues, especially the eyes, kidneys, heart, blood vessels, and nerves. The intention of using ML in healthcare is to increase the diagnostic accuracy and effectiveness of therapy and help clinicians in their practice of patient management with improved outcomes. Disease prediction using ML is gaining significant attention for healthcare

Methods: This study explores the use of six supervised machine learning algorithms to predict diabetes using the Pima Indians Diabetes Database (PIDD). Several well-known ML techniques were implemented and compared, including Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT). The performance of the models was assessed using prediction accuracy.

Results: The results of study indicate Support Vector Machine (SVM) outperformed the other models. SVM obtained a prediction accuracy of 74%, outperforming the other algorithms.

Conclusion: The findings suggest that machine learning methods can significantly improve early diabetes prediction. SVM provides best predictive ability for the given dataset. This study demonstrates that machine learning–based systems can assist healthcare professionals in making early diagnoses and informed decisions, thereby helping to prevent serious complications associated with diabetes.".

Implication: This work can help healthcare facilities understand the usefulness and use of machine learning algorithms in early diabetes prediction.

NPRC Journal of Multidisciplinary Research
Volume 2, Issue 14, December, 2025 Pages: 1-15
ISSN: 3059-9148 (Online)
DOI: 10.3126/nprcjmr.v2i14.86923

Keywords: diabetes prediction, machine learning, Pima Indians Diabetes Database, Support Vector Machine, healthcare

1.Introduction

Diabetes is recognized as a major global healthcare issue that is increasing at a rapid and alarming rate (Chandra Sen et al., 2020). It is considered one of the deadliest chronic diseases, arising from a group of metabolic disorders in which the body either does not produce sufficient insulin or cannot effectively make use of the insulin it generates (Ibrahim & Adnan, 2021). When a person has abnormally high blood glucose levels because of either inadequate insulin manufacturing or an inappropriate cell response to insulin, they have diabetes mellitus, one of the metabolic illnesses.

To manage non communicable illnesses, more research is needed. One of the chronic illnesses that will be discussed in this essay is diabetes mellitus (DM).

Diabetes mellitus, or DM, is a metabolic disease characterized by elevated blood sugar levels. In this case, insulin transports blood sugar into cells, where it is stored or used to create energy. Recent research indicates that 80% of type-2 diabetes cases can be avoided with early detection. Pregnant women with excessive blood sugar are diagnosed with gestational diabetes (Prabhu & Selvabharathi, 2019).

1.1 Types of Diabetes

Diabetes is a long-lasting health problem that changes how your body controls blood sugar, also called glucose. There are different kinds of diabetes, each with its own reasons and features.

Type 1-Diabetes (T1D)

It is the common kind of diabetes and is identified by the body's inadequate production of insulin. The illness can strike at any age, although children and teenagers are the ones who get it most often (Sarwar et al., 2018). In this type of diabetes, pancreas will produce insulin that helps human organs to energize through sugar level in the blood cells. But there may be a chance that pancreas might be producing little amount of insulin or no insulin. Insulin injections are commonly used for controlling. Type-1 diabetes is common in any aged people but it most affects the people among under age 30. Particularly if the patient's heredity having Type-1 diabetes will lead to higher risk. Statistically below 10% of the people impacted by this particular form of diabetes.

Type 2-Diabetes (T2D)

It is the most prevalent kind of diabetes and is distinguished by the body's insufficient synthesis of insulin. All age groups are affected, and patients frequently show signs of obesity, overweight, urination, etc., which are associated with the Insulin resistance (Mujumdar & Vaidehi, 2019). Historically, the adults are most commonly affected by Type 2 diabetes. Statistics revealed that betwixt 90-95 percent of the people will be affected by type 2 diabetes. Diet through weight management and exercise are the common ways to control Type 2 diabetes.

NPRC Journal of Multidisciplinary Research
Volume 2, Issue 14, December, 2025 Pages: 1-15
ISSN: 3059-9148 (Online)
DOI: 10.3126/nprcjmr.v2i14.86923

Gestational Diabetes Mellitus (GDM)

The kind of diabetes that causes pregnant women to have hyperglycemia. This diabetes increases the risk of type-2 diabetes in both the mother and the fetus. In general, the pregnant women will have Gestational diabetes who never had a diabetes in their lifetime. The glucose level will be high when the women get pregnancy. The baby has higher glucose level at the time of pregnancy. Changes in hormone will also leads to high glucose level in blood that affects the action of insulin.

Prediabetes (PD)

Genetic abnormalities that result in increased insulin production are the cause of this form of diabetes., side effects of chemicals, or increase in other hormonal levels in the body. Lifestyle habits and demographic factors are examined and reported the main indicators that play an important role to control and manage Type-2 Diabetes Mellitus. Diet and exercise play an important role to avoid or manage the T2DM, it can reduce the complications of even those people who are at high risk of being involved towards disease

1.2 Motivation of the study

a. Diabetes statistics in India

India is the country with the second-largest diabetic population in the world, after China, with an estimated 77 million people. In India, 17% of the global population has diabetes. and, according to calculations made in October 2018, comprises 17.5% of the world's population (India). The TFPR editorial (TFPR Editorial, 2020) predicts that by 2045, there will be 134 million. 72.96 million adults (those over 20) in India have diabetes out of the country's total population. Among these patients, 10.9-14.2% are from urban areas and 3.0-7.8% are from rural areas (INSIGHTSIAS, 2019).

b. Statistics of diabetes in Nepal

Approximately 7.7% of adults in Nepal (aged 20–79) are living with diabetes, amounting to around 1.26 million people in 2024. A nationwide meta-analysis reports an overall diabetes prevalence of about 8.5% among adults, with prediabetes affecting roughly 9.2%. Back in 2011, around 488,000 adults were estimated to have diabetes, whereas by 2024, this number is projected to rise to about 1.3 million, more than two and a half times higher than in 2011. If current trends persist, the population of adults with diabetes could reach 2.4 million by 2050 (Dhimal et al., 2019).

1.3 Problem Statements

Diabetes is a long-term health issue that's becoming more common worldwide and can lead to serious problems over time. Catching it early and preventing it is key to helping people stay healthier. Machine learning is a smart way to look at big sets of medical data, which can improve how we predict diabetes and spot the main causes. Since different machine learning tools perform differently, it's important to check how well they work to pick the best ones for doctors to use. Finding the main risk factors for diabetes helps doctors and patients prevent and manage the disease better. Machine learning is now widely used to predict diabetes. The PIMA

NPRC Journal of Multidisciplinary Research
Volume 2, Issue 14, December, 2025 Pages: 1-15
ISSN: 3059-9148 (Online)
DOI: 10.3126/nprcjmr.v2i14.86923

Indians Diabetes Dataset is commonly used for this purpose, and methods such as logistic regression, support vector machines, and neural networks are applied. Among them, ensemble methods usually give more accurate results. The multimodal data along with ML and NN should be used in future for better results.

Objective

1. To analysis the performance of supervised ML for diabetes prediction on PIDD dataset.
2. To identify the most effective machine learning algorithm for diabetes prediction.

Research Questions

1. How do supervised machine learning algorithms perform in predicting diabetes on the PIMA Indians Diabetes Dataset (PIDD)?
2. Which supervised machine learning algorithm achieves the best performance for diabetes prediction on the PIMA?

2. Literature Review

Hasan et al. (2020) proposed a data preprocessing framework for diabetes prediction that improves data quality through outlier detection, feature selection, and missing value handling. The study tested several machine learning models, including decision trees, random forests, k-nearest neighbors, naïve Bayes, AdaBoost, XGBoost, and multilayer perceptrons. In addition, the authors introduced an ensemble classifier that combined the outputs of several models through a weighted voting mechanism, thereby enhancing prediction accuracy.

(Kumari, Kumar, & Mittal, 2021) suggested an ensemble method to predict diabetes. PID dataset was used for the experiment. As base learners AdaBoost, SVM, LR, RF, NB, Bagging, GB, XGBoost, and CatBoost were used. RF, LR, and NB were used to obtain the final result by using soft voting. 79% accuracy was obtained by this method.

Alaa and Al-bakry (2021) conducted a study to diagnose diabetes using classification algorithms on PIMA dataset and found LR have highest accuracy of 94%.

Gonzalez et al. (2021) combined XGBoost with Random Forest (RF) for diabetes prediction. Their study highlighted that selecting the most important features improved model performance and could help in personalized diabetes management.

Kumari et al. (2021) tested several machine learning models, including Logistic Regression, RF, SVM, NB, AdaBoost, Gradient Boosting, CatBoost, XGBoost, and Bagging. Using the PIMA dataset, their approach achieved 79.04% accuracy, 73.48% precision, 71.45% recall, and 80.6% F1-score.

Joshi and Dhakal (2021) used Decision Tree and Logistic Regression algorithms to forecast diabetes on the PID dataset using classification tree for feature selection and obtained an accuracy of 78.26%.

Using the same dataset, Barik, Mohanty, Mohanty, and Singh (2021) applied XGBoost and Random Forest for diabetes prediction. The study creates the accuracy of 71% and 74% by RF and XGBoost respectively.

According to Mushtaq et al. (2022) proposed a system to predict diabetes using ML algorithms

NPRC Journal of Multidisciplinary Research
Volume 2, Issue 14, December, 2025 Pages: 1-15
ISSN: 3059-9148 (Online)
DOI: 10.3126/nprcjmr.v2i14.86923

at two stages. The dataset was initially balanced using Tomek, IQR, and the synthetic minority oversampling technique (SMOTE). The suggested effort produced an accuracy of 82% using the Pima Indian Diabetic (PID) dataset.

Rawat et al.(2022*)* suggested ML algorithms such as NB, SVM, neural network (NN), Adaboost, KNN, and Linear SVM to predict diabetes. When it comes to accuracy, NN performs better than others

More recently, Gundogdu (2023) proposed a hybrid approach for diabetes diagnosis that integrates multiple linear regression (MLR), Random Forest (RF), and XGBoost (XG) using questionnaire-based data. In this study, MLR-RF was used for feature selection, and XGBoost handled classification on hospital records of 520 individuals from Sylhet, Bangladesh (200 control and 320 diabetes cases). The method showed high reliability, achieving 99.2% accuracy, an AUC of 99.3%, and a prediction time of 0.048 seconds.

El-Sofany et al. (2024) developed a mobile-based system to predict diabetes risk using multiple machine learning techniques. XGBoost with SMOTE oversampling demonstrated the best results when tested on a private dataset and the PIMA Indians Diabetes Dataset (PIDD), obtaining 97.4% accuracy and an F1-score of 0.95 on the private dataset and 83.1% accuracy with an F1-score of 0.76 on the combined datasets.

Bateja, Dubey, and Bhatt (2024) explored the use of machine learning with electronic health records (EHRs) and drug recommendation systems to support personalized diabetes care. They used IDBSCAN with MapReduce for data preprocessing and trained several classifiers for timely predictions and recommendations.

3. Research Methodology

3.1 Research Design

This study adopted a quantitative and predictive research design using supervised machine learning techniques to predict the occurrence of diabetes. The objective of the study is to develop and evaluate classification models that can accurately predict a person have diabetic and non-diabetic based on clinical attributes.

3.2 Dataset Description

The analysis and prediction of diabetes are performed on the Pima Indian diabetes dataset which was downloaded from the UC Irvine machine learning repository, consist of 768 records with a female aged 21 and above. This dataset is female-centric only. The dataset comprises of numerically-valued eight attributes with the class label as diabetic or non-diabetic. It includes 268 diabetic instances and 500 non-diabetic instances The outcome is binary, with 0 representing the non-diabetic and 1 indicating diabetic patient.

3.3 Data Preprocessing

To enhance the dataset's quality, data preparation was done. As raw data is prone to missing values noise, and inconsistency, making it consistent need to be pre-processed, reflecting the data quality and thereby providing reliable and accurate results. Data cleaning, transformation, integration and reduced data size are different data pre-processing steps.

NPRC Journal of Multidisciplinary Research
Volume 2, Issue 14, December, 2025 Pages: 1-15
ISSN: 3059-9148 (Online)
DOI: 10.3126/nprcjmr.v2i14.86923

3.4 Model Development

The dataset was distributed into a Training Set (70%) of the total. Model evaluation on unseen data was conducted using the Testing Set (30%). The model's ability to generalize to new, unseen data is ensured by appropriate dataset splitting. Six supervised machine learning classifiers were implemented which are discussed in below.

3.5 **Machine Learning Algorithm**

1.LogisticRegression(LR)

Logistic regression is a statistical technique that employs a logistic function to model a binary dependent variable. It has extensive applications in biomedical research to model disease risk and other binary outcomes, especially when the dependent variable is dichotomous or the relationship between the dependent variable and independent variables is nonlinear. Logistic regression estimates the probability of an event occurring based on a set of predictor variables and offers insights into the impact of each predictor variable on the likelihood of the event taking place.

2.NaïveBayes(NB)

In statistics, naive Bayes are simple probabilistic classifiers that apply Bayes' theorem. This theorem is based on the probability of a hypothesis, given the data and some prior knowledge. The naive Bayes classifier assumes that all features in the input data are independent of each other.

3.k-NearestNeighbors(KNN)

The K-Means clustering algorithm is a widely used and straightforward analytical technique that involves selecting a training set and a predetermined number of clusters (k) to identify. The algorithm then groups items in the training set into clusters based on their similarity, which is often determined by measuring their distance from each other using metrics like Euclidean distance

4.SupportVectorMachine(SVM)

To distinguish between cases with and without diabetes, SVM creates an ideal decision boundary, or hyperplane.. It is effective for high-dimensional medical data and performs well in handling complex, non-linear relationships using kernel functions.

5.DecisionTree(DT)

Decision trees are machine learning models that provide a high level of interpretability by allowing data to be stratified or segmented. These models enable the continuous splitting of data based on specific parameters until a final decision is reached. it is used in both classification and regression problems

6.RandomForest(RF)

The decision tree algorithm has a significant flaw, which is its tendency to overfit. Overfitting leads to complex models with high variance that have good accuracy during training but poor generalization to other datasets. To address this issue, the random forest algorithm uses bagging, an extension of a technique that reduces model variance by averaging a set of observations (we will explain it later). The random forest combines the predictions from multiple decision trees to produce a single output

NPRC Journal of Multidisciplinary Research
Volume 2, Issue 14, December, 2025 Pages: 1-15
ISSN: 3059-9148 (Online)
DOI: 10.3126/nprcjmr.v2i14.86923

3.6 Model Evaluations

Confusion Matrix: Confusion matrix is used to describe how well the models performed. A confusion matrix for a binary classification problem is typically a 2x2 table with the following entries:

1. True Positives (TP): The number of cases correctly predicted as the positive class (e.g., correctly predicting patients has disease).
2. True Negatives (TN): The number of cases correctly predicted as the negative class (e.g., correctly finding patients has no diabetes).
3. False Positives (FP): The number of cases incorrectly predicted as the positive class.
4. False Negatives (FN): The number of cases incorrectly predicted as the negative class

Performance metrics:

Accuracy (ACC). The calculation of accuracy involves dividing the total number of accurate predictions by the total number of instances in the dataset. in this study It shows the percentage of patients, with diabetic or not

Accuracy =(TP+TN)/(TP+TN+FP+FN)

Precision: The precision is determined by dividing the number of correctly predicted positive instances by the total number of positive instances. It measures how well the model prevents false positives.

Precision=TP/(TP+FP)

Recall (Sensitivity): The number of accurate positive predictions divided by the total number of positives is how recall is calculated. It is also known as sensitivity or true positive rate (TPR).

Recall=TP/(TP+FN)

F1 score: The F1-score is the harmonic mean of precision and recall, providing a balance between these two metrics. It is calculated as twice the product of precision and recall divided by their sum. The F1-score gives an overall measure of a model's prediction quality.

F1 score = 2*(Precision*Recall)/(Precision+Recall)

ROC/AUC Curve: It indicates how effectively the model can distinguish between different classes. The model predicts classes denoted by 0 as value 0 and classes denoted by 1 as value 1 if the AUC is high. Analogously, a high AUC value indicates that the model is more effective, allowing us to differentiate between patients with and without illness.

.

Kappa value: the kappa statistic measures how well the instances are classified by the ML classifier and classify the labelled data as ground truth while controlling for the anticipated accuracy of a random classifier. It examines how effective a classifier is for a specific dataset.

$$Kappa=\frac{i0-ie}{1-ie}$$ Where i_0 is overall accuracy and i_e is a measure of the agreement between the model predictions and the actual class values as if happening by chance.

ROC curve: is calculated by contrasting the true positive rate (TPR) against the false positive rate (FPR) at various threshold levels and efficiently divides the signal from the noise.

NPRC Journal of Multidisciplinary Research
Volume 2, Issue 14, December, 2025 Pages: 1-15
ISSN: 3059-9148 (Online)
DOI: 10.3126/nprcjmr.v2i14.86923

Mean absolute error: it reflects the gap between the original and predicted values as determined by averaging the absolute difference across the data set.
RMSE: it is a prominent approach to evaluating the error in the model for predicting statistical data. RMSE scores between 0.0 and 0.5 which implies that the model can accurately predict the data.
RAE: it is a method of evaluating the effectiveness of a predictive model. It is expressed as a ratio, contrasting mean errors to trivial errors.
Root relative squared error (RRSE): it is a basic indicator that provides an idea of how well a model performs. Furthermore, it is a variation of the relative squared error (RSE).
Matthews correlation coefficient: it examines categorization quality by accounting for true and false positives and negatives. In this 1 represents a perfect prediction, 0 reflects no better than a random prediction, and -1 indicates an absolute conflict between prediction and observation.

4. Results

4.1 Descriptive statistics of dataset

Supervised machine learning methods were chosen and used to predict diabetes in order to provide a thorough and equitable study of the algorithms. 10-fold cross-validation was used in the investigation. The following sections discuss each model's performance.
. Table 1 presents the statistical summary of the dataset's attributes

Table 1: *Descriptive analysis of PIDD dataset*

Statistics	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DPF	Age
Count	768	768	768	768	768	768	768	768
Mean	3.84	121.60	72.20	26.60	118.68	32.4	0.47	33.24
Std	3.36	30.40	12.10	9.60	93.08	6.80	0.33	11.76
Min	0.00	44.00	24.00	7.00	14.00	18.20	0.07	21.00
25%	1.00	99.70	64.00	20.50	79.70	27.50	0.24	24.00
50%	3.00	117.00	72.00	23.00	79.70	32.0	0.37	29.00
75%	6.00	140.25	80.00	32.00	127.25	36.60	0.62	41.00
Max	17.00	199.00	122.00	99.00	846.00	67.10	2.42	81.00

The findings indicate that 268 individuals with diabetes, or 34.9% of the total subjects, and 500 non-diabetic participants, or 65.1%.

NPRC Journal of Multidisciplinary Research
Volume 2, Issue 14, December, 2025 Pages: 1-15
ISSN: 3059-9148 (Online)
DOI: 10.3126/nprcjmr.v2i14.86923

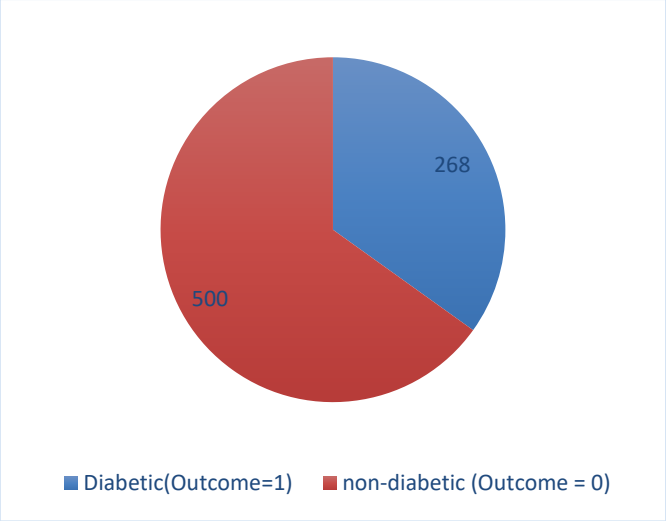

Figure 1: Outcome of dataset

4.2 Performance of classification Analysis

Six machine learning classifiers—Naïve Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR) are compared in Table 2. and assessed using common categorization criteria on the PID dataset. Support Vector Machine (SVM) had the greatest accuracy of all the models (74.3%), closely followed by Naïve Bayes (72.6%) and Logistic Regression (74.0%). These findings show that both probabilistic and linear classifiers work well with the PID dataset. Both SVM and LR achieved the greatest values (0.74) in terms of precision and recall. Additionally, Naïve Bayes demonstrated consistent performance with recall and precision values of 0.72.

Table 2: *Performance of classification algorithm*

Model	Accuracy	Precision	Recall	ROC area	MCC	Kappa Value
NB	72.6	0.72	0.72	0.81	0.45	0.45
KNN	66.1	0.67	0.66	0.65	0.30	0.32
SVM	74.3	0.74	0.74	0.74	0.48	0.48
DT	71.8	0.71	0.71	0.71	0.43	0.43
RF	64.9	0.65	0.65	0.64	0.47	0.29
LR	74.0	0.74	0.74	0.83	0.47	0.48
NB	72.6	0.72	0.72	0.81	0.45	0.45

Logistic regression had the highest ROC value (0.83), followed by Naïve Bayes (0.81), according to the ROC area (AUC) analysis. Strong discriminative capability is indicated by these results, which show that these models can successfully differentiate between positive and negative classes across a range of decision thresholds. KNN and RF, on the other hand,

NPRC Journal of Multidisciplinary Research
Volume 2, Issue 14, December, 2025 Pages: 1-15
ISSN: 3059-9148 (Online)
DOI: 10.3126/nprcjmr.v2i14.86923

generated relatively lower ROC values (0.65 and 0.64, respectively), indicating a weaker separation of the classification boundary.

Matthews Correlation Coefficient (MCC), useful for handling unbalanced datasets, SVM achieved highest MCC score 0.48, performing better than by RF and LR both have 0.47. This indicates that SVM has the most reliable overall prediction. Kappa values for SVM, LR, KNN and RF are 0.48,0.48 ,0.32 and 0.29 respectively.

The study found that SVM outperforms the others (Table 2) in terms of accuracy, the error rates for each classifier were covered in Table 3. For best outcomes, the error rate should be as low as possible. RAE is a number between 0 and 1, with 0 denoting a good fit for the dataset and 1 denoting a bad fit (Kangra & Singh, 2023).

Table 3: *Error rate analysis for classifier*

	MAE	RMSE	RAE	RRSE	Time(sec)
NB	0.3	0.42	0.61	85.7	0.01
KNN	0.33	0.58	0.67	116.1	0
SVM	0.25	0.5	0.51	101.3	0.09
DT	0.34	0.47	0.68	95.9	0.08
RF	0.33	0.41	0.66	83.9	0.46
LR	0.33	0.4	0.66	81.9	0.08

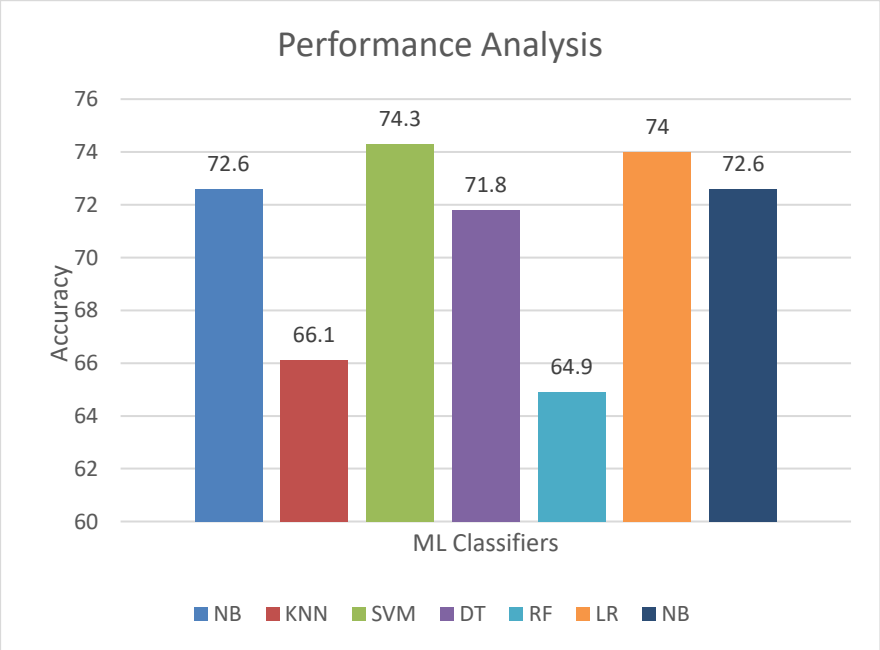

Figure 2: Comparison of Accuracy Performance

NPRC Journal of Multidisciplinary Research
Volume 2, Issue 14, December, 2025 Pages: 1-15
ISSN: 3059-9148 (Online)
DOI: 10.3126/nprcjmr.v2i14.86923

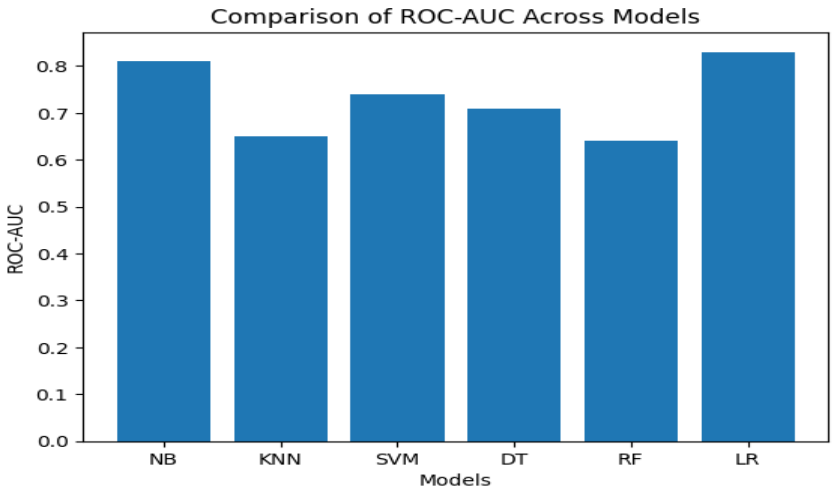

Figure 3: Comparisons of ROC-AUC across classifiers

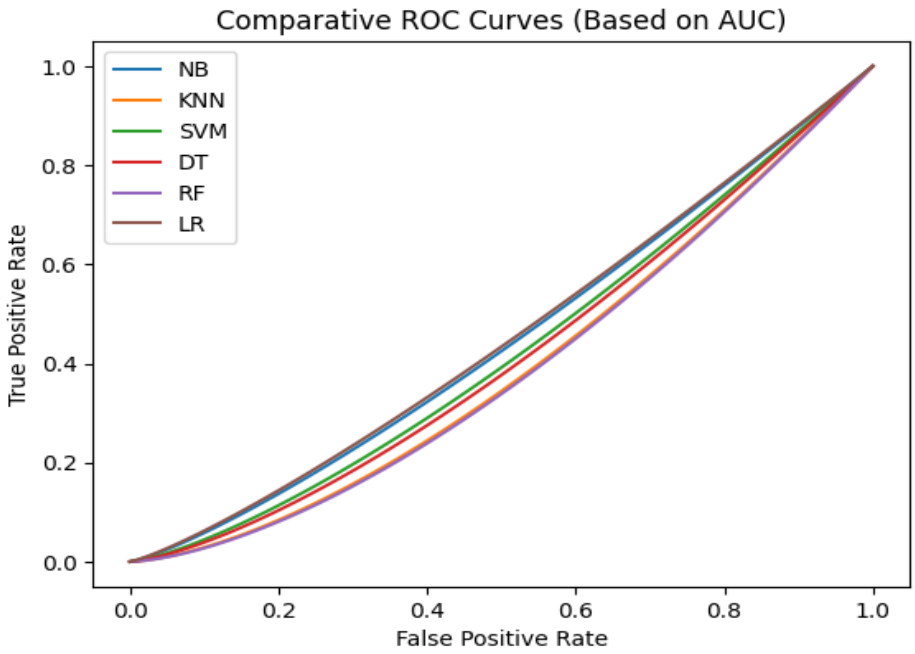

Figure 4: Comparisons of ROC curve

5. Discussion

The results of the study reveal that SVM and Logistic Regression outperform other classifiers on classification tasks. SVM's excellent performance is due to its ability to produce optimal separating hyperplanes, which is particularly helpful when class boundaries are fairly complex. Logistic regression is especially helpful in applications where probabilistic interpretation and threshold selection are critical because of its higher ROC area (0.83), which shows that it offers the best trade-off between sensitivity and specificity. Naïve Bayes also demonstrated competitive performance, particularly in terms of ROC area (0.81).

KNN and Random Forest performed significantly worse, particularly in Kappa statistics and

NPRC Journal of Multidisciplinary Research
Volume 2, Issue 14, December, 2025 Pages: 1-15
ISSN: 3059-9148 (Online)
DOI: 10.3126/nprcjmr.v2i14.86923

the ROC area. KNN's lower performance may be due to sensitivity to feature scaling and noise, whereas RF's mediocre results may indicate overfitting or insufficient parameter adjustment. In terms of accuracy, MCC, Kappa, and ROC metrics, SVM is the most reliable classifier for the PID dataset whereas with Logistic Regression second in performance.

ROC–AUC Analysis

The ROC analysis provides understanding of model's ability to differentiate through classes. From above shown in Figure 3, Logistic Regression, Naïve Bayes and SVM have ROC-AUC values 0.83, 0.81 and 0.74 respectively. A higher AUC indicates better separability between positive and negative classes. Strong ROC performance shows how robust and appropriate logistic regression is for probabilistic decision-making systems. Weaker discrimination ability is indicated by the comparatively lower AUC values of Random Forest (0.64) and KNN (0.65). The ROC analysis confirms that Logistic Regression offers the best trade-off between sensitivity and specificity, making it particularly suitable for real-world deployment where threshold flexibility is essential

MCC and Cohen's Kappa Interpretation

Matthews Correlation Coefficient (MCC) and Cohen's Kappa provide more reliable evaluation for imbalanced datasets. SVM achieved the highest MCC value (0.48), confirming its robustness and balanced predictive capability. Logistic Regression and Random Forest followed closely with MCC values of 0.47. The alignment of MCC, Kappa, and ROC–AUC metrics strengthens the conclusion that SVM and LR provide the most stable and trustworthy classification outcomes.

6. Conclusions

The majority of people worldwide suffer from diabetes, a common illness. Diabetes can create various issues; thus it must be identified early. This study provides a comprehensive analysis and discussion of diabetes prediction models, focusing on classification algorithms NB, KNN, SVM, DT, RF, and LR. The SVM showed the highest predictive accuracy of 74%. The model was using the PID diabetes and implemented using Python. It can be concluded that SVM and LR outperform in terms of accuracy for PIDD and for ROC area LR performs better. This research may aid healthcare concerns in the early detection of diabetes, saving doctors time and effort. Logistic regression's outstanding ROC-AUC performance highlights its interpretability and discriminative power two essential characteristics for diagnostic and decision-support systems.

Transparency Statement: The author confirms that this study has been conducted with honesty and in full adherence to ethical guidelines.
Data Availability Statement: Author can provide data.
Conflict of Interest: The author declares there is no conflicts of interest.
Authors' Contributions: The author solely conducted all research activities i.e., concept, data collecting, drafting and final review of manuscript.

NPRC Journal of Multidisciplinary Research
Volume 2, Issue 14, December, 2025 Pages: 1-15
ISSN: 3059-9148 (Online)
DOI: 10.3126/nprcjmr.v2i14.86923

References

Alaa, F., & Al-bakry, A. M. (2021). Diagnosis of diabetes using machine learning algorithms. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.07.196

Azad, C., Bhushan, B., Sharma, R., Shankar, A., Singh, K. K., & Khamparia, A. (2022). Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus. *Multimedia Systems, 28*(4), 1289–1307. https://doi.org/10.1007/s00530-021-00817-2

Barik, S., Mohanty, S., Mohanty, S., & Singh, D. (2021). Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques. In *Smart Innovation, Systems and Technologies* (Vol. 153, pp. 399–409). Springer. https://doi.org/10.1007/978-981-15-6202-0_41

Bateja, R., Dubey, S. K., & Bhatt, A. K. (2024). Diabetes prediction and recommendation model using machine learning techniques and MapReduce. *Indian Journal of Science and Technology, 17*(26), 2747–2753. https://doi.org/10.17485/IJST/v17i26.530

Dharmarathne, G., Jayasinghe, T. N., Bogahawaththa, M., Meddage, D., & Rathnayake, U. (2024). A novel machine learning approach for diagnosing diabetes with a self-explainable interface. *Healthcare Analytics, 5*, 100301. https://doi.org/10.1016/j.health.2024.100301

El-Sofany, H., El-Seoud, S. A., Karam, O. H., El-Latif, Y. M. A., & Taj-Eddin, I. A. T. F. (2024). A proposed technique using machine learning for the prediction of diabetes disease through a mobile app. *International Journal of Intelligent Systems, 2024*, 1–13. https://doi.org/10.1155/2024/6688934

Farooqui, N. A. R., & Tyagi, A. (2018). Prediction model for diabetes mellitus using machine learning techniques. *International Journal of Computer Science and Engineering, 6*(3), 292–296. https://doi.org/10.26438/ijcse/v6i3.292296

Febrian, M. E., Ferdinan, F. X., Sendani, G. P., Suryanigrum, K. M., & Yunanda, R. (2023). Diabetes prediction using supervised machine learning. *Procedia Computer Science, 216*, 21–30. https://doi.org/10.1016/j.procs.2022.12.107

Gowthami, S., Venkata Siva Reddy, R., & Ahmed, M. R. (2024). Exploring the effectiveness of machine learning algorithms for early detection of type 2 diabetes mellitus. *Measurement: Sensors, 31*, 100983. https://doi.org/10.1016/j.measen.2023.100983

INSIGHTSIAS. (2019, October 11). *National diabetes and diabetic retinopathy survey*. https://www.insightsonindia.com/2019/10/11/national-diabetes-and-diabetic-retinopathy-survey/

Ismail, L., Materwala, H., Tayefi, M., Ngo, P., & Karduck, A. P. (2022). Type 2 diabetes with artificial intelligence machine learning: Methods and evaluation. *Archives of Computational Methods in Engineering, 29*(1), 313–333. https://doi.org/10.1007/s11831-021-09582-x

Joshi, R. D., & Dhakal, C. K. (2021). Predicting type 2 diabetes using logistic regression and machine learning approaches. *International Journal of Environmental Research and Public Health, 18*(14), 7346. https://doi.org/10.3390/ijerph18147346

NPRC Journal of Multidisciplinary Research
Volume 2, Issue 14, December, 2025 Pages: 1-15
ISSN: 3059-9148 (Online)
DOI: 10.3126/nprcjmr.v2i14.86923

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

Kumari, K. S., & Bhargavi, K. (2021). Performance analysis of diabetes mellitus using machine learning techniques. *Turkish Journal of Computer and Mathematics Education, 12*(6), 225–230.

Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering, 2*, 40–46. https://doi.org/10.1016/j.ijcce.2021.01.001

Mallika, C., & Selvamuthukumaran, S. (2021). A hybrid Crow Search and Grey Wolf Optimization technique for enhanced medical data classification in diabetes diagnosis system. *International Journal of Computational Intelligence Systems, 14*(1). https://doi.org/10.1007/s44196-021-00013-0

Mushtaq, Z., Ramzan, M. F., Ali, S., Baseer, S., Samad, A., & Husnain, M. (2022). Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques. *Hindawi, 2022*(Special Issue). https://doi.org/10.1155/2022/6521532

Nagabushanam, P., Jayan, N. C., Antony Joel, C., & Radha, S. (2021). CNN architecture for diabetes classification. In *2021 3rd International Conference on Signal Processing and Communication (ICPSC)* (pp. 166–170). IEEE. https://doi.org/10.1109/ICSPC51351.2021.9451724

Patil, R., Sharvari, T., & Nirmal, R. (2021). Hybrid ANFIS-GA and ANFIS-PSO based models for prediction of type 2 diabetes mellitus. In *Advances in Intelligent Systems and Computing* (Vol. 1227, pp. 11–23). Springer.

Patil, R., Tamane, S., Rawandale, S. A., & Patil, K. (2022). A modified mayfly-SVM approach for early detection of type 2 diabetes mellitus. *International Journal of Electrical and Computer Engineering, 12*(1), 524–533. https://doi.org/10.11591/ijece.v12i1.pp524-533

Patra, R., & Khuntia, B. (2021). Analysis and prediction of Pima Indian diabetes dataset using SDKNN classifier technique. *IOP Conference Series: Materials Science and Engineering, 1070*(1), 1–14. https://doi.org/10.1088/1757-899X/1070/1/012059

Pethunachiyar, G. A. (2020). Classification of diabetes patients using kernel based support vector machines. In *2020 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 22–25). IEEE. https://doi.org/10.1109/ICCCI48352.2020.9104185

Prabhu, P., & Selvabharathi, S. (2019). Deep belief neural network model for prediction of diabetes mellitus. In *2019 3rd International Conference on Imaging, Signal Processing and Communication (ICISPC)* (pp. 138–142). IEEE. https://doi.org/10.1109/ICISPC.2019.8935838

Pradhan, R., Aggarwal, M., Maheshwari, D., Chaturvedi, A., & Sharma, D. (2020). Diabetes mellitus prediction and classifier comparative study. In *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control* (pp. 133–139). IEEE. https://doi.org/10.1109/PARC49193.2020.236572

NPRC Journal of Multidisciplinary Research
Volume 2, Issue 14, December, 2025 Pages: 1-15
ISSN: 3059-9148 (Online)
DOI: 10.3126/nprcjmr.v2i14.86923

Rajendar, S., Thangaraj, R., Palanisamy, J., & Kaliappan, V. K. (2020). Comparative analysis of classifier models for the early prediction of type 2 diabetes. *International Journal of Advanced Science and Technology, 29*(7), 2184–2194.

Rajeswari, S. V. K. R., & Ponnusamy, V. (2021). Prediction of diabetes mellitus using machine learning. *Annals of the Romanian Society for Cell Biology, 25*(5), 17–20.

Rawat, V., Joshi, S., Gupta, S., Singh, D. P., & Singh, N. (2022). Machine learning algorithms for early diagnosis of diabetes mellitus: A comparative study. *Materials Today: Proceedings, 56*(Part 1), 502–506. https://doi.org/10.1016/j.matpr.2022.02.172

Samreen, S. (2021). Memory-efficient, accurate and early diagnosis of diabetes through a machine learning pipeline employing crow search-based feature engineering and a stacking ensemble. *IEEE Access, 9*, 134335–134354. https://doi.org/10.1109/ACCESS.2021.3116383

Shambharkar, S. S., Moon, P. S., & Bainalwar, P. A. (2023). Machine learning-based approach for early detection and prediction of chronic diseases. In *Proceedings of the 1st DMIHER International Conference on Artificial Intelligence in Education and Industry 4.0 (IDICAIEI)* (pp. xx–xx). IEEE. https://doi.org/10.1109/IDICAIEI58380.2023.10406914

Sharma, A., Guleria, K., & Goyal, N. (2021). Prediction of diabetes disease using machine learning model. In *Lecture Notes in Electrical Engineering* (Vol. 733, pp. 683–692). Springer. https://doi.org/10.1007/978-981-33-4909-4

Talukder, M. A., Islam, M. M., Uddin, M. A., Kazi, M., Khalid, M., Akhter, A., & Moni, M. A. (2024). Toward reliable diabetes prediction: Innovations in data engineering and machine learning applications. *Digital Health*. Advance online publication. https://doi.org/10.1177/20552076241271867

Tan, Y., Chen, H., Zhang, J., Tang, R., & Liu, P. (2022). Early risk prediction of diabetes based on GA-Stacking. *Applied Sciences, 12*(2), 632. https://doi.org/10.3390/app12020632

TFPR Editorial. (2020). Diabetes is a pandemic in India. But the Sugar Association wants people to consume more! *The Future of Public Relations*.

Tripathi, G., & Kumar, R. (2020). Early prediction of diabetes mellitus using machine learning. In *ICRITO 2020 - IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)* (pp. 1009–1014). IEEE. https://doi.org/10.1109/ICRITO48877.2020.9197832