

Comparative Performance of Random Forest and Multilayer Perceptron Models for Diabetes Classification Using the Pima Indian Diabetes Dataset

*¹Adama Asmau Sani, ¹Isah Olawale Mustapha, ¹Abdullateef Adekunle Olorukooba, ²Muhammad Abubakar Ndagi, ³Yunus Abdulmuqtadir and ¹Bello-Sulayman Omotosho Yusuf

¹Department of Computer Science, Faculty of Computing, Engineering and Technology, Al-Hikmah University, Ilorin, Kwara State, Nigeria.

²Ibrahim Badamasi Babangida University Lapai, Niger State, Nigeria.

³University of Ilorin, Kwara State, Nigeria.

*Corresponding authors' email: adamalami251@gmail.com

ABSTRACT

Diabetes has become one of the most severe public health issues around the globe and is especially severe in developing nations. Aging leads to an increase in population, as well as impact rate. If timely diagnosis is not achieved, diabetes can significantly increase the prevalence of microvascular and macrovascular complications that can lead to increased costs of care and mortality. Commonly believed, none of these tests, namely Hemoglobin A1c (HbA1c), the Oral Glucose Tolerance Test (OGTT) or the Fasting Plasma Glucose (FPG) is a predictive or predisposing test. Making a model that can foresee these potentially fatal diseases is crucial for identifying those at risk before they become very sick. Using the Pima Indians Diabetes Dataset of 768 with 9 variables, this research assesses diabetes prediction using MLP Neural Network (MLP-NN) and Random Forest (RF) models. Several methods for feature engineering and data preparation were used to improve the model's performance. We validated our model using a three-fold stratified cross validation technique. This study uses ROC-AUC, F1-Score, Accuracy, and Brier for evaluating the performance of the model. With an accuracy of 76.43%, F1-Score of 63.28%, Brier 15.62% and ROC-AUC of 83.86%. However, after the simulation of this model if design and implemented accordingly, it will predict accurately and be reliable based on the results.

Received: 02 Sep. 2026

Accepted: 05 Oct. 2026

Published: 08 Oct. 2026

Keywords:

Diabetes prediction; Random Forest; MLP Neural Network; machine learning; Clinical decision support

INTRODUCTION

Diabetes is a chronic metabolic disease, may have two possible causes insufficient insulin production or insulin dysfunction. It disrupts the body's glucose balance and destroys blood vessels and neurons when it's uncontrolled. Diabetes can lead to amputation of the lower limbs, heart disease, stroke and microvascular complications such as neuropathy and retinopathy. These problems take a heavy toll in terms of the global burden of disease, mortality and health care expenditures as per the study of Ganie et al. (2023) and Tasin et al. (2023). Populations are becoming older and growing at a faster rate, which adds to the hidden burden of individuals living with diabetes. The metabolic and behavioral variables that increase the likelihood of developing diabetes are major reasons why many individuals with the disease go undetected (Rahman et al., 2025).

With the advent of increasingly powerful computers and cutting edge technologies, the machine learning (ML) paradigm has emerged as a powerful instrument for disease risk prediction. According to Afsaneh et al. (2022), ML has the ability to uncover nonlinear correlations in clinical data sets that conventional statistical approaches fail to detect. Due to their ability to mimic complex feature interactions among variables such as age, BMI, insulin level, glucose concentration, and insulin level, ensemble and neural approaches have been thoroughly investigated for structured medical data (Mohsen et al., 2023).

A study by Andrade-Arenas & Yactayo-Arias, (2025) that uses Decision Tree (DT), Naive Bayes (NB), K-Nearest

Neighbours (KNN), Logistic Regression (LR), Random Forest (RF), XGBoost, Artificial Neural Network (ANN), Support Vector Machine (SVM) and AdaBoost (AB) on Pima Indian dataset. The data was balanced through SMOTE technique and different test set evaluation measures were used to evaluate the test sets, such as accuracy, precision, recall, F1 score, AUC-ROC and AUPRC, using a cross validation procedure. While both the RF and the AdaBoost showed good accuracy for cross-validation, the AdaBoost also showed good results on the other metrics and in this regard can be considered robust: 0.791 accuracy, 0.797 precision, 0.786 recall, 0.788 F1; The RF at 0.816 accuracy, 0.804 precision, 0.866 recall, 0.828 F1. Results demonstrated that the ensemble methods are superior in consistently predicting diabetes in this study, which are ideal for clinical decision support and advocated for future studies that validate on a variety of different, heterogeneous and other physiologic data such as information regarding gender, ethnicity and age which are suitable for the general population of the real world. On the other hands, this research uses the same dataset (Pima Indians Diabetes Dataset) from UCI machine learning repository for the study, but comparing the performance of the Random Forest and Multilayer perceptron neural network model.

The goal of this research is to evaluate the performance of Random Forest and MLP Neural Network models using standard classification metrics and explainable AI Techniques. To achieve this, certain objective must be met. To develop diabetes prediction models using Random Forest

and MLP Neural Network models, integrate SHAP to explain model predictions and assess clinical applicability and to identify the most influential features contributing to diabetes prediction using SHAP explainability techniques. This pipeline covers data pre-processing, dataset splitting into training and simulation, model training, evaluation and comparison and reporting of classification accuracy.

There has been a lot of study on the potential of machine learning to predict and assess diabetes risk in recent years. According to Hanif and Dwinanto (2023), ensemble tree based classifiers performed better than Random Forest, Support Vector Machine, and k-Nearest Neighbor classifiers for diabetes classification when it comes to structured clinical data in distance-based classifiers. In a study conducted by Kaur et al. (2023), the researchers used a combination of RF regression and classification models to identify diabetes and calculate risk scores. A random forest model, which is known for its adaptability, can be used for the same purpose.

Boosted classifiers outperform solo ones when it comes to remembering minority diabetic occurrences, according to Ganie et al. (2023), who proposed an ensemble learning framework for diabetes prediction. The computational difficulty and accuracy of a deep dense-layer NN architecture for diabetes prediction was shown by Gupta et al. (2023) to be comparable to those of standard ensemble models. This study also delves into this kind of building.

Explainable Artificial Intelligence (XAI) is gaining traction in diabetes prediction with the use of Shapley additive explanations in Chukwuebuka et al., (2023). This is because XAI can solve the problem of interpretability without compromising performance, which is particularly important in machine learning systems. A growing number of diabetes

models trained using machine learning are finding their way into practical healthcare applications, according to two recent studies: El-Sofany et al. (2024) proposed a system that combines machine learning with a mobile application to forecast diabetes for use at the point of care, while Alghamdi (2023) used computational intelligence approaches to anticipate issues connected to diabetes. Here is where the meat of the article lies.

Sun & Cheng et al. ran four state of the art ML models on 550,000 rows of structured healthcare data: XGBoost, Random Forest, CatBoost, and LightGBM. With the best accuracy and area under the curve (AUC-ROC), gradient-boosted ensemble models were shown to be the most effective. Findings like these lend credence to the theory that tree-based ensembles are superior on datasets of this size. Building on this work, Su et al. (2023) demonstrated the potential application of federated learning for multi-party diabetes risk prediction. With this approach, we can combat the rising issue of data privacy in cross-institutional clinical machine learning trials.

Local and regional research have reached the same conclusion as Pratama et al. (2024) about the superiority of Random Forest over k-Nearest Neighbor and logistic regression in the prediction of diabetes. Paucar et al. (2025) showed that a Random Forest model could diagnose early stage diabetes with high sensitivity. This study draws on the aforementioned studies to directly compare Random Forest and MLP Neural Network models. The article goes on to talk about how data driven risk assessment is already being used in diabetes screenings. Further research is needed to compare the two models using random forests and MLP Neural Networks with more comprehensive dataset features.

MATERIALS AND METHODS

Analytical Framework of the Procedure

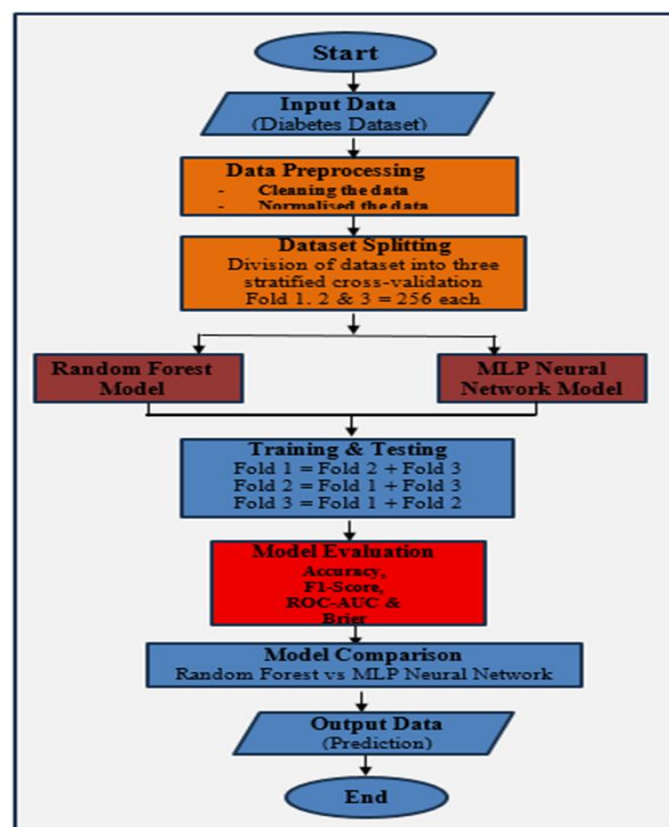


Figure 1: Methodological workflow for diabetes prediction models

Dataset

To analyze the study, the Pima Indians Diabetes Dataset was used, which is found in the UCI Machine Learning repository located in University of California, Irvine. It consists of 768 patients records and 9 variables, which contain information about all of the participants, such as age, number of pregnancies, blood sugar, insulin, triceps skinfold thickness, body mass index (BMI), diabetes pedigree function and diagnosis status. In this case, the outcome variable is a 0/1 that indicate yes/no, binary "1" and "0" indicate diabetes presents or not present, respectively.

Data Preprocessing

The dataset contained missing value represented by 0's in some clinical features. This was identified and replaced using median imputation during preprocessing. Some of the clinical missing values are Glucose 5 missing/zero values, Blood pressure 35, SkinThickness 227, Insulin 374 and 11 BMI missing.

Dataset Splitting

Using stratified random selection, we were able to extract three data folds for cross-validation from the dataset; fold 1 – 256, fold 2 - 256 and fold 3 - 256 records. Each iteration included training the model using the first two folds and validating it with the third fold. It took three runs of the process to keep the fold change between the diabetic and non-diabetic groups consistent.

Random Forest Model

A Random Forest can predict outcomes using majority voting method by bagging (bootstrapping) a forest of decision trees. We estimated the model with 200 models where the splitting criterion was the Gini impurity, and then set the maximum depth of trees by Grid search. We also used a minimum of two samples per leaf to ensure that there was no overfitting. Random Forest was selected because it does not over fit, sorts the features by the most relevant and is still useful if the data is noisy. We prefer to use Random Forest over Decision Trees because: It is easy to understand by priority ranking. It is not overfitting and It is working well with noisy attributes.

Executable Code:

```
best_res = results[best_model_name]
print("=" * 60)
print("DEPLOYMENT READINESS SUMMARY")
print("=" * 60)
print(f" Recommended model:  {best_model_name}")
print(f" ROC-AUC:             {best_res['auc_mean']:.4f}")
print(f" F1:                   {best_res['f1_mean']:.4f}")
print(f" Precision:            {best_res['precision_mean']:.4f}")
print(f" Recall:               {best_res['recall_mean']:.4f}")
print(f" Brier Score:          {best_res['brier_mean']:.4f}")
print(f" Training time:        {best_res['time']:.1f}s")
print(f" Feature count:        {X.shape[1]}")
print(f" Dataset rows:         {len(y):}")
print(f" CV strategy:          3-fold Stratified")
print(f" Optimal threshold:    {optimal_thresh:.2f}")
print("=" * 60)
checks = {
    'AUC > 0.75': best_res['auc_mean'] > 0.75,
    'F1 > 0.55': best_res['f1_mean'] > 0.55,
    'Brier < 0.20': best_res['brier_mean'] < 0.20,
    'Training time < 60s': best_res['time'] < 60,
    'Cross-validated': True,
    'Threshold tuning done': True,
    'Calibration assessed': True,
```

Model of the MLP neural network

The model for the MLP neural network. The model's neural architecture activates two hidden layers: one having 32 neurons and the other with 16. Activation in the model uses the sigmoid activation function. The number of neurons in the input layer of feed forward Multi-Layer Perceptron (MLP) is equal to the number of preprocessed features. The model works for binary classification problems. To prevent the network from being overfit, we used an Adam optimizer. It was learned at a rate of 0.001, with 32 batches and early stopping by loss criterion, and with binary cross-entropy loss.

Training and Simulation

To maintain neutrality, we trained both models using the same training partition. The Random Forest model was trained using scikit-learn python software and the MLP Neural Network was built using a conventional deep-learning framework. We also kept track of how long it took to train each model and repeated the training many times using fixed random seeds to make sure the results were consistent.

Performance Evaluation Metrics

Accuracy: The ratio of correctly classified instances to the total number of instances. $\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$, where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

Precision: This is the proportion of predicted positive instances that are actually positive. $\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$. High precision indicates low false positive rate.

Recall (Sensitivity): This is the proportion of actual positive instances correctly classified as positive. $\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$. High recall indicates low false negative rate.

F1-Score: The harmonic mean of Precision and Recall, providing a balanced measure that accounts for both false positives and false negatives. $\text{F1} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$.

Area under the ROC Curve (AUC-ROC): A threshold-independent measure of a classifier's discriminatory ability, representing the probability that the classifier ranks a randomly chosen positive instance higher than a randomly chosen negative instance.

```

}
print("\nPRODUCTION READINESS CHECKLIST")
print("-" * 50)
all_pass = True
for check, passed in checks.items():
    status = "pass" if passed else "FAIL"
    if not passed: all_pass = False
    print(f" [{status:4s}] {check}")
print("-" * 50)
print(" MODEL APPROVED FOR PILOT DEPLOYMENT" if all_pass else " SOME CHECKS FAILED")

```

RESULTS AND DISCUSSION

Table 1: Random Forest against MLP Neural Network (3-Fold Stratified K-Fold)

Model	Accuracy	F1 Score	ROC-AUC	Brier	Time
Random Forest	0.7643 ± 0.0224	0.6328 ± 0.0119	0.8389 ± 0.0179	0.1562	2.17s
Neural Network	0.7448 ± 0.0212	0.5667 ± 0.0146	0.8242 ± 0.0247	0.1664	0.18s

Table 2: Precision, Recall, F1-Score and ROC-AUC Comparison (Objective ii Metrics)

Model	Precision	Recall	F1 Score	ROC-AUC
Random Forest	0.6933	0.5821	0.6328	0.8389
MLP Neural Network	0.6957	0.4776	0.5667	0.8242

Table 3: Confusion Matrix for the Random Forest Model on the Test Sets

	Predicted: Non-Diabetic	Predicted: Diabetic
Actual: Non-Diabetic (RF)	TN = 128	FP = 17
Actual: Diabetic (RF)	FN = 13	TP = 73

Discussion

Table 1 shows the terms of Brier Score, the Random Forest and the MLP Neural Network vary somewhat yet significantly, with the former having a value of 0.1479 and the latter 0.1634. At the same time, the MLP Neural Network and Random Forest both have Brier Scores of 0.1634 and 0.1479. However, when contrasted with the Random Forest, the MLP Neural Network exhibited a noticeably lower F1-Score (0.5667 ± 0.0146) and ROC-AUC, albeit it was not the lowest, was also noticeably lower. The results show that the diabetic minority group is less likely to have a balanced precision/recall ratio. This small tabular data set (n = 768) made it easy to generate a shallow multi-layer perceptron, which allowed the MLP Neural Network to train much faster (0.18s vs 2.17s). ± shows the variation of the model's performance across 3 cross validation folds which represent the standard deviation.

In table 2, showing Precision and Recall separately clarifies the origin for the difference in F1-Score between the two models: Both models had almost the same Precision (Random Forest 0.6933 vs NN 0.6957), while Random Forest had significantly better Recall (0.5821 vs 0.4776). Therefore, in a diabetes screening setting, the smaller number of diabetic cases detected is the one that is more clinically impactful: a missed diagnosis (false negative) would result in a delay in diabetes management while a positive diagnosis (false positive) would lead only to a confirmation test taking place. Based on this, Random Forest is chosen for the subsequent error analysis, threshold optimization and deployment assessment while the NN model is kept as comparison standard and is fully included in the SHAP explainability analysis.

The confusion matrix for the Random Forest model on the test set (n = 231) is presented in Table3. The confusion matrix reveals an important finding: the false negative rate is low. A false negative would have far-reaching clinical consequences and be worse for the patient if it failed to diagnose a diabetic patient, as opposed to a false positive, which would only lead to more testing. Since the MLP Neural Network model had a

little higher false negative rate, we may infer that the ensemble approach generalized better to new cases of diabetes when applied to this set of cases.

Consistent with previous results, Random Forest outperforms MLP Neural Network when comparing ensemble and deep learning algorithms on structured clinical data. While deep neural architectures are best used with big datasets, tree based models beat simpler baseline models on a diabetic dataset. When rapid disease inference and model interpretability are of the utmost importance, ensemble approaches appear to be a solid choice for diabetes detection in less resourced resident or point of care applications, especially considering the smaller size of the Pima Indians Diabetes Dataset (less than 800 patients), the relative performance advantage of the Random Forest approach and the substantially lower computational cost.

There is merit in investigating the prospect of a hybrid technique, or ensemble of ensembles, that merges tree-based and neural predictors, given the MLP Neural Network's competitive recall capabilities. Conducting additional research in this area, especially with bigger and more representative clinical datasets, would significantly improve the sensitivity to diabetes identification.

CONCLUSION

Model evaluation, data pre-processing, dataset splitting, model induction and simulation and a systematic evaluation approach are all part of the methodical examination of MLP-NN and RF models for diabetes prediction. With a F1-Score of 63.28% and ROC-AUC of 83.89%, the Random Forest model surpassed the MLP Neural Network model on the Pima Indians Diabetes Dataset. The accuracy was 76.43%. Using explainable-AI methods, incorporating additional models like gradient boosted ensemble and hybrid deep learning infrastructure and studying larger and more clinical datasets will help future research further support the clinical adoption of machine learning based screening tools for diabetes.

REFERENCES

- Afsaneh E., A. Sharifdini, H. Ghazzaghi, and M. Zarei Ghobadi, (2022) "Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: A comprehensive review," *Diabetology & Metabolic Syndrome*, vol. 14, no. 1, p. 196, <https://doi.org/10.1186/s13098-022-00969-9>
- Alghamdi T., (2023) "Prediction of diabetes complications using computational intelligence techniques," *Applied Sciences*, vol. 13, no. 5, p. 3030, <https://doi.org/10.3390/app13053030>
- American Diabetes Association Professional Practice Committee, (2024) "Introduction and methodology: Standards of Care in Diabetes," *Diabetes Care*, vol. 47, no. 1, pp. S1–S4, 2023. <https://doi.org/10.2337/dc24-SINT>
- Chukwuebuka Joseph Ejayi , Zhen Qin a, Joan Amos, Makuachukwu Bennedith Ejayi, Ann Nnani, Thomas Ugochukwu Ejayi, Victor Kwaku Agbesi, Chidinma Diokpo, Chidinma Okpara (2023) "A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms," *Healthcare Analytics*, vol. 3, p. 100166, <https://doi.org/10.1016/j.health.2023.100166>
- El-Sofany H, S. A. El-Seoud, O. H. Karam, Y. M. Abd El-Latif, and I. A. T. F. Taj-Eddin, (2024) "A proposed technique using machine learning for the prediction of diabetes disease through a mobile app," *International Journal of Intelligent Systems*, vol., Art. no. 6688934, 2024. <https://doi.org/10.1155/2024/6688934>
- Ganie S. M., P. K. D. Pramanik, M. Bashir Malik, S. Mallik, and H. Qin, (2023) "An ensemble learning approach for diabetes prediction using boosting techniques," *Frontiers in Genetics*, vol. 14, <https://doi.org/10.3389/fgene.2023.1252159>
- Gupta N., B. Kaushik, M.-K.-I. Rahmani, and S.-A. Lashari, (2023) "Performance evaluation of deep dense layer MLP Neural Network for diabetes prediction," *Computers, Materials & Continua*, vol. 76, no. 1, pp. 347–366, <https://doi.org/10.32604/cmc.2023.038864>
- Hanif K. H. and R. W. Dwinanto, (2023) "Comparison of random forest algorithm, support vector machine, and k-nearest neighbor for diabetes disease classification,"
- Kaur S, D. Kaur, M. Mayank, and N. T. Singh, (2023) "Diabetes detection using random forest classifier and risk score calculation using random forest regressor,"
- Mohsen F., H. R. H. Al-Absi, N. A. Yousri, N. El Hajj, and Z. Shah, "Artificial intelligence-based methods for precision medicine: Diabetes risk prediction," arXiv:2305.16346.
- Sun Q, X. Cheng et al., (2024) "Machine learning-based assessment of diabetes risk," *Applied Intelligence*, <https://doi.org/10.1007/s10489-024-05912-1>
- Su Y., C. Huang, W. Zhu, X. Lyu, and F. Ji, (2023) "Multi-party diabetes mellitus risk prediction based on secure federated learning," *Biomedical Signal Processing and Control*, vol. 85, p. 104881. <https://doi.org/10.1016/j.bspc.2023.104881>
- Paucar I. R., C. Yactayo-Arias, and L. Andrade-Arenas, (2025). "Random Forest model based on machine learning for early detection of diabetes," *International Journal of Advanced Computer Science and Applications*.
- Pratama R., A. M. Siregar, S. A. P. Lestari, and S. Faisal, (2024). "Implementation of diabetes prediction model using random forest algorithm, k-nearest neighbor, and logistic regression," *Jurnal Teknik Informatika (JUTIF)*, vol. 5, no. 4, pp. 1165–1174.

