Sigma J Eng Nat Sci, Vol. 44, No. 3, pp. 1816–1825, June, 2026

Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2026.2065

Research Article

Diabetes onset prediction using random forest: A machine learning approach with the Pima Indians diabetes dataset

Jessie R. PARAGAS[1,*], Kent Claire Apple Joy M. PALLOMINA[2], Alexis Luke G. BARLOMENTO[1]

[1]*Department of Information Technology, Eastern Visayas State University, Leyte, 6500, Philippines*
[2]*Department of Information Technology, Visayas State University, Leyte, 6521, Philippines*

ARTICLE INFO

Article history
Received: 15 December 2024
Revised: 13 February 2025
Accepted: 27 March 2025

Keywords:
Diabetes; Machine Learning; PIMA Indians Diabetes Dataset; Random Forest

ABSTRACT

Diabetes, a chronic metabolic disorder, has affected millions of people worldwide, thus it is important to develop accurate predictive models for early intervention and improved patient prognosis. This paper aims to introduce a predictive model for diabetes onset using the Pima Indians Diabetes Dataset and the random forest algorithm, a machine learning approach that handles complex, nonlinear data well. The model was built with the goal of optimizing accuracy, reliability and interpretability for clinical applications. Data normalization and class imbalance management with the synthetic minority over-sampling Technique were among the data pre-processing steps that resulted in a substantial improvement in model performance. The random forest model exhibited good performance in predicting diabetes with an accuracy of 80%, precision of 77% and recall of 68%. The feature importance analysis revealed that glucose levels, body mass index and age are the most important predictors. To improve the model interpretability and clinical utility, the predictions were interpreted using the Shapley Additive Explanations, thereby allowing the healthcare professionals to trust and utilize the model for risk assessment. The results underscore the potential of machine learning, particularly the random forest algorithm in aiding early diagnosis of diabetes and assisting healthcare providers to identify high-risk individuals for timely intervention.

Cite this article as: Paragas JR, Pallomina KCAJM, Barlomento ALG. Diabetes onset prediction using random forest: A machine learning approach with the Pima Indians diabetes dataset. Sigma J Eng Nat Sci 2026;44(3):1816–1825.

INTRODUCTION

Diabetes is a long-lasting metabolic condition marked by high blood sugar levels, impacting more than 529 million people around the globe [1,2]. Factors such as lifestyle and genetics drove the rapid rise in type 2 diabetes, placing immense pressure on healthcare systems globally, mainly due to complications like cardiovascular disease [3,4]. The increasing prevalence of diabetes resulted in severe health outcomes, including renal failure, visual impairment, and amputations, which had a profound impact on public health, healthcare costs and the quality of life of individuals [5,6]. To mitigate this burden, it was crucial to implement

*Corresponding author.
*E-mail address: jessie.paragas@evsu.edu.ph

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

effective management and prevention strategies to decrease mortality and morbidity and improve overall well-being [7,8].

However, even with the progress on diabetes management, the current prediction approaches are often not accurate enough, as they are not able to consider the individual difference in the treatment responses. Traditional approaches such as logistic regression and decision trees, failed to model the complex interactions among a large number of health indicators, and thus lower the prediction accuracy. Machine learning approaches, especially the random forests, provided a solution by using large datasets and modeling nonlinear relationships, thus improved the prediction performance [3,9].

Recent studies [7] and [8] have shown the effectiveness of machine learning algorithms including random forests in signaling diabetes. Accurate predictive models and prompt intervention resulted in a significant decrease in complications and better patient outcomes. Random forest models demonstrated high accuracy and reliability in dealing with high-dimensional, non-linear data using data such as EHRs and genetic information. This study suggests that the random forest algorithm could offer more accurate and personalized predictions of diabetes onset compared to conventional models, such as logistic regression and decision trees.

The study used Pima Indians Diabetes Dataset (PIDD) one of the most used datasets in diabetes study to develop a random forest based predictive model [10,11]. The PIDD contains important health indicators such as glucose levels and body mass index (BMI). The health metrics in the dataset made it an ideal candidate for random forest as it can infer complex relationships between features and improve the accuracy of the prediction. The objectives of this study were as follows: 1) to develop a robust predictive model for the onset of diabetes using random forest algorithms, 2) to optimize the model for better accuracy and reliability, and 3) to ensure the interpretability and transparency of the model for clinical settings. The goal of this study was to develop personalized approaches to diabetes prediction to improve prediction accuracy and to make personalized interventions to improve overall health outcomes.

REVIEW OF RELATED LITERATURE

Recent literature on diabetes has been concerned with the application of machine learning algorithms for the improvement of diagnosis, management and prediction of the disease [12]. Various techniques such as decision trees, support vector machines (SVMs), neural networks and ensemble methods have been used on large datasets to study and discover patterns related to diabetes risk factors and complications [13-15]. Studies have shown that logistic regression and random forest algorithms are able to effectively predict the onset of diabetes by studying demographic, lifestyle and genetic data [16].

Machine learning algorithms have been promising to diagnose diabetes especially random forests [9,17]. Random forests were particularly beneficial to handle large datasets, to recognize complex non-linear relationships between different health indicators and to improve the prediction accuracy [18]. Random forests outperformed traditional models such as logistic regression that yielded around 70% accuracy in similar studies. This finding is in line with the goal of this study to build a robust predictive model with optimal accuracy and reliability [10,19,20].

Random forests had many advantages but they also had a major difficulty which is their interpretability. The lack of transparency was a hindrance to apply it in a clinical setting since health professionals needed to know why such predictions were made [21]. To overcome these challenges, methods such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) were invented to give insights into how much specific features contributed to the individual prediction [22-24]. These methods improved the transparency by clarifying how much each feature contributed to the prediction [25]. For example, SHAP values helped better understanding of the decision-making process of the model by illustrating the influence of each feature on the prediction, leading to increased trust from healthcare professionals [26]. By including these techniques, the study sought to ensure that the model's predictions were not only accurate, but also explainable and actionable in clinical settings [27].

Another improvement in the current approaches was the handling of imbalanced datasets, which can significantly affect the model's performance. The PIDD used in this work had a class imbalance of 600 nondiabetic and 168 diabetic instances [11,28]. Cost sensitive learning and Synthetic Minority Over-sampling Technique (SMOTE) were widely used to handle the imbalance [29,30]. SMOTE increased the recall, which allowed the model to identify positive cases of diabetes better. In this work, SMOTE was used to balance the dataset to optimize the model for high accuracy and reliability [31,32].

PIDD was also one of the most common datasets used in diabetes research and served as the foundation of our analysis. It included key health indicators such as age, BMI, and glucose level that are well suited to the random forest's capacity to model complex feature interactions. Recent studies using PIDD demonstrated that machine learning approaches can achieve high accuracy in predicting diabetes which reinforced the relevance of this dataset to our work [33,34]. Focusing on a dataset with established links to diabetes risk factors, our study sought to build on the existing literature while providing valuable insights to the growing field of diabetes prediction.

MATERIALS AND METHODS

Data Description

The Pima Indians Diabetes Dataset (PIDD) from the National Institute of Diabetes and Digestive and Kidney Diseases was used in the present study to analyze the high prevalence of diabetes among the Pima Indian population [11]. The dataset had 768 records with nine attributes, including eight predictor variables, such as the number of pregnancies, plasma glucose concentration after a 2-hour oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin, body mass index (BMI), diabetes pedigree function and age [35,36], and the ninth variable denoting the target for the presence (1) or absence (0) of diabetes [37]. The unique genetic and lifestyle factors among the Pima Indian population contributed significantly to a high incidence of diabetes, which is why this dataset was particularly relevant.

The PIDD was processed using Python which is a versatile programming language that is widely used in data analysis and machine learning. The main libraries used in this study were Pandas for data manipulation , NumPy for numerical operations and Scikit-learn for machine learning model development. Before model development, extensive exploratory data analysis (EDA) was performed which included calculation of descriptive statistics and visualization of distributions to understand the characteristics of the data. Initial insights showed that 34.9 % of the records indicated diabetes, indicating the balance of the dataset.

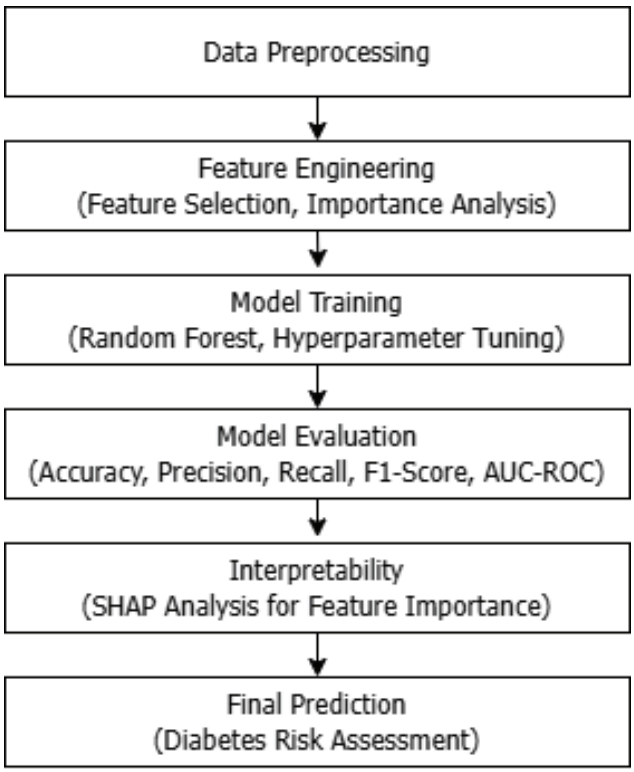

Figure 1. Architecture of the implemented model.

Figure 1 presents the overall architecture of the proposed diabetes prediction framework using the Random Forest algorithm. The workflow outlines the key stages involved in the development of the predictive model, beginning with data preprocessing and feature engineering, followed by model training and evaluation. To enhance transparency and clinical applicability, the framework incorporates SHAP-based interpretability before generating the final diabetes risk prediction. The architecture provides a structured overview of the methodology adopted in this study, while the specific processes and outcomes of each stage are discussed in the succeeding sections.

Data Preprocessing

Data preprocessing involved various techniques such as normalization, imputation, and outlier detection. o improve the performance of the models, the continuous variables were normalized using the StandardScaler from scikit-learn to make the features, such as glucose concentration and BMI, comparable in scale. Normalization is necessary in order not to allow any component to dominate the model due to the differences in scale [38,39]. Missing values were treated using the SimpleImputer module to replace them with the mean or median of the feature. The distribution of the features was visualized using Matplotlib and Seaborn, which was helpful to identify and deal with extreme values that could affect the model performance. Outliers were identified using methods such as the IQR (Interquartile Range).

Statistical Analysis

Descriptive statistics were performed to describe the data set and analyze the relationship between health attributes and onset of diabetes [40]. Mean and median and standard deviation were calculated for all variables. Correlation analysis was done with the Pearson or Spearman correlation coefficients to test the strength and direction of the relationship between predictor variables and the target variable (presence of diabetes) [41]. Pearson coefficients were used for normally distributed data and Spearman coefficients for non normally distributed variables [42]. Further significance testing was done by t-tests and ANOVA to test whether the differences in predictor variables were statistically significant between diabetic and nondiabetic groups at a significance level of $p<0.05$ [43].

Selection of a Machine Learning Model

For this study, the random forest algorithm was selected as the machine learning model and was implemented using the Scikit-learn library in Python [44]. This library was a comprehensive platform for the development and evaluation of machine learning models with utilities for training, testing and tuning [41,45]. This model alleviated the limitations of traditional prediction techniques by considering the complex interactions of multiple health indicators

contained in the PIDD. The random forest model alleviated these limitations by aggregating the predictions of a multitude of decision trees which reduced overfitting and increased robustness. This was especially important for medical datasets like PIDD where individual responses to health indicators varied.

Random forests also gave some insight into the importance of features which allowed us to identify the most compelling health metrics that could be used to predict diabetes onset. Random forests classify tasks by building multiple decision trees and aggregating their predictions , avoiding overfitting and increasing robustness to noisy data [46,47]. The RandomForestClassifier module was used to develop the predictive model, taking advantage of the algorithm's ensemble nature to achieve higher classification accuracy. This ensemble approach was able to capture interactions among variables, which is important in medical datasets where the development of diabetes is influenced by factors such as glucose levels and BMI. Random forests also provided feature importance rankings, which provided insight into the variables that were strong predictors of diabetes [48]. The rankings were visualized with the visualization tools provided by scikit-learn , which underlined influential variables such as glucose levels and BMI. This interpretability was important for healthcare professionals to be able to identify interventions for individuals at risk.

However, random forests may present computational challenges in the hyperparameter optimization [49]. Other algorithms, such as SVMs, were considered but not used because of the complexity of tuning the parameters, such as the kernel functions and the regularization parameters [50]. The choice of random forests was based on the balanced performance, interpretability, and robustness to PIDD. The reason to choose the algorithm was based on the fact that it can model complex nonlinear relationships of large datasets, which is the reason for the selection of random forests for this study, which is a critical method for dealing with high-dimensional medical data such as PIDD. Random forests also provided a trade-off between accuracy and interpretability through the feature importance rankings. The rankings provided healthcare professionals with actionable insights by identifying key predictors of diabetes such as glucose levels and BMI. While other models like SVMs can sometimes provide higher accuracy, they are less interpretable and hence random forests are more appropriate for clinical applications where interpretability is important.

Model Training and Hyperparameter Tuning

Model training and hyperparameter tuning were extensively performed using Python and Scikit-learn. The PIDD was split into training (70%), validation (15%) and test (15%) sets using the train_test_split function in Scikit-learn, ensuring that the model is evaluated without bias. The model learnt different combinations of features and target labels from the training set and the validation set was used to evaluate the performance of the model and tune the hyperparameters during training [51].

The researchers employed Scikit-learn's GridSearchCV and RandomizedSearchCV to fine-tune hyperparameters. These utilities provided an efficient search of the pre-defined hyperparameter space by sampling combinations of parameters, such as the number of trees (n_estimators), the maximum depth of a tree (max_depth), and the minimum number of samples required to split an internal node (min_samples_split). The authors used K-fold cross-validation (k=5) to ensure the model's performance was consistent across the different data splits, minimizing the possibility of overfitting [52,53]. The performance of the random forest model on the validation and test sets was assessed using a set of model evaluation metrics available in Scikit-learn, including accuracy, precision, recall, and the F1-score, which provide an indication of its capacity to generalize to unseen data during the training process [54].

Data Interpretation

The authors analyzed the clinical significance of the random forest model results. The accuracy metrics, precision and recall, were analyzed to determine the model's effectiveness in accurately predicting patients at risk of diabetes [55]. High precision means that the model made accurate predictions thus minimizing false positives and unnecessary interventions. High recall means that the model was able to identify most of the real cases of diabetes which is crucial for early diagnosis and treatment [56].

The model pinpointed major predictors such as glucose levels and BMI and offered practical insights for medical practitioners within clinical environments. The feature importance evaluation suggested that the most important markers for diabetes onset were glucose level, BMI and age thus giving the healthcare practitioners defined targets for intervention. The implementation of this predictive model into electronic health record (EHR) systems allowed healthcare practitioners to provide real-time risk assessments for proactive management of patients' predisposition to onset of diabetes. This met the study's aim of ensuring that the model's predictions are both accurate and actionable in clinical settings enabling early intervention and personalized treatment. Healthcare practitioners could use these insights to guide their interventions towards high-risk patients thus leading to improved early diagnosis and targeted treatment.

Evaluation Metrics

Different performance metrics provided by Scikit-learn were used for the evaluation process. In particular, the accuracy_score, precision_score, recall_score and F1_score metrics were used to quantify the classification performance of the model. The researcher also generated the ROC curves and Area Under the Curve (AUC) scores using the roc_auc_score function of Scikit-learn to evaluate the

ability of the model to differentiate between diabetic and nondiabetic patients [57].

The confusion_matrix function of Scikit-learn was used to generate confusion matrices for visual representation of true positives, true negatives, false positives and false negatives [58]. The analysis helped identify areas of model misclassification and guide future model modifications [59]. The entire analysis pipeline from data preprocessing to model evaluation was performed in Python, with substantial use of Jupyter Notebook environment for code development, visualization and documentation [56].

RESULTS AND DISCUSSION

Model Training and Performance

The analysis was performed using PIDD, which included 768 records with important medical and demographic features for diabetes prediction. To ensure a comprehensive evaluation, the dataset was split into training (70%), validation (15%), and testing (15%) subsets. The authors built the random forest model based on these training records, with primary features like BMI, glucose levels, and age being significant in diabetes prediction.

Hyperparameter tuning was carried out by grid search and 5-fold cross validation to optimize model parameters such as number of trees (n_estimators), max tree depth (max_depth), and minimum samples to split (min_samples_split). This ensured the model generalized sufficiently to ignore data. This is important in clinical applications, where prediction accuracy can have serious healthcare implications.

Accuracy, Precision, Recall, and F1-Score

The accuracy score of 80% in Table 1 indicated the model's overall ability to correctly classify cases of diabetes. The precision score of 77% implied that the model was generally correct in the prediction of diabetes, while the recall score of 68% indicated that the model missed some diagnoses (false negatives). The F1 score of 72%, which takes into account both precision and recall, highlighted the need to improve the sensitivity of the model to true positives, which is important for reducing missed cases of diabetes that could result in serious health consequences.

Table 1. Performance metrics of the random forest model for diabetes prediction

Metric	Score
Accuracy	0.80
Precision	0.77
Recall	0.68
F1-score	0.72
AUC-ROC	0.84

Misclassification Analysis

As shown in confusion matrix in Figure 2, the model classified 92 nondiabetic cases as true negatives and 31 diabetic cases as true positives. However, the model misclassified 18 diabetic subjects as nondiabetic (false negatives) and 13 nondiabetic subjects as diabetic (false positives).

False negatives were particularly problematic in clinical settings as it meant missed diabetes diagnoses which could lead to the untreated progression of the disease, whereas false positives could lead to unnecessary interventions. These misclassifications demonstrated the need for the models to be further refined through feature engineering or tuning to improve sensitivity (recall) without significantly reducing precision.

Effect of Handling Imbalanced Data

To handle the class imbalance (600 nondiabetic and 168 diabetic instances), SMOTE was used to oversample the minority class (diabetic cases). The model had a lower recall before applying SMOTE, indicating that it had difficulties identifying positive diabetes cases. The recall increased to 68% after applying SMOTE, which shows the importance of handling data imbalance in reducing the bias towards the majority class as shown in Figure 3. Applying SMOTE to balance the dataset was important to prevent the model to be biased towards the majority class (nondiabetic cases) which is important to maximize the model's reliability. SMOTE improved the recall of the model especially in correctly identifying true-positive diabetes cases, which directly helped to achieve the goal of developing a robust and reliable predictive model. It was important that the model could identify the high-risk patients without a substantial increase in false positives to make it useful clinically.

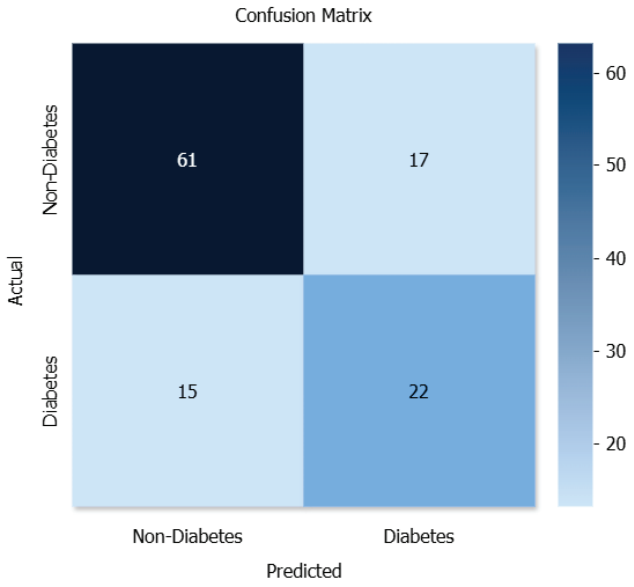

Figure 2. Confusion matrix of PIDD based on the random forest model.

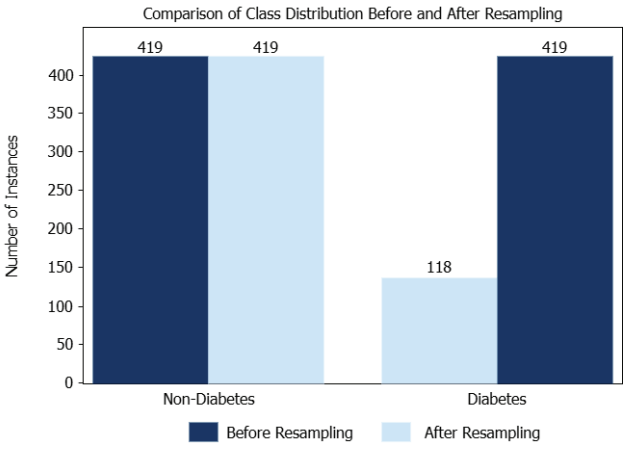

Figure 3. Class distribution before and after applying SMOTE to balance diabetic and nondiabetic instances.

However, the precision is slightly decreased which shows the trade-off between these two metrics when working with imbalanced datasets.

Analysis of the model before and after SMOTE showed that precision slightly lowered (80% to 77%) and recall increased significantly (60% to 68%) meaning the model was more likely to find diabetes cases without a significant increase in false positives. This improvement was important to properly find high-risk patients and receive timely intervention.

Feature Importance and Clinical Relevance

The feature importance analysis in Figure 4 the identified age, glucose levels and BMI as the features that had the biggest impact in predicting the onset of diabetes which is in line with known clinical risk factors. These insights were useful for health care professionals to direct targeted interventions such as physicians to target those with high glucose and BMI for preventive measures or early intervention.

The model also revealed other significant predictors such as number of pregnancies and diabetes pedigree function that impact the risk of diabetes. The results proved that the potential of machine learning to find complex relationships among health parameters that might be ignored by traditional models.

Moreover, the interpretability of the random forest model could have been further improved using SHAP-based approaches. SHAP values could have brought more insight into how a particular feature was influencing individual predictions, bridging the gap between black-box machine learning and transparent clinical decision making. Such transparency would be key in establishing trust and encouraging uptake in healthcare settings.

Comparison with Existing Models

The random forest outperformed traditional models such as logistic regression. Similar studies have reported accuracy in the order of 70% for logistic regression, while the random forest achieved an accuracy of 80%, showing its ability to model nonlinear interactions between features, which proved essential for complex medical datasets such as PIDD, as shown in Figure. 5.

Other machine learning models (such as SVM and neural networks) were also considered by the researchers but not implemented due to the higher complexity in parameter tuning and interpretability issues. Nevertheless, future studies could have benefited from a more extensive comparison with these models to explore trade-offs between performance and interpretability.

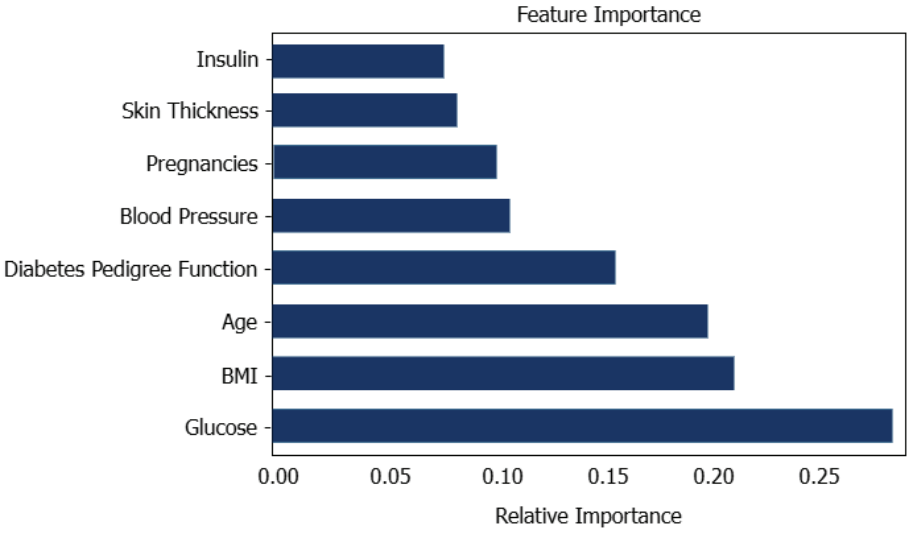

Figure 4. Most important features for diabetes prediction like glucose levels and BMI.

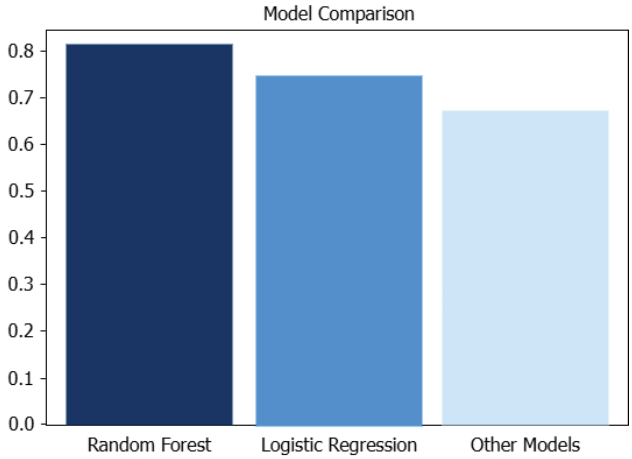

Figure 5. Comparison of the random forest model's accuracy with traditional models.

Receiver Operating Characteristics Curve

The ROC curve in Figure. 6 shows the trade-off between the model's sensitivity (true positive rate) and specificity (true negative rate) at different points. The AUC-ROC score of 0.84 indicated excellent discrimination, further confirming the model's readiness for real-world clinical applications where accurate discrimination between diabetic and nondiabetic patients is critical.

Such performance indicated that researchers could use the model with confidence to detect diabetes early, especially in high-risk populations such as the Pima Indians. However, further refinement, perhaps with ensemble or hybrid methods, could have elevated these metrics further, ensuring both sensitivity and precision were optimized.

The results of the random forest model for diabetes prediction with PIDD has important implications in healthcare applications. The model exhibited an impressive ability to discriminate between diabetes and nondiabetes cases with 80% accuracy and 77% precision. This level of performance highlights the potential of the model as a decision-support tool, helping healthcare providers in early identification of patients who are at risk, an essential component for timely interventions.

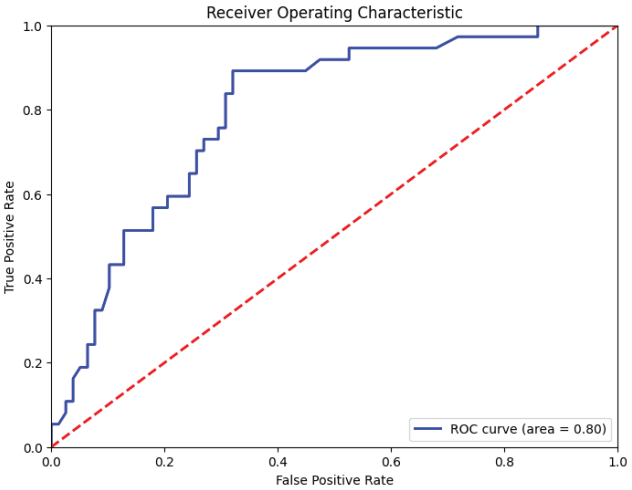

Figure 6. ROC curve for the random forest model.

The model achieved a recall of 68% and still misclassified some cases, so it had to be balanced with the model's overall strengths. The analysis of misclassification showed that the model correctly classified 92 nondiabetic cases, which confirmed its reliability. Future work could have considered the adjustment of the classification threshold based on clinical priorities to address the false negatives identified, thus improving the sensitivity without increasing false positives too much.

Feature importance analysis revealed that age, glucose levels, and BMI were significant predictors of diabetes onset. These findings supported targeted clinical interventions and allowed health care providers to focus screening on patients with these risk factors. Incorporation of the predictive model into electronic health record systems would enable healthcare professionals to conduct real-time risk assessments and deliver timely preventative care for high-risk patients.

CONCLUSION

This study proposed a promising predictive model to predict the onset of diabetes using the random forest algorithm dedicated to PIDD. The model achieved an accuracy of 80%, precision of 77% and recall of 68%, demonstrating its ability to distinguish diabetic and non-diabetic cases. Researchers found that levels of glucose, BMI and age were significant predictors, emphasising the importance of these factors in predicting risk for diabetes.

The results indicate that the model could be a powerful tool in clinical decision making to recognize and intervene early with at risk patients. Moreover, the research suggested that the sensitivity needs to be enhanced to avoid false negatives which future researchers could address by adjusting the classification thresholds or improving the model.

The study was also performed on random forest algorithm but it would be interesting to explore other machine learning techniques to predict diabetes. Future research may explore hybrid approaches that combine random forest strengths with other machine learning algorithms to improve interpretability and accuracy. For example, the combination of ensemble methods or deep learning models, with transparent predictions through interpretability tools such as SHAP, could have produced more robust models for clinical application. Also, extending the study to larger and more diverse datasets would have increased the model's ability to generalize to broader populations, thus increasing reliability and clinical relevance.

Finally, the use of machine learning model, random forest algorithm, for diabetes prediction demonstrated a

promising future for personalized healthcare. This study offered useful insights by highlighting the interpretability and clinical significance of the model that could result in improved diabetes control and better health outcomes for patients. The implementation of these results in clinical practice could have resulted in more proactive and efficient healthcare interventions for people at risk for diabetes.

ACKNOWLEDGEMENTS

The researchers thank God Almighty for providing them the strength to complete the paper. The researchers express their sincere gratitude to the National Institute of Diabetes and Digestive and Kidney Diseases for providing the PIDD resource that was vital to this research. Lastly, the researchers thank Visayas State University Alangalang and Eastern Visayas State University Main Campus for their unfailing support and encouragement throughout the journey.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

[1] Hossain MJ, Al-Mamun M, Islam MR. Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. Health Sci Rep 2024;7:e2004. [CrossRef]

[2] Bansal T. Nutritional strategies for diabetes prevention: Exploring the role of diet and lifestyle choices. Nutr Food Process 2023;6:1-4. [CrossRef]

[3] IDHS: Williams semi-annual report #8. Available at: https://www.dhs.state.il.us/page.aspx?item=78778 Accessed on Oct 11, 2024.

[4] Bansal T. Nutritional strategies for diabetes prevention: Exploring the role of diet and lifestyle choices. Nutr Food Process 2023;6:1-4. [CrossRef]

[5] Oikonomou EK, Khera R. Machine learning in precision diabetes care and cardiovascular risk prediction. Cardiovasc Diabetol 2023;22:259. [CrossRef]

[6] Mohsen F, Al-Absi HRH, Yousri NA, El Hajj N, Shah Z. A scoping review of artificial intelligence-based methods for diabetes risk prediction. NPJ Digit Med 2023;6:197. [CrossRef]

[7] International Diabetes Federation. IDF diabetes atlas. 10th ed. Brussels, Belgium: International Diabetes Federation; 2021

[8] World Health Organization. Diabetes. Available at: https://www.who.int/news-room/fact-sheets/detail/diabetes Accessed on Oct 8, 2024.

[9] Introduction: Standards of medical care in diabetes-2021. Diabetes Care 2021;44(Suppl 1):S1-S2. [CrossRef]

[10] Fregoso-Aparicio L, Noguez J, Montesinos L, García-García JA. Machine learning and deep learning predictive models for type 2 diabetes: A systematic review. Diabetol Metab Syndr 2021;13:148. [CrossRef]

[11] Badawy M, Ramadan N, Hefny HA. Healthcare predictive analytics using machine learning and deep learning techniques: A survey. J Electr Syst Inf Technol 2023;10:40. [CrossRef]

[12] Kaggle: Your machine learning and data science community. Available at: https://www.kaggle.com/ Accessed on Oct 9, 2024.

[13] Schonlau M, Zou RY. The random forest algorithm for statistical learning. Stata J 2020;20:3-29. [CrossRef]

[14] Mangkunegara IS. The use of artificial intelligence in disease diagnosis: A systematic literature review. J Adv Health Inform Res 2023;1:142-156. [CrossRef]

[15] Kothuri SN, Tanusree M, Chaitanya NLS, Chaitanya SMK. Precision agriculture advisor. Research Square 2024. doi: 10.21203/rs.3.rs-4677379/v1 [Epub ahead of print][CrossRef]

[16] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017;15:104-116. [CrossRef]

[17] Sarifuddin EF. Investing in soft commodities - Latest business and travel news. Available at: https://www.jvidusun.co.id/investing-in-soft-commodities/ Accessed on Oct 11, 2024.

[18] Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak 2011;11:51. [CrossRef]

[19] Reddy SS, Sethi N, Rajender R. A comprehensive analysis of machine learning techniques for

incessant prediction of diabetes mellitus. Int J Grid Distrib Comput 2020;13.
[20] Griffin LP, Casselberry GA, Hart KM, Jordaan A, Becker SL, Novak AJ, et al. A novel framework to predict relative habitat selection in aquatic systems: Applying machine learning and resource selection functions to acoustic telemetry data from multiple shark species. Front Mar Sci 2021;8:631262. [CrossRef]
[21] Sharma T, Shah M. A comprehensive review of machine learning techniques on diabetes detection. Vis Comput Ind Biomed Art 2021;4:30. [CrossRef]
[22] Zhu T, Li K, Herrero P, Georgiou P. Deep learning for diabetes: A systematic review. IEEE J Biomed Health Inform 2021;25:2744-2757. [CrossRef]
[23] Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. Wiley Interdiscip Rev Data Min Knowl Discov 2020;10:e1379. [CrossRef]
[24] Bopche R, Gustad LT, Afset JE, Ehrnström B, Damås JK, Nytrø Ø. Inhospital mortality, readmission, and prolonged length of stay risk prediction leveraging historical electronic health records. Health Informatics Apr 16. Preprint. doi: 10.1101/2024.04.15.243058752024 [CrossRef]
[25] SA, SR. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. Decis Anal J 2023;7:100230. [CrossRef]
[26] Kabir J, Chakraborty A, Mahmood AA, Chakaraborty A. Exploring explainable artificial intelligence: A comparative analysis of interpretability techniques. Int J Adv Res Comput Commun Eng 2024;13:1-6. [CrossRef]
[27] Li J, Zhang Y, He S, Tang Y. Interpretable mortality prediction model for ICU patients with pneumonia: Using Shapley additive explanation method. BMC Pulm Med 2024;24:447. [CrossRef]
[28] Pallomina KCAJM, Brosas DG. Aligning educational outcomes with industry demands: An automated competency assessment tool based on the PSF-AAI framework for analytics and AI careers. In: 2025 IEEE 16th Control and System Graduate Research Colloquium (ICSGRC); Shah Alam, Malaysia. 2025. p. 42-46. [CrossRef]
[29] Wang J, Wiens J, Lundberg S. Shapley Flow: A graph-based approach to interpreting model predictions. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics. PMLR; 2021. p. 721-729. Available at: https://proceedings.mlr.press/v130/wang21b.html
[30] Pallomina KCAJM, Brosas DG. Integrating skills competency assessment and academic performance for career pathway mapping in analytics and AI. In: 2025 Eighth International Conference on Vocational Education and Electrical Engineering (ICVEE); Surabaya, Indonesia. 2025. p. 15-19. [CrossRef]
[31] Houston L, Yu P, Martin A, Probst Y. Heterogeneity in clinical research data quality monitoring: A national survey. J Biomed Inform 2020;108:103491. [CrossRef]
[32] He Z, Tang X, Yang X, Guo Y, George TJ, Charness N, et al. Clinical trial generalizability assessment in the big data era: A review. Clin Transl Sci 2020;13:675-684. [CrossRef]
[33] Yadav P, Sharma SC, Mahadeva R, Patole SP. Exploring hyper-parameters and feature selection for predicting non-communicable chronic disease using stacking classifier. IEEE Access 2023;11:80030-80055. [CrossRef]
[34] Alnowaiser K. Improving healthcare prediction of diabetic patients using KNN imputed features and tri-ensemble model. IEEE Access 2024;12:16783-16793. [CrossRef]
[35] Anwar NHK, Saian R. Predictive accuracy for two diabetes datasets using Ant-Miner algorithm. 2020;9.
[36] Abnoosian K, Farnoosh R, Behzadi MH. Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. BMC Bioinformatics 2023;24:337. [CrossRef]
[37] Sarkar PJ, Pawar DS. Enhancing early detection and prediction of diabetes mellitus in patients of Indian origin through rigorous machine learning techniques with comprehensive models evaluation. Intell Syst Appl Eng 2024;12:634-639.
[38] Sreelatha S, Shivashetty V. Proactive cervical cancer risk assessment using data-driven analytics. IAES Int J Artif Intell 2024;13:4301-4311. [CrossRef]
[39] Raschka S, Mirjalili V. Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. Birmingham, UK: Packt Publishing; 2019.
[40] Robillard JM, Sporn AB. Static versus interactive online resources about dementia: A comparison of readability scores. Gerontechnology 2018;17:29-37. [CrossRef]
[41] Durmus M. A hands-on introduction to essential Python libraries and frameworks (with code samples). Independently Published.
[42] Orliaguet L, Ejlalmanesh T, Humbert A, Ballaire R, Diedisheim M, Julla JB. Interferon regulatory factor 5 represses oxidative respiration in human and murine macrophages by inhibition of mitochondrial matrix protein GHITM. Research Square 2022. Preprint. doi: 10.21203/rs.3.rs-477956/v1 [CrossRef]
[43] Bae H, Park SY, Kim CE. A practical guide to implementing artificial intelligence in traditional East Asian medicine research. Integr Med Res 2024;13:101067. [CrossRef]
[44] Kim E, Yang SM, Jung DH, Kim HY. Differentiation between Weissella cibaria and Weissella confusa using machine-learning-combined MALDI-TOF MS. Int J Mol Sci 2023;24:11009. [CrossRef]

[45] Uddin MZ. Machine learning and Python for human behavior, emotion, and health status analysis. 1st ed. Boca Raton: CRC Press; 2024. [CrossRef]

[46] Tania M. Automated classification and trend analysis of large language model survey papers using machine learning and natural language processing techniques. Song Z, Ke K. Prediction for CET-4 based on random forest. Procedia Comput Sci 2023;228:429-437. [CrossRef]

[47] He L, Levine RA, Fan J, Beemer J, Stronach J. Random forest as a predictive analytics alternative to regression in institutional research. Pract Assess Res Eval 2018;23:1-16.

[48] Sarker IH. Machine learning: Algorithms, real-world applications and research directions. SN Comput Sci 2021;2:160. [CrossRef]

[49] Kong Y, Yu T. A deep neural network model using random forest to extract feature representation for gene expression data classification. Sci Rep 2018;8:16477. [CrossRef]

[50] Zand N. nillzand/mushroom-classification-decision-tree-visualization. Python. 2023 Apr 7 [cited 2024 Oct 11]. Available from: https://github.com/nillzand/mushroom-classification-decision-tree-visualization

[51] F1 score [Internet]. FasterCapital.. Available at: https://fastercapital.com/keyword/f1-score.html Accessed on Oct 11, 2024.

[52] Ait Tchakoucht T, Elkari B, Chaibi Y, Kousksou T. Random forest with feature selection and K-fold cross validation for predicting the electrical and thermal efficiencies of air-based photovoltaic-thermal systems. Energy Rep 2024;12:988-999. [CrossRef]

[53] Sarkar O, et al. Multi-scale CNN: An explainable AI-integrated unique deep learning framework for lung-affected disease classification. Technologies 2023;11:134. [CrossRef]

[54] Nguyen LP, et al. The utilization of machine learning algorithms for assisting physicians in the diagnosis of diabetes. Diagnostics 2023;13:2087. [CrossRef]

[55] Jose R, Syed F, Thomas A, Toma M. Cardiovascular health management in diabetic patients with machine-learning-driven predictions and interventions. Appl Sci 2024;14:2132. [CrossRef]

[56] Arcipreste BM. Product complaint understanding using NLP techniques.

[57] Anantrasirichai N, et al. Adaptive-weighted bilateral filtering and other preprocessing techniques for optical coherence tomography. Comput Med Imaging Graph 2014;38:526-539. [CrossRef]

[58] Chong KM. Privacy-preserving healthcare informatics: A review. ITM Web Conf 2021;36:04005. [CrossRef]

[59] Wirth FN, Meurers T, Johns M, Prasser F. Privacy-preserving data sharing infrastructures for medical research: Systematization and comparison. BMC Med Inform Decis Mak 2021;21:242. [CrossRef]