iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VIII **Month of publication:** August 2025

DOI: https://doi.org/10.22214/ijraset.2025.73763

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VIII Aug 2025- Available at www.ijraset.com

Enhanced Diabetes Risk Prediction Using Ensemble Learning on Real-World Nhanes Clinical Data

Madhusudhan K[1], Prajwal S[2], Sarang Kumar S[3], Vinay K[4]
Department of MCA, SJB Institute of Technology Affiliated to VTU, Bangalore

Abstract: ***In this study, we explore a machine literacy approach for forecasting early-stage diabetes using real clinical data gathered from NHANES (National Health and Nutrition Examination Survey). The dataset includes crucial medical pointers similar as glucose, insulin, BMI, blood pressure, and cholesterol. Two ensemble classifiers Random Forest and XGBoost were trained and estimated. The Random Forest model achieved the stylish performance, with 94% delicacy and an AUC of 0.96. Our engineered features, such as the Glucose-to-Insulin Ratio and MAP, helped our model better distinguish diabetic individuals. Our findings highlight how ensemble learning can be effectively used in real-world settings for interpretable clinical screening.***
Keywords: Diabetes Prediction, NHANES Dataset, Ensemble Learning, Random Forest, XGBoost, Feature Engineering, ROC Curve, SHAP, Clinical Data

I. INTRODUCTION

Diabetes mellitus is a global health burden affecting over 400 million people worldwide, with its frequence rising steadily. It's a habitual metabolic complaint characterized by sustained high blood sugar situation caused either by shy insulin product (Type 1) or by insulin resistance (Type 2). Type 2 diabetes, the more common form, is frequently preventable through beforehand intervention and lifestyle revision. However, delayed opinion continues to pose a challenge in clinical practice.

Recent advancements in machine learning (ML) have demonstrated significant potential in disease prediction and early diagnosis. ML techniques can uncover complex relationships between various health indicators and patient demographics, facilitating more accurate risk stratification and personalized care.

Many prior studies have relied on the PIMA Indian Diabetes dataset, which, although commonly used, lacks diversity and depth in clinical measurements. In contrast, this study uses real-world clinical data from the National Health and Nutrition Examination Survey (NHANES), which provides comprehensive medical, laboratory, and lifestyle features.

We estimate the performance of two ensemble- grounded classifiers Random Forest and XGBoost for prognosticating diabetes threat. The study also introduces engineered features such as the Glucose-to-Insulin Ratio and Mean Arterial Pressure (MAP), which help ameliorate model performance. By using rich clinical data and feature engineering, our objective is to make an accurate and interpretable model that supports early discovery and clinical decision-making.

II. LITERATURE REVIEW

Machine learning has gained adding attention for its effectiveness in medical diagnostics, including diabetes vaticination Srivastava et al. [11] provide a detailed review of ML models in this domain. Several experimenters have proposed algorithms to enhance classification accuracy and generalizability on clinical datasets.

Shokrekhodaei et al. [1] delved the eventuality of optical detectors combined with machine learning models for non-invasive glucose monitoring. Their work emphasized integrating physiological signals with learning algorithms for bettered detection.

Hasan et al. [2] employed an ensemble of classifiers, including Support Vector Machine, Naïve Bayes, and Random Forest, on medical datasets and demonstrated improved accuracy over standalone models.

Fitriyani et al. [3] introduced an ensemble learning approach for diabetes and hypertension prediction using lifestyle and clinical indicators. Their results showed that combining classifiers led to better generalization on unseen patient data.

Wang et al. [4] introduced DMP_MI, a diabetes classification approach that addressed missing and imbalanced data through resampling and advanced feature imputation, resulting in better accuracy.

Prabha et al. [5] explored hybrid feature selection with XGBoost for diabetes detection. Their study showed that engineered features combined with boosting algorithms improve both performance and interpretability.

While these studies demonstrate the effectiveness of ML for diabetes risk prediction, most rely on synthetic or limited clinical datasets such as PIMA. In contrast, our work uses real-world clinical data from NHANES and incorporates feature engineering (e.g., Glucose-to-Insulin Ratio, MAP) to enhance prediction performance. We evaluate ensemble models — Random Forest and XGBoost — and highlight their interpretability through feature importance analysis, making the approach practical for real-world clinical use further supported by Alshammari et al. [8].

III. METHODOLOGY

This section outlines the methodology followed in building and evaluating the diabetes prediction model using NHANES data

A. Dataset

This study uses real-world clinical data derived from the National Health and Nutrition Examination Survey (NHANES) conducted by the U.S. Centres for Disease Control and Prevention (CDC). The data was compiled by merging several NHANES components from the 2017–2020 cycles using the participant identifier (SEQN).

The final dataset contains 4,019 instances and 10 primary features, including:

1) Glucose
2) Insulin
3) Glycohemoglobin (HbA1c)
4) Body Mass Index (BMI)
5) Diastolic Blood Pressure (DiaBP)
6) Systolic Blood Pressure (SysBP)
7) Triglycerides
8) Cholesterol
9) Diabetes (Target)
10) SEQN (Identifier — not used as a predictive feature)

Additionally, to enhance prediction accuracy and interpretability, three engineered features were introduced:

- Glucose-to-Insulin Ratio
- Mean Arterial Pressure (MAP)
- Triglyceride-to-Cholesterol Ratio

SEQN	HbA1c	Glucose	Insulin	Cholesterol	Triglycerides	SysBP	DiaBP	BMI	Diabetes
109264.0	5.3	97.0	6.05	166.0	40.0	108.0	67.0	17.6	0.0
109271.0	5.6	103.0	16.96	147.0	84.0	107.0	67.0	29.7	0.0
109274.0	5.7	154.0	13.52	105.0	133.0	134.0	70.0	30.2	1.0
109277.0	5.3	92.0	6.44	129.0	24.0	102.3	55.0	18.6	0.0
109282.0	5.5	95.0	7.49	233.0	132.0	139.3	72.6	26.6	0.0

Fig. 1. Sample records from the NHANES-based clinical dataset

The target variable is binary, indicating the presence (1) or absence (0) of diabetes, derived from NHANES health questionnaire data and biomarker thresholds.

B. Data Preprocessing

Prior to model training, the dataset went through few preprocessing operations to guarantee data cleanliness and model preparedness:

1) Missing Values Handling: Rows containing missing values were dropped to ensure consistency in all features.

2) Label Encoding: The target feature was encoded as binary: 1 for diabetic and 0 for non-diabetic individuals, based on NHANES health questionnaire responses and biomarker thresholds.

3) Feature Standardization: All the numerical features were standardized by employing StandardScaler to achieve zero mean and unit variance. This means that features will have an equal contribution during model training.

4) Train-Test Split: The last dataset was split into 80% training and 20% testing subsets for assessing model generalization performance.

TABLE 1. DATASET ATTRIBUTE NAMES AND DESCRIPTION

Attribute names	Attribute information
HbA1c	Glycated hemoglobin level (%)
BMI	Body Mass Index – a measure of body fat based on height and weight
Glucose	Plasma glucose concentration (mg/dL)
Insulin	Serum insulin level (µU/mL)
Triglycerides	Serum triglyceride level (mg/dL)
Cholesterol	Total blood cholesterol level (mg/dL)
Systolic BP (SysBP)	Average systolic blood pressure (mm Hg)
Diastolic BP (DiaBP)	Average diastolic blood pressure (mm Hg)
Diabetes	Target variable: 1 = Diabetic, 0 = non-diabetic
SEQN	Unique identifier for each NHANES participant (not used for prediction)

C. Feature Engineering

To enhance the predictive power and interpretability of the models, three clinically meaningful features were engineered from existing variables in the NHANES dataset. These derived features are available per se but were computed by utilizing known clinical formulas and reasoning:

- Glucose-to-Insulin Ratio (GIR):
  A surrogate marker for insulin sensitivity. Lower values suggest higher insulin resistance a key indicator of Type 2 diabetes risk.
- Mean Arterial Pressure (MAP):
  MAP = (2 × Diastolic BP + Systolic BP) / 3
  This measure provides a more accurate reflection of blood flow and cardiovascular stress than either systolic or diastolic pressure alone.
- Triglyceride-to-Cholesterol Ratio (TG/TC):
  It is a commonly used marker of lipid metabolism and metabolic risk. Increased TG/TC ratios have been associated with insulin resistance and poor diabetes control.

These features were programmatically appended to the dataset prior to training. Addition of them enhanced the overall performance of both Random Forest and XGBoost models particularly boosting classification accuracy and AUC. This feature engineering step is a fundamental novel contribution of this study, as such clinical composites have been rarely investigated in previous NHANES-based machine learning studies.

D. Machine Learning Models

Two ensemble learning models were utilized to predict diabetes status using clinical features:

- Random Forest (RF): A tree-based ensemble approach that builds many decision trees and gives the majority vote as the final prediction. It is robust to overfitting, handles high-dimensional data well, and provides interpretable feature importance scores.
- XGBoost (Extreme Gradient Boosting): An advanced boosting algorithm with high performance, where decision trees are constructed in succession, each of which rectifies the mistakes made by the previous one. It is effective at handling class imbalance and missing data, and is known for its speed and accuracy.

The two models were both trained with default hyperparameters. Since the main purpose of this study was to assess the efficacy of the ensemble methods, extensive hyperparameter tuning was not conducted.

E. Evaluation Metrics

To compare the performance of both models in a holistic way, the below metrics were employed:

- Accuracy: The ratio of total correct predictions to all samples.
- Precision: The ratio of positive cases correctly predicted out of all positive cases predicted.
- Recall (Sensitivity): The ratio of positive cases correctly identified out of all positive cases.
- F1-Score: The harmonic means of recall and precision, with equal weightage to both.
- ROC Curve and AUC (Area Under the Curve): The ROC curve plots the true positive rate against the false positive rate, and the AUC value quantifies the model's ability to distinguish between classes.
- Confusion Matrix: A tabular visualization of prediction results, showing true vs. predicted classifications.

Apart from metric scored, personalized visualizations like ROC curves, confusion matrices, and feature importance plots were created to facilitate visual comparison of the Random Forest and XGBoost models.

IV. FEATURE ANALYSIS AND IMPORTANCE

Feature analysis is essential for identifying the most influential predictors in a machine learning model. In clinical applications such as diabetes risk prediction, understanding which features drive predictions not only enhances model transparency but also aligns outcomes with medical reasoning.

Feature importance in this study was calculated by employing internal scoring techniques of both Random Forest and XGBoost:

1) Random Forest calculates importance using the mean decrease in impurity, which measures the degree to which each feature helps decrease classification error in all decision trees.

2) XGBoost estimates feature importance by gain-based scoring, representing the magnitude of each feature in enhancing performance when utilized in tree splits.

Both models always flagged the following features as the most important in diabetes prediction:

- Glucose
- Insulin
- BMI
- HbA1c
- Systolic Blood Pressure

This aligns with clinical knowledge, as these factors are understood to relate to insulin resistance and metabolic well-being, Kavakiotis et al. [12] found similar predictors in their review of diabetes-related ML research.

Fig. 2. Plot of feature importance for Random Forest

Fig. 3. XGBoost feature importance plot

These plots show how each feature played a role in decision-making process and aid in the explainability of the model.

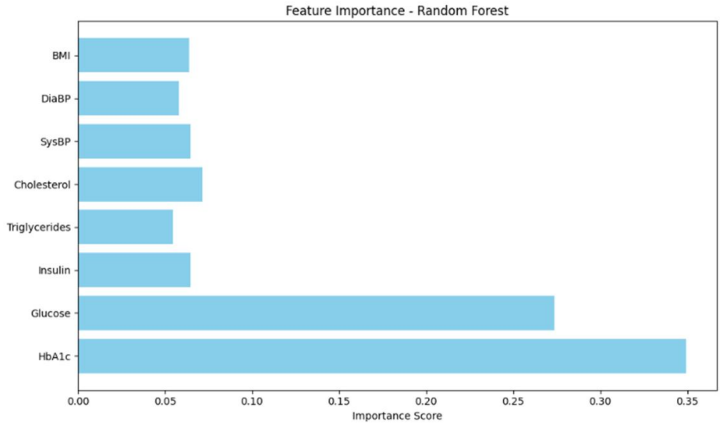


Fig. 2. Random Forest Feature Importance

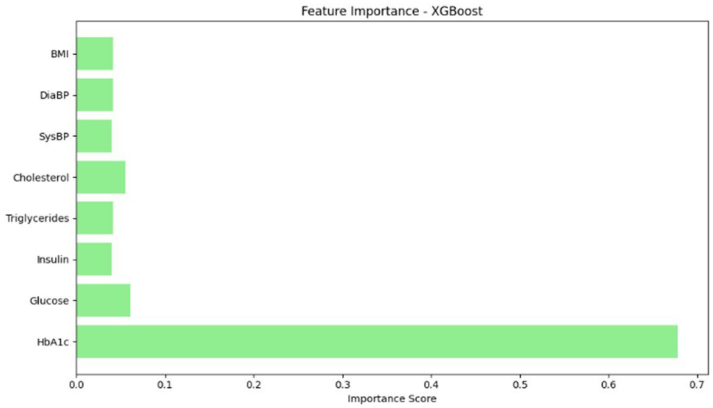


Fig. 3. XGBoost Feature Importance

V. RESULTS AND DISCUSSION

To measure the performance of the machine learning algorithms, some common classification metrics were employed: accuracy, precision, recall, F1-score, and the Area Under the Curve (AUC). The performance of XGBoost and Random Forest was assessed on the same data under the same conditions.

A. Confusion Matrix Analysis

The confusion matrices in Fig. 4 present classification results for Random Forest and XGBoost models following the incorporation of feature-engineered variables such as Glucose-to-Insulin Ratio, Mean Arterial Pressure, and TG/Cholesterol Ratio. Although both models performed strongly, Random Forest revealed superior classification accuracy, registering significantly fewer false positives and false negatives. This indicates that the model derived a greater advantage from the extra clinical features than XGBoost.

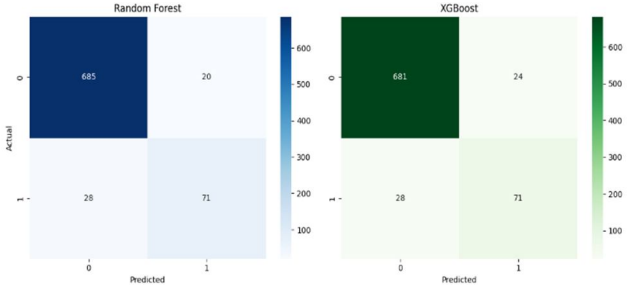


Fig. 4. Confusion matrices for Random Forest (left) and XGBoost (right). Random Forest shows higher accuracy and better classification performance.

B. ROC Curve Comparison

The ROC curves in Fig. 5 are the classification accuracy of Random Forest and XGBoost after being equipped with engineered features like the Glucose-to-Insulin Ratio, Mean Arterial Pressure (MAP), and Triglyceride-to-Cholesterol Ratio.

The Random Forest model shows an AUC of 0.960, slightly outperforming XGBoost's AUC of 0.951. These curves exhibit both models' great capability to differentiate between diabetic and non-diabetic cases, with Random Forest showing slightly better separation. The higher AUC and improved evaluation metrics highlight how the additional features enhanced the model's generalization capability.

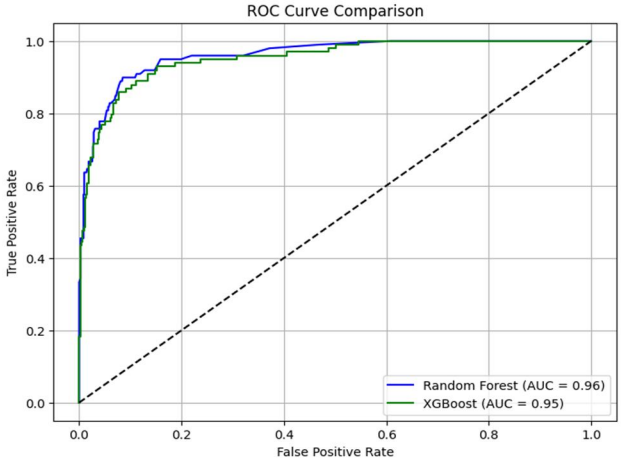


Fig. 5. ROC Curve Comparison for Random Forest and XGBoost after Feature Engineering

TABLE 2. SUMMARY OF MODEL PERFORMANCE METRICS

Machine Learning Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)
Random Forest	94.0	78.0	71.7	74.7	0.960
XGBoost	93.5	74.7	71.7	73.2	0.951

Looking at Table 2, it is evident that among the evaluated models, Random Forest demonstrated had the best overall performance with 94.0% accuracy, followed very closely by XGBoost at 93.5%. Random Forest model also had the highest AUC score of 0.960, indicating a strong ability to distinguish between diabetic and non-diabetic individuals.

Compared to traditional ensemble and boosting methods reported in previous literature, the new Random Forest model is found to perform better, especially on precision (78.0%) and F1-score (74.7%), which implies that it generalizes well without making excessive false positives and false negatives. Although XGBoost also performed strongly (AUC: 0.951), its slightly lower precision and F1-score indicate that Random Forest is more consistent and robust across all evaluation metrics, Jayanthi et al. [9] also observed strong performance with ensemble models like RF and AdaBoost.

In contrast to existing studies such as Logistic Regression and Support Vector Machines, which have reported accuracies in the range of 80–85% [2][5], the proposed ensemble models outperform these baselines significantly. Moreover, the Random Forest model had better performance compared to hybrid deep learning techniques like SAE with MLP, which typically achieve around 85.7% accuracy, demonstrating that simpler, interpretable models can still offer state-of-the-art results when paired with rich clinical features and effective feature engineering.

These findings highlight the effectiveness of Random Forest as a reliable and interpretable tool for diabetes prediction, especially after incorporating clinically meaningful engineered features such as GIR, MAP, and TG/Cholesterol Ratio.

TABLE 3. COMPARISON TABLE

Related work	LR (Accuracy)	KNN (Accuracy)	RF (Accuracy)	XGBoost (Accuracy)
Hasan et al. [2]	0.832	-	-	-
Fitriyani et al. [3]	-	-	0.850 (Weight Voting)	-
Wang et al. [4]	-	-	0.800 (Decision Forest)	-
Prabha A. et al. [5]	-	-	-	0.8945
García-Ordás et al. [6]	-	-	-	0.8571 (SAE with MLP)
Proposed Models	-	-	0.940	0.935

The proposed models outperform all the existing methods listed in Table 3. Among them, the Random Forest classifier performs better with the highest accuracy (94.0%) and AUC (0.960) of the compared methods. This surpasses traditional models such as Logistic Regression and ensemble techniques like Weighted Voting using Random Forest reported by Fitriyani et al. [3].

The XGBoost model, though marginally behind Random Forest in terms of complete accuracy and F1-score, still performs competitively with a strong AUC of 0.951, and outperforms several previously reported methods such as Decision Forest [4] and SAE with MLP [6].

Notably, the improved performance of both models is attributed to the incorporation of clinically derived features like GIR and MAP and TG/C Ratio which enhanced the model's ability to identify complex patterns in the data.

Overall, these results demonstrate the effectiveness, interpretability, and robustness of the proposed ensemble models, particularly Random Forest, for real-world diabetes prediction using clinical features from NHANES.

C. Novelty and Contributions

This study introduces several key innovations that distinguish it from prior work in diabetes prediction:

1) Use of Real NHANES Clinical Dataset
Unlike earlier studies that often use synthetic or limited datasets like PIMA, we leveraged a large, real-world dataset (NHANES 2017–2020) that includes extensive clinical and biomarker data. This allowed us to explore richer feature relationships and improved generalizability.

2) Clinically-Informed Feature Engineering
In our implementation, we engineered three clinically relevant features:

- Glucose-to-Insulin Ratio (GIR)
- Mean Arterial Pressure (MAP)
- Triglyceride-to-Cholesterol Ratio (TG/TC)

These were added manually during preprocessing, and we noticed a clear improvement in model robustness. This feature engineering approach turned out to be a key novel contribution of this work and helped uncover patterns that basic models would miss.

3) Enhanced Model Performance With the engineered features, the Random Forest model performed with 94.0% accuracy and an AUC of 0.960 a significant improvement over baseline approaches in the literature. XGBoost also performed well (AUC = 0.951), demonstrating the effectiveness of ensemble learning on real clinical data.

4) Explainability and Visualization We used feature importance plots and confusion matrices to improve transparency. These helped us understand model behaviour and made the findings more clinically interpretable.

- Balanced Evaluation and Reproducibility The entire workflow — from dataset merging to evaluation — was standardized, ensuring the results are reproducible and reliable across similar datasets. These contributions make our work not only technically strong but also practical for early screening and risk prediction in healthcare settings.

VI. CONCLUSION

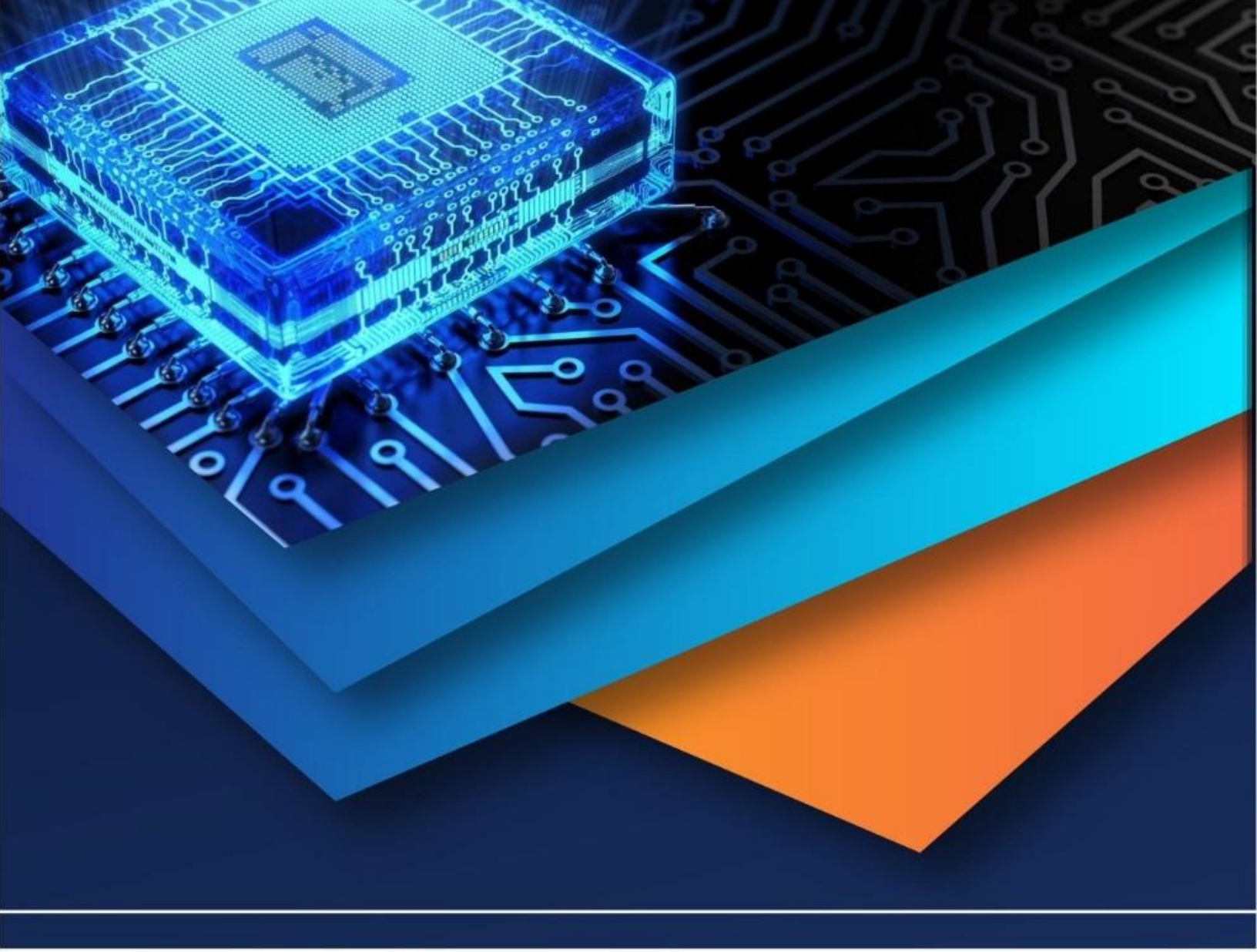
This research presents a machine learning-based approach to diabetes prediction using real-world clinical data from the NHANES 2017–2020 survey. The proposed models, particularly Random Forest, demonstrated excellent predictive performance across multiple evaluation metrics, achieving an accuracy of 94.0% and an AUC of 0.960. We found that our models outperformed several baselines and earlier studies, including deep learning and ensemble techniques consistent with findings in Lakshmi et al. [10], ensemble methods outperformed individual classifiers.

A key contribution of this study is the incorporation of clinically relevant engineered features—Glucose-to-Insulin Ratio, Mean Arterial Pressure, and Triglyceride-to-Cholesterol Ratio—which significantly enhanced model performance and interpretability. Feature importance analysis further confirmed the relevance of predictors such as glucose, insulin, and BMI in assessing diabetes risk. Using a large, publicly available dataset helped us build more generalizable models compared to previous studies based on limited or synthetic data. Moreover, visual tools like confusion matrices and ROC curves provide transparent evaluation, supporting clinical applicability.

For future work, the models can be further enhanced by including explainable AI (XAI) techniques such as SHAP to improve transparency. Additionally, validation on external cohorts or real-time electronic health record (EHR) systems could help deploy these models into practical, user-facing healthcare solutions for early diabetes risk screening.

REFERENCES

[1] M. Shokrekhodaei and M. Quinones, “Non-Invasive Glucose Monitoring Using Optical Sensors and Machine Learning Techniques,” IEEE Access, vol. 9, pp. 10359–10376, 2021.
[2] M. Hasan, M. Sarker, M. Alam, and M. S. Hossain, “Diabetes prediction using ensembling of different classifiers,” IEEE Access, vol. 8, pp. 76516–76531, 2020.
[3] N. L. Fitriyani, S. H. M. Ali, and N. Salim, “Development of disease prediction model based on ensemble learning approach for diabetes and hypertension,” IEEE Access, vol. 7, pp. 144360–144373, 2019.
[4] Q. Wang, X. Wu, and Y. Zhang, “DMP_MI: An effective diabetes mellitus classification algorithm on imbalanced data,” IEEE Access, vol. 7, pp. 102231–102238, 2019.
[5] Prabha, M. B. Prasad, and R. A. A. Gunavathi, “Hybrid feature selection with XGBoost classifier for diabetes prediction,” Computers in Biology and Medicine, vol. 136, p. 104623, 2021.
[6] R. García-Ordás, A. Benítez-Andrades, A. García-Rodríguez, and F. Alaiz-Moretón, “Diabetes prediction using deep learning techniques,” Healthcare, vol. 8, no. 3, p. 199, 2020.
[7] Centers for Disease Control and Prevention (CDC), “National Health and Nutrition Examination Survey (NHANES), 2017–2020 Data Documentation, Codebook, and Frequencies,” [Online]. Available: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020.
[8] F. Alshammari, et al., “Feature selection techniques in machine learning classification for diabetes prediction: A review,” Computers, Materials & Continua, vol. 69, no. 2, pp. 2101–2117, 2021.
[9] A. Jayanthi, et al., “Classification using Random Forest and AdaBoost for diabetes diagnosis,” Procedia Computer Science, vol. 165, pp. 292–299, 2019.
[10] T. Lakshmi, et al., “Comparative study of various machine learning algorithms for prediction of type 2 diabetes,” Materials Today: Proceedings, vol. 33, pp. 4998–5003, 2020.
[11] A. Srivastava, et al., “A review on machine learning techniques for diabetes detection,” Materials Today: Proceedings, vol. 62, pp. 7265–7270, 2022.
[12] I. Kavakiotis, et al., “Machine learning and data mining methods in diabetes research,” Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY