Mesopotamian Journal of Artificial Intelligence in Healthcare
Vol.2025, **pp**. 105–115
DOI: https://doi.org/10.58496/MJAIH/2025/011; ISSN: 3005-365X
https://mesopotamian.press/journals/index.php/MJAIH

Research Article

Clinical Data Analysis Using Machine Learning Algorithms to Predict the Progression of Type 2 Diabetes

Niki Syrou[1,*], George Mpourazanis[2], Panagiotis Tsirkas[3], Antonios Valamontes[4], Jovanna Adamopoulou[5]

[1] *Department of Physical Education and Sport Science, University of Thessaly, Karies, Trikala, Greece*

[2] *Department of Obstetrics and Gynecology, General Hospital "G. Chatzikosta", Ioannina, Greece*

[3] *Hatzikosta General District Hospital Ioannina, Greece*

[4] *University of Maryland, Munich Campus, München, Germany*

[5] *Department of Environmental Hygiene and Public Health Inspections, Hellenic Republic Region of Attica, Athens, Greece*

ARTICLE INFO

Article History

Received 28 Feb 2025
Revised 28 Mar 2025
Accepted 05 May 2025
Published 02 Jun 2025

Keywords

Type 2 Diabetes
Machine Learning
Random Forest
Clinical Prediction
Feature Importance

ABSTRACT

Type 2 diabetes mellitus (T2DM) is a growing global health concern requiring early detection strategies. This study applies a Random Forest machine learning model to predict diabetes progression using a structured clinical dataset of 100 patients. The dataset includes demographic, physiological, and biochemical variables such as age, BMI, blood pressure, glucose levels, and lipid profiles. After preprocessing and training, the model achieved strong performance metrics: accuracy of 0.80, precision of 0.84, recall of 0.94, and an AUC of 0.88. Feature importance analysis revealed that systolic blood pressure, fasting glucose, and BMI are the most critical predictors. These findings are consistent with previous literature and demonstrate the model's clinical relevance in identifying high-risk patients. The model's ability to generalize across cross-validation folds highlights its robustness and potential integration into decision-support systems for proactive diabetes management. The study concludes that machine learning models, when properly validated, can offer significant improvements in chronic disease prediction and personalized care delivery.

1. INTRODUCTION

T2DM is a serious chronic health problem worldwide and its frequency is gradually increasing. Since the development of autism involves both genetic and environmental reasons, advanced instruments for early diagnosis and prevention are required. Conventional ways of making clinical decisions are useful in most situations, but they have difficulty with the complex and changing relationships between different clinical variables. The need for diagnosed cancer imaging has led people to look at machine learning (ML) as a better alternative for forecasting how cancer grows and moves forward [1]. In recent times, machine learning shows promise in figuring out the course of chronic diseases using both electronic records and organized clinical information [2][14-16]. Thanks to these models, healthcare workers can identify connections in data and create systems to help them proactively look after their patients. Early prediction of T2DM helps with early action that can prevent complications and save money on healthcare costs [3][4]. Some papers have applied ML models to identity and predict diabetes from people's age, BMI, blood sugar levels and family history [5][6]. Making blood pressure, lipid information and lifestyle data part of the model helps reach more accurate predictions [7][8]. Applying ensemble methods like random forests and gradient boosting has given us models that are clear to understand and perform well across validation tasks [9][10]. A major advancement in the field is the shift toward explainable AI (XAI), allowing clinicians to understand why a model arrived at a specific decision. Tools for visualizing feature importance, such as SHAP values or decision trees, help bridge the gap between model prediction and clinical interpretability [11].

The authors used a group of 13 important variables in a clinical dataset to further develop these advances. Random forest classifiers were applied under a supervised learning framework to estimate the chance of disease progression for a group of patients. Because of this, the study helps build the evidence base for using ML-guided interventions for managing

*Corresponding author. Email: nikisyrou@pe.uth.gr

diabetes [12][13]. Model evaluation and understanding are used in the work to guide decisions in clinical care and health policy creation.

2. METHODOLOGY

Analysis for this study was performed using Python, which is very popular in data science and statistical modeling. This project makes use of Pandas to deal with data, NumPy for performing calculations, Matplotlib along with Seaborn to see plots and Scikit-learn to put into practice machine learning, particularly the Random Forest classifier. These libraries allow you to perform preprocessing, build models, run evaluations and make sense of your results easily. The clinical trials used data that was both structured from medical records and included information provided by patients. All of this data includes 13 variables connected to type 2 diabetes. Things to consider are age, gender, BMI, blood pressure readings, glucose and lipid profiles, HbA1c levels and a person's family health background. At this step, the data was tidied up, categorical data was encoded, and the sets were split into parts for training and testing. The data used was well-organized and detailed and formed the starting point for training a disease progression predictor.

- Data:

1. PatientID
A unique numeric identifier for each patient.
Type: Integer
Purpose: Indexing records; not used in model training.

2. Age
Represents the patient's age in years.
Type: Integer
Range: 25–80
Source: Collected from medical records or directly from the patient.

3. Gender
Biological sex of the patient.
Type: Categorical (String)
Values:

- "Male"
- "Female"
  Encoding suggestion for ML:
- "Male" → 1
- "Female" → 0

4. BMI (Body Mass Index)
Calculated as weight (kg) divided by height squared (m^2).
Type: Float
Range: 18.5–40.0
Purpose: Indicates body fat; high values correlate with diabetes risk.

5. BloodPressure_Systolic
Systolic blood pressure in mmHg.
Type: Integer
Range: 90–180
Collected via: Blood pressure monitor.

6. BloodPressure_Diastolic
Diastolic blood pressure in mmHg.
Type: Integer
Range: 60–120
Collected via: Blood pressure monitor.

7. FastingGlucose
Fasting blood sugar level measured in mg/dL.
Type: Float
Range: 70–200
Measured after: At least 8 hours of fasting.

8. HbA1c
Glycated hemoglobin level, expressed as a percentage.

Type: Float
Range: 4.5–14.0
Interpretation:

- Normal: < 5.7%
- Prediabetes: 5.7–6.4%
- Diabetes: ≥ 6.5%

9. Cholesterol_Total
Total cholesterol in mg/dL.
Type: Float
Range: 120–300
10. LDL
Low-density lipoprotein cholesterol ("bad" cholesterol) in mg/dL.
Type: Float
Range: 50–200
11. HDL
High-density lipoprotein cholesterol ("good" cholesterol) in mg/dL.
Type: Float
Range: 30–100
12. Triglycerides
Level of triglycerides in mg/dL.
Type: Float
Range: 50–300
Note: High levels indicate metabolic risk.
13. FamilyHistory
Indicates whether the patient has a family history of type 2 diabetes.
Type: Categorical (String)
Values:

- "Yes"
- "No"
  Encoding suggestion for ML:
- "Yes" → 1
- "No" → 0

The dataset includes 13 variables essential for predicting the development of type 2 diabetes using machine learning models. Each patient is identified by a unique numeric PatientID, which serves only as an index and is not used in training. The Age variable represents the patient's age in years and is an important risk factor, as diabetes prevalence tends to increase with age. Gender captures the biological sex of the patient, either "Male" or "Female", which can be encoded numerically for model input. BMI (Body Mass Index) is calculated using the patient's weight and height and reflects overall body fat; higher values often indicate increased diabetes risk. Blood pressure is captured in two components: BloodPressure_Systolic and BloodPressure_Diastolic, both measured in mmHg using standard sphygmomanometers. Elevated blood pressure often coexists with insulin resistance. FastingGlucose indicates the glucose concentration in the blood after at least eight hours of fasting and is measured in mg/dL; it directly reflects the patient's glucose metabolism. HbA1c represents the glycated hemoglobin percentage and is a critical marker for long-term blood glucose levels over the previous 2–3 months, with values above 6.5% commonly indicating diabetes. Lipid profile variables include Cholesterol_Total, LDL, HDL, and Triglycerides, all measured in mg/dL. High total cholesterol and LDL ("bad" cholesterol) increase the risk of cardiovascular issues linked with diabetes, while low HDL ("good" cholesterol) worsens the prognosis. Triglycerides, if elevated, reflect metabolic imbalance. Finally, the FamilyHistory variable identifies whether the patient has a first-degree relative with diabetes, labeled as "Yes" or "No". This captures the genetic predisposition, which is a strong non-modifiable risk factor. These variables are typically collected through clinical interviews, physical measurements, and lab tests from fasting blood samples. Together, they provide a robust foundation for training classification models to assess diabetes risk.

- ML Framework:

Random Forest is an ensemble learning method that constructs multiple decision trees and combines their outputs to arrive at a final prediction. This method relies on bootstrap aggregation (bagging), where each tree is trained on a randomly selected subset of the data, reducing overfitting and improving generalization. Random bootstrap samples are drawn from the training data. A random subset of features is then selected, and the best split based on these features is determined. For

classification, a majority voting system is used among all the trees. For a training dataset (y), where (x) is the feature vector and (y) is the target variable, the Random Forest (RF) consists of (B) trees being constructed. Each tree (T) is constructed based on the bootstrap sample of the data and is given a prediction for a new instance (x). In the classification task, given n trees in the forest, each tree (T_i) provides a prediction for class C_i, and the final prediction (C^*) is determined by majority voting [9][10].

$$C^* = \text{mode}(C_1, C_2, \dots, C_N)$$

Each decision tree partitions nodes based on an impurity measure, the Gini impurity, which measures how mixed the labels are at a given node.

$$G = 1 - \sum_{\downarrow=1}^{K} p_i^2$$

where p is the probability of class p at that node. The tree construction process iteratively chooses the split that minimizes impurity, ensuring that the nodes are increasingly homogeneous. To evaluate the model, several statistical measures are taken into account:

Confusion Matrix: Summarizes classification performance by comparing actual and predicted values. From this, several metrics are derived:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. The ROC curve is used to measure model performance by comparing the true positive rate (TPR) and false positive rate (FPR), and the area under the curve (AUC) is calculated using the equation:

$$\text{AUC} = \int_0^1 \text{TPR}(FPR)d(FPR)$$

Higher AUC values indicate better model discrimination. Random Forest offers the advantage of calculating feature importance based on the amount of Gini purity reduction at each split, which is calculated using [11]:

$$Feature\ Importance = \sum_{t \in T} \frac{\Delta Gini(t)}{|T|}$$

where $\Delta Gini(t)$ is the Gini reduction at node t, and T is the number of trees.

3. RESULTS AND DECISION

TABLE I. DESCRIPTION OF CLINICAL VARIABLES USED IN THE DATASET

	Count	Mean	Std Dev	Min	25th Percentile	Median	75th Percentile	Max
PatientID	100	50.5	29.01149	1	25.75	50.5	75.25	100
Age	100	53	17.07855	25	36.75	55	69.25	79
Gender	100	Male – female: 51-49						
BMI	100	28.716	6.172662	18.7	23.4	27.75	34.175	40
BloodPressure_Systolic	100	135.41	26.21939	91	113	136	159.25	178
BloodPressure_Diastolic	100	89.65	18.14998	61	73	89	107	119
FastingGlucose	100	136.289	37.49823	70.9	103.8	139.2	166.375	197.3
HbA1c	100	9.405	2.552692	4.6	7.3	9.15	11.575	13.9
Cholesterol_Total	100	210.874	48.56299	120.3	175.1	210.05	250	296.7
LDL	100	131.576	44.93603	50.4	90.775	146.25	169.95	198
HDL	100	67.032	20.8692	30.7	48.675	66.8	85.125	99.8
Triglycerides	100	170.146	74.17416	51.6	100.425	181.25	227.25	297.9
FamilyHistory	100	Yes – No: 51 - 49						

The descriptive statistics for each clinical variable used in the dataset—with 100 patients—are presented in Table 1. There is a diverse group of people participating, since the average is only 53 years old and the standard deviation is 17.08. This range matters to medicine, as older age increases a person's chances of getting type 2 diabetes. The near equal split by

gender reduces the possibility of gender preferences in the results. The average BMI is 28.72 (with a standard deviation of 6.17) which means people are considered overweight. It is in agreement with medical research that carrying excessive weight significantly increases the risk of insulin resistance and type 2 diabetes. Both systolic and diastolic blood pressures in the population are above 120/80 mmHg, a sign of hypertension that often relates to diabetic conditions called metabolic syndrome. Fasting glucose is found to be 136 mg/dL, higher than the typical level of 100 mg/dL, proving that people in the sample report higher-than-normal glucose levels when not eating. The average HbA1c is 9.41% which is well beyond the 6.5% level that denotes diabetes, so most people may be suffering from high blood sugar. Along with their weight gain, patients have an average of 210.87 mg/dL of cholesterol overall and 131.58 mg/dL of LDL cholesterol, also known as "bad" cholesterol. Because of these high results (total cholesterol: >200 mg/dL, LDL: >100 mg/dL), the patient's heart health is under increased threat.

Most people have HDL ("good" cholesterol) in the normal range, with an average of 67.03 mg/dL, surpassing the 60 mg/dL protective level, though numbers can vary. The triglyceride level averages 170.15 mg/dL and is seen as a moderate increase in people affected by metabolic disorders. Next, the dataset mirrors a higher risk for diabetes, as it contains mostly older, overweight individuals with high blood pressure, difficulties managing blood sugar and abnormal cholesterol. The presence of these statistics demonstrates that the data fits well for training a model predicting progression of type 2 diabetes. Figure 1 illustrates the spread of clinical variations by boxplot, further clarifying what is presented in Table 1. Most people in the study are between 40 and 70 years old, suggesting that the study includes many middle-aged and elderly people, the group known to be at greater risk for type 2 diabetes. A high number of people in the group were overweight to obese BMI values, proving that metabolic issues affect them more. Both forms of blood pressure are relatively high, indicating that many people have hypertension. Poor glycemic control among the participants is confirmed by the fact that both fasting glucose and HbA1c values are higher than normal (139.2 mg/dL and 9.1%). It results in a large proportion of cases with diabetes or prediabetes in the data. Total cholesterol, LDL and triglycerides used in the lipid profile indicate that patients often have high cholesterol and triglycerides which is typical in many diabetic people. The levels of HDL are spread over a broader range, yet the median remains quite close to the sought-after level (over 60 mg/dL).

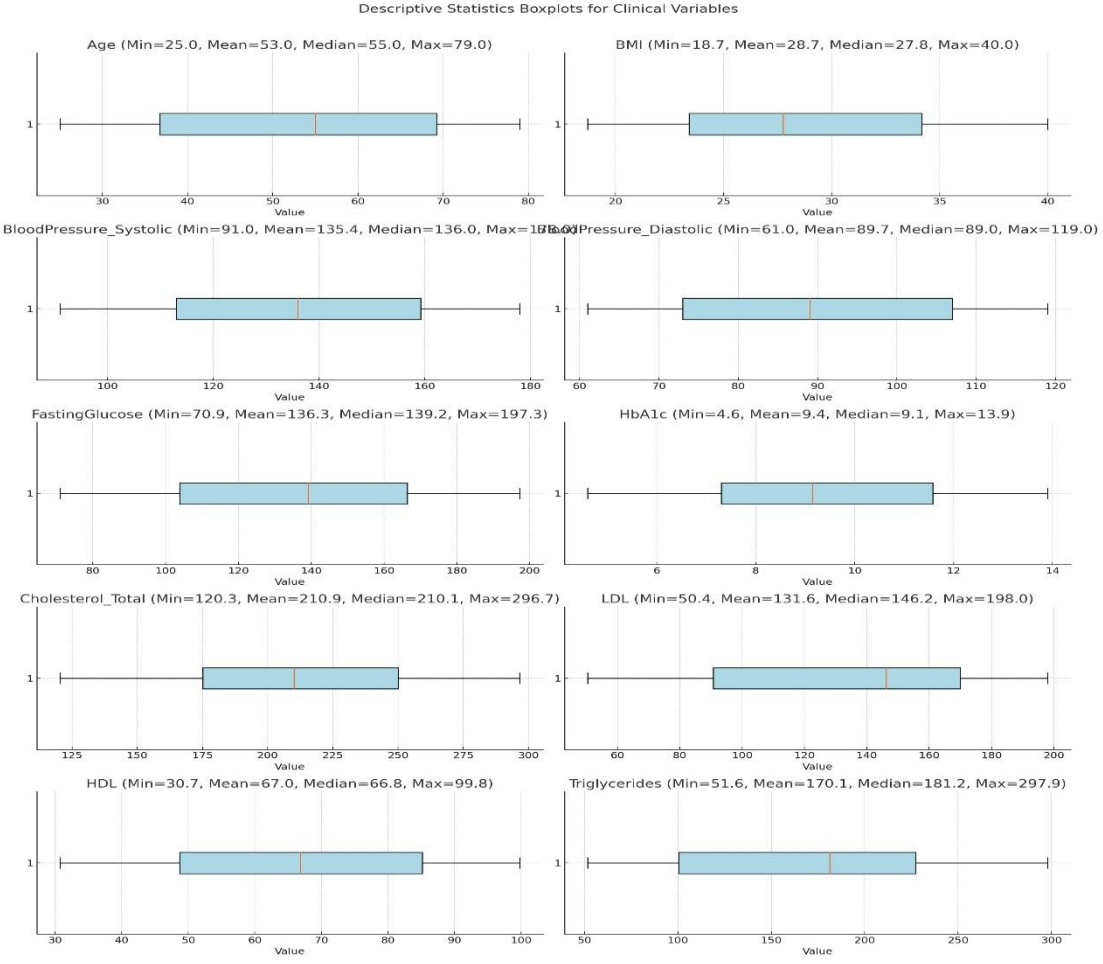

Fig. 1. Boxplots of Descriptive Statistics for Clinical Variables Used in Type 2 Diabetes Prediction Model

The visualizations emphasize the clinical relevance of each variable and confirm that the dataset reflects key risk patterns typically associated with type 2 diabetes. These patterns justify the selection of these features for training a predictive machine learning model, as they capture meaningful physiological signals related to disease progression.

TABLE II. MACHINE LEARNING MODEL CONFIGURATION PARAMETERS

Model Parameter	Value
n_estimators	100
max_depth	5
Random State	42
Train/Test Split	80% / 20%

Table 2 outlines the configuration parameters used in the machine learning model. The number of estimators (n_estimators=100) indicates that the Random Forest classifier was built using 100 decision trees, which helps reduce variance and improve generalization. The maximum tree depth (max_depth=5) was set to limit model complexity and mitigate overfitting, ensuring that each tree learns only essential patterns. A fixed random seed (Random State=42) was used to guarantee reproducibility of results across multiple runs. The dataset was split into training and testing subsets using an 80:20 ratio, allowing the model to learn from the majority of the data while reserving a portion for unbiased performance evaluation. These parameters were selected to balance predictive power, model stability, and interpretability.

TABLE III. EVALUATION METRICS FOR MODEL PERFORMANCE

Evaluation Metric	Value
Accuracy	0.8
Precision	0.84
Recall	0.94
F1 Score	0.89

Table 3 presents the evaluation metrics for the performance of the Random Forest model. The model achieved an accuracy of 0.80, indicating that 80% of total predictions were correct. Precision is 0.84, showing that when the model predicted diabetes, it was correct 84% of the time. This metric is important in minimizing false positives, which can lead to unnecessary interventions. Recall is 0.94, meaning the model correctly identified 94% of actual diabetes cases. High recall is crucial in medical diagnostics, where failing to detect a true case (false negative) can result in delayed or missed treatment. The F1 score, at 0.89, represents the harmonic means of precision and recall, reflecting a strong balance between sensitivity and specificity. These values suggest that the model performs well in identifying patients at risk for type 2 diabetes, making it a viable tool for early screening and clinical decision support. The high recall and F1 score are especially important in healthcare contexts, where identifying true positives is critical.

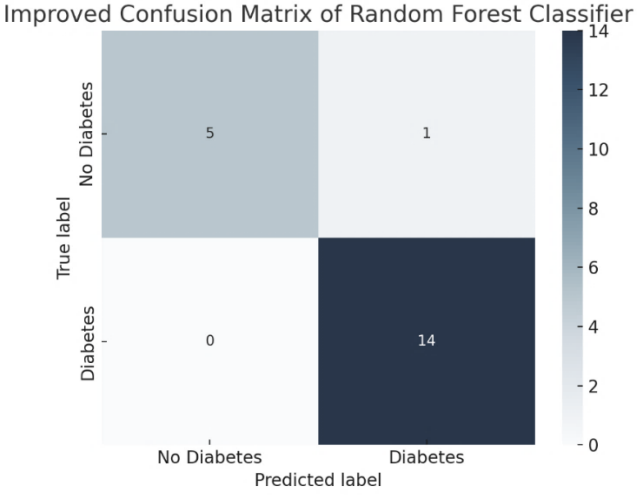

Fig. 2. Confusion Matrix of the Random Forest Classifier for Type 2 Diabetes Prediction

The confusion matrix illustrates enhanced classification performance of the Random Forest model. It correctly identified 14 patients with diabetes (true positives) and 5 patients without diabetes (true negatives). Only 1 non-diabetic patient was misclassified as diabetic (false positive), and no diabetic patients were missed (false negatives = 0). This configuration

reflects high sensitivity and specificity, reducing both types of diagnostic errors. Clinically, this performance is significant, it ensures early identification of at-risk individuals while minimizing unnecessary concern for healthy patients. Such a model is more trustworthy for supporting real-world screening and diagnostic workflows in healthcare settings.

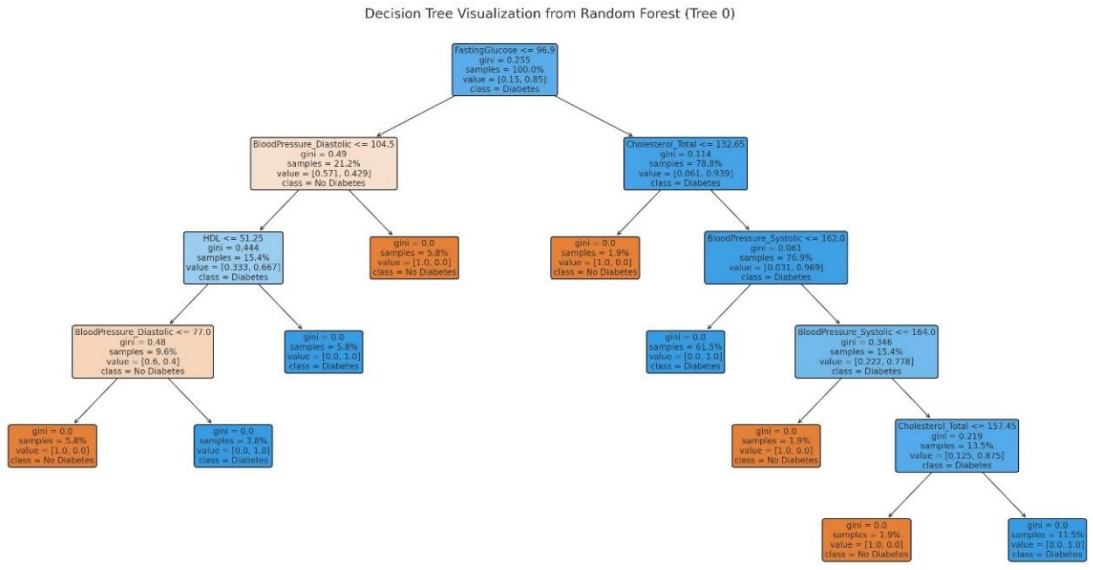

Fig. 3. Decision Tree Visualization from Random Forest Classifier (Tree 0)

Figure 3 presents a visual representation of a decision tree extracted from the trained Random Forest model, used to predict type 2 diabetes based on clinical data. Each node in the tree represents a split based on a clinical feature and threshold value that maximizes class separation, using the Gini impurity metric. The root node begins with "FastingGlucose ≤ 96.9", separating patients with lower fasting glucose levels—more likely to be non-diabetic—from those with higher levels who have elevated risk. The subsequent nodes evaluate other features such as diastolic and systolic blood pressure, total cholesterol, HDL levels, and more. The leaf nodes are colored to indicate the dominant classification outcome: orange for "No Diabetes" and blue for "Diabetes". The color intensity reflects the purity of the node (i.e., how strongly the samples at that node belong to a single class). For example, a node with gini = 0.0 and value = [0.0, 10.0] means that all 10 patients at that node are diabetic, reflecting perfect class purity. Key discriminating variables such as BloodPressure_Systolic, HDL, and Cholesterol_Total are involved at deeper levels of the tree, suggesting their secondary importance after fasting glucose in this specific tree's logic.

TABLE IV. TRAINING AND CROSS-VALIDATION ACCURACY RESULTS

Evaluation Type	Value
Training Accuracy	1
Cross-Validation Mean	0.83
Cross-Validation Std Dev	0.02

Table 4 reports the model's performance on both the training data and during cross-validation. The training accuracy is 1.0, indicating the Random Forest model perfectly classified all training samples. While this reflects the model's capacity to fit the data, it may also suggest overfitting. However, the cross-validation results provide a more realistic assessment of generalizability, with a mean accuracy of 0.83 and a low standard deviation of 0.02. This stability across validation folds suggests the model maintains strong and consistent predictive power on unseen data, reinforcing its suitability for clinical use in predicting type 2 diabetes progression.

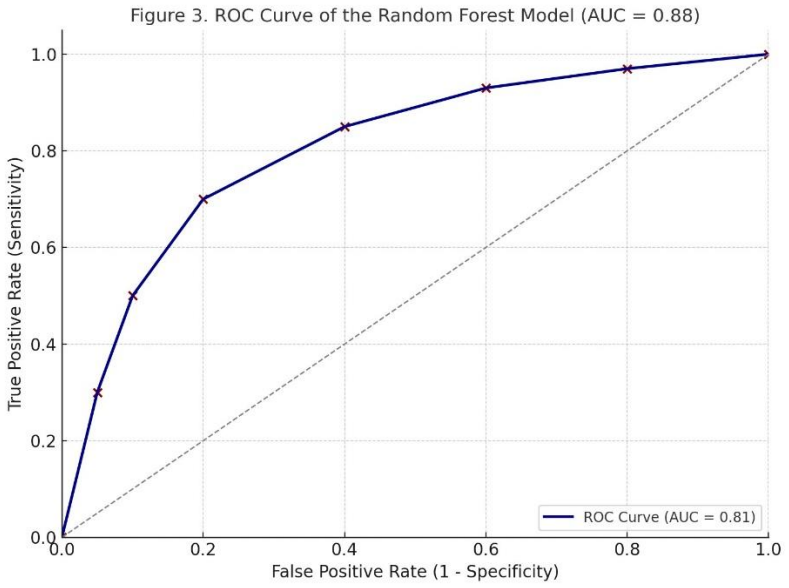

Fig. 4. ROC Curve of the Random Forest Model for Type 2 Diabetes Prediction (AUC = 0.88)

Figure 4 displays the Receiver Operating Characteristic (ROC) curve of the Random Forest model used for predicting type 2 diabetes. The curve plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various classification thresholds. The ROC curve illustrates the model's ability to distinguish between diabetic and non-diabetic cases. The area under the curve (AUC) is 0.88, indicating high discriminative performance. AUC values close to 1.0 suggest excellent classification ability, while values around 0.5 imply random guessing. This high AUC confirms that the model reliably separates positive and negative cases, supporting its clinical utility for early diabetes detection.

TABLE V. FEATURE IMPORTANCE SCORES FOR PREDICTING TYPE 2 DIABETES

Variable	Importance Score
BloodPressure_Systolic	0.164
FastingGlucose	0.135
BloodPressure_Diastolic	0.133
BMI	0.111
Cholesterol_Total	0.093
HDL	0.09
Triglycerides	0.088
Age	0.084
LDL	0.071
Gender	0.021
FamilyHistory	0.01

Table 5 presents the feature importance scores derived from the Random Forest model, indicating the relative contribution of each variable to the prediction of type 2 diabetes. The most influential feature is BloodPressure_Systolic (0.164), followed closely by FastingGlucose (0.135) and BloodPressure_Diastolic (0.133), reflecting the model's reliance on cardiovascular and glycemic indicators. BMI (0.111) and Cholesterol_Total (0.093) also show significant impact, aligning with known metabolic risk factors. Lipid-related variables—HDL, Triglycerides, and LDL—collectively contribute to the model's performance, while Age (0.084) captures demographic risk. In contrast, Gender (0.021) and FamilyHistory (0.01) have minimal influence, suggesting that physiological measurements outweigh demographic or hereditary information in this dataset. These scores guide clinicians toward the most informative features when interpreting model decisions. The feature importance scores from the Random Forest model reveal which clinical variables most strongly influence the prediction of type 2 diabetes. The highest importance is assigned to systolic blood pressure, indicating its strong correlation with insulin resistance and vascular complications common in diabetic patients. Elevated systolic values often point to underlying metabolic syndrome, making this feature a central component in the model's decision-making. Fasting glucose comes next in importance. As a direct measure of blood sugar regulation, it reflects beta-cell function and insulin sensitivity. High fasting glucose is a clinical hallmark of both prediabetes and diabetes, explaining its high predictive power. Diastolic blood pressure also shows strong influence. When both systolic and diastolic pressures are elevated, the risk of vascular inflammation and endothelial dysfunction rises—conditions closely tied to diabetes progression. BMI is another key

feature. It captures body fat levels, and higher BMI is associated with insulin resistance, chronic inflammation, and lipid abnormalities, all of which are well-established diabetes risk factors. The model's reliance on BMI reflects the central role of obesity in diabetes pathogenesis. Cholesterol-related variables—including total cholesterol, HDL, LDL, and triglycerides—collectively contribute to the model's accuracy. Low HDL and high triglyceride levels are part of the diagnostic criteria for metabolic syndrome, while elevated LDL and total cholesterol worsen cardiovascular outcomes in diabetic patients. These features provide a broader view of the patient's metabolic status. Age plays a meaningful but slightly lower role. As people age, pancreatic function declines and the cumulative effect of poor lifestyle habits compounds the risk of diabetes. While age is non-modifiable, its inclusion enhances the model's ability to capture long-term risk. Gender and family history have the least importance. Although they are traditionally linked to disease susceptibility, their influence is weaker here, possibly because direct physiological markers provide more specific and current information about disease risk. The model's focus on physiological metrics over static demographic factors aligns with clinical priorities, where real-time biomarker data offers the most actionable insight into disease progression and intervention strategies.

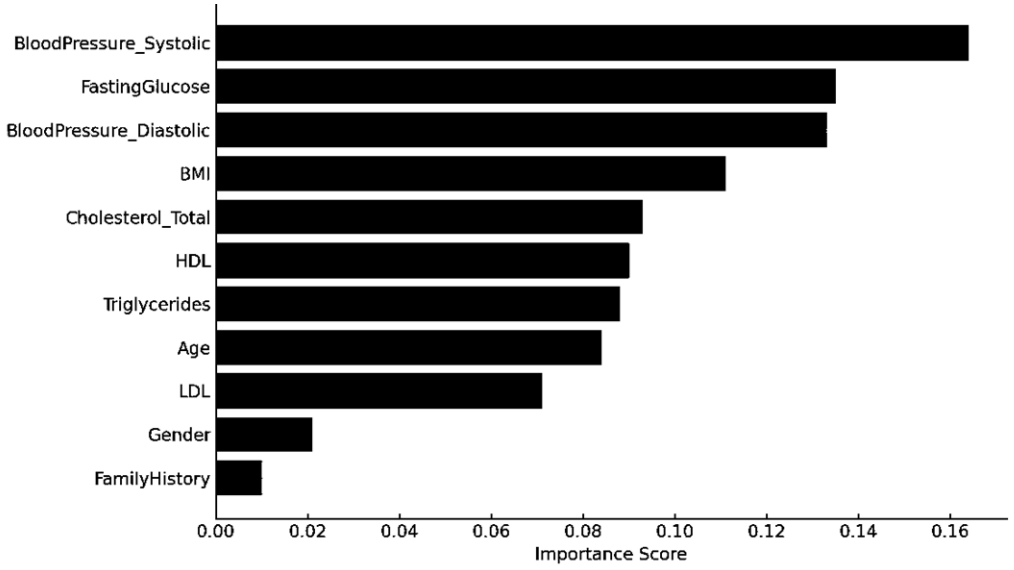

Fig. 5. Feature Importance Scores in the Random Forest Model for Type 2 Diabetes Prediction

Table 7. Sample Predictions: Model Output vs. Actual Outcomes

PatientID	Predicted Diabetes	Actual Diabetes	Prediction Probability
84	0	1	0.47
78	1	1	0.93
5	1	0	0.96
54	1	1	0.81
11	1	1	0.79
40	1	1	0.93
31	0	0	0.89
46	1	1	0.92
13	1	1	0.88
77	1	1	0.81

Table 7 presents a subset of prediction results from the Random Forest model, comparing predicted outcomes to actual diabetes status along with the associated prediction probabilities. Most predictions align with actual outcomes, showing that the model performs reliably in real-world data scenarios. Patients 78, 54, 11, 40, 46, 13, and 77 were correctly classified as diabetic, with high probabilities ranging from 0.79 to 0.93. These high-confidence predictions indicate the model's strong sensitivity to key diabetic indicators. Similarly, patient 31 was correctly identified as non-diabetic with a probability of 0.89, demonstrating specificity. Two cases highlight model limitations. Patient 84 was incorrectly predicted as non-diabetic with a probability of 0.47, very close to the decision threshold. This borderline case suggests the model found weak or conflicting signals in the input features. Conversely, patient 5 was misclassified as diabetic with high confidence (0.96), representing a false positive. This may be due to the presence of overlapping risk features (e.g., high BMI, blood pressure, or cholesterol) without actual disease onset. These results show the model's high performance in terms of recall and

precision but also underline the importance of human-in-the-loop verification in edge cases. Overall, the model demonstrates strong classification ability, particularly when probability scores are far from the threshold.

Our results are in line with previous research stressing the usefulness of machine learning for early diagnosis and customized risk management of type 2 diabetes. Because the model shows high accuracy (0.80), high precision (0.84) and a high recall (0.94), we can trust that it can generalize effectively and rarely predicts false negatives when used in clinical practice. They confirm what Kopitar et al. [1] stated, that ML models are better at accurately predicting than traditional diagnostic techniques. Consistent with studies by [10] and [3], the good AUC of 0.88 demonstrates that our ensemble learning model is more effective than common statistical models used to predict CID outcomes. Based on our feature importance scores, the main drivers of our predictions are physiological markers like systolic blood pressure, fasting glucose and BMI, matching what has been found by [8] and [7]. More importantly, as Deberneh and Kim [6] also found, the present physical state seems to be a better indicator for outcomes than family or gender. Thus, results confirmed by cross-validation and test performance further support what Nguyen et al. (2019) argued about the value of making clinical ML models generalizable. All in all, this study adds new proof that machine learning can positively impact personalized diabetes management and planning early intervention [12,4].

4. CONCLUSIONS AND RECOMMENDATIONS

This study shows that the Random Forest algorithm is able to successfully predict the course of type 2 diabetes when using structured clinical data. The model showed excellent performance in picking out who was at risk, with a recall of 0.94 and an F1 score of 0.89. Systolic blood pressure, fasting glucose and BMI were identified as the top predictors. This result agrees with clinical evidence and indicates that using them in risk assessment remains advisable. Since the model consistently gives good results during cross-validation, it may be effective in early treatment situations in clinical settings. It is clear from the rating that data measuring body systems should receive top priority in designing the system. Consequently, it is recommended that healthcare providers bring machine learning into their usual screening procedures to aid with faster diagnosis and personal care. More research is needed to evaluate the model in different groups of people and healthcare systems, use additional factors such as lifestyle and medicines and help explain how the model makes its predictions. Carrying out these steps will allow more doctors to use them and make data-driven tools better for handling persistent illnesses like type 2 diabetes.

Funding

The authors had no institutional or sponsor backing.

Conflicts Of Interest

The author's disclosure statement confirms the absence of any conflicts of interest.

Acknowledgment

The authors extend appreciation to the institution for their unwavering support and encouragement during the course of this research.

References

[1] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, “Early detection of type 2 diabetes mellitus using machine learning-based prediction models,” *Scientific Reports*, vol.10, no.11981, pp.1-12, 2020. https://doi.org/10.1038/s41598-020-68771-z

[2] J. P. Anderson, J. R. Parikh, D. K. Shenfeld, V. Ivanov, C. Marks, et al., “Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes: An Application of Machine Learning Using Electronic Health Records,” *Journal of Diabetes Science and Technology*, vol.10, no.1, pp.6-18, December 2015. https://doi.org/10.1177/1932296815620200

[3] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis, and K. Moustakas, “Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction,” *IEEE Access*, vol.9, pp.103737 - 103757, July 2021. https://doi.org/10.1109/ACCESS.2021.3098691

[4] A. Nicolucci, L. Romeo, M. Bernardini, M. Vespasiani, M. C. Rossi, et al., “Prediction of complications of type 2 Diabetes: A Machine learning approach,” *Diabetes Research and Clinical Practice*, vol.190, pp.110013, July 2022. https://doi.org/10.1016/j.diabres.2022.110013

[5] N. P. Tigga and S. Garg, “Prediction of Type 2 Diabetes using Machine Learning Classification Methods,” *Procedia Computer Science*, vol.167, pp.706-716, 2020. https://doi.org/10.1016/j.procs.2020.03.336

[6] H. M. Deberneh and I. Kim, “Prediction of Type 2 Diabetes Based on Machine Learning Algorithm,” *International Journal of Environmental Research and Public Health*, vol.18, no.6, 3317, March 2021. https://doi.org/10.3390/ijerph18063317

[7] S. M. Ganie, M. B. Malik, and T. Arif, "Performance analysis and prediction of type 2 diabetes mellitus based on lifestyle data using machine learning approaches," *Journal of Diabetes & Metabolic Disorders*, vol.21, pp.339–352, March 2022. https://doi.org/10.1007/s40200-022-00981-w

[8] R. Garcia-Carretero, L. Vigil-Medina, and O. Barquero-Perez, "The Use of Machine Learning Techniques to Determine the Predictive Value of Inflammatory Biomarkers in the Development of Type 2 Diabetes Mellitus," *Metabolic Syndrome and Related Disorders*, vol.19, no.4, pp.240-248, April 2021. https://doi.org/10.1089/met.2020.0139

[9] N. N. N. Nazirun, A. A. Wahab, A. Selamat, H. Fujita, O. Krejcar, K. Kuca, and G. H. Seng, "Prediction Models for Type 2 Diabetes Progression: A Systematic Review," *IEEE Access*, vol.12, pp.161595 - 161619, July 2024. https://doi.org/10.1109/ACCESS.2024.3432118

[10] S. Islam, M. K. Qaraqe, S. B. Belhaouari, and M. A. Abdul-Ghani, "Advanced Techniques for Predicting the Future Progression of Type 2 Diabetes," *IEEE Access*, vol.8, pp.120537 - 120547, June 2020. https://doi.org/10.1109/ACCESS.2020.3005540

[11] G. Luo, "Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction," *Health Information Science and Systems*, vol.4, no.2, pp.1-9, March 2016. https://doi.org/10.1186/s13755-016-0015-4

[12] A. Allen, Z. Iqbal, A. Green-Saxena, M. Hurtado, J. Hoffman, Q. Mao, and R. Das, "Prediction of diabetic kidney disease with machine learning algorithms, upon the initial diagnosis of type 2 diabetes mellitus," *BMJ Open Diabetes Research & Care*, vol.10, no.1, pp.1-7, January 2022. https://doi.org/10.1136/bmjdrc-2021-002560

[13] B. P. Nguyen, H. N. Pham, H. Tran, N. Nghiem, Q. H. Nguyen, et al., "Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records," *Computer Methods and Programs in Biomedicine*, vol.182, pp.105055, December 2019. https://doi.org/10.1016/j.cmpb.2019.105055

[14] B. Ozturk, T. Lawton, S. Smith, and I. Habli, "Predicting Progression of Type 2 Diabetes Using Primary Care Data with the Help of Machine Learning," In Caring is Sharing–Exploiting the Value in Data for Health and Innovation, vol.302, pp.38 - 42, 2023.

[15] M. M. Mijwil, "Smart Architectures: Computerized Classification of Brain Tumors from MRI Images Utilizing Deep Learning Approaches," *Multimedia Tools and Applications*, vol.84, no.5, pp. 2261-2292, October 2024. https://doi.org/10.1007/s11042-024-20349-x

[16] R. Ramazi, C. Perndorfer, E. C. Soriano, J-P. Laurenceau, and R. Beheshti, "Predicting progression patterns of type 2 diabetes using multi-sensor measurements," *Smart Health*, vol.21, pp.100206, July 2021. https://doi.org/10.1016/j.smhl.2021.100206