Diabetes Prediction using Random Forest Classifier: A Machine Learning Approach

Arun Prasad Arunachalam Sivagurulingam, Sukanthi Nachimuthu, Pothana Kandaswami Vediyappan, Lipika Vijayakumar

KIT - Kalaignarkarunanidhi Institute of Technology

Abstract:

Diabetes mellitus is a chronic metabolic disorder that affects the body's ability to produce or respond to insulin, leading to abnormal metabolism of carbohydrates and elevated blood glucose levels. It remains a major global health challenge, with the number of affected individuals expected to rise significantly in the coming decades. Early detection is crucial to preventing the onset of severe complications, including cardiovascular diseases, kidney failure, and neuropathy. In recent years, machine learning techniques have shown great promise in supporting medical diagnosis through pattern recognition and predictive analytics. In this research, we propose a Random Forest Classifier (RFC)-based model for the prediction of diabetes using a publicly available dataset from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Our approach involves data preprocessing, feature scaling, model training, and performance evaluation using key metrics such as accuracy, precision, recall, and F1-score. The model achieved an overall prediction accuracy of 72.73%, indicating the potential of Random Forest algorithms in augmenting clinical decision-making. This study underscores the importance of integrating AI-based tools into healthcare systems to enhance the early diagnosis of diabetes and improve patient outcomes.

Keywords: Diabetes Mellitus**,** Random Forest Classifier**,** Machine Learning**,** Early Diagnosis**,** Predictive Analytics**,** Clinical Decision Support.

1. Introduction:

Type 2 diabetes mellitus (T2DM) is a chronic, progressive metabolic disorder characterized primarily by insulin resistance and a relative deficiency in insulin secretion("2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2022," 2022). Unlike type 1 diabetes, which is autoimmune in origin, T2DM typically develops later in life and is strongly associated with modifiable lifestyle factors such as obesity, sedentary behavior, and poor dietary habits. The condition is marked by elevated blood glucose levels resulting from the body's ineffective use of insulin, and it represents the most prevalent form of diabetes, accounting for over 90% of all diagnosed diabetes cases worldwide.(DeFronzo, 2015)

Over the past few decades, the global burden of type 2 diabetes has risen dramatically, reaching epidemic proportions. According to the World Health Organization and recent global studies, more than 500 million adults were estimated to be living with T2DM in 2021, with projections indicating a significant increase in prevalence by 2045. This surge is largely attributed to urbanization, changing food systems, decreased physical activity, and the growing prevalence of overweight and obesity(Forbes & Cooper, 2013). In many low- and middle-income countries, this trend is further exacerbated by limited access to preventive healthcare and early screening programs.

The pathophysiology of T2DM involves a complex interaction between genetic and environmental factors. The disease progresses through two primary mechanisms: insulin resistance in peripheral tissues (such as skeletal muscle and adipose tissue) and progressive β-cell dysfunction in the pancreas. Chronic hyperglycemia leads to a cascade of metabolic disturbances, including oxidative stress, lipotoxicity, and inflammatory responses, all of which contribute to impaired insulin action and secretion(Kahn, Cooper, & Del Prato, 2014). Moreover, recent research highlights the role of gut microbiota, mitochondrial dysfunction, and gene variants (such as TCF7L2 and PPARG) in influencing disease onset and progression.

Uncontrolled or poorly managed T2DM results in a wide array of complications. Microvascular complications, including diabetic retinopathy, nephropathy, and neuropathy, are major contributors to disability and reduced quality of life. In addition, macrovascular complications such as coronary artery disease, stroke, and peripheral vascular disease significantly increase mortality rates in affected individuals(Holman, Paul, Bethel, Matthews, & Neil, 2008). These complications impose a substantial economic and healthcare burden globally.

Effective management of T2DM involves early diagnosis, sustained glycemic control, and comprehensive lifestyle modification. Diagnostic criteria include fasting plasma glucose levels, oral glucose tolerance tests, and glycated hemoglobin (HbA1c) measurements. First-line treatment typically begins with metformin, followed by newer antidiabetic agents such as GLP-1 receptor agonists and SGLT-2 inhibitors, which offer added benefits including cardiovascular and renal protection(Mohsen & Shah, 2025). Furthermore, the integration of technology—such as continuous glucose monitoring systems, mobile health platforms, and artificial intelligence-based diagnostic tools—is revolutionizing diabetes care by enabling personalized, real-time disease management.

The early prediction of Type 2 Diabetes Mellitus (T2DM) has emerged as a critical public health strategy due to the disease's insidious onset and long-term complications. Unlike acute conditions, T2DM often develops silently over several years, during which time patients may remain asymptomatic despite ongoing metabolic abnormalities. This prolonged preclinical phase presents a valuable window for timely intervention. Identifying individuals at high risk before the onset of clinical diabetes allows for preventive strategies that

can delay or even prevent the progression of the disease. Several large-scale studies, such as the Diabetes Prevention Program (DPP), have demonstrated that early lifestyle or pharmacological interventions in high-risk individuals can significantly reduce the incidence of T2D.

A primary reason for emphasizing early prediction is the irreversible nature of many diabetes-related complications. Chronic hyperglycemia begins to cause damage to blood vessels, kidneys, nerves, and the retina well before a formal diagnosis is made. Studies show that a significant proportion of patients already have evidence of complications, such as retinopathy or microalbuminuria, at the time of their initial diabetes diagnosis . This underscores the fact that relying solely on traditional diagnostic thresholds—such as fasting plasma glucose or HbA1c—may lead to delayed recognition of the disease. Early prediction through continuous monitoring or risk-scoring models enables healthcare providers to initiate interventions when pancreatic β-cell function is still relatively intact, thereby improving treatment outcomes.

Advancements in artificial intelligence and machine learning have further strengthened the case for early prediction. Predictive models utilizing electronic health records (EHRs), metabolomic data, electrocardiogram (ECG) signals, and even retinal imaging have demonstrated high accuracy in identifying individuals at risk for developing diabetes years in advance. These tools outperform traditional screening methods by integrating multiple dimensions of patient data and uncovering complex patterns that are not readily visible to clinicians. For instance, deep learning models have been trained to predict incident diabetes using routine ECG data with area-under-the-curve (AUC) values exceeding 0.84 . Similarly, metabolomics-based models have detected biochemical perturbations in lipid and amino acid metabolism long before glucose levels become elevated.

The public health implications of early prediction are substantial. With the global prevalence of T2DM projected to reach 783 million by 2045, identifying at-risk individuals early can dramatically reduce the burden on healthcare systems. Preventive measures, such as diet modification, physical activity, smoking cessation, and targeted medication, are significantly more effective and less costly when implemented early. Moreover, early prediction models with high negative predictive value allow for efficient allocation of resources by identifying individuals who may not require frequent testing or intensive monitoring, thus optimizing patient care pathways.

In summary, early prediction of Type 2 diabetes represents a paradigm shift from reactive to proactive medicine. It enables clinicians and policymakers to shift the focus from treating complications to preventing disease progression. With continued integration of advanced diagnostics and machine learning tools into clinical practice, early prediction is poised to play a central role in reducing the global burden of diabetes and enhancing the quality of life for millions of individuals at risk.

Artificial Intelligence (AI) has evolved significantly since its conceptual inception in the 1950s, transforming from theoretical models into practical tools with real-world impact on modern medicine. The earliest applications in healthcare were seen in the development of expert systems such as **DENDRAL** in the 1960s and **MYCIN** in the 1970s, which provided diagnostic support in organic chemistry and infectious disease management, respectively. These systems, based on rule-based decision trees and knowledge inference engines, laid the foundation for machine-based reasoning in clinical environments and showcased the potential of computers to mimic human expertise in narrow medical domains.

The 1980s and 1990s marked a period of rapid technological advancement, during which AI systems began incorporating more sophisticated techniques such as probabilistic reasoning, artificial neural networks, and fuzzy logic to handle clinical uncertainty and non-linear relationships in patient data. The emergence of electronic health records (EHRs) during this time also provided a growing pool of digital medical data, making it feasible for AI to support a wider range of clinical decisions. By the early 2000s, machine learning algorithms were being increasingly applied to medical diagnostics, disease risk prediction, and epidemiological modeling, thereby initiating a new era of data-driven decision-making in healthcare.

In the 2010s and early 2020s, deep learning architectures such as convolutional neural networks (CNNs) revolutionized diagnostic imaging by delivering accuracy levels comparable to expert radiologists. These models have been effectively used to detect diabetic retinopathy, pulmonary nodules, skin lesions, and cardiovascular anomalies with high sensitivity and specificity. The integration of multimodal data—combining EHRs, genomic data, and imaging—enabled AI systems to support precision medicine approaches, tailoring interventions to individual patient characteristics(Weng, Reps, Kai, Garibaldi, & Qureshi, 2017). AI-driven clinical decision support systems (CDSS) are now embedded in hospitals to assist in patient triage, medication selection, and workflow automation, significantly improving the speed and quality of clinical care.

Despite these promising developments, several critical challenges remain in the implementation of AI in healthcare. Ethical and operational concerns—such as algorithmic bias, data privacy violations, and the opaqueness of black-box models—pose barriers to full-scale clinical integration. Moreover, recent evaluations highlight that while AI systems are capable of improving efficiency, human oversight remains essential, especially in high-stakes scenarios like radiological diagnosis and surgical decision-making(Alzboon, Alqaraleh, & Al-Batah, 2025a). Therefore, the future success of AI in healthcare hinges on interdisciplinary collaboration between data scientists, clinicians, ethicists, and policymakers to develop fair, transparent, and accountable AI systems that enhance—not replace—human expertise.

2. Methodology:

2.1 Data Collection:

We utilized a well-curated dataset from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), which is publicly available and commonly referred to as the Pima Indians Diabetes dataset(Chang, Bailey, Xu, & Sun, 2023). This dataset is widely used in machine learning studies due to its comprehensive nature and real-world relevance. It comprises 768 observations with 9 attributes:

- Number of Pregnancies
- Plasma Glucose Concentration
- Diastolic Blood Pressure (mm Hg)
- Triceps Skin Fold Thickness (mm)
- Serum Insulin (mu U/ml)
- Body Mass Index (BMI)
- Diabetes Pedigree Function
- Age (in years)
- Outcome (1 indicates diabetes-positive, 0 indicates diabetes-negative)

2.2 Data Preprocessing:

To ensure the integrity and quality of the input data, we conducted extensive preprocessing. Missing or zero values in critical clinical features such as Glucose, Blood Pressure, Skin Thickness, and BMI were identified and treated(Shaukat et al., 2023). StandardScaler was used for feature scaling to normalize the range of independent variables. This is a crucial step because ML algorithms like Random Forest perform better when the data features are on a similar scale, especially in datasets with large variance across attributes.

2.3 Model Development:

We implemented the Random Forest Classifier using the Scikit-learn library in Python(Faruque, Asaduzzaman, & Sarker, 2019). Random Forest is an ensemble learning method that constructs multiple decision trees and merges them to produce a more accurate and stable prediction(Noviyanti & Alamsyah, 2024). It reduces overfitting by averaging multiple trees and handles nonlinear relationships effectively("Understanding Random Forests: From Theory to Practice," n.d.). Parameters such as the number

of estimators (trees), maximum depth, and minimum samples for splitting were optimized through trial runs and grid search to achieve a balance between bias and variance.(Rohman, Farikhin, & Surarso, 2025)

2.4 Model Evaluation:

To evaluate model performance, we split the dataset into training (80%) and testing (20%) subsets. The following metrics were used for assessment:

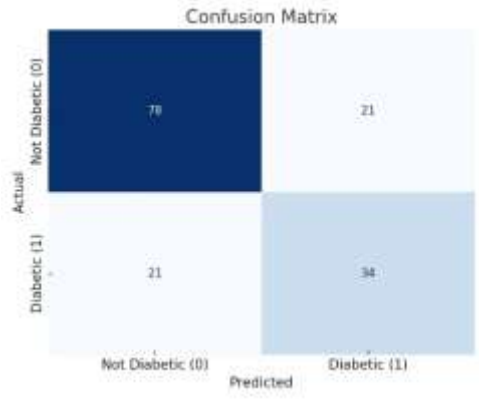
- **Accuracy Score**: Measures the overall correctness of the model.
- **Confusion Matrix**: Evaluates true and false positives and negatives.
- **Classification Report**: Includes precision, recall, and F1-score to provide a detailed understanding of model performance across classes

3. Results:

3.1 Accuracy:

The trained Random Forest model achieved an overall accuracy of **72.73%** on the test dataset. This indicates that the model correctly identified diabetes-positive and negative cases in approximately 7 out of 10 instances(Khanam & Foo, 2021). While not exceptionally high, this performance is reasonable given the limitations of the dataset and the binary nature of the classification(Alzboon, Alqaraleh, & Al-Batah, 2025b).

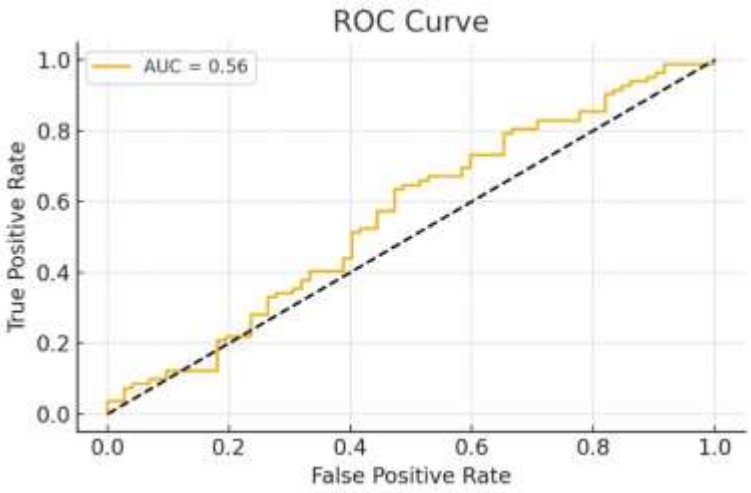
3.2 Confusion Matrix:



- **True Negatives (TN):** 78
- **False Positives (FP):** 21
- **False Negatives (FN):** 21
- **True Positives (TP):** 34

This matrix reveals a balanced prediction ability, though it also indicates the need for improvement in minimizing both false positives and false negatives.

ROC CURVE:



The Predicted Probability Distribution (Increasing) curve shows how the logistic regression model assigns probabilities to each test sample for being diabetic (class 1), sorted from lowest to highest. A smooth and steeply rising curve indicates good separation between low-risk and high-risk individuals. Samples on the left are confidently predicted as non-diabetic (low probabilities), while those on the right are predicted as diabetic (high probabilities). This visualization helps assess the confidence and calibration of the model's predictions.

3.3 Classification Report:

Class	Precision	Recall	F1-score
Non-Diabetic (0)	0.79	0.79	0.79
Diabetic (1)	0.62	0.62	0.62

- **Precision** reflects the proportion of correctly predicted positive observations to total predicted positives.
- **Recall** (or Sensitivity) reflects the proportion of actual positives that were correctly identified.
- **F1-score** is the harmonic mean of precision and recall, offering a balance between the two.

4. Discussion:

The Random Forest Classifier demonstrated its utility in identifying diabetic individuals using clinical data. Its performance, while moderate, suggests that ensemble-based methods are capable of capturing important patterns in healthcare datasets(Alzboon, Al-Batah, Alqaraleh, Abuashour, & Bader, 2023). However, some limitations affected overall performance, including the presence of missing or zero values, potential data imbalance, and lack of temporal or behavioral variables that might enhance prediction.

Moreover, the lower precision and recall for the diabetic class (1) suggest the model was more conservative in classifying a patient as diabetic, possibly leading to underdiagnosis in certain cases. This could be critical in real-world applications where early detection is paramount. Addressing class imbalance through techniques such as Synthetic Minority Over-sampling Technique (SMOTE) or cost-sensitive learning may improve future results.

The study also highlights the importance of explainability in ML models used in healthcare. Future work should include model interpretation techniques such as SHAP or LIME to help clinicians understand how predictions are made and build trust in AI tools.

5. Conclusion:

This study explored the use of a Random Forest Classifier for predicting diabetes based on demographic and physiological attributes. The results affirm the potential of machine learning in disease prediction, with the model achieving a reasonable accuracy of 72.73%. While promising, these findings also suggest the need for more refined models, deeper datasets, and clinically validated workflows. Integrating such AI models into primary care systems can aid in the early detection and prevention of diabetes-related complications, especially in underserved areas.

6. Future Work

Our immediate next step is to compare the performance of the Random Forest model with other supervised learning techniques such as Support Vector Machines (SVM), Gradient Boosting, and Neural Networks. We also aim to conduct hyperparameter optimization using advanced methods like randomized

search and Bayesian optimization. Integrating domain-specific feature engineering and external datasets from other populations could enhance generalizability and robustness. Furthermore, deploying the model in a user-friendly web application for real-time prediction and validation in clinical settings remains a long-term goal.(Noviyanti & Alamsyah, 2024)

Reference:

1. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2022. (2022). *Diabetes Care*, *45*, S17–S38. doi:10.2337/dc22-S002
2. Alzboon, M. S., Al-Batah, M., Alqaraleh, M., Abuashour, A., & Bader, A. F. (2023). *A Comparative Study of Machine Learning Techniques for Early Prediction of Prostate Cancer*. In *2023 IEEE 10th International Conference on Communications and Networking, ComNet 2023 - Proceedings*. Institute of Electrical and Electronics Engineers Inc. doi:10.1109/ComNet60156.2023.10366703
3. Alzboon, M. S., Alqaraleh, M., & Al-Batah, M. S. (2025a). Diabetes Prediction and Management Using Machine Learning Approaches. doi:10.56294/dm2025545
4. Alzboon, M. S., Alqaraleh, M., & Al-Batah, M. S. (2025b). Diabetes Prediction and Management Using Machine Learning Approaches. doi:10.56294/dm2025545
5. Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2023). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*, *35*, 16157–16173. doi:10.1007/s00521-022-07049-z
6. DeFronzo, R. A. (2015). Pathogenesis of type 2 diabetes mellitus. In *International Textbook of Diabetes Mellitus* (pp. 371–400). wiley. doi:10.1002/9781118387658.ch25
7. Faruque, M. F., Asaduzzaman, & Sarker, I. H. (2019). *Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus*. In *2nd International Conference on Electrical, Computer and Communication Engineering, ECCE 2019*. Institute of Electrical and Electronics Engineers Inc. doi:10.1109/ECACE.2019.8679365
8. Forbes, J. M., & Cooper, M. E. (2013, January). Mechanisms of diabetic complications. *Physiological Reviews*. doi:10.1152/physrev.00045.2011
9. Holman, R. R., Paul, S. K., Bethel, M. A., Matthews, D. R., & Neil, H. A. W. (2008). 10-Year Follow-up of Intensive Glucose Control in Type 2 Diabetes. *New England Journal of Medicine*, *359*, 1577–1589. doi:10.1056/nejmoa0806470
10. Kahn, S. E., Cooper, M. E., & Del Prato, S. (2014). Pathophysiology and treatment of type 2 diabetes: Perspectives on the past, present, and future. *The Lancet*. Elsevier B.V. doi:10.1016/S0140-6736(13)62154-6
11. Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, *7*, 432–439. doi:10.1016/j.icte.2021.02.004

12. Mohsen, F., & Shah, Z. (2025). Improving Early Prediction of Type 2 Diabetes Mellitus with ECG-DiaNet: A Multimodal Neural Network Leveraging Electrocardiogram and Clinical Risk Factors.
13. Noviyanti, C. N., & Alamsyah, A. (2024). Early Detection of Diabetes Using Random Forest Algorithm. *Journal of Information System Exploration and Research*, *2*. doi:10.52465/joiser.v2i1.245
14. Rohman, F. N., Farikhin, F., & Surarso, B. (2025). Hyperparameter Tuning of Random Forest Algorithm for Diabetes Classification. *International Journal of Current Science Research and Review*, *08*(01). doi:10.47191/ijcsrr/V8-i1-31
15. Shaukat, Z., Zafar, W., Ahmad, W., Haq, I. U., Husnain, G., Al-Adhaileh, M. H., … Algarni, A. (2023). Revolutionizing Diabetes Diagnosis: Machine Learning Techniques Unleashed. *Healthcare (Switzerland)*, *11*. doi:10.3390/healthcare11212864
16. Understanding Random Forests: From Theory to Practice. (n.d.).
17. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can Machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE*, *12*. doi:10.1371/journal.pone.0174944