www.jcimcr.org

JCIMCR OPEN ACCESS Journal of Clinical Images and Medical Case Reports

ISSN 2766-7820

Research Article

Open Access, Volume 6

Machine learning-based diabetes prediction: A comprehensive study on predictive modeling and risk assessment

Yalda Ghazizadeh[1]*; Shirin Salehi[2]; Mirhamid Mirsaeid Ghazi[3]

[1]Student Research Committee, School of Pharmacy, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

[2]Faculty of Pharmacy, University of Iran Medical Science, 123 University Avenue, Tehran 56789, Iran.

[3]School of Business Administration, Lakehead University, Thunder Bay, Ontario, Canada.

***Corresponding Author:** Yalda Ghazizadeh
Student Research Committee, School of Pharmacy,
Shahid Beheshti University of Medical Sciences,
Tehran, Iran.
Email: kiarashbabaie1990@gmail.com

Received: Apr 06, 2025
Accepted: Apr 28, 2025
Published: May 05, 2025
Archived: www.jcimcr.org
Copyright: © Ghazizadeh Y (2025).
DOI: www.doi.org/10.52768/2766-7820/3578

Abstract

Background: The hallmark of diabetes, a long-term metabolic disease, is hyperglycemia brought on by insulin failure. Its rising prevalence around the world emphasizes the necessity of precise and effective forecast techniques. By examining medical data, Machine Learning (ML) has demonstrated potential in the diagnosis and prognosis of diabetes.

Objective: Using 768 female individuals, all aged 21 years or older, residing near Phoenix, Arizona, attempts to assess how well different machine learning classifiers predict diabetes. The goal is to identify the best model while resolving issues with feature selection and data preparation.

Methods: Analysis was done on six different machine learning methods. The dataset was preprocessed using feature standardization, missing value management, and Recursive Feature Elimination (RFE) to choose the best features. The model's performance was assessed using the following metrics: accuracy, sensitivity, specificity, positive predictive value, and negative predictive value.

Results: The models with the best accuracy were Random Forest (RF) (84%), Support Vector Machine (SVM), and Logistic Regression (82%). Decision trees and Naïve Bayes performed competitively, but marginally worse. The results imply that using very easy preprocessing methods, traditional machine learning models may produce accurate diabetes predictions.

Conclusion: With their capacity to balance interpretability and accuracy, traditional machine learning models—Random Forest and SVM in particular-show great promise for diabetes prediction. Conventional ML models are still useful for clinical applications because of their transparency and simplicity of use, even if sophisticated deep learning techniques can improve prediction. To increase forecast accuracy and generalizability, future studies should investigate hybrid techniques that combine deep learning and conventional models.

Keywords: Diabetes prediction; PIMA Indian dataset; Diabetes diagnoses; Machine learning.

Introduction

Diabetes is a metabolic condition that affects how well a person's body processes blood glucose, also referred to as blood sugar. It is typified by hyperglycemia, which is brought on by deficiencies in either insulin action or production, or both [1,2]. Diabetes has emerged as a global public health emergency. As of 2019, the International Diabetes Federation estimates that 463 million people globally have diabetes [3]. It is expected to reach 578 million (10.2%) by 2030 and 700 million (10.9%) by 2045 due to its fast-rising incidence [4]. Predictive analysis is a method that uses a variety of machine learning algorithms, data mining, and statistical methods to analyze past and present data to predict future occurrences. By implementing predictive analysis to healthcare data, important decisions and predictions can be made. Predictive analytics applies machine learning techniques to diagnose diseases as accurately, improve patient care, optimize assets, and improve clinical results [5]. The application of machine learning to diabetes prediction has been the subject of several studies. For example, a thorough assessment of machine learning's use in diabetes research was carried out by Kavakiotis [6] et al. who emphasized its value in decision support and predictive analytics. Comparing several categorization models, Sisodia [7] et al. discovered that ensemble approaches, such as Random Forest, performed better than conventional algorithms. Even though machine learning has showed promise, issues including feature selection, data quality, and model interpretability still exist. A machine learning-based system for diabetes classification was suggested by Feng [8] et al. who used methods like Synthetic Minority Over-sampling Technique with Edited Nearest Neighbors (SMOTEENN) and Generative Adversarial Networks (GANs) to solve issues with feature analysis, class imbalance, and data preparation. Their method showed the potential of cutting-edge AI models to enhance diabetes prediction and obtained great accuracy (96.27% for binary classification). Predictive medicine leverages advanced bioinformatics, genomics, and artificial intelligence to assess an individual's risk of developing diseases and tailor preventive or therapeutic strategies accordingly [9-13]. By analyzing genetic variations, biomarkers, and patient history, predictive models can identify predispositions to conditions such as cancer, cardiovascular diseases, and neurodegenerative disorders [14-18]. High-throughput sequencing and machine learning algorithms enable early diagnosis, prognosis estimation, and personalized treatment plans based on a patient's molecular profile [19-22]. Additionally, pharmacogenomics-a key component of predictive medicine-optimizes drug selection and dosage by predicting individual responses to medications, reducing adverse effects and enhancing treatment efficacy [23]. This approach is revolutionizing healthcare by shifting from a reactive to a proactive model, ultimately improving patient outcomes and reducing medical costs. This study uses data from 768 female individuals, all aged 21 years or older, residing near Phoenix, Arizona, to evaluate how well different machine learning algorithms predict diabetes. Using several methods, we want to identify the model that produces the best accurate predictions and investigate possible enhancements for further studies.

Methods

Participants: This study focused on 768 female individuals, all aged 21 years or older, residing near Phoenix, Arizona, with data collected by the National Institute of Diabetes and Digestive and Kidney Diseases. Diabetes diagnosis followed the World Health Organization's criteria, which classify an individual as diabetic if their plasma glucose level reaches or exceeds 200 mg/dL (11.1 mmol/L) two hours after a glucose load during a survey examination.

Study features: Each patient record includes the following features.

- Pregnancies – Number of times pregnant.
- Glucose – Plasma glucose concentration during 2 h in an oral glucose tolerance test.
- Blood Pressure – Diastolic blood pressure (mm Hg).
- Skin thickness – Triceps skinfold thickness (mm), indicative of body fat percentage.
- Insulin – 2-hour serum insulin level (mu U/ml).
- BMI – Body mass index, a measure of body fat based on weight and height.
- Diabetes pedigree function – A score indicating genetic predisposition to diabetes.
- Age – Age of the patient (years).

Statistical analysis

This research includes various processes such as data preprocessing, data normalization, feature selection, and evaluating the results (Figure 1). Data preparation was done to deal with missing values, standardize features, and eliminate outliers before model training. To resolve missing values, we used a hybrid strategy that used mean and median imputation to preserve the dataset's distribution while reducing potential bias. Identifying and managing outliers is critical for reliable data analysis. The Insulin variable in the original dataset included a substantial number of outliers that persisted even after missing values were imputed; hence, these outliers were deleted. Several machine learning models were used in this study, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR) to train models for predicting individuals with diabetes [24-26]. The dataset is divided 80/20% into training and test sets. 10-fold cross-validation was utilized to validate the performance of the learning model [27].

Data availability: You can download the data by https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Results

Diabetic individuals were 37 years old on average, while not diabetic individuals were 31 years old on average. Pregnancies, glucose, skin thickness, insulin, and BMI show significant differences between diabetic and non-diabetic groups, emphasizing their significance in diabetes prediction. (Table 1) shows a summary of characteristics of each group.

With highest Sensitivity (0.74), Specificity (0.90), Accuracy (0.84), PPV (0.80), and NPV (0.87), the Random Forest (RF) model exhibits the best overall performance. This implies that

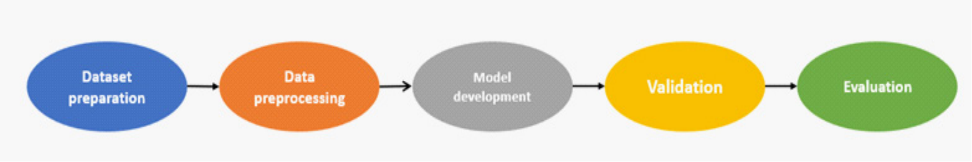

Figure 1: Overview of the proposed methodology.

Table 1: Shows a correlation heatmap that depicts the correlations between the dataset's distinct attributes. The data shows that glucose levels had the largest positive link with diabetes outcome, whereas other parameters like as BMI, age, and Diabetes Pedigree Function have moderate correlation. In contrast, factors like as skin thickness and insulin levels had less correlation with the outcome.

Feature	Non-Diabetic	Diabetic	p-value
Pregnancies (Mean (SD))	3.30(3.02)	4.87(3.74)	<0.001
Glucose (Mean (SD))	109.98(26.14)	141.26(31.94)	<0.001
Blood Pressure (Mean (SD))	68.18(18.06)	70.82(21.49)	0.072
Skin Thickness (Mean (SD))	19.66(14.89)	22.16(17.68)	0.038
Insulin (Mean (SD))	68.79(98.87)	100.34(138.69)	<0.001
BMI (Mean (SD))	30.30(7.69)	35.14(7.26)	<0.001
Diabetes Pedigree Function (Mean (SD))	0.43(0.30)	0.55(0.37)	<0.001
Age (Mean (SD))	31.19(11.67)	37.07(10.97)	<0.001

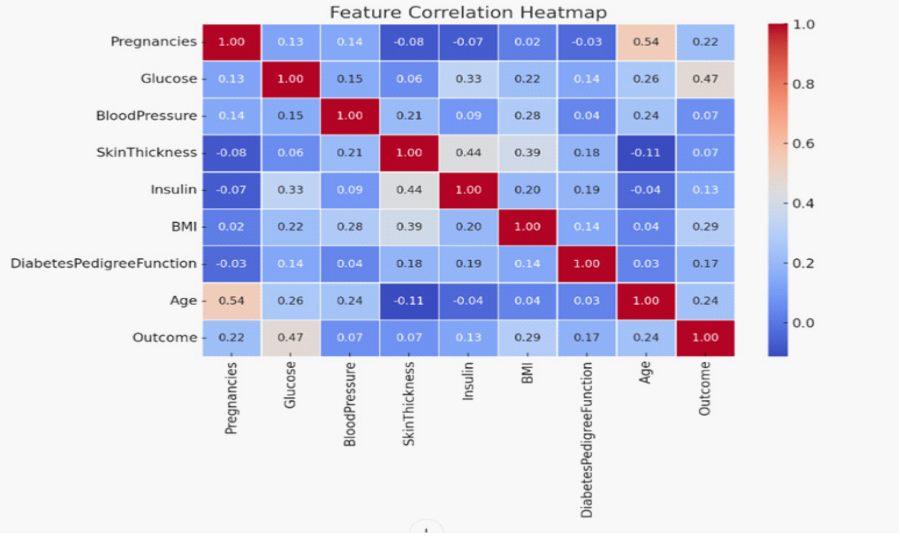

Figure 2: Feature correlation heatmap.

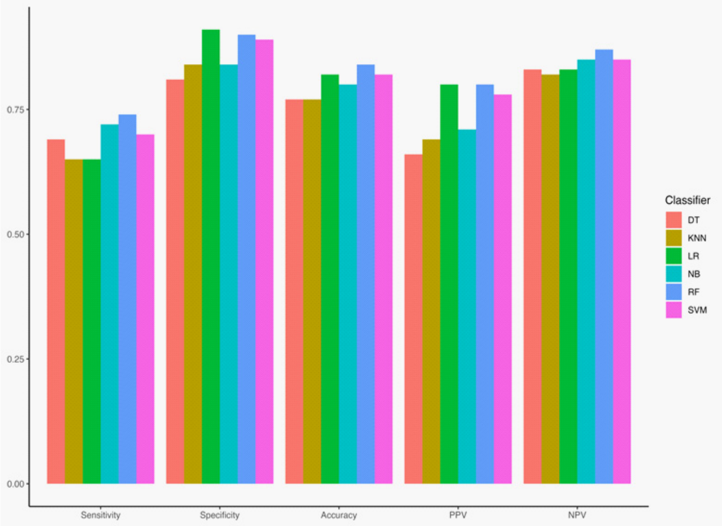

Figure 3: Evaluation metrics of different machine learning models in diabetes prediction.

RF is highly capable of accurately classifying both positive and negative situations. With an accuracy of 0.82, SVM and Logistic Regression (LR) both demonstrate excellent performance. LR has greater Specificity (0.91), whereas SVM has a higher Sensitivity (0.70) than LR (0.65). With an accuracy of 0.80 and a balance between sensitivity (0.72) and specificity (0.84), Naïve Bayes (NB) performs a bit worse than SVM. With an accuracy of 0.77, a lower PPV, and a modest sensitivity, Decision Tree (DT) and KNN perform the worst.

Discussion/conclusion

Yahyaoui et al. suggested a diabetes predicting framework leveraging machine learning and deep learning methodologies [28]. They applied RF, SVM, and convolutional neural network to identify and diagnose diabetes patients. The outcomes indicated that the RF model surpassed deep learning and support vector machine SVM techniques, attaining a total accuracy of 83.67%. Sharma et al. Used approaches such as NB [28], LR, decision tree, and artificial neural network for diabetes prediction. Among these strategies, LR yielded the highest precision of 80.43% in identifying whether a patient has diabetes or not. Haritha et al. leveraged the PIMA dataset with a KNN classifier and the Cuckoo fuzzy KNN algorithm [30], obtaining an accuracy of 81.00%. Patra and Kuntia [31] introduced a Standard Deviation KNN (SDKNN) algorithm for diabetes categorization on the PIDD dataset. The model applied the standard deviation of KNN attributes to determine the distance between training and testing data, achieving an accuracy of 83.76% for the enhanced weighted SDKNN. By combining Generative Adversarial Networks (GANs) and Synthetic Minority Over-sampling Technique with Edited Nearest Neighbors (SMOTEENN), recent studies, such the one by Feng et al. (2023), have shown even greater classification accuracy (96.27%). Although these techniques enhance model performance, it's crucial to understand that they also increase computing complexity and necessitate the use of bigger training datasets. On the other hand, our work shows that conventional machine learning models may still perform quite competitively even in the absence of advanced data augmentation or deep learning methods [32]. Even though this study concentrated on conventional machine learning models, combining convolutional and Artificial Neural Networks (ANNs and CNNs) might improve prediction accuracy, especially for bigger datasets. An approach to diabetes risk assessment that is more dynamic and individualized may be possible by using real-time health monitoring data from wearable technology (such as smartwatches and glucose monitors). The generalizability of machine learning models for diabetes prediction would be enhanced by using datasets that contain people from different ethnic origins. Physicians may be able to make better judgments by integrating machine learning-based diagnostic assistance systems with Electronic Health Records (EHRs). This might ultimately result in earlier identification and better patient outcomes. Our study demonstrates that conventional machine learning models continue to be quite successful in predicting diabetes and provide notable benefits in terms of accessibility and interpretability. Our findings demonstrate that Random Forest and SVM are appropriate for clinical applications, they can achieve good prediction accuracy with comparatively little preprocessing. In the end, this work adds insightful information to the continuing investigation of diabetes prediction powered by machine learning. Even if deep learning has the potential to lead to more developments, our results confirm that conventional machine learning techniques are still essential for medical diagnosis. To develop more reliable, comprehensible, and clinically useful models for diabetes prediction and treatment, future research should try to integrate the advantages of both conventional and deep learning approaches.

References

1. Diagnosis and classification of diabetes mellitus. Diabetes Care, 2014; 37(1): 81-90.

2. Butt UM, et al. Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications. Journal of Healthcare Engineering, 2021; 2021(1): 9930985.

3. Latest figures show 463 million people now living with diabetes worldwide as numbers continue to rise. Diabetes Res Clin Pract. 2019; 157: 107932.

4. Wang X, et al. Thiazolidinedione derivatives as novel GPR120 agonists for the treatment of type 2 diabetes. RSC Advances. 2022; 12(10): 5732-5742.

5. Mujumdar A, V Vaidehi. Diabetes Prediction using Machine Learning Algorithms. Procedia Computer Science. 2019, 165: 292-299.

6. Kavakiotis I, et al. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J. 2017; 15: 104-116.

7. Sisodia D, DS Sisodia. Prediction of Diabetes using Classification Algorithms. Procedia Computer Science. 2018; 132: 1578-1585.

8. Feng X, Y Cai, R Xin. Optimizing diabetes classification with a machine learning-based framework. BMC Bioinformatics. 2023; 24(1): 428.

9. Sadeghnezhad E, et al. Cross talk between energy cost and expression of Methyl Jasmonate-regulated genes: from DNA to protein. Journal of Plant Biochemistry and Biotechnology. 2019; 28: 230-243.

10. Samandari Bahraseman MR, et al. The use of integrated text mining and protein-protein interaction approach to evaluate the effects of combined chemotherapeutic and chemopreventive agents in cancer therapy. Plos one. 2022; 17(11): 0276458.

11. Shiralipour A, et al. Identifying key lysosome-related genes associated with drug-resistant breast cancer using computational and systems biology approach. Iranian Journal of Pharmaceutical Research: IJPR, 2022; 21(1): 130342.

12. Soltanyzadeh M, et al. Clarifying differences in gene expression profile of umbilical cord vein and bone marrow-derived mesenchymal stem cells; A comparative in silico study. Informatics in Medicine Unlocked. 2022. 33: 101072.

13. Zareei S, et al. PeptiHub: A curated repository of precisely annotated cancer-related peptides with advanced utilities for peptide exploration and discovery. Database. 2024; 2024: 092.

14. Houri H, et al. High prevalence rate of microbial contamination in patient-ready gastrointestinal endoscopes in Tehran, Iran: An alarming sign for the occurrence of severe outbreaks. Microbiology Spectrum. 2022; 10(5): e01897-22.

15. Kharaghani AA, et al. High prevalence of Mucosa-Associated extended-spectrum β-Lactamase-producing Escherichia coli and Klebsiella pneumoniae among Iranain patients with inflammatory bowel disease (IBD). Annals of Clinical Microbiology and Antimicrobials. 2023; 22(1): 86.

16. 16. Khorsand, B., et al., Alpha influenza virus infiltration prediction using virus-human protein-protein interaction network. Mathematical Biosciences and Engineering, 2020. 17(4): p. 3109-3129.

17. Khorsand B, A Savadi, M Naghibzadeh. SARS-CoV-2-human protein-protein interaction network. Informatics in medicine unlocked. 2020; 20: 100413.

18. Khorsand B, et al. Overrepresentation of Enterobacteriaceae and Escherichia coli is the major gut microbiome signature in Crohn's disease and ulcerative colitis; a comprehensive metagenomic analysis of IBDMDB datasets. Frontiers in cellular and infection microbiology. 2022; 12: 1015890.

19. Haghzad T, et al. A computational approach to assessing the prognostic implications of BRAF and RAS mutations in patients with papillary thyroid carcinoma. Endocrine; 2024; 86(2): 707-722.

20. Khorsand B, A Savadi, M Naghibzadeh. Comprehensive host-pathogen protein-protein interaction network analysis. BMC bioinformatics, 2020; 21: 1-22.

21. Khorsand B, A Savadi, M Naghibzadeh. Parallelizing assignment problem with DNA strands. Iranian Journal of Biotechnology, 2020; 18(1): 2547.

22. Razavi SA, et al. Metabolite signature of human malignant thyroid tissue: A systematic review and meta-analysis. Cancer Medicine. 2024; 13(8): 7184.

23. Khorsand B, et al. OligoCOOL: A mobile application for nucleotide sequence analysis. Biochemistry and Molecular Biology Education. 2019; 47(2): 201-206.

24. Hourfar H, et al. Machine Learning-Driven Identification of Molecular Subgroups in Medulloblastoma via Gene Expression Profiling. Clinical Oncology. 2025: 103789.

25. Khorsand B, et al. Enhancing ischemic stroke management: Leveraging machine learning models for predicting patient recovery after Alteplase treatment. Brain Injury. 2025: 1-7.

26. Khorsand B, et al. Enhancing the accuracy and effectiveness of diagnosis of spontaneous bacterial peritonitis in cirrhotic patients: A machine learning approach utilizing clinical and laboratory data. Advances in Medical Sciences. 2025; 70(1): 1-7.

27. Hesami Z, et al. Microbiota as a State-of-the-art Approach in Precision Medicine for Pancreatic Cancer Management: A Comprehensive Systematic Review. iScience. 2025: 112314.

28. Yahyaoui A, et al. A decision support system for diabetes prediction using machine learning and deep learning techniques. in 2019 1st International informatics and software engineering conference (UBMYK). 2019.

29. Sharma A, K Guleria, N Goyal. Prediction of diabetes disease using machine learning model. in International Conference on Communication, Computing and Electronics Systems: Proceedings of ICCCES 2020. 2021.

30. Haritha R, DS Babu, P Sammulal. A hybrid approach for prediction of type-1 and type-2 diabetes using firefly and cuckoo search algorithms. International Journal of Applied Engineering Research. 2018; 13(2): 896-907.

31. Patra R. Analysis and prediction of Pima Indian Diabetes Dataset using SDKNN classifier technique. in IOP Conference series: materials science and engineering. 2021.

32. Irankhah L, et al. Analyzing the performance of short-read classification tools on metagenomic samples toward proper diagnosis of diseases. Journal of bioinformatics and computational biology. 2024. 22(5): 2450012.