AI-Based Predictive Model for Accurate Diabetes Detection

Deepali Singh[1], Megha Singh[2]

[1,2]Department of engineering and technology, Sharda University, India

deepalisingh3184@gmail.com[1], meghasingh.1510@gmail.com[2]

Abstract

Diabetes mellitus is a chronic metabolic disorder that has millions of people in the entire world being admitted in hospitals across the globe due to insulin deficiency or resistance. Due to that, it is among the major causes of severe complications, such as cardiovascular disease, kidney failure, neuropathy and vision loss, including Diabetic Retinopathy. The most important step is early and correct diagnosis so as to stop the spread of the disease and avoid these debilitating consequences of the disease. Conventional care procedures can be invasive, inconvenient, time-consuming and imaginative besides being confined depending on the region of concern, especially the underserved which implies that having scalable thoughtful solutions that automatically initiate health care interventions is vital. Artificial Intelligence (AI) and Machine Learning (ML) have promised to transform screening processes in a way that will lead to enhanced patient outcomes by enabling examination of large, complicated datasets to validate the presence of subtle patterns, and make appropriate predictions.1 The following paper will synthesize studies into several types of AI-driven attempts, including those models based on organized clinical data, imaging-based diagnostics (retinal and tongue images), and real-time monitoring (wearable technology). The discussion will entail the discussion of the performance of various ML and Deep Learning algorithms, what makes the process of sound data preprocessing another critical factor, and why Explainable AI (XAI) is necessary to build trust and clinical reception. The results show that there are meaningful accuracy gains and powerful features over diabetes prediction and show that AI can play an increasingly destructive role in diabetes management.

Keywords: *diabetes prediction, machine learning, deep learning, voice biomarker, ECG, explainable AI, XAI, hybrid ensemble.*

1. Introduction

Diabetes mellitus is one of the global health emergencies where millions of people on the planet have been affected. Being a chronic and exhibiting continuously elevated blood sugar levels, this persistent and metabolic disorder occurs as a result of either the inability of the human body to produce or actually use insulin. The International Diabetes Federation stated that there were 536.6 million individuals age between 20 and 79 that had diabetes in 2021 with the projections reaching the mark of 783.2 million answering these criteria by 2045. This is a highly significant group of the global population in 2021 that is attributable to the implications of the unprotected type of diabetes and that a whole spectrum of complications to spurt include cardiovascular disease, renal failure, nephropathy, and blindness. Diabetic Retinopathy (DR) is the most important microvascular complication of diabetes in the modern world, and this complication follows a gradual process, where the mild microaneurysms proceed to serious proliferative DR; formation of new aberrant blood vessels. Diabetes besides having severe health consequences is a colossal economic burden. One illustration is that the United States alone spent 237 billion on the treatment of diabetes alone in 2017 and an estimate of the productivity discharge involving the disease occurred through an estimation of ninety billion dollars.

Even though the common methods of the diabetes diagnosis and complications offer a lower level, they have far reached limitations. DR screening through manual examination by ophthalmologists is often not only time consuming, but also inter-rater unreliable. Accessibility to specialized medical practitioners particularly in rural or underserved locations presents a significant challenge and slows down diagnosis and other treatment outcomes. Even clinical visits and screening diabetes using blood diagnostic tests and medical history have proven not enough to generate a warning sign. Moreover, it is intrusion in terms of diabetes and losses in the policies of popularizing screening and coronary intervention into health systems.

The artificial intelligence (AI) and machine learning (ML) are thus the examples of the really disruptive solutions due to the appearance of the described challenges. Other possibilities of using such technologies are the processing of major and complex data, as well as the detection of slight variations in trends and exceptionally precise predictions, which profoundly affect the nature of the screening process and improve patient care. The fact that the artificial intelligence will be learning and creating on the background of the new stimuli and that it will integrate the healthcare structures comparatively easily does also bring some exciting pros to the old approach. Artificial intelligence architectures are expeditious, inexpensive, and potentially life-inspiring since the quantartment and lucid precision in diagnosis.

Through the spotting of appropriate diabetes models using AI in the present day, this research paper has tried to review, synthesize, and analyze the existing situation over a broad spectrum. It addresses different ways of information, strictly discusses the greatest successes of the techniques, and comes to terms with the current problems. Diabetes and AI in healthcare background. The following pages will outline the background of the topic of the article, provide the full literature review of various AI/ML approaches.

2. Research Methodology

An elaborate combination of machine learning algorithms, deep learning structures, and detailed data processing methods have been used to create Artificial Intelligence-based forecast models in diabetes diagnosis. These models are thoroughly tested and they utilize standard measures so as to prove their strength, consistency, and comparability.

Conventional machine learning has become central in diabetes prediction because it is easy to interpret, computationally efficient, and can easily deal with complex data. Logistic Regression is an easy and effective vanilla model but it can be constrained by the assumption of linearity. Non-parametric method such as K-Nearest Neighbours (KNN) is flexible and it also can be applied in data imputation, but it is unstable to parameter selection and computationally expensive with large data. The Support Vector Machines (SVMs) work well when dealing with a high dimension task and are advantageous in the use of kernel functions such as the Radial Basis Function (RBF) which make them suitable in the complex problem of predicting the HbA1c values. Decision Trees are visually interpretable and can be used to model non-linear relationships but can have the problem of overfitting. Random Forest is an implementation of an ensemble of decision trees that enhances the model stability and precision in the case of minimizing the overfitting. XGBoost is one of the gradient boosting methods that have gained popularity because of its efficiency, speed, and effective predictive capabilities.

Architectures of deep learning, mainly the Convolutional Neural Networks (CNNs) have transformed diabetes detection using images. CNNs can automatically extract hierarchical features of raw image data and hence would be suitable to detecting diabetic retinopathy as well as diagnosing diabetes using tongue images. The use of transfer learning, i.e., applying a pre-trained model such as ResNet or Inception on a domain-specific problem dramatically improves results in the case of limited data. The CNN model highly used in non-invasive tongue image analysis comprises basic traditional steps, namely; image acquisition, preprocessing, convolutional layer and pooling layer to extract the features, fully connected layers and lastly a sigmoid classifier that produces binary results (diabetic or non-diabetic).

The other step that is equally essential is that of preprocessing and feature engineering step since the provided input data are clean, balanced, and relevant. The missing values can be dealt with via KNN imputation, or deleting of a record which contains more missing entries than it has non-missing values, which keeps the data in integrity. Any outliers will be handled using techniques such as one-class SVM and Isolation Forest as a way of removing tremendously outside data points that may corrupt training. Oversampling the minority class with SMOTE, ADASYN, and recent hybrid algorithms SMOTE-ENN methods alleviates class imbalance present in medical datasets, and more importantly also removes noisy samples belonging to the majority class. The scaling of features so that their contribution in weight is uniform with faster convergence is possible by converting them using Min Max normalization or standardization. The process of feature selection can improve the implementation of models as it preserves only the features that matter applying the Chi-Square tests, correlation analysis as well as Principal Component Analysis (PCA) and optimization techniques such as Genetic Algorithms (GA). GA approaches the feature selection problem as the problem of optimization, where it iteratively searches a set to contain the most informative sample therefore it works best in the high dimensional spaces. The preprocessing procedures are generally applied in a row and follow a given order, imputation, then outlier elimination, then class proportioning, and finally feature selection to be guaranteed that the succeeding stage is better than the former one.

Finally, there is a set of evaluation measures upon which the model performance is ascertained. Accuracy is a crude scale that measures the correctness of the predictions whereas Precision and Recall measure the exactness and completeness of positive predictions respectively. The F1- score is a harmonic mean of Precision and Recall, which is best suited to an imbalanced dataset. AUC-ROC is meant to quantify the ability of a model to distinguish between the classes using this threshold and AUC-PR is dedicated to Precision-Recall trade-offs. Specificity gives the ability of the model to recognize healthy people correctly. The Matthews Correlation Coefficient (MCC) provides a balanced score, taking into account every major way to evaluate the confusion matrix and is therefore more trustworthy than plain accuracy at determining the imbalanced datasets. In regression-based prediction such as prediction of the level of HbA1c, statistics such as Mean Squared Error (MSE) are applied as a measure of averaged squared difference between predicted and real values.

3. Results and Discussion

Although AI has demonstrated tremendous potential in transforming the management of diabetes, there are various barriers to its implementation and success. Quality and availability of data is one of the key challenges. Medical data is usually full of missing values, imbalanced classes, and noise. Furthermore, it is hard to train strong AI models because of privacy issues and legal limitations, and the data available is of low quality and annotated. A very important challenge is model interpretability. With the developments in XAI, most models are black boxes, and clinicians cannot easily trust and interpret their predictions. Ensuring that explanations are relevant and of clinical value is a complex problem, which must be investigated at all times and it means the presence of data scientists and healthcare staff.

Issues of discrimination and equality are apposite. Because the datasets to which the models are applied are not representative, AI can be biased and offer inaccurate and incorrect predictions, particularly in minority populations. Close curation of data, algorithmic fairness and active monitoring can eliminate such biases. Also, regulatory and ethical concerns also have to be considered. The use of AI models in medical facilities should occur in a highly secured setting as may be written in FDA and compliance to laws of data safety including HIPAA and GDPR. The ethical concern on patient consent, patient data ethical ownership, and accountability shall also have to be mentioned. The other colossal issue is media generalization. Trained models may not apply well in other qualities of the clinical environment or population. In order to be relevant when applied into practice, it is also worthwhile to make sure the models can be generalized with regards to other demographics and health care settings. Potential constraints of datasets due to context are also another challenge. The majority of models are developed on the suitability of purely numerical or categorical data without regard to the circumstantial variables, such as family history, drug use and lifestyle parameters. Information of this nature would be extremely handy in terms of enhancing the quality and suitability of predictions that would have been delivered after integration.

Notwithstanding such, we can say that the further future development of AI in diabetes care is rather productive, and there are several current tendencies and areas of research. One of them is the multimodal data. Imagining, genetic and wearable sensor data can help provide more comprehensive information about the health of a patient compared to structured data. This is perhaps one available approach of rendering AI models useful and better. The second trend option is the real-time monitoring of smart sensors and separate objects. These technologies have transformed this whereby health can be checked anytime and abnormalities can be known early and a foetus can be taken care of beforehand and diabetes management can even take a step towards proactive and personal care. The difference in healthcare provision can be bridged by augmenting the supply of AI-driven diagnostics that utilizes mobile health applications and telemedicine platforms to distant and underserved regions. The development of the model and its practical testing based on research and large-scale experiments is important to promote trust, as well as to ensure the safety and effectiveness of AI models.

Clinical scientists, regulators and experts ought to collaborate in an attempt to establish standardized evaluation processes and guidelines. To learn how the diseases progress and how effective the treatment is, patient longitudinal studies might be employed. It is gaining popularity in AI-assisted teleophthalmology (detecting diabetic retinopathy). The implementation of AI models in the teleophthalmology process could enhance the effectiveness of diagnoses and other elements related to the screening process, particularly in regions with insufficient access to specialized services.

4. Formatting Tables

Authors	Year	Techniques Used
Sehgal et al.	2022	KNN, SVM, Gradient Boosting
Das and Ahmed	2023	SVM + Chi-Square + SMOTE
Ahmed et al.	2023	XGBoost, Random Forest + SHAP, LIME
Teimoory & Keyvanpour	2025	Calibrated Random Forest + LIME
Alapati et al.	2024	ResNet-50 Deep Learning Classifier
Naz et al.	2024	BiLSTM (Deep Learning)
Site et al.	2023	XGBoost + Multisensor Fusion
Zhu et al.	2023	Attention-based RNN + Edge Computing

Bataineh et al.	2025	XGBoost + SHAP
Khattri et al.	2023	Random Forest, Gradient Boosting + ROS
Devi & Karthik	2023	SVM, KNN, XGBoost + RF Ensemble

Table 1: Authors and Techniques Overview

Authors	**Dataset**	**Key Features / Methods**	**Accuracy**
Sehgal et al.	PIMA Diabetes Dataset	Flask deployment, comparative study of ML models	75% (KNN)
Das and Ahmed	PIMA Diabetes Dataset	Isolation Forest for outliers, SHAP for interpretability	99.58%
Ahmed et al.	Sylhet Diabetes Dataset	Emphasis on XAI-based trust and transparency	99.4%
Teimoory & Keyvanpour	PIMA Dataset	RFE, IQR outlier removal, ADASYN balancing	88%
Alapati et al.	Gestational Diabetes Dataset	CNN for non-linear patterns in pregnancy-related data	95%
Naz et al.	Custom Medical Dataset	Sequence modeling for improved recall and precision	90%
Site et al.	Glucose, ECG, ACC Sensors	Sensor-based model with 3-sensor integration for prediction	98.2%
Zhu et al.	CGM Clinical Dataset	IoMT deployment, real-time prediction and app-based visualization	High AUC
Bataineh et al.	Multimodal Clinical & Survey Data	Prediction of diabetic emotional distress with high transparency	96.14%
Khattri et al.	Diabetes Classification Dataset	Med Assist Bot for aiding novice practitioners	87%
Devi & Karthik	PIMA Dataset	Ensemble stacking, SMOTE+ENN balancing, LIME explainability	97%

Table 2 : Comparative Summary

Approach	**Data Type**	**Technique**	**Accuracy**
Structured Clinical Data	Tabular (e.g., Pima, Sylhet, Schorling)	SVM, XGBoost, Random Forest, Stacked Ensemble	Up to 99.58% (Schorling)
Image-Based Models	Retinal Fundus Images	CNN, Transfer Learning	AUC 0.93–0.95
Non-Invasive Approaches	Tongue Images, Wearable Sensor Data	CNN, ML classifiers	74%–93%
Preprocessing Techniques	Structured Clinical Data	KNN Imputation, SMOTE, ADASYN, Isolation Forest	Improved model performance

Feature Selection	Structured Clinical Data	Genetic Algorithm, Chi-Square, PCA	Boosted accuracy (e.g., RF from 84.5% to 90.3%)
Ensemble & Hybrid Models	Structured Clinical Data	Stacking (KNN, SVM, XGBoost + RF meta-model)	Up to 98% (Pima)
Deep Learning Models	Images & Multimodal Data	CNNs, Autoencoders	High performance
XAI Integration	All (esp. Structured)	SHAP, LIME	Not model-specific

Table 3: AI Approaches in Diabetes Prediction

Approach	Strengths	Limitations
Structured Clinical Data	High accuracy, effective with feature selection and preprocessing	Needs proper handling of missing data and class imbalance
Image-Based Models	Automatic feature extraction, high sensitivity/specificity	Requires large labeled datasets, high computational resources
Non-Invasive Approaches	Accessibility, continuous monitoring	Data quality from wearables may vary; less proven clinically
Preprocessing Techniques	Enhances data quality and balance	Adds complexity and computational overhead
Feature Selection	Reduces dimensionality, improves interpretability	Risk of discarding relevant but subtle features
Ensemble & Hybrid Models	Combines strengths of multiple models, reduces overfitting	More complex to implement and interpret
Deep Learning Models	Learns complex patterns, suitable for large/unstructured datasets	Requires large datasets, high training time
XAI Integration	Enhances interpretability, builds clinical trust	Needs careful explanation design for medical use

Table 4 : Strengths and Weaknesses of Methods

.

4.1. Formatting Figures

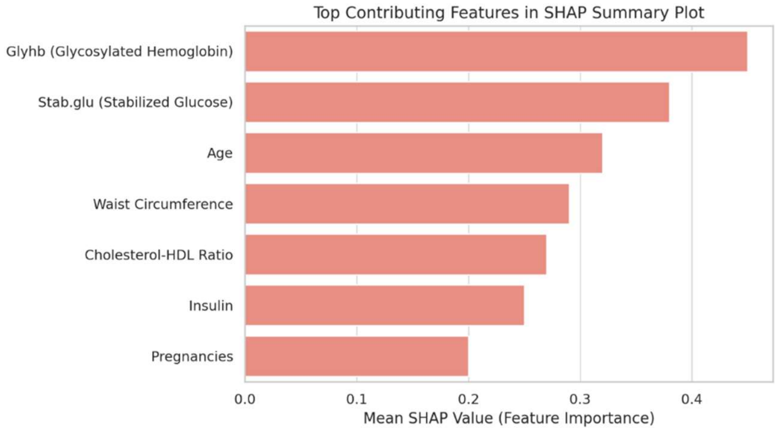

Figure 1: Features that contribute in SHAP

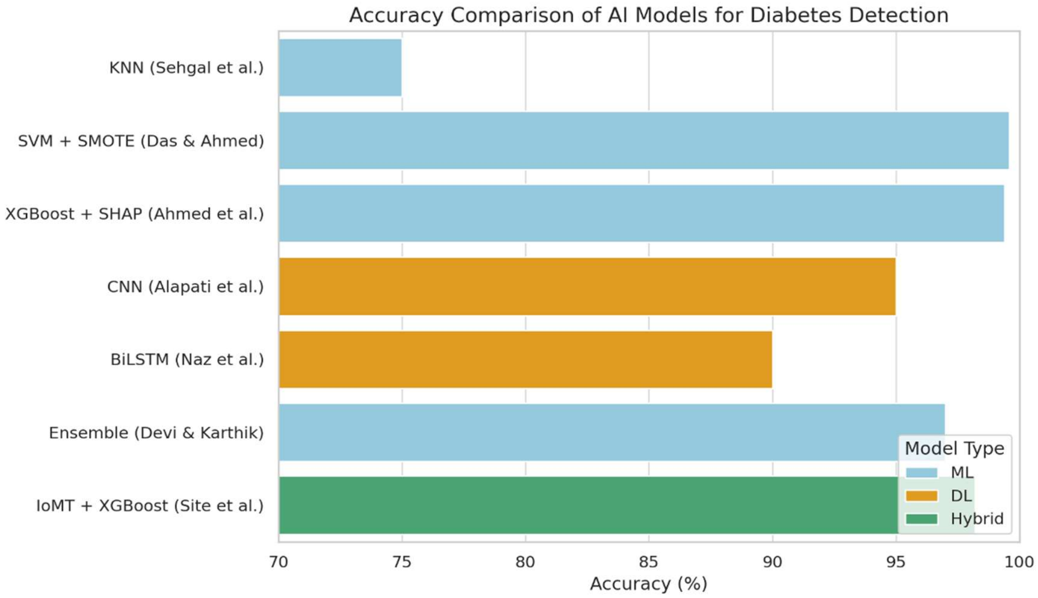

Figure 2: Comparison of different AI Models

5. Conclusions

The introduction of Artificial Intelligence and the concept of Machine Learning into the medical sector is a rather impressive phenomenon that opens opportunities and provides groundbreaking solutions to the complex of endemic problems that the global population faces with a chronic disease and a series of complications that have to do with tomorrow vision. Broadly, the prevailing analysis and discussion in this paper is a pointer to how AI-based predictive models can transform the medical management of diabetes.

Such models are always highly accurate, with typical prediction accuracies of 98-99% on clinical data predictions, and 93-95% on image-based diagnostics, when carefully designed data preparation techniques are applied (e.g. KNN imputation), outliers are removed (e.g.

OCSVM, iForest), and unbalanced class distributions are balanced (e.g. SMOTE, ADASYN, SMOTE+ENN).

However, more significantly, the need of the so-called Explainable AI (XAI), or SHAP and LIME, cannot be upheld. Such techniques would be invaluable in de-black boxing the highly complex AI models, and hence creating confidence in such models, as well as in actualizing these technologies helpful and applicable to the clinical setting.1 The XAI will render AI models as predictive tools; the XAI would be applied to provide meaningful results on what option was made by these models and what are the specific qualities that these models require to make predictions effective.

But lastly, AI-assisted predictive models will have the potential to transform the experience of diabetes screening, dramatically increase the prevalence of diabetes diagnosis and care, and eventually enhance the quality of life of affected individuals of the disease who may end up blind and acquiring other dreadful perhaps irreparable complications.1 AI will be a critical burden in helping the entire world fight the disease burden of diabetes. This dissenting influence has shifted health care to a new, easier and fairer screening and care and given health care the future where the diagnosis at the earliest stage will result in the improved treatment of the patient and an overall lower global distrust of diabetes.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Applications of AI in Predicting Drug Responses for Type 2 Diabetes. Garg S., Kitchen R., Gupta R., Pearson E. JMIR Diabetes, 2025.

[2] Let Curves Speak: A Continuous Glucose Monitor based Large Sensor Foundation Model for Diabetes Management. Luo J., Kumbara A., Shomali M. et al. arXiv, 2024.

[3] DiabetesNet: A Deep Learning Approach to Diabetes Diagnosis. Zhang Z., Ahmed K. A., Hasan M. R., Gedeon T., Hossain M. Z. arXiv, 2024.

[4] Enhanced Diabetes Detection and Blood Glucose Prediction Using TinyML-Integrated E-Nose and Breath Analysis: A Novel Approach Combining Synthetic and Real-World Data. Bioengineering, 2024.

[5] A Proposed Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App. El-Sofany, International Journal of Intelligent Systems, 2024.

[6] Diagnosing Diabetes using Machine Learning-based Predictive Models. Procedia Computer Science, 2024.

[7] Diabetes Detection Based on Health Conditions Using Advanced Learning Algorithm. ICAIS 2024.

[8] Comparative Analysis of Logistic Regression, SVM, XGBoost, and Random Forest Algorithms for Diabetes Classification. Jurnal Teknologi Sistem Informasi dan Aplikasi, 2024.

[9] A comprehensive review of machine learning techniques on diabetes detection. Visual Computing for Industry, Biomedicine, and Art, 2021.

[10] Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. (Private medical records, 2013-2018) – ML model for predicting T2D occurrence in next year.

[11] Predictive models for diabetes mellitus using machine learning techniques. Lai H., Huang H., Keshavjee K. et al. BMC Endocrine Disorders, 2019.

[12] Supervised Machine Learning based Ensemble Model for Accurate Prediction of Type 2 Diabetes. Akula R., Nguyen N., Garibay I. arXiv, 2019.

[13] Machine learning for the diagnosis of early-stage diabetes using temporal glucose profiles. Lee W. S., Jo J., Song T. arXiv, 2020.

[14] Sharma, H., Kumar, P., & Sharma, K. (2025). Smart Waste Management with IoT: An Optimized Triple Memristor Hopfield Neural Network Approach. International Journal on Smart & Sustainable Intelligent Computing, 2(1), 52-64.

[15] Gusain, N. (2025). Cardiovascular Disease Prediction through Machine Learning: A Comparative Study of Ensemble Techniques. *Revolutionary Advances in Computing and Electronics: An International Journal*, 27-40.

[16] Secure and Privacy-Preserving Automated Machine Learning Operations into End-to-End Integrated IoT-Edge-AI-Blockchain Monitoring System for Diabetes Mellitus Prediction. Hennebelle A., Ismail L., Materwala H. et al. arXiv, 2022.

[17] Srivastava, A., & Sharma, H. (2024). AI-driven environmental monitoring using Google Earth Engine. In IoT Sensors, ML, AI and XAI: Empowering A Smarter World (pp. 375-385). Cham: Springer Nature Switzerland.

[18] Bhola, A., Shrivastava, G., Sharma, H., & Kumar, P. (2025, February). Harnessing Digital Innovations for Sustainable Agriculture in India: Technology-Driven Smart Farming Framework. In International Conference On Innovative Computing And Communication (pp. 501-512). Singapore: Springer Nature Singapore..

[19] Sharma, H., Kumar, A., & Kumar, G. (2025). Privacy-Enhanced Federated Learning Framework for Intrusion Detection in Smart IoT Environments. Revolutionary Advances in Computing and Electronics: An International Journal, 15-25..

[20] Enhancing Early Detection and Prediction of Diabetes Mellitus in Patients of Indian Origin through Rigorous Machine Learning Techniques with Comprehensive Models Evaluation. Sarkar J., Pawar S. IJISAE, 2024