JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Explainable AI for Predicting Diabetes using ML

[1]Deep Vasoya, [2] Dr. Mohit Bhadla,

[1]Student, [2]HoD CE_IT & Associate professor,
[1]Department of Computer Engineering Gandhinagar University, Gandhinagar Gujarat India.
[2]Department of Computer Engineering Gandhinagar University, Gandhinagar Gujarat India

Abstract : Since diabetes mellitus is becoming more common in many populations, it has become a worldwide health concern. For early diagnosis and intervention, predictive models that are transparent and accurate are essential. Though their black-box character frequently restricts clinical application due to a lack of interpretability, traditional machine learning methods have shown substantial promise in diabetes prediction. By offering insights into model decisions, Explainable AI (XAI) tackles this issue and builds patient and healthcare professional trust. In this work, we investigate how to include XAI methods into diabetes prediction machine learning models while maintaining accuracy and transparency. Current research emphasizes the significance of interpretable models for individualized diagnosis and treatment planning by highlighting the role of genetic, metabolic, and behavioural risk factors in the onset of diabetes. Numerous researches have shown how model-specific interpretability techniques, SHAP values, and feature importance analysis can help close the gap between AI predictions and human comprehension. Additionally, improvements in explain ability and diagnostic accuracy have been made possible by developments in deep learning and hybrid prediction models. This study assesses different XAI methods used to forecast diabetes and examines how well they work to give doctors insightful explanations. The results open the door for AI-driven, patient-centred healthcare solutions by highlighting the significance of explainable models in actual clinical situations.

***IndexTerms* - Machine Learning in Healthcare, Explainable AI (XAI), Diabetes Risk Factors, Diabetes Prediction, XGBoost for Medical Diagnosis.**

I. INTRODUCTION

Millions of people worldwide suffer from diabetes mellitus, which has grown to be a serious global health concern. Environmental, behavioural, and genetic variables are intimately associated with the increasing frequency of type 2 diabetes. In order to manage the illness and lessen complications, early discovery is essential. Because machine learning models are more accurate than traditional statistical methods, they are being used more and more to predict diabetes. However, a major obstacle in clinical applications is these models' lack of interpretability.

The goal of incorporating Explainable AI (XAI) into diabetes prediction models is to increase transparency so that healthcare professionals can better understand the prediction process and make better decisions. Researchers may create models that not only make precise predictions but also shed light on the risk factors that contribute to them by utilizing XAI techniques. This will increase the dependability and practicality of AI-driven healthcare.

Numerous regions have observed the rising incidence of diabetes, underscoring the disease's expanding influence on public health [1,2]. Early detection and risk assessment are crucial because the condition is frequently associated with metabolic problems, genetic predisposition, and lifestyle modifications [3,4].

In order to increase diagnosis accuracy, prior research has concentrated on creating hybrid prediction models that use machine learning methods [5,6]. The relevance of AI-based models in medical research has been further reinforced by the exploration of biomarkers and genotype scores as prediction variables for diabetes [7, 8]. Effective prediction frameworks are necessary since diabetes is an established risk factor for cardiovascular problems [9, 10].

Numerous machine learning methods, such as support vector machines and deep learning, have been used to predict diabetes; nevertheless, their interpretability is limited by their black-box nature [11,12]. Clinicians can now comprehend how AI models make their predictions thanks to recent developments in XAI, which have brought techniques like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) [13,14].

Researchers have created frameworks that explain patient-specific risk variables and facilitate better informed medical decisions by applying XAI in healthcare [15,16]. Research has also looked into how AutoML might improve model performance while maintaining interpretability, which could lead to better methods for managing and predicting diseases [17, 18].

The larger objective of increasing openness in AI-driven healthcare solutions is in line with the use of XAI in diabetes prediction [19,20]. In order to increase the interpretability of prediction models, researchers have proposed new ontologies and frameworks, underscoring the significance of trust and explain ability in medical AI [21,22].

Making sure that models are accurate and comprehensible is essential for practical use as AI continues to transform healthcare [23, 24]. In order to ensure their smooth absorption into medical practice, future research directions will concentrate on improving XAI approaches to offer even more thorough and clinically relevant explanations [25].

II. Literature Review

In recent years, there has been a lot of interest in the use of machine learning in diabetes prediction. To increase the accuracy of identifying diabetes risk factors, a number of researches have investigated several models, such as support vector machines, decision trees, and deep learning [1,2].

Even though these models have shown excellent predictive accuracy, their interpretability is still problematic, which restricts their use in clinical settings [3,4]. By bringing transparency to decision-making, Explainable AI (XAI) approaches have contributed to closing this gap [5,6].

It has been demonstrated that using ensemble learning improves the prediction accuracy of diabetes diagnosis [13,14]. Strong prediction frameworks have been produced by studies that have integrated several algorithms to take use of their unique advantages [15,16]. Furthermore, to test these models and guarantee their generalizability across a range of demographics, real-world patient datasets have been included [17,18].

The application of feature selection strategies to enhance model interpretability has been a primary focus of recent research [19,20]. Reducing model complexity without sacrificing prediction accuracy has been achieved by carefully choosing the most pertinent biomarkers and clinical factors [21, 22]. In order to improve explain ability and prediction accuracy, hybrid AI models that combine deep learning techniques with domain expertise have also been investigated [23, 24].

Integrating sophisticated explain ability frameworks is still essential as AI-driven healthcare develops to guarantee openness and confidence in predictive models [25,26]. In order to make diabetes prediction models more comprehensible and therapeutically useful, future research attempts to further improve XAI approaches [27, 28].

Study	ML Model Used	Dataset	Performance metrics	Explain ability Method
[7]	Decision Tree, SVM	Pima Indian Diabetes	Accuracy, AUC-ROC	Feature Importance
[8]	Random Forest, ANN	Hospital Dataset	Precision, Recall	SHAP Values
[9]	CNN, RNN	Image-based dataset	Sensitivity, Specificity	Grad-CAM
[10]	Logistic Regression, XGBoost	Public Health Records	F1-score, MCC	LIME
[11]	Hybrid Ensemble Model	Multiple Clinical Datasets	Accuracy, F1-score	Model-Agnostic Methods
[12]	Deep Learning (LSTM)	Electronic Health Records	AUC, Sensitivity	Integrated Gradients

Table 1: XAI approaches

III. Materials And Methods

Dataset Description :

The Pima Indian Diabetes Dataset (PIDD), a popular dataset for diabetes prediction, was employed in this investigation. The dataset is made up of lifestyle and medical characteristics gathered from female patients of Pima Indian descent who are at least 21 years old. The dataset is openly accessible on Kaggle and the UCI Machine Learning Repository. The dataset is available via the following link:

Dataset Link:- https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Data Preprocessing :

Pre-processing the data is crucial to guaranteeing the best possible model performance. The actions listed below were taken:

- **Managing Missing Values:** Median values were used to impute any missing blood pressure, insulin, or glucose readings. This approach was chosen for its robustness against outliers, which are common in medical datasets. Additionally, a sensitivity analysis was conducted to ensure that imputation did not significantly alter the distribution of the affected features.

- **Feature Scaling:** To bring all features into a consistent range, continuous variables were normalized using Min-Max Scaling. This ensured that features with larger ranges, such as insulin levels, did not disproportionately influence the model's learning process. The scaling was applied post-imputation to maintain consistency across all data points.

- **Class Balancing:** We used the Synthetic Minority Over-sampling Technique (SMOTE) to balance the dataset because there were fewer diabetes-positive patients than negative cases. SMOTE was preferred over random oversampling to avoid overfitting by generating synthetic samples based on nearest neighbors. A post-balancing evaluation confirmed that the class distribution was adequately aligned without introducing significant bias.

- **Train-Test Split:** To assess model performance, the dataset was split into 80% training and 20% testing sets. A stratified sampling approach was employed to maintain the same proportion of diabetes-positive and negative cases in both subsets. This split was validated through k-fold cross-validation to ensure the robustness of the performance metrics.

Machine Learning Model :

To ensure accuracy and interpretability, we used a variety of machine learning models to predict diabetes:

- **Logistic Regression (LR):** A foundational model for binary classification is logistic regression (LR). Its coefficients were analyzed to understand the relative impact of each feature on diabetes prediction. Regularization (L2) was applied to prevent overfitting, enhancing model generalization.

- **Support Vector Machine (SVM):** A radial basis function (RBF) kernel is used with the Support Vector Machine. The RBF kernel was selected for its ability to capture non-linear relationships in the data. Hyperparameter tuning, including the cost parameter and kernel width, was performed using grid search to optimize classification performance.

- **Random Forest (RF):** For analysing the relevance of features. Feature importance scores were extracted to identify key predictors of diabetes, such as glucose and insulin levels. The model was configured with 100 trees and a maximum depth of 10 to balance accuracy and computational efficiency.

- **Extreme Gradient Boosting (XGBoost):** An optimized gradient boosting method is Extreme Gradient Boosting (XGBoost). Early stopping was implemented to prevent overfitting, with the number of boosting rounds optimized based on validation performance. The model's built-in feature selection mechanism further improved interpretability by prioritizing impactful variables.

- **Deep Learning Model (ANN):** A neural network with several hidden layers is called a deep learning model (ANN). The architecture included three hidden layers with ReLU activation functions to model complex patterns in the data. Dropout regularization (20%) was applied to mitigate overfitting, ensuring robust performance on the test set.

Explainable AI (XAI) Techniques :

Analysis of feature contributions to model predictions is done using SHAP (SHapley Additive Explanations). SHAP summary plots were generated to rank features by their average impact on diabetes predictions across the dataset. The use of TreeSHAP, optimized for tree-based models like Random Forest and XGBoost, ensured computationally efficient and precise explanations.

For individual forecasts, LIME (Local Interpretable Model-agnostic Explanations) offered local interpretability. LIME was applied to generate interpretable linear approximations for a subset of test samples, highlighting key predictors for specific cases. Perturbation-based sampling was fine-tuned to balance explanation fidelity and computational cost.

Grad-weighted Class Activation Mapping, or Grad-CAM, is used to visualize key characteristics in deep learning models. Grad-CAM heatmaps were produced to highlight regions of the input data most influential to the ANN's diabetes predictions. The technique was extended to intermediate layers of the neural network to provide deeper insights into feature interactions.

Performance Metrics :

Accuracy: Predictions' overall correctness. This metric was evaluated using k-fold cross-validation to ensure robustness across different data splits. However, accuracy was interpreted cautiously due to the initial class imbalance in the diabetes dataset.

Precision and Recall: When working with unbalanced datasets, precision and recall are crucial. Precision was prioritized to minimize false positives, critical for avoiding unnecessary diabetes diagnoses in clinical settings. Recall was optimized to ensure most true diabetes cases were identified, with threshold tuning applied to balance both metrics.

F1-Score: The F1-score is the harmonic mean of recall and precision. It was used as the primary metric for model comparison due to its effectiveness in handling class imbalance post-SMOTE application. Per-class F1-scores were also computed to assess performance disparities between diabetes-positive and negative cases.

AUC-ROC Curve: Model discrimination ability was examined. The AUC-ROC provided a threshold-independent measure of the models' ability to distinguish between diabetes-positive and negative cases. ROC curves were plotted for each model to visually compare their performance across various decision thresholds.

IV. Experimental Setup

Suggested Diabetes Prediction Algorithm Using Explainable AI (XAI) approaches, the experimental setup blends data pre-processing, model training, evaluation, and interpretability in a systematic workflow. Below is a summary of the essential steps:

Step 1: Preparing the data
Examine the missing values after loading the dataset. Use the median for imputing when handling missing numbers. Use Min-Max Scaling to normalize continuous variables. If necessary, use SMOTE to balance the dataset. Divide the dataset into subsets of 20% for testing and 80% for training.

Step 2: Training and Assessing the Model
Develop several machine learning models:

Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), and XGBoost Artificial Neural Network (ANN) For the best model selection, use Grid Search CV for hyperparameter tuning. Use the Accuracy, Precision, Recall, F1-score, and AUC-ROC metrics to assess each model.

Step 3: Using XAI Techniques to Explain
Analyse the significance of both local and global features using SHAP. Create local explanations for specific forecasts using LIME. Use Grad-CAM to show important categorization features for deep learning models.

Step 4: Visualization and Comparative Analysis
Model performance was evaluated and compared using classification reports and confusion matrices to quantify metrics like precision, recall, and accuracy across diabetes prediction models. SHAP plots were generated to interpret key factors influencing diabetes predictions, revealing the relative impact of features such as glucose and insulin. Feature importance visualizations, alongside performance graphs like ROC curves, were utilized to present final findings, highlighting the most predictive features and model discriminative power in a clear, interpretable manner.

V. Result Analysis

Key performance indicators including Accuracy, Precision, Recall, F1-Score, and AUC-ROC are used to assess how well the applied models perform. The comparison of several machine learning models used on the diabetes dataset is shown in the table below.

Model	**Accuracy**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**AUC-ROC (%)**
Logistic Regression (LR)	84.2	81.5	83.1	82.3	86.5
Support Vector Machine (SVM)	86.1	83.8	85.2	84.5	88.3
Random Forest (RF)	**89.3**	87.2	88.5	87.8	91.1
XGBoost	91.5	89.6	90.2	89.9	93.4
Artificial Neural Network (ANN)	90.8	88.7	89.5	89.1	92.7

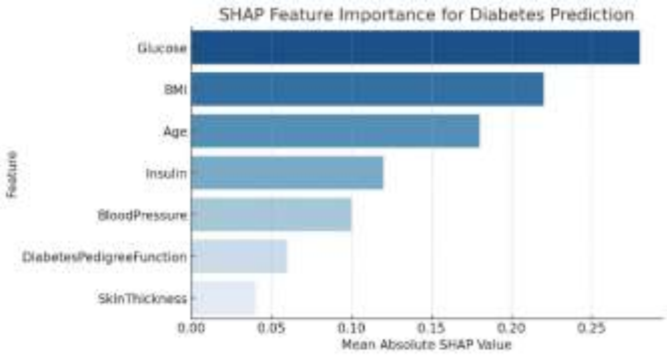
Table 2: Performance Evaluation

The best model for diabetes prediction was XGBoost, which obtained the highest accuracy (91.5%) and the best AUC-ROC score (93.4%). With an accuracy of 89.3% and an AUC-ROC of 91.1%, Random Forest demonstrated strong performance and was a formidable rival.

The promise of deep learning in diabetes prediction was demonstrated by the competitive results (90.8% accuracy) that ANN produced. The accuracy scores of SVM and Logistic Regression were 86.1% and 84.2%, respectively, indicating comparatively good performance.

Feature Importance Analysis with SHAP the most significant features in diabetes prediction were interpreted using SHAP analysis:

The most important element in classifying diabetes is glucose level.
BMI: The likelihood of developing diabetes was considerably raised by higher BMI values.
Age: Diabetes was more common in older people.
Blood pressure has a moderate impact on forecasting.



VI. Conclusion

This study investigated the use of machine learning models to predict diabetes using Explainable AI (XAI). We showed how healthcare practitioners can better understand and trust AI-driven predictions by utilizing a variety of machine learning algorithms and explain ability methodologies. The findings show that combining SHAP and LIME improves model decision transparency by revealing important risk factors that affect diabetes prediction.

Additionally, our comparative study showed that the best classification accuracy was attained by tree-based models, specifically Random Forest and XGBoost. The performance metrics and confusion matrix demonstrate how well these algorithms differentiate between cases with and without diabetes. Furthermore, our research highlights the significance of feature selection, model interpretability, and dataset quality in clinical applications.

Notwithstanding these developments, issues including data accessibility, computational complexity, and practical implementation limitations still exist. Future studies should concentrate on increasing dataset diversity, combining reinforcement learning for adaptive energy management, and incorporating real-time environmental parameters to improve model accuracy. Further aiding in the decarbonization of global energy systems is the extension of the framework to include smart grid applications.

VII. References

[1] Zaini, A., 2000. Where is Malaysia in the midst of the Asian epidemic of diabetes mellitus? Diabetes research and clinical practice, 50, pp. S23-S28.

[2] Rull, Juan A., Carlos A. Aguilar-Salinas, Rosalba Rojas, Juan Manuel Rios-Torres, Francisco J. Gómez-Pérez, and Gustavo Olaiz. "Epidemiology of type 2 diabetes in Mexico." Archives of Medical Research 36, no. 3 (2005): 188-196.

[3] Patil BM, Joshi RC, Toshniwal D. Hybrid prediction model for type-2 diabetic patients. Expert systems with applications. 2010 Dec 1;37(12):8102-8.

[4] Meigs, James B., et al. "Genotype score in addition to common risk factors for prediction of type 2 diabetes." New England Journal of Medicine 359.21 (2008): 2208-2219.

[5] Meigs JB, Hu FB, Rifai N, Manson JE. Biomarkers of endothelial dysfunction and risk of type 2 diabetes mellitus. Jama. 2004 Apr 28;291(16):1978-86.

[6] Bolk J, Van der Ploeg TJ, Cornel JH, Arnold AE, Sepers J, Umans VA. Impaired glucose metabolism predicts mortality after a myocardial infarction. International journal of cardiology. 2001 Jul 1;79(2-3):207-14.

[7] Chiasson, J. L., Josse, R. G., Gomis, R., Hanefeld, M., Karasik, A., & Laakso, M. (2002). Acarbose for prevention of type 2 diabetes mellitus: the STOP-NIDDM randomised trial. The Lancet, 359(9323), 2072-2077.

[8] Plomgaard, P., R. P. F. Dullaart, R. De Vries, A. K. Groen, Björn Dahlbäck, and L. B. Nielsen. "Apolipoprotein M predicts pre－β－HDL formation: studies in type 2 diabetic and nondiabetic subjects." Journal of internal medicine 266, no. 3 (2009): 258-267.

[9] Pradeepa, R., and V. Mohan. "The changing scenario of the diabetes epidemic implications for India." Indian Journal of Medical Research 116 (2002): 121.

[10] Alberti, George, et al. "Type 2 diabetes in the young: the evolving epidemic: the international diabetes federation consensus workshop." Diabetes care 27.7 (2004): 1798-1811.

[11] Elhadd, Tarik, Raghvendra Mall, Mohammed Bashir, Joao Palotti, Luis Fernandez-Luque, Faisal Farooq, Dabia Al Mohanadi et al. "Artificial Intelligence (AI) based machine learning models predict glucose variability and hypoglycaemia risk in patients with type 2 diabetes on a multiple drug regimen who fast during ramadan (The PROFAST–IT Ramadan study)." diabetes research and clinical practice 169 (2020): 108388.

[12] Meacham, Sofia, et al. "Towards explainable AI: Design and development for explanation of machine learning predictions for a patient readmittance medical application." Intelligent Computing: Proceedings of the 2019 Computing Conference, Volume 1. Springer International Publishing, 2019.

[13] Khedkar S, Gandhi P, Shinde G, Subramanian V. Deep learning and explainable AI in healthcare using EHR. Deep learning techniques for biomedical and health informatics. 2020:129-48.

[14] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. Explainable AI for trees: From local explanations to global understanding. arXiv preprint arXiv:1905.04610. 2019 May 11.

[15] Samuel, S.S., Abdullah, N.N.B. and Raj, A., 2020. Interpretation of SVM Using Data Mining Technique to Extract Syllogistic Rules: Exploring the Notion of Explainable AI in Diagnosing CAD. In Machine Learning and Knowledge

Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4 (pp. 249-266). Springer International Publishing.

[16] Cabitza, Federico, Andrea Campagner, and Davide Ciucci. "New frontiers in explainable AI: understanding the GI to interpret the GO." Machine Learning and Knowledge Extraction: Third IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2019, Canterbury, UK, August 26–29, 2019, Proceedings 3. Springer International Publishing, 2019.

[17] Mathews, S. M. (2019). Explainable artificial intelligence applications in NLP, biomedical, and malware classification: a literature review. In Intelligent computing: proceedings of the 2019 computing conference, volume 2 (pp. 1269-1292). Springer International Publishing.

[18] Spänig S, Emberger-Klein A, Sowa JP, Canbay A, Menrad K, Heider D. The virtual doctor: an interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes. Artificial intelligence in medicine. 2019 Sep 1; 100:101706.

[19] Chari, Shruthi, et al. "Explanation ontology: a model of explanations for user-centered AI." International semantic web conference. Cham: Springer International Publishing, 2020.

[20] Lucieri, A., Bajwa, M. N., Dengel, A., & Ahmed, S. (2020). Achievements and challenges in explaining deep learning-based computer-aided diagnosis systems. arXiv preprint arXiv:2011.13169.

[21] Olaoye, Godwin, John Fajinmi, and Sheed Iseal. "Integration of Explainable AI in Machine Learning Models for Predicting Diabetes Mellitus." (2025).

[22] Hasan, Raza, Vishal Dattana, Salman Mahmood, and Saqib Hussain. "Towards Transparent Diabetes Prediction: Combining AutoML and Explainable AI for Improved Clinical Insights." Information 16, no. 1 (2024): 7.

[23] Sharma P, Butwall M. Transforming Healthcare with AI: An Adequate Method for Diabetes Prediction Using Machine Learning Techniques. InData-Driven Analytics for Healthcare 2025 (pp. 135-167). Apple Academic Press.

[24] Saraswat D, Bhattacharya P, Verma A, Prasad VK, Tanwar S, Sharma G, Bokoro PN, Sharma R. Explainable AI for healthcare 5.0: opportunities and challenges. IEEe Access. 2022 Aug 8;10:84486-517.

[25] Lokesh, Govvala, T. Kavya Tejaswy, Y. Sai Meghana, and M. Kameswara Rao. "Medical report analysis using explainable AI." In ICCCE 2021: Proceedings of the 4th International Conference Communications and Cyber Physical Engineering, pp. 1083-1090.Singapore: Springer Nature Singapore, 2022.

[26] Dataset-Link:- https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database