Explainable and Fair AI-Based Models for Early Diabetes Risk Prediction: A Review

***Kavita Kavita[1], Gurbinder Singh Brar[2], Amandeep Singh[3] ,Kamalpreet Kaur [4]**

[1]Student ,Department of Computer Science and Engineering
Lovely Professional University, Phagwara, Punjab, India
[2]Professor ,Department of Computer Science and Engineering
Lovely Professional University, Phagwara, Punjab, India
[3]Associate Professor ,Department of Computer Science and Engineering
Lovely Professional University, Phagwara, Punjab, India
[4]Assistant Professor ,Department of Computer Science and Engineering
Lovely Professional University, Phagwara, Punjab, India
[1]*anko12kohil@gmail.com,
[2]maatibrar@gmail.com,
[3]amandeep.32664@lpu.co.in,
[4]Kamalpreet.cse98@gmail.com

ABSTRACT

It has been clearly observed that Type 2 Diabetes Mellitus, which is also known as T2DM, is rising tremendously on a global scale, which not only needs accuracy to detect at an early stage, but also requires sound ethical practices. As per traditional statistical methods, it seems difficult to handle the complicated and non-linear patterns that are in the large multimodal clinical datasets. In contrast to this, artificial intelligence models are better at achieving predictive performance, but it lacks in terms of fairness and explainability. It is crystal clear that in the medical field, it becomes necessary to gain the trust of either the patient or the doctors. So, in this review, I have reviewed some research papers from 2015 to 2025 that include separately explainable AI (XAI) and fairness-based AI models for early Diabetes risk prediction. Yet both explainable AI and fairness AI-based models have not used till now, and this review paper consists of the combination of both AI models to elevate the interpretability and equality in terms of genders and ethnicity. It has already seen that some basic machine learning algorithms, such as Random Forest (RF) and Cat-boost, utilising Electronic Health Record (EHR) data, are achieving the predictive accuracy of 0.99. Along with that, if 38% of SHAP and 26% of LIME usage can increase clinical trust. In this paper, the work of the mitigation framework in the Smart User Interface is highlighted, which successfully diminishes the Equal Opportunity Difference for different age and BMI up to 18 percentage points. Moreover, this review concludes by recommending XAI-fairness frameworks to make sure the AI systems predict accurately and fairly.

Keywords: - Explainable Artificial Intelligence (XAI), Fair Artificial Intelligence, SHAP, LIME, Early Diabetes Risk Prediction, Type 2 Diabetes Mellitus (T2DM), Clinical Decision Support Systems

I. INTRODUCTION

1.1 Predictive role of AI

Nowadays, Diabetes is growing rapidly. This disease is affecting millions of people all over the world. Behind the multi-morbid complications, diabetes mellitus (DM) is one of the serious reasons. Furthermore, this costs a lot to human to decrease its side effects. AI is widely used worldwide, especially in health sectors, as the number of diabetic patients is increasing because it was not predicted early, with the reason why, and some of them were incorrectly biased in terms of ethnicity and gender (especially for pregnant women). In this modern era, AI can handle complicated, large datasets, and at times it has improved its decision-making abilities as well [1]. These datasets contain non-linear relationships among the factors of this disease, including age, genes and lifestyle.

A. Agnihotri et al. (eds.), *Proceedings of the Conference on Bridging Engineering Disciplines with AI and Machine Learning (BEDAIML 2026)*, Advances in Intelligent Systems Research 209,
https://doi.org/10.2991/978-94-6239-697-5_36

1.2 Accuracy, Explainability, and Fairness

The balance of all three parameters while predicting the early risk of diabetes becomes crucial. Implementing AI into clinical decision support systems will enable models to understand the concept behind XAI and fair AI. It is also true that the black box shows more accuracy, but it is also clear that it lacks transparency. It is understood that some models use retrospective EHR data, which learns from the patient's history records. Still, these records are biased in terms of the fact that different people had different treatments [6]. When AI depends on biased data, it simply means that there will be unfair differences in healthcare outcomes. So, the demand for accuracy, fairness and interpretability is ethically good for everyone because without accuracy, it creates trust issues.

This was the introduction part, and the upcoming sections will be: II. Related works: methodologies and ethical foundations, III. Comparative analysis, IV. Conclusion

1.3 Scope of Review

This review paper covers some research papers that mainly focused on the work of XAI and fair AI models from the past ten years, from 2015 to 2025 [2]. [2] To find out the knowledge gaps and ongoing trends towards its clinical implementation, it organises the literature on three cores, and those are methodology prediction, XAI techniques and fairness framework.

2. RELATED WORKS: METHODOLOGIES AND ETHICAL FOUNDATIONS

2.1 Predictive Modelling

Through review analysis, it is sceptical that the dominance of some popular machine learning algorithms like random forest (RF) and extreme gradient boosting (XGBoost) is far better for their performance and efficiency in managing the EHR data. The measuring model can predict things very well, and that is none other than the Area Under the Curve (AUC). It mostly gets higher scores between 0.81 and 0.99 [5]. By analysing the review, one vital finding is that the shortage of external validation and the lack of model testing determine predicted probability accuracy.

2.2 Explainable AI for Transparent Understanding

Explanation methods can work in any kind of tools, where SHAP and LIME are two of them [1]. Prevalence of SHAP (38%) and LIME (26%)

- **SHAP (SHapley Additive exPlanations):** In most cases, for explanations, SHAP (SHapley Additive exPlanations)becomes the first choice to consider because of its feature attributions that perfectly explain the working of each input variable to predict the outcome.

- **LIME (Local Interpretable Model-agnostic Explanations):** Different patients are treated differently, and even specific medicines are given to the concerned patient as per the stage of their disease levels. Here, LIME is suitable for such cases, where individualized explanations are needed. For the patient, it is essential for clinical supervision.

2.3 The Algorithmic Fairness Deficit

It cannot be denied that equality on ethnicity and gender has become a necessity, but it is also true that in the field of fairness-based AI models, somehow, there is still a shortage of research for the combination of fairness models with XAI. Fairness includes large groups of genders and subgroups like pregnant women [12]. In fairness, only similar individuals are treated similarly [3].

The given important indicators are required for an equitable forecast [4], [5]:

- **Demographic Parity (DP):** It is based on various genders, races, and age and it should have the same proportion of people predicted as "high risk" by the model. It is not looking at the correct or wrong prediction because it treats each group equally, and in positive predictions.

- **Equal Opportunity Difference (EOD):** It is sensitive across different demographic groups. It works on identifying only actual high-risk patients in each group. It also makes sure not to overlook the minority groups.

3. COMPARATIVE ANALYSIS

3.1 Literature review of Base Research

Table 1 provides an overview of earlier studies done by different researchers on fair and explainable AI for diabetes prediction.

3.2 Comparison of different model performance

Table 2 shows performance and efficient data of high-performing fairness and XAI models.

3.3 Framework Visualisation

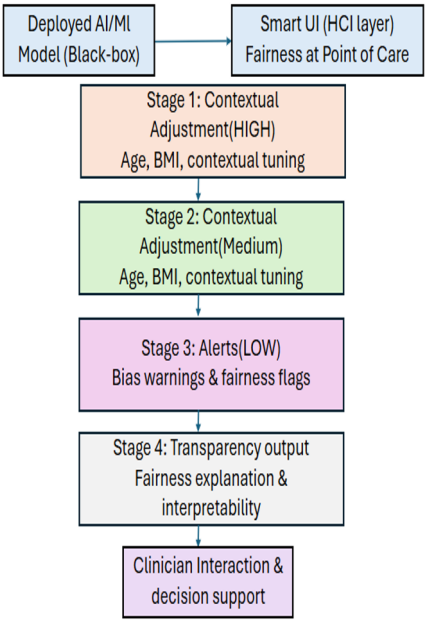

Fig 1: Smart UI–Based Fairness Framework

One good solution is fairness implementation through the Smart UI Framework in deployed systems [6]. This method does not need black box model adjustments. It simply uses the Human-Computer Interaction (HCI) layer to implement fairness as shown in Fig. 1.

- **Concept:** The main concept is the layered defense mechanism, which is used for preventing algorithmic bias at the point when the care is demonstrated.
- **Representation:** Four stages of sequential or parallel interventions are: 1) Contextual Adjustment Tools, which are of high-risk. 2) Dynamic Risk Visualizations, which are medium-risk. 3) Alert Solutions, which are of low-risk.4) Additionally, the system depicts how transparent the result of the fairness model is [10].
- **Function:** With the use of this useful framework, doctors can easily interact with this model to understand its working and not only this, but also, they can make some necessary adjustments regarding the age and the BMI to lessen the unfairness for themselves.

3.4 Explainable AI Model along with a Chatbot Assistant

- **Concept:** Mostly, it is seen that **it** becomes difficult to understand for most of the patients, as SHAP values are way too technical for them. So, combining both AI and chatbot[7] with LLM helps to convert the complex explanation into the personalized one, understandable and actionable advice for patients as shown in Fig. 2.

Table 1: Summary and comparison to make AI models for diabetes prediction

Authors	Year	Technique Used	Strengths	Errors/Limitations
[6] Tavangar et al.	2025	Smart UI (HCI Layer), Random Forest (RF)	Reduce fairness gaps between the groups of age and BMI with the help of general fairness tools	It needs clinical validation from several external sites
[8] Mohsen et al.	2023	Scoping Review (40 studies), ML/Multimodal	Use multiple data together to perform better for the prediction of type 2 diabetes by the help of careful analysis of different models	Did not test their models on external data to predict real outcomes
[3]Chinta et al.	2025	Review: Bias Taxonomy (Pre-, In-, Post-processing)	Finds effective way to make fair biased AI models in healthcare	It has shortage of quantitative analysis for diabetes performance and major focus is on general healthcare
[7]Maimaitijiang et al.	2025	CatBoost, SHAP, LLM Chatbot, SMOTE	It has AUC of 0.99. Here XAI is done with LLM chatbots to advice the patient individually.	Performance was not prioritized for multi-metric fairness
[9] Gosak et al.	2022	Systematic Review (11 articles), ML/NN	Patients with multiple health concerns were predicted correctly for diabetes-related complications	Results were fluctuating and procedure were also not well reported (e.g., they claim AUC=0.99)

[4] Liu et al.	2025	Scoping Review: Clinical AI Fairness	Found two main problems: first, less study in fairness in diabetes and second is not proper examination of the type of biased data	Whole attention was on fairness for groups but often ignores fairness for individual patients
[5]Nazirun et al.	2024	Systematic Review, ML/DL	Created explainable model for doctors and random forest worked best	Limited participants and did not test models across different ethnic groups
[2] Ueda et al.	2024	Review: Fairness, Bias, and Recommendations (FAIR Statement)	Explained different biased AI and suggested fairer and more transparent AI	Gave general suggestions on using AI in medical imaging and radiology
[10] Alkhanbouli et al.	2025	Systematic Review: XAI Methods (SHAP, LIME)	Did not draw attention on diverse dataset and used SHAP and LIME tools for the purpose of explanation	Did not cover LLM as explanation method
[11] Vinh & Byeon	2024	Review: XAI in Diabetes Diagnosis	Encouraging XAI to promote health apps and telemedicine	Dependency on limited number of datasets might not work for everyone

Table 2: Analyse the comparison between fairness and explainable AI models

Model	**Accuracy (%)**	**F1-Score**	**AUC**	**Key Observation**
[7]CatBoost (with SHAP/LLM)	99.2	0.99	0.99	It has built-in explainability, where logistic regression performed AUC 0.918
[6]RF/Smart UI (Fairness Adjusted)	91.83 (Adjusted)	0.8637 (Adjusted)	N/A	A sharp increase in fairness model with improved accuracy: for age EOD declines from 0.35 to 0.25 and EOD decreased from 0.56 to 0.38 for BMI
[9] GBoost (Ensemble for Complications)	N/A	N/A	0.847±0.081	Best performing models for the patients who had multiple health issues with serious diabetic complications
[6] SVM (Gestational Diabetes/SMOTE)	75.1	0.89	0.79	SMOTE is used to look at the imbalanced data for minority class pregnant women

- **Representation:** Firstly, the user enters the input. Secondly, ML predictions will take place, where SHAP XAI will work. Thirdly, LLM (BioMistral-7B) will be applied. Finally, personalised recommendations will be offered to the user via a chatbot assistant.

- **Function:** It interprets complex AI outputs in terms of clinically meaningful way. Furthermore, it communicates by converting technical data into natural language so that the patient can directly understand the advice given by the chatbot, or doctors can use this chatbot assistant for further treatment of the concerned patient.

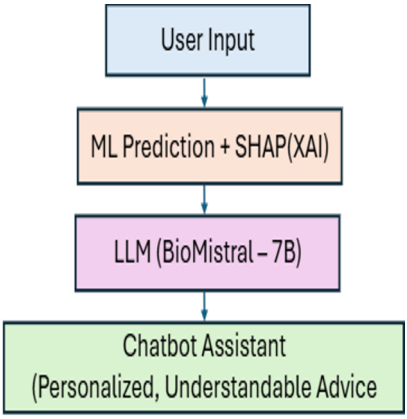

Fig 2: Explainable AI with Chatbot Assistant

4. CONCLUSION

Especially, tree-based machine learning algorithms that are trained on EHR data show excellent accuracy to predict the diabetes risk factors at an early stage. By SHAP and LIME explainability tools, it encourages crucial clinician trust. Most of the research has not even used a fair AI model across a variety of groups in terms of ethnicity and age. This simply shows that even if it is accurate, it is still biased. The latest studies were reporting mainly overall accuracy rather than calibration, which means ignoring whether the predicted risk matches with actual outcomes or not. Testing was done on a few different populations, and a lack of reliability was ensured to ensure predictions are fair across groups by using EOD or demographic parity.

REFERENCES

1. T. Vinh and H. Byeon, “Towards Transparent Diabetes Prediction: Unveiling the Factors with Explainable AI,” International Journal of Engineering Trends and Technology, vol. 72, no. 5, pp. 26–35, 2024.
2. D. Ueda, T. Kakinuma, et al., “Fairness of artificial intelligence in healthcare: review and recommendations,” Japanese Journal of Radiology, 2024.
3. S. Chinta, Z. Wang, et al., “AI-driven healthcare: Fairness in AI healthcare: A survey,” PLOS Digital Health, 2025.
4. M. Liu, Y. Ning, et al., “A scoping review and evidence gap analysis of clinical AI fairness,” npj Digital Medicine, vol. 8, no. 1, 2025.
5. Nazirun et al., “Systematic Review on Explainable AI in Diabetes,” Medical Informatics, 2024.
6. A. Tavangar, Z. Arefzadeh, and A. Asghari, “Enhancing Fairness in Diabetes Prediction Systems through Smart User Interface Design,” 2025.
7. E. Maimaitijiang, M. Aihaiti, and Y. Mamatjan, “An Explainable AI Framework for Online Diabetes Risk Prediction with a Personalised Chatbot Assistant,” Electronics (Switzerland), vol. 14, no. 18, 2025.

8. F. Mohsen, H. Al-Absi, et al., “A scoping review of artificial intelligence-based methods for diabetes risk prediction,” npj Digital Medicine, 2023.
9. L. Gosak, K. Martinovic, et al., “Artificial intelligence-based prediction models for individuals at risk of multiple diabetic complications: A systematic review of the literature,” Journal of Nursing Management, 2022.
10. R. Alkhanbouli, H. Matar Abdulla Almadhaani, et al., “The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions,” BMC Medical Informatics and Decision Making, 2025.
11. T. Vinh and H. Byeon, “Towards Transparent Diabetes Prediction: Unveiling the Factors with Explainable AI,” International Journal of Engineering Trends and Technology, vol. 72, no. 5, pp. 26–35, 2024.