International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V72I5P103

Volume 72 Issue 5, 26-35, May 2024

Original Article

Towards Transparent Diabetes Prediction: Unveiling the Factors with Explainable AI

Tran Quang Vinh[1], Haewon Byeon[2*]

[1,2]*Department of Digital Anti-aging Healthcare (BK-21), Inje University Medical Big Data Research Center, Inje University, South Korea.*

[2]*Corresponding Author : bhwpuma@naver.com*

Received: 11 October 2023 Revised: 23 March 2024 Accepted: 27 April 2024 Published: 26 May 2024

Abstract *- Automatic diabetes prediction using machine learning and Explainable AI (XAI) has emerged as a promising approach for early detection and improved patient outcomes. This study investigates the current landscape of XAI research in diabetes diagnosis. The paper examines the transition from basic machine learning algorithms to complex deep learning models, emphasizing the importance of data quality and data preprocessing for accurate and interpretable results, particularly when dealing with tabular data from medical records. The integration of XAI techniques allows us to understand how these models arrive at their predictions, fostering trust and transparency. Despite these advancements, limitations remain. The generalizability of findings based on limited datasets needs further exploration through studies using more diverse data sources and real-world clinical settings. Additionally, the potential of XAI in diabetes management can be further enhanced by integrating these models with mobile applications and Internet of Things (IoT) sensor technology, paving the way for personalized and continuous monitoring. In conclusion, XAI research in diabetes prediction holds immense potential for improving healthcare delivery. By addressing current limitations and exploring new avenues of research, XAI can empower healthcare professionals and patients in the fight against diabetes.*

Keywords *- Diabetes, Explainable Artificial Intelligence, LIME, SHAP, Deep Learning.*

1. Introduction

In simple terms, Diabetes Mellitus (DM) is a group of conditions affecting how the body regulates blood sugar. These chronic disorders are mainly characterized by consistently high blood sugar levels [1]. Often, the root cause lies in problems with insulin production, insulin effectiveness, or both [1]. According to the World Health Organization (WHO), diabetes has become a much more significant cause of death since 2000, with a documented 70% increase in its contribution to mortality [2]. In 2019 alone, diabetes and related kidney complications are estimated to have caused two million deaths [3]. While research on diabetes continues, with scientists exploring better treatments and even potential cures, there is currently no universally accepted cure for the disease. Diabetes comes in two main forms: Gestational Diabetes (GD), type 1 (T1DM), and type 2 (T2DM). While all three raise blood sugar, their causes differ. T1DM is an autoimmune disease where the body attacks insulin-producing cells. It typically strikes in childhood [4] and requires lifelong insulin injections. In contrast, T2DM arises from insulin resistance or insufficient production. Often developing in adults [5], it can be managed with diet, exercise, and sometimes medication. Scientists emphasize that physical activity, along with medication and dietary changes, is a cornerstone of preventing and managing diabetes [6]. However, early detection remains crucial in today's healthcare landscape. Identifying diabetes early can not only prevent complications but also reduce the risk of developing other chronic diseases like kidney problems, heart attacks, and strokes. Given the large number of potential patients, efficiently using resources to predict the disease becomes essential. This is where various prediction and classification tools developed by researchers come into play.

The COVID-19 pandemic significantly accelerated the adoption of telemedicine and telehealth technologies [7]. This growth, coupled with advancements in patient empowerment tools like Artificial Intelligence [8] (AI), Electronic Health Records (EHRs), mobiles, and wearable devices [9], has fostered a trend towards self-management and self-diagnosis [10], [11], particularly among diabetes patients. Patient data collected by wearable devices or EHRs can be leveraged with promising tools such as Deep Learning (DL) or Machine Learning (ML) models to enhance early diabetes detection [9], [12], provide individualized medical assistance, and support advanced analytics [13]. Despite their potential, applying ML and DL models in clinical settings presents unique challenges. A significant challenge lies in the opacity of these models.

The inner workings and how they arrive at decisions are often shrouded in a'black-box'. This lack of transparency can make it difficult for both patients and medical professionals to trust the recommendations provided fully. Furthermore, data quality, bias, and limited data size can all impact the accuracy and reliability of these models. To address these concerns, eXplainable AI (XAI) emerges as a critical tool [14].

XAI techniques are crucial for understanding how diabetes prediction models reach diagnoses. This transparency builds trust among patients and medical professionals, allowing them better to grasp the reasoning behind the output of the model. Moreover, XAI tools can help identify and mitigate potentially biased training data, promoting fairer healthcare decisions. As there is no comprehensive review on XAI application in diabetes diagnosis, this paper explores how XAI techniques can enhance these aspects, paving the way for responsible application in clinical practice and improved patient outcomes. Furthermore, we delve into potential future research directions within the domain of XAI for diabetes prediction, focusing on advancements in both explainability methods and the models themselves.

2. Materials and Methods

Our investigation into existing research XAI for diabetes diagnosis using ML and DL models relied on two main literature databases: PubMed and Google Scholar. Starting with PubMed, a vast repository of biomedical literature boasting over 36 million entries [15], we conducted a search using the keywords "Diabetes," "Explainable," "Artificial Intelligence.", and "AND" operator. This initial search yielded 200 articles. However, upon closer examination, only two studies directly connected to XAI methods in diagnosing diabetes with ML/DL models remained after excluding research on unrelated topics like foot ulcers, life satisfaction, and imaging-based AI diagnosis.

To gain a more comprehensive understanding, we replicated the search strategy using Google Scholar, another prominent academic search engine. By manually evaluating the retrieved results, we identified nine additional studies relevant to XAI applications in diabetes diagnosis with ML/DL models.

2.1. Categorization of Explainable Artificial Intelligence Methods

The field of XAI focuses on lifting the veil on how AI models arrive at their decisions. Various categorization schemes have been proposed to classify XAI methods, highlighting different aspects of their functionality. These XAI methods are not restricted to a single category. In fact, a particular method might fit into several categories depending on its unique features. Figure 1 illustrates the categorization of XAI methods.

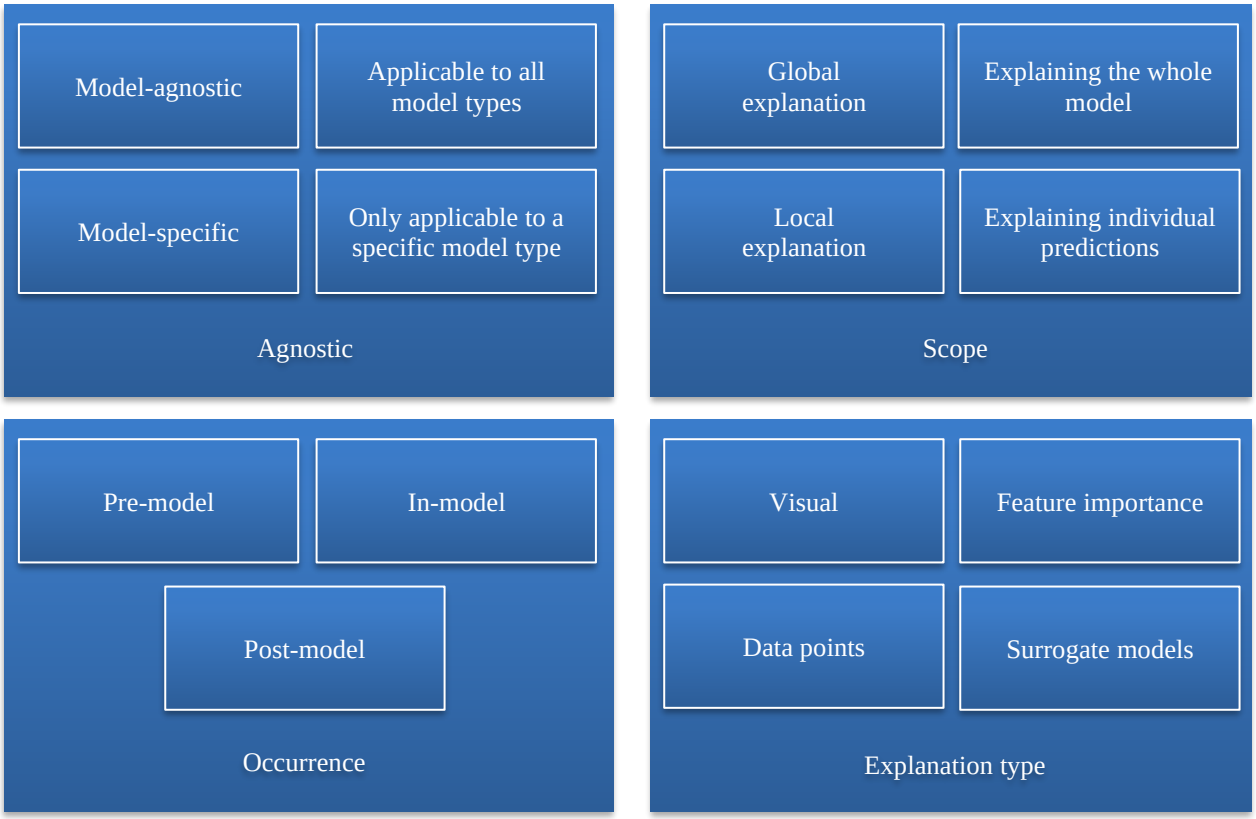

Fig. 1 XAI methods categorization

One distinction separates model-specific and model-agnostic methods. Methods designed for specific models take advantage of their internal workings and settings to explain their outputs. On the other hand, model-agnostic methods are versatile and can be applied to explain predictions from any model, regardless of its internal structure.

Local Interpretable Model-Agnostic Explanations [14] (LIME) and SHapley Additive exPlanations [16] (SHAP) are prominent examples of model-agnostic methods for explaining ML and DL predictions. Because they do not rely on the specific inner workings of a model, these methods can be used to explain predictions from a much broader range of models, making it easier to understand how these models arrive at their decisions. Another categorization focuses on the scope of explanation. Global methods provide insights into the overall behaviour of a model. They aim to explain how different features generally contribute to the model's performance across all predictions. Conversely, local methods delve deeper into explaining individual predictions. They pinpoint the specific features that were most influential in a particular model output. A third categorization considers the stage of model development when XAI methods are applied. Pre-model methods are data exploration tools used independently of a specific model.

They can help researchers understand the data itself before model building commences. Principal Component Analysis [17] is a noteworthy example of a pre-model XAI technique. In-model methods are integrated directly into the model architecture, fostering inherent interpretability by design. Lastly, post-model methods are employed after a model is trained. They examine the trained model's decision-making process to gain insights into what it learned from the data.Finally, XAI methods can be categorized based on their underlying functionalities. Surrogate methods utilize simpler models to mimic the behaviour of complex models. By interpreting the simpler model's decisions, we can gain insights into the complex behaviour of the model. For instance, decision trees [18], [19] are common types of surrogate methods. Visualization methods, on the other hand, employ visual representations like charts or graphs to explain specific aspects of a model's decision-making process. Understanding these different categorization schemes is crucial for researchers and practitioners to select appropriate XAI methods for their specific needs. The choice of method depends on the type of model being investigated, the desired level of explanation (global versus local), and the stage of model development. The following section will delve into the application of popular XAI methods in the context of diabetes diagnosis.

2.2. Commonly Used XAI Method in Diabetes Diagnosis

2.2.1. SHAP

SHAP [16] is a popular XAI technique used in diagnosing diabetes. It employs a unique approach based on game theory to assign importance to each feature in the data set, explaining how they influence the model's predictions [20]. According to the Shapley value, a feature's contribution to the model's output is calculated by considering its influence across all possible combinations of other features (as shown in Equation 1). This contribution is then weighted and summed up.

$$\phi_j(val) = \sum_{S \subset \{1,\dots,p\}\{j\}} \frac{|S|!\,(p-|S|-1)!}{p!}(val(S \cup \{j\} - val(S)) \quad (1)$$

The equation (Equation 1) calculates the Shapley value by considering all possible feature subsets (S). For each subset (S), the model's prediction is calculated with only those features included (denoted by x in the equation). The number of features in the model is represented by p. Equation 2 details how this prediction is then averaged across all possible feature combinations that are not included in the current subset (S):

$$val_x(S) = \int \hat{f}(x_1, \dots, x_p) d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X)) \quad (2)$$

2.2.2. LIME

LIME interprets individual predictions by creating a simpler, linear model function $g(z')$ around a specific prediction. This local model function approximates the behavior of the complex model $f(h_x(z'))$ in the vicinity of that prediction.

LIME uses interpretable data points (e.g., turning on/off words in text data or keeping/replacing superpixels in images) to explain the complex model's decision. The mapping function $x = h_x(x')$ converts these interpretable points back to the original input format.

LIME then optimizes a function L to find the best linear model function $f(h_x)$ that explains the complex model's behavior g for that specific prediction x based on the interpretable data point x'. LIME minimizes the following objective Equation 3:

$$\xi = \underset{g \in G}{argmin}\, L(f, g, \pi_{x'}) + \Omega(g) \quad (3)$$

2.2.3. Feature Importance with ELI5

ELI5 (Explain Like I'm 5) is a Python library that tackles the challenge of interpretability in artificial intelligence models. It provides valuable information such as feature importance rankings and model weights. ELI5 reveals which of these features were most crucial for the model's decision. Additionally, ELI5 might utilize permutation importance, a technique that shuffles a single feature's values and observes the impact on the model's performance. A significant performance drop suggests that the shuffled feature played a vital role in the original predictions. Figure 2 (by Özkur et al., 2020) [21] likely offers a schematic diagram illustrating ELI5's inner workings. Similarly, Figure 3 (from a separate study [22]) might showcase how ELI5 can be used to examine feature importance and weights in a diabetes classification task.

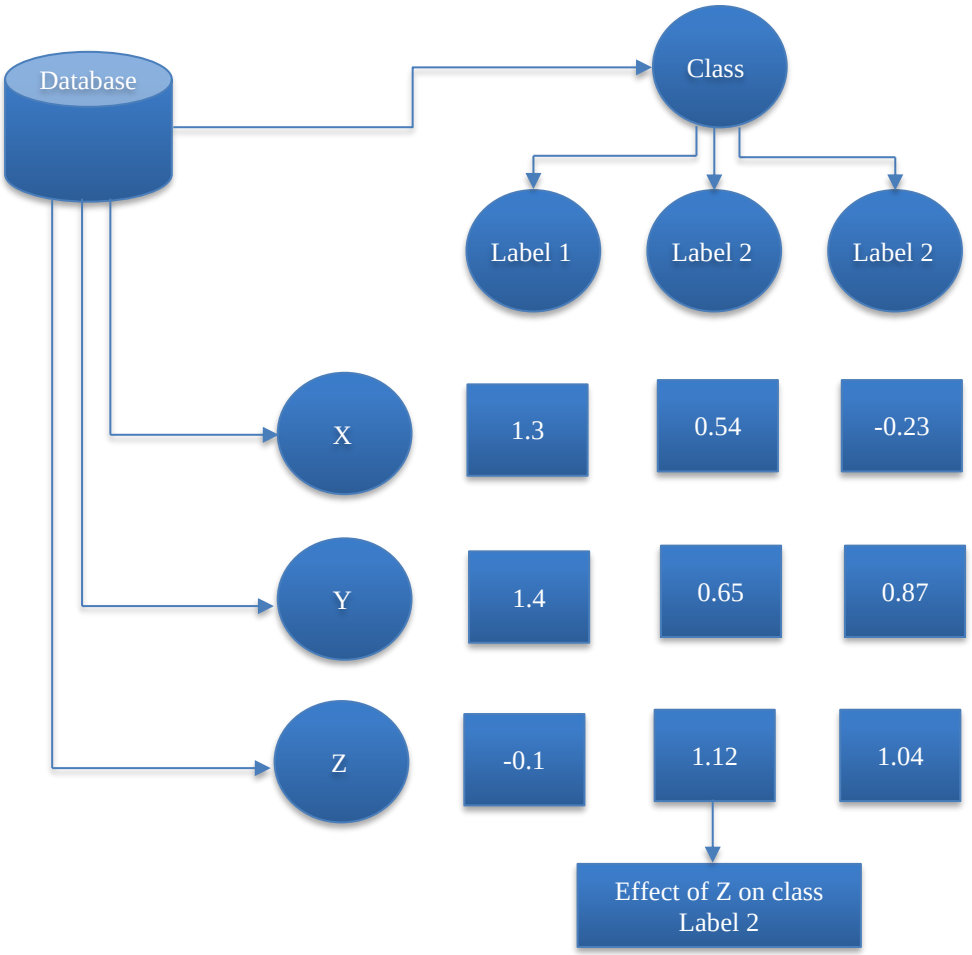

Fig. 2 Schematic diagram of ELI5 [21]

y=Diabetic (probability **0.518**) top features

Contribution[?]	Feature
+0.346	<BIAS>
+0.126	Age
+0.104	Glucose level
+0.009	BP
-0.010	BMI
-0.011	Thickness of Skin
-0.017	Diabetes pedigree function
-0.029	Body Insulin

Fig. 3 Schematic diagram of XGBoost

2.2.4. Quantum lattice (QLattice)

Broløs et al. (2021) introduced the QLattice, a supervised Machine Learning Algorithm inspired by Richard Feynman's Path integral formulation [23]. Unlike conventional black-box models, the QLattice prioritizes interpretability in its predictions. This is achieved by employing symbolic regression, a technique that seeks to recover a mathematical formula explicitly outlining the relationships between features and the target variable. The core concept underlying the QLattice leverages Feynman Path Integrals. This theoretical framework allo`ws for the exploration of numerous potential paths (models) to solve a given problem. Similarly, the QLattice explores a variety of candidate models represented by QGraphs (Figure 4).

These QGraphs define the connections between features and the target variable through mathematical interactions. The QLattice iteratively evaluates these QGraphs based on their fit to the provided data. By selecting the best-performing models, the QLattice refines its search, converging towards an optimal model that effectively captures the underlying relationships.

2.2.5. Anchor

Anchor is a model-agnostic system introduced in 2018 by Ribeiro et al. [24] that explains the behavior of complex models with high-precision rules, representing local, “sufficient” conditions for predictions.

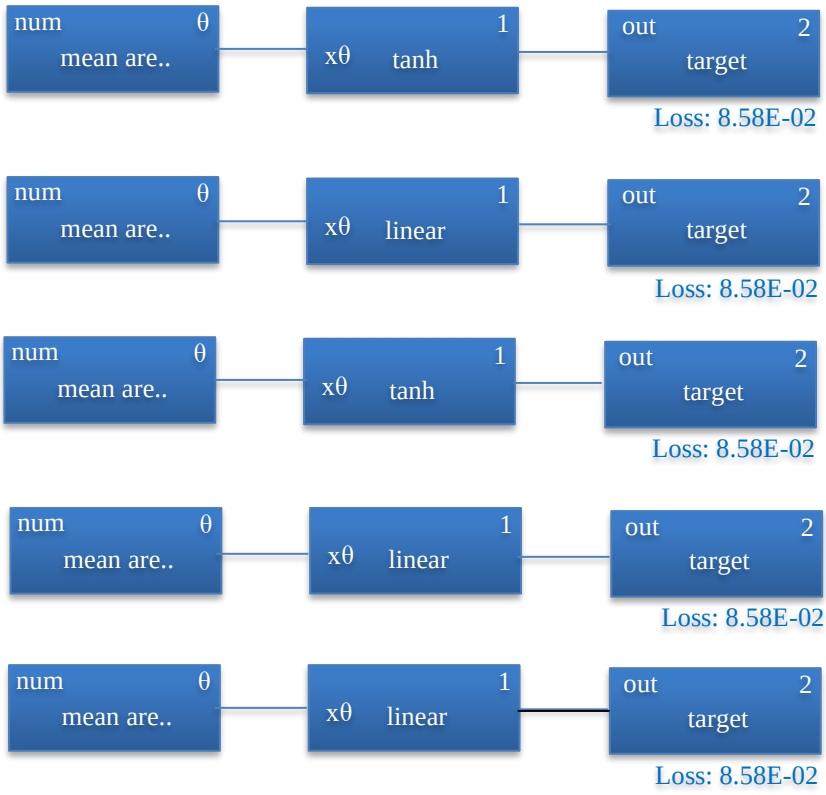

Fig. 4 Qgraph

Anchors function by identifying localized, "sufficient" conditions within the data that lead to specific predictions. Unlike the complex inner workings of the model itself, anchors are expressed as simple rules, akin to spotlights illuminating the key data points driving a particular prediction. This clarity empowers users to gain insights into the model's decision-making process.

We consider a black-box model f that maps an instance x from an input space X to a prediction y in the output space Y. Local model-agnostic interpretability focuses on explaining a specific prediction f(x) for a particular instance x. Here, $D(z|A)$denotes the conditional distribution of an instance z given a set of feature predicates A that define the anchor rule.

An anchor A is deemed sufficient if, for instances where the anchor holds $A(x) = 1$, the model's prediction is likely to remain consistent across samples drawn from the conditional distribution $D(z|A)$. Mathematically, this is expressed as Equation 4:

Where:
$\mathbb{E}$ denotes the expectation operator
$\boldsymbol{\tau}$ represents the desired precision threshold (typically set to a high value, e.g., 0.95)
$[\mathbb{1}_{f(x)=f(z)}]$ is an indicator function that equals $\mathbb{1}$ if f(x) and f(z) are the same prediction, and 0 otherwise

In essence, an anchor guarantees that changes to irrelevant features (those not included in the anchor rule A) are unlikely to alter the model's prediction for instances where the anchor applies. This focus on local fidelity ensures high-precision explanations.

2.3. Data Collection in Reviewed Studies

Tabular data, organized in tables and databases, is the backbone of countless applications. It is the most common data format, offering a wealth of information for analysis and prediction. The recent surge in advancements within biotechnology and health sciences has significantly increased the production of tabular data [25]. This includes genetic data and clinical information on diabetes patients stored in a tabular format within EHRs, clinical laboratory results, and data collected from wearable devices. Given its prevalence, this study will focus specifically on analyzing tabular data.

$$\mathbb{E}_{D(z|A)}\left[\mathbb{1}_{f(x)=f(z)}\right] \geq \tau, A(x) = 1 \quad (4)$$

Many researchers have used the PIMA Indian Diabetes (PID) dataset [26] along with various Machine Learning models for their studies. This dataset is known for its accuracy and privacy protection, and it is publicly available from the University of California, Irvine [26]. The PID dataset includes 768 data points, each containing information on eight different characteristics (detailed in Table 1).

Notably, the dataset focuses on a binary outcome: whether a patient has diabetes or not. '0' represents no diabetes, while '1' signifies the presence of the disease. In this study, we aim to compare and analyze these previous research efforts that have utilized the PID dataset. Researchers have utilized large datasets to study various medical conditions. One such dataset is MIMIC-III (Medical Information Mart for Intensive Care), developed by MIT, which contains detailed information on over 53,000 patients admitted to intensive care units [27]. This rich resource offers a wide range of data points, including demographics, vital signs, medications, and lab tests (over 4,700 measurements and 390 different tests). However, MIMIC-III is not specifically designed for diagnosing specific illnesses like GD.

Table 1. Attributes utilized in the PIMA dataset

Predictor Variables	Description
Pregnancies	Number of times pregnant
Glucose	Glucose levels during oral glucose tolerance test
Diastolic Blood Pressure	Diastolic blood pressure (mmHg)
Triceps Skinfold Thickness	Triceps skinfold thickness (mm)
Insulin	Hourly serum insulin levels (μIU/ml)
BMI	Body Mass Index (calculated from weight and height)
Diabetes Pedigree Function	A function describing diabetes family history
Age	Age in years

Table 2. Recent research on explainable AI for diabetes diagnosis

Article	Dataset	Models	XAI method	F1 Score	Accuracy	Other metrics
Guha et al. (2020) [18]	ESDRPD dataset (520 patients)	Random Forest	SHAP/LIME/ ELI5	0.95	95%	Precision: 0.95 Recall: 0.94 AUC: 0.98
Vakil et al. (2021) [33]	ESDRPD dataset (520 patients)	Random Forest	SHAP	0.99	99.0%	Recall: 0.99 Precision: 0.99
Kibria et al. (2022) [39]	PIMA	Ensemble of XGBoost and Random Forest	LIME/SHAP	0.89	90.0%	Precision: 0.88 Recall: 0.89 AUC: 0.95
Du et al. (2022) [42]	Data from PEARS study [45]	SVM	SHAP	N/A	75.1%	AUC: 0.79 AUC-PR: 0.49
Joseph et al. (2022) [43]	PIMA dataset and ESDRPD dataset	Bayesian-Optimized TabNet	SHAP/LIME/ TabNet/ ELI5	0.88	92.2%	Precision: 0.86 Specificity: 0.95
Vishwarupe et al. (2022) [22]	Local Pune dataset (1367 patients)	Random Forest	SHAP/LIME/ ELI5	N/A	82.23%	N/A
El-Rashidy et al. (2023) [28]	MIMIC III dataset (16,354 pregnant women)	DNN	SHAP	0.94	95.7%	Precision: 0.95 Recall: 0.89 AUC: 0.94
Curiafra et al. (2023) [38]	Dhaka dataset (306 patients)	XGBoost	LIME	1.0	100%	Precision: 1.0
Tasin et al. (2023) [46]	PIMA	XGBoost with ADASYN	LIME/SHAP	0.81	88.5%	Precision: 0.82 Recall: 0.80
Dharmarathne et al. (2024) [47]	Public diabetes dataset [48]	XGBoost	SHAP	0.65	77.0%	Precision: 0.6 Recall: 0.73 AUC: 0.82
Vivek Khanna et al. (2024) [29]	Public diabetes dataset (133 pregnant women)	Ensemble stack	SHAP/LIME/ ELI5/Qlattice/ Anchor	N/A	96.0%	Precision: 0.99 Sensitivity: 0.95

For GD diagnosis, researchers turn to more targeted datasets. El-Rashidi et al.'s research, for example, used data specifically collected from pregnant women [28]. Their dataset included information on 16,354 women and focused on 20 relevant features during the gestational period (between 6 and 26 weeks), such as BMI, age, and glucose levels. A smaller size of data has been used with 133 pregnant women with 18 diagnosed GD-positive in Vivek Khanna et al. research [29]. Other relevant datasets for diabetes research include the Early-Stage Diabetes Risk Prediction Dataset (ESDRPD) from Sylhet Diabetes Hospital, a local medical survey in Pune [22] and the Dhaka dataset, both collected through patient surveys [30] and physician-approved [31]. These datasets encompass 520, 1367, and 306 patients, respectively.

3. Results and Discussion

Table 2 summarizes the search methodology employed and the findings obtained from both databases. Guha et al. (2020) [18] were among the first to explore interpretable machine learning models for diabetes diagnosis. Using a dataset of Bangladeshi patients (ESDRPD), they employed techniques like SHAP plots, feature importance, and LIME to understand how the models arrived at their predictions.

Their interpretable Random Forest (RF) model [32] achieved an impressive accuracy of 95% and an Area Under the Curve (AUC) of 0.98. The model also revealed that polyuria (excessive urination) and polydipsia (excessive thirst) are key contributors to diabetes risk. However, the study does have limitations. The focus on a specific dataset raises concerns about generalizability to other populations with different demographics or risk factors. Furthermore, their research shows the lack of preprocessing techniques for missing values, data imbalance, and methods. These shortcomings could potentially affect not only the model's performance but also the XAI methods' results. These are valid points that could be further emphasized.

Vakil et al. (2021) [33] conducted a similar analysis on the same ESDRPD dataset. They used SHAP and several ML algorithms, including XGBoost [34] (XGB), decision trees [35], support vector machines (SVMs) [36], and K-Nearest Neighbors (KNNs) [37]. Their RF model achieved the highest accuracy (99%) and also identified polyuria as the most important feature for predicting diabetes. However, their paper does not mention incorporating data preprocessing techniques like handling missing values or imbalanced classes. This raises concerns about the potential for overfitting in their model. The issue, as mentioned earlier, was also observed in the investigation conducted by Curia et al. (2023) [38], wherein they implemented XGB in conjunction with LIME on the Dhaka dataset encompassing 306 patients. Their findings achieved a perfect accuracy of 100%, alongside an F1-score and Area Under the Curve (AUC) of 1.0. Kibria et al. (2022) [39] conducted a comparative study utilizing the PIMA dataset. Preprocessing techniques included missing value imputation and min-max scaling. Because of the high variance of the PIMA dataset, the researchers employed the Synthetic Minority Oversampling Technique & Edited Nearest Neighbors (SMOTETomek) [40] method alongside five-fold cross-validation to mitigate overfitting. Six machine learning algorithms were initially evaluated: AdaBoost Classifier (ADA) [41], RF, XGB, SVMs, and Logistic Regression (LR). The two models with the best performance, RF and XGB, were then combined using a weighted ensemble approach with soft voting to create a more robust diagnostic model. This approach achieved an impressive accuracy of 90% and an F1 score of 0.89. Additionally, glucose was identified as the most influential feature impacting the model's predictions by both SHAP and LIME. While the proposed approach offers an effective, reliable, and explainable diabetes prediction model, further research is necessary to ensure the generalizability of the findings. The model's performance on more diverse datasets and its effectiveness in real-world clinical settings need to be evaluated. Another recent work by Tasin et al. (2023) also utilized the Pima dataset but employed Adaptive Synthetic Algorithm (ADASYN) with XGBoost. They achieved an accuracy of 88.5% and implemented XAI with LIME and SHAP. Notably, they developed a mobile application for automatic diabetes prediction, showcasing the potential for real-world application.

While previous research focused on T2DM prediction, Du et al. (2022) [42] conducted a similar study specifically for GD in pregnant women. Their ML-based system (CDSS) tackles the challenge of imbalanced data (more non-GD cases) and uses techniques like SMOTE for better analysis. Additionally, the system incorporates SHAP values to explain its reasoning for each feature, enhancing trust and understanding. Their model, employing algorithms like SVM, achieved an accuracy of 75% and an AUC of 0.79 in predicting GD risk. Importantly, they went beyond theory and implemented the model in a web server for academic use, demonstrating its potential clinical application. However, limitations remain, including the need for validation on more diverse populations and further accuracy improvement. Despite these, this research represents a significant step towards incorporating ML for GD prediction in clinical settings.

Existing research has primarily relied on ML models for early-stage diabetes detection. Joseph et al.’s study [43] presents a different approach with an interpretable TabNet [44] DL model tuned via Bayesian optimization (BO-TabNet). The proposed BO-TabNet achieved high accuracy (over 92%) on two datasets and offered interpretability through a combination of the model's attention mechanism and LIME/SHAP tools. It identified insulin and polyuria as key features for diabetes classification in the respective datasets. However, limitations exist. The datasets used have issues like missing data, outliers, class imbalance, and potential biases.

Additionally, the model classifies diabetes as a binary outcome. To improve upon this work, future research should explore more diverse datasets, investigate techniques to address data limitations, and consider incorporating multi-class outcomes for a more comprehensive approach. Furthermore, exploring alternative methods for hyperparameter optimization and feature selection, along with investigating other interpretable models, could lead to even more robust and generalizable models for early-stage diabetes detection.

A recent study by El-Rashidy et al. (2023) [28] unveiled a promising framework using fog computing to predict GD in pregnant women. This system merges the strengths of cloud and fog computing for efficient data processing, delivering real-time results. It functions in three layers: sensors worn by the pregnant woman collect vital signs (IoT layer), the data is analyzed to predict GD risk using a Deep Neural Network (DNN) model interpretable with SHAP (fog layer), and the processed data is securely stored for further analysis (cloud layer). Studies using the MIMIC III dataset, a massive collection of data for GD prediction, show promising results with 95.7% accuracy and an AUC of 0.94. While this framework holds immense potential, further research is necessary. Security and privacy concerns need to be addressed.

Additionally, the model's generalizability to diverse populations and its ability to handle a larger patient volume requires investigation. Overall, this framework utilizing DNNs, SHAP, and fog computing has the potential to revolutionize GD and diabetes prediction and significantly improve healthcare delivery for pregnant women.

XAI research in diabetes has evolved from basic ML algorithms to complex DL models. These models leverage techniques like LIME and SHAP to provide valuable insights into their decision-making processes. However, achieving successful XAI applications in diabetes prediction, especially with tabular data, hinges on effective data preprocessing for both traditional ML and DL approaches. Poorly formatted data directly hinders model performance and its ability to learn meaningful patterns.

Preprocessing addresses common issues in tabular data, such as missing values, outliers, and inconsistencies. This ensures clean data for accurate predictions and allows XAI techniques to provide clear explanations of the model's reasoning.

Additionally, techniques like SMOTE and ADASYN can help balance imbalanced datasets, leading to fairer and more interpretable models. Furthermore, the research is not limited to just the models themselves. There is a growing interest in integrating these models with mobile applications, IoT sensors, and cloud computing. This suggests a move towards real-time prediction and potentially remote patient monitoring.

4. Conclusion

In conclusion, this review has explored the promising path of XAI research in diabetes diagnosis. While there is still room for growth compared to other healthcare applications, the field is making significant strides. The burgeoning availability of big data and EHRs presents a wealth of opportunities for further research. By incorporating diverse data types and leveraging explainable AI techniques, researchers can develop more robust and interpretable models for diabetes prediction.

This can ultimately lead to earlier diagnoses, better treatment decisions, and improved patient outcomes. Additionally, the integration of these AI models with mobile applications and sensor technology holds immense potential for personalized and continuous diabetes management. As XAI research in diabetes continues to evolve, we can anticipate a future where AI plays a crucial role in empowering both healthcare professionals and patients in the fight against this chronic disease.

Acknowledgement

The Basic Science Research Program supports this research through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF- RS-2023-00237287, NRF-2021S1A5A8062526) and local government-university cooperation-based regional innovation projects (2021RIS-003)

References

[1] Astrid Petersmann et al., “Definition, Classification and Diagnosis of Diabetes Mellitus,” *Experimental and Clinical Endocrinology & Diabetes*, vol. 127, no. S01, pp. S1–S7, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[2] The Top 10 Causes of Death, World Health Organization, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[3] Diabetes, World Health Organization, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[4] Mark A. Atkinson, George S. Eisenbarth, and Aaron W. Michels, “Type 1 Diabetes,” *The Lancet*, vol. 383, no. 9911, pp. 69–82, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[5] Sudesna Chatterjee, Kamlesh Khunti, and Melanie J. Davies, “Type 2 Diabetes,” *The Lancet*, vol. 389, no. 10085, pp. 2239–2251, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[6] John P. Kirwan, Jessica Sacks, and Stephan Nieuwoudt, "The Essential Role of Exercise in the Management of Type 2 Diabetes," *Cleveland Clinic Journal of Medicine*, vol. 84, pp. S15–S21, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[7] Julia Shaver, "The State of Telehealth Before and After the COVID-19 Pandemic," *Primary Care: Clinics in Office Practice*, vol. 49, no. 4, pp. 517–530, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Fei Jiang et al., "Artificial Intelligence in Healthcare: Past, Present and Future," *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230-243, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[9] Şefki Kolozali et al., "Explainable Early Prediction of Gestational Diabetes Biomarkers by Combining Medical Background and Wearable Devices: A Pilot Study with a Cohort Group in South Africa," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 4, pp. 1860–1871, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Mirza Mansoor Baig et al., "Early Detection of Prediabetes and T2DM Using Wearable Sensors and Internet-of-Things-Based Monitoring Applications," *Applied Clinical Informatics*, vol. 12, no. 1, pp. 1–9, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Marleen Olde Bekkink et al., "Early Detection of Hypoglycemia in Type 1 Diabetes Using Heart Rate Variability Measured by a Wearable Device," *Diabetes Care*, vol. 42, no. 4, pp. 689–692, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] Toshita Sharma, and Manan Shah, "A Comprehensive Review of Machine Learning Techniques on Diabetes Detection," *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, no. 3, pp. 1-16, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Ciro Rodriguez-León et al., "Mobile and Wearable Technology for the Monitoring of Diabetes-Related Parameters: Systematic Review," *JMIR mHealth and uHealth*, vol. 9, no. 6, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, pp. 1135–1144, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[15] About, PubMed, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/about/

[16] Scott M. Lundberg, and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768-4777, 2017. [Google Scholar] [Publisher Link]

[17] Michael Greenacre et al., "Principal Component Analysis," *Nature Reviews Methods Primers*, vol. 2, no. 100, pp. 1–21, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Anshuman Guha, "Building Explainable and Interpretable Model for Diabetes Risk Prediction," *International Journal of Engineering Research and Technology*, vol. 9, no. 9, pp. 1037-1042, 2020. [Google Scholar] [Publisher Link]

[19] Shichao Jia et al., "Visualizing Surrogate Decision Trees of Convolutional Neural Networks," *Journal of Visualization*, vol. 23, pp. 141–156, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[20] Robert J. Aumann, and Sergiu Hart, *Handbook of Game Theory with Economic Applications*, Elsevier Science, vol. 2, pp. 1-818, 1992. [Google Scholar] [Publisher Link]

[21] Cem Özkurt, "Combining Chaotic Transformations and Machine Learning Algorithms: Evaluating Explainable Artificial Intelligence Model Performance," *Research Square*, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Varad Vishwarupe et al., "Explainable AI and Interpretable Machine Learning: A Case Study in Perspective," *Procedia Computer Science*, vol. 204, pp. 869-876, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[23] Kevin René Broløs et al., "An Approach to Symbolic Regression Using Feyn," *arXiv*, pp. 1-18, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, pp. 1527-1535, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[25] Ioannis Kavakiotis et al., "Machine Learning and Data Mining Methods in Diabetes Research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104-116, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[26] Pima Indians Diabetes Database, Kaggle, 2016. [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[27] Alistair E.W. Johnson et al., "MIMIC-III, A Freely Accessible Critical Care Database," *Scientific Data*, vol. 3, no. 1, pp. 1-9, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[28] Nora El-Rashidy et al., "Utilizing fog Computing and Explainable Deep Learning Techniques for Gestational Diabetes Prediction," *Neural Computing and Applications*, vol. 35, no. 10, pp. 7423-7442, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[29] Varada Vivek Khanna et al., "Explainable Artificial Intelligence-Driven Gestational Diabetes Mellitus Prediction using Clinical and Laboratory Markers," *Cogent Engineering*, vol. 11, no. 1, pp. 1-19, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[30] Karthick Kanagarathinam, "*Early Stage Diabetes Risk Prediction Dataset*," IEEEDataSet, 2021. [CrossRef] [Publisher Link]

[31] Sayed Asaduzzaman et al., "Dataset on Significant Risk Factors for Type 1 Diabetes: A Bangladeshi Perspective," *Data in Brief*, vol. 21, pp. 700-708, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[32] Leo Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [CrossRef] [Google Scholar] [Publisher Link]

[33] V. Vakil et al., "Explainable Predictions of Different Machine Learning Algorithms Used to Predict Early Stage Diabetes," *arXiv*, pp. 1-15, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[34] Tiani Chen, and Carloas Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[35] V. Kishore Ayyadevara, *Decision Tree*, Pro Machine Learning Algorithms, pp. 71-103, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[36] M.A. Hearst et al., "Support Vector Machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998. [CrossRef] [Google Scholar] [Publisher Link]

[37] Padraig Cunningham, and Sarah Jane Delany, "k-Nearest Neighbour Classifiers: A Tutorial," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–25, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[38] Francesco Curia, "Explainable and Transparency Machine Learning Approach to Predict Diabetes Develop," *Health and Technology*, vol. 13, no. 5, pp. 769–780, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[39] Hafsa Binte Kibria et al., "An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI," *Sensors*, vol. 22, no. 19, pp. 1-37, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[40] Gustavo E.A.P.A. Batista, Ronaldo C. Prati, and Maria Carolina Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20-29, 2004. [CrossRef] [Google Scholar] [Publisher Link]

[41] Robert E. Schapire, *Explaining AdaBoost*, Empirical Inference, pp. 37–52, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[42] Yuhan Du et al., "An Explainable Machine Learning-Based Clinical Decision Support System for Prediction of Gestational Diabetes Mellitus," *Scientific Reports*, vol. 12, no. 1, pp. 1-14, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[43] Lionel P. Joseph, Erica A. Joseph, and Ramendra Prasad, "Explainable Diabetes Classification using Hybrid Bayesian-Optimized TabNet Architecture," *Computers in Biology and Medicine*, vol. 151, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[44] Sercan Ö. Arik, and Tomas Pfister, "TabNet: Attentive Interpretable Tabular Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 6679-6687, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[45] Maria A. Kennelly et al., "Pregnancy Exercise and Nutrition with Smartphone Application Support: A Randomized Controlled Trial," *Obstetrics & Gynecology*, vol. 131, no. 5, pp. 818–826, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[46] Isfafuzzaman Tasin et al., "Diabetes Prediction using Machine Learning and Explainable AI Techniques," *Healthcare Technology Letters*, vol. 10, no. 1–2, pp. 1–10, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[47] Gangani Dharmarathne et al., "A Novel Machine Learning Approach for Diagnosing Diabetes with a Self-Explainable Interface," *Healthcare Analytics*, vol. 5, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[48] Diabetes Dataset, kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/mathchi/diabetes-data-set