International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Review of Explainable Machine Learning Frameworks for Early Diabetes Prediction Using Gradient Boosting and SHAP Analysis

Abhishek Sharma[1], Dr. Atul Barve[2]

[1]M.Tech Scholar, Computer Science and Engineering, Prestige Institute of Engineering Management and Research
[2]Professor, Computer Science and Engineering, Prestige Institute of Engineering Management and Research

Abstract
Due to rising rates of complications and costs of treatment, diabetes mellitus represents one of the most rapidly increasing chronic metabolic conditions in the world and is a major concern of public health policy. A growing body of research indicates that diabetes prevalence has grown significantly over the last two to three decades and that this trend is particularly evident in low- and middle-income countries [1], [2]. Therefore, early detection of diabetes is important for providing an opportunity for preventive intervention and slowing disease progression. Traditional methods for diagnosing diabetes are based on laboratory-based clinical tests that are typically conducted after metabolic anomalies have appeared [17], [18].
Recently, researchers have been applying machine learning (ML) to diabetes prediction using structured clinical data to demonstrate better predictive performance than traditional statistical models [27], [28]. While many high performing ML models operate as "black boxes," they do not provide a clear explanation about the factors that contributed to their predictions. A primary concern for clinical application of these models is that the lack of transparency limits their ability to be trusted by clinicians, and ultimately, limits the potential for widespread use of these tools for making clinical decisions [5], [40].
The purpose of this review paper is to present a comprehensive and systematic review of machine learning-based approaches for predicting the onset of diabetes early in the disease process with a focus on ensemble learning techniques and explainable AI (XAI). In addition to exploring datasets that were commonly used for machine learning, data preprocessing strategies, class balance methods, and predictive model techniques that were reported in prior studies, the review will explore predictive modeling and evaluation metric techniques. The review also focuses on gradient boosting-based models because of their high predictive performance on structured clinical data [8], [9] and SHAP (SHapley Additive exPlanations), a widely accepted XAI method based on cooperative game theory [12].

Keywords: Diabetes Mellitus, Machine Learning, Explainable Artificial Intelligence, Gradient Boosting, SHAP, Healthcare Analytics, Clinical Decision Support Systems.

I. INTRODUCTION

Diabetes Mellitus, as a chronic metabolic condition, is due to either an impaired ability to secrete Insulin (or) reduced Sensitivity to Insulin (or) an impaired ability to do both. If left untreated, diabetics can expect a high level of complications including Cardiac Disease, Renal Failure, Neuropathies, Retinopathy, and Early Mortality. Diabetes Mellitus is becoming increasingly common as shown by International Health Institutions who report an alarming increase in the number of cases occurring throughout all demographics [1], [2]. The increase in Diabetics has also put an increasing burden on healthcare through the cost associated with managing the disease, the cost of Chronic Care, and Loss of Productivity.

Early Detection and Risk Assessment of potential Diabetics is critical to prevent Diabetes Complications. The Pre-Diabetic Phase is a High-Risk Phase of Diabetes that allows patients to be identified and treated before developing the disease [17], [18]. Current methods of detecting diabetes are generally Reactive rather than Proactive. Blood Glucose Challenge Tests and Hemoglobin A1C Test are primarily performed when there is an existing Physiological Deficit [28]. Early detection and risk analysis are essential since they can prevent complications associated with diabetes. Additionally, measuring Hb1c levels may not be effective since it is mainly a post-detection method.

Fig. 1. Conceptual overview of rising diabetes prevalence, limitations of traditional diagnosis, and explainable machine learning–based early prediction.

The rising incidence of diabetes, together with its drawbacks in traditional risk prediction methods, underscores the importance of early and clear risk prediction processes, as conceptually outlined in Fig. 1.

A. Role of Machine Learning in Diabetes Prediction

The growing number of electronic health records available has led to various applications of machine learning in disease prediction. The capabilities of machine learning algorithms to handle multiple variables in recognizing complex patterns have been difficult to achieve using conventional methods [27], [32]. Logistic regression analysis was used extensively in early work as it is easy to interpret [6].

Although logistic regression helps to explain individual variables affecting the outcome of disease prediction models, it cannot handle complex interactions as it is limited to linear associations.

These drawbacks can be overcome by the use of advanced machine learning algorithms such as decision trees, support vector machines, random forest, and gradient boosting for the prediction of diabetes, proposed in [7], [27]. Ensemble learning techniques, especially gradient boosting, have demonstrated enhanced predictive capabilities through the incorporation of several weak models for the suppression of bias and variance [8]. XGBoost and associated boosting platforms have garnered much attention because of their efficiency associated with the management of the structured health data [9], [47].

B. The Challenge of Interpretability in Clinical AI

Despite their good predictive power, a majority of machine learning algorithms, especially ensemble methods, as well as deep learning algorithms, have faced criticism for their opacity. These algorithms have generally been regarded as black-box systems since they are able to make predictions with their processes not entirely understandable [5], [42]. In medical settings, a medical algorithm could fail to be transparent, hence raise serious concerns with regard to trust, accountability, and ethics [40].

Clinical decision-making not only needs accurate predictions but also explanations that are aligned with medical knowledge. It is essential for physicians to comprehend why a patient has been assigned a high-risk group so that they can verify the results and rely on them for decisions. Nowadays, there are greater demands on transparency and explainability requirements for AI systems that work in the medical field [26], [41].

C. Explainable Artificial Intelligence in Healthcare

Explainable Artificial Intelligence (XAI) has now become a prominent area of concern as a fix to the interpretability of complex machine learning models. The purpose of XAI approaches is to provide explanations to human experts in the predictions made by the model through the use of inputs [10], [11]. Also, in the field of health care, the use of explainability has been significant for understanding the predictions of the models [20].

Among the different XAI methodologies, SHapley Additive exPlanations, abbreviated as SHAP, have gained popularity because of their sound theoretical basis under cooperative game theory and capacity for ‘global and local’ explanation generation [12]. SHAP provides each feature with a value for a specific explanation regarding a model’s prediction, allowing clinicians to determine a patient-specific factor. Various studies have found that explanations using SHAP correlate aptly with known diabetic risk factors, improving trust and usability [24], [50].

D. Motivation and Scope of This Review

Despite the fact that machine learning–based diabetes prediction has been explored in many research works, the literature remains fragmented across different datasets, algorithms, evaluation metrics, and explainability approaches. In many cases, numerous research projects have focused primarily upon predictive accuracy while offering little or no analysis towards the clinical use of such systems as well as their ability to be interpreted [27], [36]. In addition, there has been an increasing trend towards applying explainability techniques only after the model has been developed, as opposed to developing them concurrently during model development and testing. This literature review addresses both of these areas of need through a synthesized analysis of existing research concerning explainable machine learning systems for predicting early onset type 2 diabetes, specifically focusing on ensemble models (i.e., particularly Gradient Boosting) and SHAP-based explainability frameworks which provide both high levels of system performance and high levels of system transparency. By synthesizing results from

multiple studies, the goal is to help researchers and practitioners develop highly reliable and understandable diabetes risk assessment systems which can be utilized in clinical settings.

II. RELATED WORK

The literature on the prediction of diabetes employing computational methods has developed extensively in the past two decades. Initially, researchers used statistical modeling; concurrently, papers on predictive analytics for diabetes prediction employing machine learning algorithms, ensemble methods, and deep learning algorithms are becoming common. The existing literature on diabetes prediction employing various computational methods concerning five broad categories is presented in this chapter, namely: (A) classical techniques, (B) machine learning techniques, (C) ensemble methods for prediction, (D) deep learning methods, and (E) XAI methods.

A. Traditional Statistical Approaches for Diabetes Prediction

Early diabetes risk prediction studies primarily relied on classical statistical models, particularly logistic regression, due to their simplicity and interpretability in medical research. In these models, disease probability is estimated as a function of clinical variables such as glucose level, BMI, age, blood pressure, and family history. Hosmer et al. demonstrated that logistic regression is effective for binary medical outcomes because of its statistical efficiency and interpretability [6].

Large-scale epidemiological studies further identified plasma glucose concentration, BMI, age, and genetic susceptibility as dominant predictors of diabetes [17], [18]. Although these models enable transparent estimation of risk factors through odds ratios, they assume linear relationships and offer a simplified representation of complex physiological processes. Moreover, traditional statistical approaches struggle to capture feature interdependencies and perform poorly on heterogeneous datasets [27], motivating the shift toward machine learning–based prediction methods.

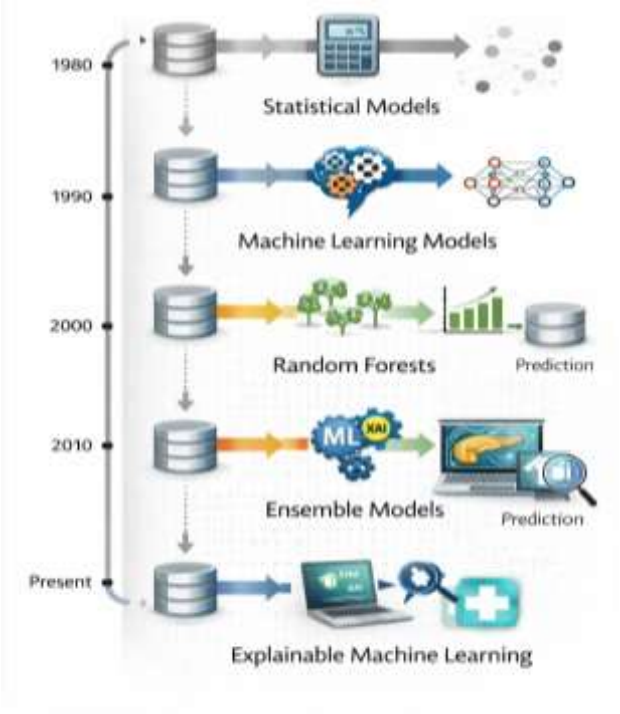

Fig. 2. Evolution of diabetes prediction approaches from statistical models to explainable machine learning frameworks.

The progression of diabetes prediction methodologies, from traditional statistical models to modern explainable machine learning frameworks, is illustrated in Fig. 2.

B. Classical Machine Learning Models in Diabetes Diagnosis

To get around the rigidity in statistical models, researchers explored classical algorithms of machine learning capable of modeling nonlinear relationships. Decision trees were among the first ML models applied to predict diabetes because of their rule-based structure and inherent interpretability. Decision trees give intuitive decision paths that easily align with clinical reasoning. However, they are very sensitive to noise and prone to overfitting, especially when trained with small datasets.

Support Vector Machines (SVMs) have also gained broad applications for diabetes classification. The application of SVM can model complex boundaries of decisions by projecting data into high-dimensional feature spaces. A number of studies presented higher accuracy in SVM than in logistic regression, particularly when non-linear kernels were applied [27]. However, the shortcomings of SVM include poor interpretability and heavy computational complexity when dealing with large datasets.

Due to their simplicity and low cost of training, some works have explored the use of k-Nearest Neighbors and Naïve Bayes classifiers as lightweight alternatives. However, these models are highly susceptible to feature scaling, noise, and assumptions about the data distribution, which eventually leads to inconsistent performance among the different studies that have been carried out in this matter [36].

Moreover, despite classical ML being more flexible than statistics-based methods, their ability to predict and provide robustness was not quite effective for actual medical application. This created a further need for methods such as ensemble learning.

C. Ensemble Learning Techniques for Diabetes Prediction

Ensemble learning has become a prominent approach in diabetes prediction due to its ability to improve accuracy and robustness by combining multiple weak learners. Random Forests, introduced by Breiman, are among the earliest ensemble methods widely adopted in healthcare analytics for classification tasks [7]. Several studies report that Random Forests outperform individual decision trees and logistic regression models, particularly for noisy and imbalanced medical datasets [27]. However, feature importance scores obtained from Random Forests provide only global insights and lack patient-level interpretability.

Gradient Boosting represents a more advanced ensemble technique in which models are trained sequentially to correct the errors of previous learners, as originally proposed by Friedman [8]. Later, XGBoost was developed to improve computational efficiency and scalability [9]. In diabetes prediction, gradient boosting models consistently achieve superior performance in terms of accuracy, F1-score, and ROC-AUC [47], [52]. Bhattacharya et al. demonstrated that boosted tree classifiers effectively capture subtle feature interactions, leading to improved diagnostic performance [47]. Despite these advantages, such models are often criticized for their black-box nature, which limits direct clinical interpretability.

D. Deep Learning Approaches in Diabetes and Healthcare Analytics

Deep learning, has greatly improved healthcare analytics; especially when predicting diseases through hierarchically organized representation of features [30][33]. It has also been used in diabetes research on EHRs, Medical Images, as well as Physiological Time Series Data. Miotto et al., developed "Deep EHR," demonstrating that a neural network could identify latent characteristics of patients' health to make predictions about their risk of disease [34]. Rajkomar et al., showed that a model trained on large scale EHR data was able to accurately predict the presence or absence of several diseases [35]. However, Deep Learning's use is impeded by structured tabular datasets such as the Pima Indians Diabetes Dataset

which require a significant amount of data to process and have low interpretability. Rudin stated that black box models in Healthcare present both Ethical and Practical Risks [40]; therefore, although Deep Learning has potential to be utilized on Multimodal Data Sets, its lack of transparency makes it less than ideal to be utilized to predict Early-Onset Diabetes with Structured Clinical Data.

E. Explainable Artificial Intelligence (XAI) in Healthcare

The complexity of ML and deep learning models has highlighted the need for explainable AI (XAI), which aims to make predictions interpretable for humans [10], [11]. In healthcare, XAI is crucial for building trust and accountability [20]. Traditional model-agnostic methods like LIME approximate complex models locally but lack global consistency [5]. SHapley Additive exPlanations (SHAP), grounded in cooperative game theory, assigns feature contributions fairly and supports both global and patient-specific explanations [12]. Studies show SHAP aligns with medical knowledge, enhancing practitioners' trust, with key factors for diabetes prediction including glucose level, BMI, age, and genetics [20], [24], [27], [47], [48].

F. Handling Clinical Data Challenges

Typical datasets in the clinical field are hindered by missing data, noisy data, and the problem of class imbalance. Diabetes datasets are prone to the problem of the majority class outstripping the minority class by a considerable margin. This problem could mislead the training process and result in false negatives, as pointed out in [15], [16].

For overcoming this problem, Chawla et al. introduced the Synthetic Minority Over-Sampling Technique, also called SMOTE, for balancing data sets by synthesizing minority class samples [15]. The technique has been widely used in studies for predicting Diabetes, and its efficiency has been verified in improving Sensitivity, Recall, without elevating Overfitting in past researches.

Dealing with missing values in the data also forms a significant challenge. Median imputation and MissForest among other methods have been employed in the literature to handle incomplete clinical records while preserving data distribution characteristics, as in [14], [37]. Proper data preprocessing indeed has been shown to impact significantly both the predictive performance and explainability.

G. Research Gaps and Motivation

Despite a considerable number of diabetes prediction studies, there are still quite a few gaps: Firstly, the existing literature has relied excessively on prediction accuracy alone while compromising much on interpretability and clinical utility. Explainability is considered in an afterthought fashion rather than as an integral part of model evaluation studies. Moreover, only a few works systematically investigate the interactions between class imbalance handling and explainability outcomes.

These limitations emphasize the need for explainable ensemble learning frameworks that combine predictive modeling and interpretability into one pipeline. A combination of gradient boosting with SHAP-based explanations assumes a very promising direction by balancing performance, transparency, and clinical relevance.

III. DATASETS USED IN DIABETES PREDICTION STUDIES

Datasets form the core of machine learning–based diabetes prediction research. Model reliability, generalizability, and clinical relevance are strongly linked with the nature and quality of data used. Currently, studies available combine publicly available benchmark datasets with real-world clinical datasets from hospitals or EHR systems. This section highlights a review of the most commonly used datasets, their characteristics, and challenges thereof.

A. Publicly Available Benchmark Datasets

The Pima Indians Diabetes Dataset (PIDD), from the University of California at Irvine's UCI Machine Learning Repository is one of the most popular datasets used in developing predictive models for diabetes due to their broad availability and ability to replicate results using other researchers. The PIDD contains data for 768 patients that includes eight clinical variables, including glucose levels, body mass index (BMI), blood pressure and age [13]. A number of studies have evaluated statistical, machine learning, ensemble and explainable AI models on the PIDD [27],[47] and they consistently found that glucose concentration, BMI, age and diabetes pedigree function were the primary predictors of diabetes risk consistent with what has been demonstrated clinically [17],[18]. The PIDD is limited in its use as a model for other populations because it only includes data from a single population and contains physiologically impossible or "impossible" zero values [14].

B. Large-Scale Survey and Repository-Based Datasets

Several studies utilize datasets derived from large health surveys such as NHANES, which offer greater population diversity and include lifestyle and demographic variables [28]. Although these datasets improve generalization, they frequently suffer from missing values, noise, and inconsistencies due to self-reported data.

C. Hospital-Based and Electronic Health Record Datasets

Hospital-based and EHR datasets provide richer clinical information, including longitudinal records, laboratory results, and medication history [34], [35]. These datasets often enable improved predictive performance but are typically proprietary and affected by privacy concerns, data heterogeneity, and high levels of missingness [25], [43].

D. Class Distribution Characteristics

Most diabetes datasets exhibit significant class imbalance, with non-diabetic cases dominating diabetic samples. This imbalance can bias predictive models toward the majority class and reduce sensitivity, leading to false-negative predictions [15], [16]. Consequently, many studies apply resampling or cost-sensitive learning techniques to mitigate this issue.

E. Dataset Limitations and Generalizability

While useful in their own right, current datasets have many of the same limitations; they are demographic-specific, contain limited samples, and provide no longitudinal or lifestyle information and are therefore limited in their ability to be used for real-world clinical applications and highlight the need for multi-site, large-scale, and externally-validated datasets that are both representative and diverse [25], [28], [36].

F. Summary

Dataset selection plays a critical role in diabetes prediction research. While benchmark datasets such as PIDD support methodological comparison, their limitations restrict clinical generalization. Future studies should prioritize large, diverse, and longitudinal clinical datasets to develop robust and deployable diabetes prediction systems.

IV. DATA PREPROCESSING AND FEATURE ENGINEERING TECHNIQUES

Effective data preprocessing and feature engineering are essential steps in developing reliable machine learning models for diabetes prediction. Clinical datasets often contain noise, missing values, redundant features, and imbalanced class distributions, all of which can negatively affect model performance and

interpretability. This section reviews commonly adopted preprocessing strategies and feature engineering techniques reported in diabetes prediction literature.

Fig. 3. Overview of the data preprocessing and feature engineering pipeline used in diabetes prediction studies.

A typical preprocessing pipeline adopted in diabetes prediction studies is illustrated in Fig. 3.

A. Handling Missing and Invalid Values

In clinical datasets, missing data can be prevalent because of an incomplete record, incorrect data collection, or a variety of other factors. Physiological implausible zeros are common in certain data elements in the Pima Indians Diabetes Dataset, including glucose, insulin, blood pressure, skin thickness, and BMI; these have been treated as missing values by researchers [13][27]. There are many ways to deal with missing data. While simple approaches such as mean or median imputation are efficient from a computational perspective and robust, median is typically used on skewed medical data [14]. More complex imputation strategies, such as k-nearest neighbors (k-NN), and ensemble imputation methods, such as MissForest, use the relationships between features to generate estimates of missing values [37]. When selecting a missing value imputation strategy for use in a healthcare application, the researcher must consider both the degree of accuracy and the degree of interpretability for each method, as well as the degree of robustness for each method.

B. Feature Scaling and Normalization

Feature scaling is an important preprocessing step for many machine learning algorithms. Clinical variables such as glucose levels, insulin concentration, and age often exist on different numerical scales. Without scaling, features with larger magnitudes may dominate the learning process, leading to biased models.

Min–Max and standardization are two of the most frequently used data normalization approaches to enhance model performance in diabetes prediction studies [32]. Standardization produces a dataset

where each feature has an average of zero and variance of one; it is especially helpful when using models that are based on distances (such as k-nearest neighbour) or linear classifiers (including logistic regression and support vector machine) [6]. On the other hand, since tree-based ensembles including random forest and gradient-boosting models are naturally scale invariant they do not require feature normalization [7] [8].

C. Feature Selection and Dimensionality Reduction

The process of feature selection is an essential element of diabetes prediction as it identifies those clinical attributes which contribute most to prediction while eliminating redundant or unimportant elements of the data set so that predictive modeling is improved with regard to its generalized applicability and reduced likelihood of overfitting and increased interpretability. The most common methodologies for performing feature selection include filter based approaches such as correlation analysis and mutual information, wrapper based methodologies such as recursive feature elimination, and embedded methodologies such as determining feature importance through tree-based models [7], [17], [18], [27], [47].

Embedded methodologies are commonly used because of their speed and ability to illustrate key predictor variables (such as glucose levels, BMI, age, and diabetes pedigree function) as well as dimensionality reduction techniques such as Principal Component Analysis (PCA); however, PCA reduces the complexity of the feature space but transforms original features into latent components and loses clinical relevance and therefore are less preferred in research related to health care [32].

D. Handling Class Imbalance

Class imbalance is a common issue in diabetes datasets, as non-diabetic cases often outnumber diabetic ones, causing models to favor the majority class and miss high-risk patients [15], [16]. Resampling techniques, especially the Synthetic Minority Oversampling Technique (SMOTE), generate synthetic minority samples to balance classes and improve recall and F1-score, particularly with ensemble models [15], [27], [47]. Alternatives include random undersampling and cost-sensitive learning, which adjust misclassification penalties to emphasize minority detection. Careful oversampling generally enhances diabetes prediction without compromising interpretability.

E. Feature Engineering for Clinical Interpretability

In addition to preprocessing, feature engineering is critical for improving both the clinical interpretability and clinical relevance of models. In many cases, derived or engineered features (e.g., age groupings, BMI categories, and aggregated risk scores) are created to make ML input data match established clinical standards and guidelines [17][18].

Additionally, feature engineering can support explainability frameworks by creating features that are more interpretable to clinicians. Studies that have incorporated SHAP-based explainability have emphasized the need for meaningful representation of features so that explanations are both clinically relevant and actionable [12] [20].

F. Summary

The reviewed literature highlights that data preprocessing and feature engineering significantly influence both predictive performance and interpretability of diabetes prediction models. Robust handling of missing values, appropriate scaling, effective feature selection, and class imbalance correction are critical prerequisites for reliable model development. These preprocessing strategies form the foundation for advanced predictive modeling techniques discussed in subsequent sections.

V. PREDICTIVE MODELING APPROACHES FOR DIABETES PREDICTION

Predictive modeling forms the core of machine learning–based diabetes risk assessment. Over the years, a wide range of algorithms have been explored to identify individuals at high risk using structured clinical data. This section reviews commonly adopted predictive modeling approaches, ranging from traditional statistical methods to advanced ensemble learning techniques, with emphasis on their suitability for diabetes prediction tasks.

A. Logistic Regression as a Baseline Model

Because of its ease of use, it's possible to interpret results from logistic regression models; also because of the mathematical basis upon which they are built, logistic regression models (which estimate disease likelihood as a function of clinical characteristics) are frequently selected as baseline models when predicting diabetes [6]. As well, studies have demonstrated that using logistic regression can help identify features that are significant predictors of diabetes, including plasma glucose levels, body mass index, age and family history of the disease [17][18].

However, logistic regression models assume a linear relationship between each variable, and therefore cannot be used to represent the complex interactions of physiology that occur among multiple factors [27]. In other words, although logistic regression models can provide an easy way to understand how various clinical features affect disease likelihood, their predictive accuracy will generally be lower than that of more advanced machine learning models when applied to diverse clinical datasets.

B. Tree-Based Models and Decision Trees

Decision Trees are popular for Diabetic Prediction due to their Rule-Based Decision Structure (similar to Clinical Reasoning) and Intuitive Decision Paths [31]. The use of Hierarchical Rules based on Threshold Values provides Inherent Interpretability by Partitioning the Feature Space. While Decision Trees offer good interpretability, Single Decision Trees are often subject to Overfitting and Instability, especially with Noisy or Limited Datasets; and even minor Data Variations can significantly alter the Tree Structure. Therefore, Decision Trees are usually utilized as Base Learners within Ensemble Methods, versus as Standalone Predictive Models [7].

C. Random Forest Models

Random Forests are ensemble models that construct multiple decision trees using bootstrapped samples and random feature selection, and aggregate their predictions through majority voting or averaging [7]. This strategy reduces variance and improves generalization compared to single decision trees.

In diabetes prediction, Random Forests have shown higher accuracy and robustness than logistic regression and standalone decision trees, particularly for noisy and imbalanced clinical datasets [27]. Although feature importance scores provide global insights into influential predictors, they lack patient-specific explanations. Consequently, while Random Forests achieve a favorable balance between performance and interpretability, their explanatory capability remains limited compared to advanced explainability frameworks.

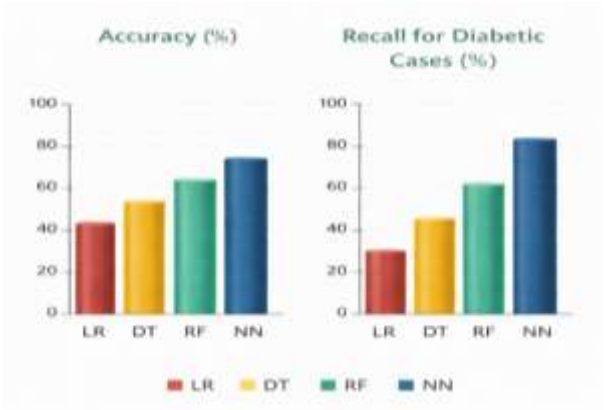

Fig. 4. Comparison of predictive modeling approaches used for diabetes prediction, from traditional statistical models to explainable ensemble learning frameworks.

A high-level comparison of predictive modeling approaches employed in diabetes prediction studies is illustrated in Fig. 4.

D. Gradient Boosting–Based Models

Gradient boosting uses an ensemble learning methodology, to produce highly accurate predictions using structured clinical data by creating decision trees sequentially, correcting prior model errors as opposed to Random Forests that create independent decision trees [8]. Modern Boosting Algorithms are built upon Friedman's Framework, including but not limited to XGBoost; both were designed to increase speed, scale and efficiency for healthcare datasets [9]. A number of studies have shown Gradient Boosting is frequently better than other Machine Learning (ML) methods when used to predict diabetes; increasing predictive accuracy, F1-Score and ROC-AUC values [47], [52]. Additionally, its ability to identify relationships between features allows it to be effective in identifying high-risk patients early in the progression of disease; however, its lack of transparency or interpretability can make it difficult to use in clinical settings.

E. Support Vector Machines and Distance-Based Models

Support Vector Machines (SVMs) are used for diabetes prediction due to their ability to model complex decision boundaries with kernel functions, often achieving competitive performance [27]. However, they require careful parameter tuning and feature scaling, and their decisions are difficult to interpret clinically. Distance-based models like k-nearest neighbors (k-NN) have also been applied, but their performance is sensitive to noise, feature scaling, and dataset size, making them less common in recent studies [36].

F. Deep Learning Models for Structured Clinical Data

Deep learning models, including fully connected deep neural networks, have been investigated for diabetes prediction, particularly in studies involving large-scale electronic health record datasets [30], [34]. These models can learn complex feature representations but require substantial amounts of data to avoid overfitting.

For structured tabular datasets such as the Pima Indians Diabetes Dataset, deep learning models often do not offer significant performance improvements over ensemble learning methods and suffer from limited interpretability [40]. Consequently, their adoption for early diabetes prediction using structured clinical data remains limited.

G. Model Comparison and Selection Criteria

While model selection is typically based on predictive performance when reviewing studies, it has been influenced by increasing numbers of different characteristics which include, but are limited to; model interpretability, model robustness, and model usability as these can be used to support clinical practices. As an example, while ensemble methods (such as Gradient Boosting) have demonstrated a superior predictive ability compared to many alternate approaches, they also offer a great deal of flexibility for using additional methods to explain their predictions and thus support their application within clinical settings [8] [12].

In addition to achieving accurate results, the requirement to select models that will provide an adequate level of transparency will continue to become a focal point of diabetes prediction research and other clinical decision support systems.As such, models employing gradient boosting along with an explainable AI framework are being increasingly selected by researchers studying recent diabetes prediction research.

H. Summary

The various forms of predictive modeling that were applied to predict diabetes are discussed here. Although logistic regression is interpretable, however, given its structured clinical data, it is outperformed by ensemble models such as gradient boosting models. The black box nature of these models further illustrate the necessity of an interpretable framework to address this issue which will be covered in the next section.

VI. EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR DIABETES PREDICTION

As machine learning models become increasingly complex, the need for transparency and interpretability has gained critical importance, particularly in healthcare applications. In diabetes prediction, explainable artificial intelligence (XAI) plays a vital role in bridging the gap between high predictive performance and clinical trust. This section reviews explainability techniques applied in diabetes prediction studies, with a particular focus on SHapley Additive exPlanations (SHAP).

Fig. 5. Explainable machine learning framework for diabetes prediction using SHAP-based global and local explanations.

The overall explainable prediction workflow adopted in recent diabetes prediction studies is illustrated in Fig. 5.

A. Importance of Explainability in Clinical Decision Support

A model's predictive accuracy is not enough to be adopted by clinicians. Clinicians need to know how a model came to its conclusions about a patient's risk level. Clinicians need to validate a model's output; identify errors or biases in a model; and inform patients of their model generated risk assessment in a fair and ethical way [20] [26].

Black box models have been shown to provide accurate predictive information; however they lack transparency which creates serious problems with accountability, equity and regulatory compliance in healthcare [40] [41]. Explainable models are now being recognized as a key component of AI-based clinical decision support systems — rather than simply an option.

B. Model-Specific vs Model-Agnostic Explainability Methods

Explainability methods can be broadly categorized into model-specific and model-agnostic techniques. Model-specific methods rely on the internal structure of a given algorithm, such as coefficient analysis in logistic regression or feature importance scores in decision trees and random forests [6], [7]. While these approaches offer global insights, they often fail to provide patient-level explanations for complex ensemble models.

Model-agnostic techniques, on the other hand, can be applied to any machine learning model regardless of its internal structure. These methods are particularly useful for explaining black-box models such as gradient boosting and deep learning architectures [11].

C. Local Interpretable Model-Agnostic Explanations (LIME)

LIME is one of the earliest model-agnostic explainability techniques proposed to explain individual predictions by approximating a complex model locally using an interpretable surrogate model [5]. LIME has been used in diabetes prediction studies to explain what features are contributing to a particular patient's predictive risk of developing diabetes so that the clinician can see where that decision boundary is for that patient.

Despite this ability to provide personalized feature explanations for each individual, LIME's explanations can also be unstable, meaning a small amount of change in input data can result in a completely different explanation. Furthermore, LIME does not have global consistency and therefore it is less likely to be useful in determining how well the model predicts for an entire population [11].

D. SHapley Additive exPlanations (SHAP)

SHapley Additive exPlanations (SHAP) have become one of the most popular tools for making healthcare models understandable. Based on cooperative game theory, SHAP assigns each feature a "contribution score" that shows how much it influences the model's prediction [12].

One of SHAP's strengths is that it provides both global and local explanations. Global explanations highlight the features that matter most across the entire dataset, while local explanations show how each feature affects an individual patient's risk. This makes it especially useful in diabetes prediction, where doctors need insights both at the population level and for specific patients [24], [48].

E. SHAP in Diabetes Prediction Studies

Many researchers have applied SHAP to diabetes prediction models and found that, in general, plasma glucose, BMI, age, and Diabetes Pedigree Function are the top predictors based on their feature influence [27], [47], [52]. Since these results are consistent with well-established medical knowledge, they add to the clinical credibility of SHAP-based explanations for these models.

SHAP provides three types of visualization (summary, dependence, and force) that provide an easy-to-interpret view into how a model behaves. The presence of high glucose levels and high BMI values generally contributes positively to predicted risk of diabetes; while lower age and a normal BMI will contribute negatively to this risk, thus providing a way to personalize the patient's risk and plan for interventions accordingly.

F. Global vs Patient-Level Explainability

Global explainability helps clinicians and researchers understand overall model behavior, identify dominant risk factors, and assess alignment with medical guidelines. Patient-level explainability, on the other hand, enables individualized analysis by revealing why a particular patient is classified as high or low risk.

In clinical practice, patient-level SHAP explanations can assist in shared decision-making by allowing healthcare providers to explain model predictions in terms of tangible clinical variables. This transparency enhances patient trust and supports ethical AI deployment [20], [26].

G. Limitations and Challenges of Explainability Methods

In addition to their advantages, however, there are also important limitations to the explainability techniques available. SHAP can be very computationally intensive when dealing with large data sets and complex models and a lack of sufficient clinical expertise in interpreting results from SHAP can lead to incorrect interpretation of SHAP outputs [41]. In addition, the post-hoc explanations generated by SHAP may only provide an approximation of how the model actually reasons about the input data rather than accurately representing how it does so. Therefore, some researchers now suggest that in order to avoid making incorrect assumptions based on the output of SHAP (and other similar) explanation tools, researchers should use these tools as a complement to formal model validation and domain specific knowledge [40] [42].

H. Summary

The reviewed literature highlights that explainable AI is essential for the successful deployment of machine learning models in diabetes prediction. Among available XAI techniques, SHAP stands out due to its theoretical soundness, consistency, and ability to provide both global and patient-level explanations. Integrating SHAP with ensemble learning models such as gradient boosting enables the development of predictive systems that are not only accurate but also transparent and clinically trustworthy.

VII. CHALLENGES, LIMITATIONS, AND OPEN RESEARCH ISSUES

Despite significant progress in applying machine learning and explainable artificial intelligence for diabetes prediction, several challenges and limitations continue to hinder the reliable deployment of these systems in real-world clinical settings. This section discusses key technical, clinical, and ethical challenges identified in the literature, along with open research issues that require further investigation.

Fig. 6. Key challenges and limitations associated with machine learning–based diabetes prediction systems.

The major challenges affecting the development and deployment of diabetes prediction models are summarized in Fig. 6.

A. Dataset-Related Challenges

The majority of limitations found in all of the diabetes prediction studies have been related to the use of extremely limited, demographically restricted datasets. Publically available benchmark datasets for diabetes predictions, including the Pima Indians Diabetes Dataset, are commonly used because of their accessability and replicability; however, these datasets are representative of only very specific (limited) demographic groups and are lacking in diversity [13], [27] which could result in poorly generalized models across a larger/ethnically diverse population.

Most datasets also fail to provide longitudinal data, only providing an instantaneous view of a patient's clinical status. Diabetes is a chronic disease that is affected by long term life style and metabolic changes; therefore, the inability of models to be able to utilize temporal data will severely limit the model's ability to make accurate predictions about how the disease will progress [28], [35].

B. Data Quality and Label Reliability

Clinical datasets often contain missing values, noise, and inconsistencies arising from measurement errors or incomplete records. While preprocessing techniques such as imputation and resampling can mitigate these issues, they may also introduce bias or distort underlying data distributions if not applied carefully [14], [37].

Another concern is label reliability. Diabetes diagnosis labels are typically derived from clinical thresholds or self-reported medical histories, which may vary across institutions and studies. Inconsistent labeling can affect model training and evaluation, leading to unreliable predictions.

C. Model Generalization and External Validation

Studies frequently report good predictive performance via cross-validation on one dataset yet rarely test for their models' external validity across other datasets. External validation is necessary to ensure that the models are robust enough to be used in real-world settings with a variety of clinical populations; since models that function well on a benchmark dataset do not necessarily perform as well as expected

in diverse clinical settings where there may exist differences between patient demographics, how data is collected, and healthcare practices.

Ensuring generalization across populations remains an open challenge and highlights the need for multi-center validation studies and standardized evaluation protocols.

D. Trade-Off Between Accuracy and Interpretability

Ensemble learning methods, which include Gradient Boosting Models (GBMs), while able to produce higher levels of prediction, by definition will also increase model complexity. Explainability techniques like SHAP provide explanations after a model has been built but cannot completely alleviate the concern that many have about the "black box" nature of GBM's [40] [42].

The literature contains continuing discussions on whether models that are inherently interpretable should be given preference to those that are complex but have post-hoc explanations especially when high-risk areas exist such as healthcare. Finding the right balance between accuracy and interpretability continues to be a significant challenge for researchers.

E. Explainability Misinterpretation and Over-Reliance

Although explainability methods enhance transparency, there is a risk of misinterpreting explanation outputs. Feature attribution scores may be incorrectly assumed to represent causal relationships, which can lead to misleading clinical conclusions if not interpreted with domain expertise [41].

Furthermore, excessive reliance on model explanations without rigorous validation may create a false sense of trust. Explainable AI systems should be viewed as decision-support tools rather than decision-makers, complementing clinical judgment rather than replacing it [20], [26].

F. Ethical, Legal, and Regulatory Considerations

The deployment of AI-driven diabetes prediction systems raises ethical and regulatory concerns related to data privacy, fairness, and accountability. Clinical datasets often contain sensitive personal information, requiring strict compliance with data protection regulations [25].

Bias in training data can lead to unfair predictions for certain demographic groups, potentially exacerbating health disparities. Ensuring fairness and transparency in model development and evaluation is therefore essential for ethical AI adoption in healthcare [41], [48].

G. Open Research Directions

The previous literature provides a number of potential avenues for future research. For instance, integration of multimodal and longitudinal patient data, creation of standardised benchmarks for evaluating models, closer coupling of model predictiveness with model explainability; and clinician-based evaluations of the utility of explainability techniques for clinicians to be able to use them in real-world clinical settings.

H. Summary

In summary, while explainable machine learning approaches show strong potential for early diabetes prediction, significant challenges related to data quality, generalization, interpretability, and ethics remain unresolved. Addressing these challenges is essential for translating research advancements into safe, reliable, and clinically deployable decision support systems.

VIII. FUTURE RESEARCH DIRECTIONS

The next step for the future of predicting diabetes through machine learning (ML) and explainable AI (XAI) is to create results that are clinical-usable, reliable and deployable in real-world settings. Temporal and longitudinal data sources (e.g., serial laboratory tests and long-term trend analysis of an

individuals' lifestyles) can enhance the ability to predict an individual's likelihood of developing diabetes and the rate of progression of their disease [28][35]. In addition, using multiple types of data (i.e., structured clinical variables combined with unstructured clinical note data; wearable sensor data; patient self-reported data regarding lifestyle habits) can provide a complete view of the health status of an individual and improve the ability to predict an individual's likelihood of developing diabetes while also providing robust explanations [30][43].

Validating models on external populations and demonstrating consistent performance across different populations and health care environments are important due to the potential variability in performance when models have been developed on a single dataset [36]. Also, there needs to be improvement in how explanations are used by clinicians, particularly as it relates to trust and decision making. Human centered studies can assist in evaluating this [20], [26] and hybrid approaches that utilize both high performance ensembles and interpretable models can decrease the need for post-hoc explanations and maintain accuracy [40], [42]. Another area of significant importance is the development of causal relationship-based explainability methods versus association based explainability methods [40], [42]. In addition to the technical challenges, there are numerous ethical and regulatory issues that must be addressed prior to deploying AI based diabetes prediction systems into healthcare settings, including data privacy, mitigating bias and ensuring compliance with relevant healthcare regulations [41], [48].

IX. CONCLUSION

As global rates of diabetes continue to rise, this study reviewed how machine learning could be used to help identify people at risk for developing diabetes earlier than traditional methods allow; it also looked at ensemble machine learning methods and how to make machine learning systems explainable so they are trustworthy enough to be used in clinical settings. Traditional medical diagnostic techniques have several limitations, including that they cannot provide a clear picture of a person's health status until after disease symptoms develop, making early detection through the use of predictive systems essential in providing proactive care. Ensemble learning models were found to have superior results compared to many other types of machine learning and statistical analysis when applied to structured clinical data. This is primarily because ensemble models can learn to capture complex nonlinear relationships between features of clinical data. Although ensemble models were shown to be among the best for predicting who will develop diabetes earlier than traditional medical methods, the high complexity of these models is currently a barrier to widespread adoption in clinical settings. Explainable artificial intelligence (XAI) methods, particularly SHAP, have emerged as a means to help clinicians understand the reasoning behind predictions from machine learning models. While XAI shows promise in improving clinician trust and usability of predictive systems in healthcare, several challenges remain. These include limited variability in training datasets, class imbalance in clinical populations, insufficient external validation of models, and ethical concerns regarding patient privacy and potential biases. Despite these limitations, explainable ensemble machine learning offers a promising approach for developing clinically relevant models for early diabetes prediction, with the potential to support proactive healthcare practices and improve patient outcomes.

REFERENCES

1. World Health Organization, *Global Report on Diabetes*, Geneva, Switzerland: WHO Press, 2016.
2. International Diabetes Federation, *IDF Diabetes Atlas*, 9th ed., Brussels, Belgium, 2019.

3. K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.
4. A. Esteva et al., "A guide to deep learning in healthcare," *Nat. Med.*, vol. 25, no. 1, pp. 24–39, 2019, doi: 10.1038/s41591-018-0316-z.
5. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. ACM SIGKDD*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
6. D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed., Hoboken, NJ, USA: Wiley, 2013, doi: 10.1002/9781118548387.
7. L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
8. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.
9. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
10. F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017, doi: 10.48550/arXiv.1702.08608.
11. R. Guidotti et al., "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, 2019, doi: 10.1145/3236009.
12. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, 2017, doi: 10.48550/arXiv.1705.07874.
13. UCI Machine Learning Repository, "Pima Indians Diabetes Dataset," 2019. [Online]. Available: https://archive.ics.uci.edu
14. R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed., Hoboken, NJ, USA: Wiley, 2019, doi: 10.1002/9781119482260.
15. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
16. T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.
17. American Diabetes Association, "Standards of medical care in diabetes—2022," *Diabetes Care*, 2022, doi: 10.2337/dc22-SINT.
18. D. M. Nathan et al., "Medical management of hyperglycemia in type 2 diabetes," *Diabetes Care*, vol. 32, no. 1, pp. 193–203, 2009, doi: 10.2337/dc08-9025.
19. E. J. Topol, *Deep Medicine*, New York, NY, USA: Basic Books, 2019.
20. A. Holzinger et al., "Explainable AI in medicine," *WIREs Data Min. Knowl. Discov.*, 2019, doi: 10.1002/widm.1312.
21. D. S. W. Ting et al., "Artificial intelligence and deep learning in ophthalmology," *Br. J. Ophthalmol.*, 2021, doi: 10.1136/bjophthalmol-2019-315061.
22. R. Islam et al., "Explainable artificial intelligence approaches: A survey," *IEEE Access*, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3173520.
23. J. Chen, L. Song, and M. Wainwright, "Learning to explain machine learning models," *Nat. Mach. Intell.*, 2021, doi: 10.1038/s42256-020-00262-3.

24. M. Tjoa and C. Guan, "A survey on explainable AI toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, 2021, doi: 10.1109/TNNLS.2020.3027314.
25. S. Beam and I. Kohane, "Big data and machine learning in health care," *JAMA*, 2018, doi: 10.1001/jama.2017.18391.
26. A. Holzinger et al., "Explainable AI methods in medical decision making," *Artif. Intell. Med.*, 2021, doi: 10.1016/j.artmed.2020.101938.
27. M. Kavakiotis et al., "Machine learning and data mining methods in diabetes research," *Comput. Struct. Biotechnol. J.*, 2017, doi: 10.1016/j.csbj.2016.12.005.
28. Z. Obermeyer and E. J. Emanuel, "Predicting the future—big data and machine learning in clinical medicine," *N. Engl. J. Med.*, 2016, doi: 10.1056/NEJMp1606181.
29. C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022.
30. Y. Bengio, I. Goodfellow, and A. Courville, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015, doi: 10.1038/nature14539.
31. J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, 1986, doi: 10.1007/BF00116251.
32. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009, doi: 10.1007/978-0-387-84858-7.
33. I. Goodfellow et al., *Deep Learning*, MIT Press, 2016.
34. S. Shickel et al., "Deep EHR," *J. Biomed. Inform.*, 2018, doi: 10.1016/j.jbi.2018.01.003.
35. J. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," *NPJ Digit. Med.*, 2018, doi: 10.1038/s41746-018-0029-1.
36. A. Rajula et al., "Comparison of ML and DL methods," *BMC Med. Res. Methodol.*, 2020, doi: 10.1186/s12874-020-00974-0.
37. S. V. Stekhoven and P. Bühlmann, "MissForest," *Bioinformatics*, 2012, doi: 10.1093/bioinformatics/btr597.
38. M. Goldstein et al., "Peeking inside the black box," *IEEE Intell. Syst.*, 2015, doi: 10.1109/MIS.2015.70.
39. R. Caruana et al., "Intelligible models for healthcare," in *Proc. KDD*, 2015, doi: 10.1145/2783258.2788613.
40. C. Rudin, "Stop explaining black box models," *Nat. Mach. Intell.*, 2019, doi: 10.1038/s42256-019-0048-x.
41. S. Barredo Arrieta et al., "Explainable AI: Concepts and surveys," *Inf. Fusion*, 2020, doi: 10.1016/j.inffus.2019.12.012.
42. Z. C. Lipton, "The mythos of model interpretability," *Commun. ACM*, 2018, doi: 10.1145/3233231.
43. R. Miotto et al., "Deep learning for healthcare," *Brief. Bioinform.*, 2018, doi: 10.1093/bib/bbx044.
44. M. Samek et al., *Explainable AI*, Springer, 2019, doi: 10.1007/978-3-030-28954-6.
45. A. Gunning, "Explainable artificial intelligence (XAI)," DARPA, 2017.
46. A. Holzinger, "The next frontier of AI," *IEEE Computer*, 2021, doi: 10.1109/MC.2020.3028656.
47. S. Bhattacharya et al., "PCA–Firefly based XGBoost for diabetes diagnosis," *Health Inf. Sci. Syst.*, 2020, doi: 10.1007/s13755-020-00117-1.
48. A. B. Arrieta et al., "XAI toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.