© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1718378

A Systematic Literature Review of Machine Learning and Explainable AI Approaches for Diabetes Prediction: Taxonomy, Gaps, and Future Directions (2020–2025)

SMARIKA SINGH[1], PROF.(DR.) RITU SINDHU[2], DR. SHIVANI SHARMA[3]
[1]*MTech Scholar, Amity School of Engineering & Technology Amity University Haryana, Gurugram, India*
[2]*Professor, Amity School of Engineering & Technology Amity University Haryana, Gurugram, India*
[3]*Assistant Professor, Amity School of Engineering & Technology Amity University Haryana, Gurugram, India*

Abstract — Diabetes mellitus is a rapidly escalating global public health crisis affecting over 537 million adults worldwide, with projections suggesting 783 million by 2045. The exponential growth of health data repositories and machine learning research has produced a vast body of literature on automated diabetes risk prediction. However, this literature suffers from recurring methodological inconsistencies — particularly over-reliance on a single benchmark dataset, data leakage from improper SMOTE application, absence of threshold optimisation for real-world imbalanced distributions, and shallow explainability that hinders clinical adoption. This paper presents a systematic literature review of 34 studies published between 2020 and 2025, covering classical machine learning, ensemble methods, deep learning architectures, and explainable AI frameworks applied to diabetes prediction. We propose a six-category taxonomy of approaches, analyse comparative performance across datasets, and identify eight specific research gaps that remain unaddressed in the existing literature. We also present our own empirical work — a SMOTE-enhanced hybrid RF+MLP stacking framework validated across Pima Indian and BRFSS 2015 datasets achieving consistent AUC of 0.808 and 0.816 respectively with 90.7% clinical recall — as a representative study addressing several identified gaps. Based on the review, we outline seven concrete future research directions essential for bridging the gap between research prototypes and clinically deployable diabetes screening tools.

Keywords — Diabetes Prediction, Machine Learning, Ensemble Methods, Explainable AI, SHAP, LIME, SMOTE, Cross-Dataset Validation, Systematic Review, Random Forest, Xgboost, Deep Learning

I. INTRODUCTION

Diabetes mellitus is among the most common diseases affecting people across the globe today and it is also among the diseases with the largest economic impacts in the twenty-first century. The IDF noted that approximately 537 million adults suffered from diabetes in 2021, which was a 16% rise from the 2019 numbers, with a projection of 783 million adult cases in 2045 [1]. It is estimated by WHO that diabetes leads to about 1.5 million deaths per year, mostly due to complications such as cardiovascular ailments, renal failure, and diabetic neuropathy [2]. This disease is common in low- and middle-income countries such as India, where approximately 77 million adults suffer from it [1].

Machine learning (ML) as a tool to predict the risk of developing diabetes mellitus has received significant interest from researchers during the last ten years. Machine learning algorithms allow modeling the complicated, non-linear interconnection between many risk factors, which is not possible using classical approaches based on risk factor thresholds [3]. However, a careful assessment of the state of art shows that methodological discrepancies, dependence on benchmarks, and inadequate interpretability analysis significantly limit the clinical applicability of such algorithms. In particular, Olusanya et al. [6] found out that data leakage due to SMOTE oversampling prior to train-test splitting is a common problem in published papers.

This paper makes four contributions. First, we systematically review 34 ML and XAI studies for diabetes prediction published between 2020 and 2025, drawn from IEEE Xplore, PubMed, Scopus, and arXiv. Second, we propose a six-category taxonomy organising approaches by methodology type. Third, we identify eight specific, evidence-based research gaps with supporting citations. Fourth, we present our own empirical study as a worked example addressing multiple identified gaps,

achieving cross-dataset AUC consistency of 0.808 (Pima) and 0.816 (BRFSS 2015) with 90.7% clinical recall through a SMOTE-enhanced stacking ensemble.

The rest of this paper is organized as follows: In Section II, we discuss the methodology adopted for the review. Section III deals with the taxonomy of approaches. Section IV provides us with comparative analysis. Research gaps are discussed in Section V. In Section VI, we discuss our empirical study.

II. REVIEW METHODOLOGY

A. Search Strategy

The search process followed by the researcher for identifying the required literature included the use of four databases which are as follows: IEEE Xplore, PubMed, Scopus, and arXiv. The key terms used for the search included: "diabetes prediction," "machine learning," "deep learning," "explainable AI," "SHAP," "LIME," "Random Forest," "BRFSS," and "Pima Indian dataset." AND, OR were the boolean operations used during the search process.

B. Inclusion and Exclusion Criteria

Eligible papers involved those which: (1) had a machine learning or deep learning approach to predict or classify diabetes; (2) contained quantitative performance measures such as accuracy, AUC, precision, recall, or F1-scores;
(3) employed open access or clinically validated data sets; and (4) were peer-reviewed published papers, conference papers, or pre-prints with confirmed performance outcomes. Ineligible papers were those which: (1) addressed the issue of diabetes treatment rather than diabetes prediction; (2) were limited to qualitative approaches and did not involve quantitative performance measures; and (3) used proprietary data sets without explanations.

C. Study Selection

From the preliminary search, 312 studies were identified. From these, 48 studies that were duplicate and 178 studies based on their title/abstract review were eliminated. Consequently, 86 full articles were examined for eligibility, and among them, 34 studies met all the inclusion criteria and became the subject of the systematic review. The selected studies are diverse in nature in terms of methodology, data, and evaluation frameworks used.

D. Data Extraction

The following details were extracted for each study: first author and publication year; algorithms used; datasets used; metrics used (accuracy, AUC-ROC, recall, F1 score); approach used for interpretability of the model; strategy used to handle class imbalance; and whether the model has been deployed in clinical practice/application. The collected data was presented in a comparison table (Table I).

III. TAXONOMY OF APPROACHES

Based on the systematic review, we propose a six-category taxonomy organising diabetes ML research by primary methodological approach. Table II summarises the taxonomy with representative studies, typical AUC ranges, and primary limitations for each category.

TABLE II: Taxonomy of ML Approaches for Diabetes Prediction

Category	Algorithms	Representative Studies	Typical AUC	Limitation
Classical ML	LR, DT, SVM, KNN, NB	Joshi [4], Naeem [9], Naaz [11]	0.75-0.85	Linear boundaries, low capacity
Ensemble	RF, XGBoost, LightGBM, Stacking	Tasin [5], Ali [8], Allani [13]	0.82-0.90	Prone to leakage errors
Deep Learning	LSTM, CNN, DNN, BiLSTM	LSTM-DNN [23], BiLSTM [24]	0.85-0.98*	Often inflated, no XAI
XAI-Integrated	RF+SHAP, LightGBM+LIME	Hasan [14], Frontiers [10]	0.82-0.90	Global-only explanation
Multi-Dataset	Hybrid across 2+ datasets	Allani [13], THIS STUDY	0.80-0.82	Rare in literature
Genomic/Advanced	CNN+RNN, Transformers	Isiaka [18]	0.90-0.95	Needs genomic data

A. Classical Machine Learning

Classic ML techniques such as Logistic Regression (LR), Decision Trees (DT), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naive Bayes (NB) form the first group of diabetes prediction methods and can be considered the most understandable ones. Thus, Joshi and Dhakal [4] used LR and DT for Pima Indians' data obtaining 78.3% accuracy and identifying glucose, BMI, and age as the main predictors. Naeem et al. [9] implemented KNN and Bernoulli Naive Bayes resulting in 80% accuracy. Additionally, Al-Batah et al. [31] performed an analysis of eight classic algorithms for a survey dataset, and neural networks reached 78.6% accuracy. Despite the interpretability and efficiency of these algorithms, they have a limitation of working under assumptions of either linear or simple non-linear decision boundaries [4], [6].

B. Ensemble and Boosting Methods

Ensemble techniques have been found to perform better than traditional single classifiers in the domain of diabetes predictions. Tasin et al. [5] achieved 81% accuracy and AUC of 0.84 in their approach that involved combining two datasets – the Pima Indian dataset and a private Bangladeshi dataset – using XGBoost with ADASYN oversampling, and this was one of the most methodologically robust approaches published yet. Ali et al. [8] introduced a stacking technique with Random Forest as the meta learner that resulted in an approximate accuracy of 84% on Pima. A stacked ensemble technique involving Bagging, LightGBM, and XGBoost [12] yielded accuracy of 92.9%. Allani [13] used the LightGBM technique on the BRFSS dataset along with SMOTE, developed a website using Dash, and included SHAP and LIME in their framework. The hybrid XGB-RF model by [16] delivered 96% accuracy and AUC of 0.990 although this figure requires deeper scrutiny. Ganie and Malik [34] have illustrated the efficacy of ensembles on datasets other than the standard Pima Indian dataset.

C. Deep Learning

Deep learning methods boast the highest accuracies according to the reviewed literature. Ensemble LSTM- DNN-CNN models with soft voting

approach [23] obtained an accuracy of about 98% for Pima. BiLSTM models with created temporal biomarkers [24] provided 93.2% accuracy on longitudinal datasets. The combination of CNN-LSTM models along with conditional generative adversarial networks [25] yielded accuracies of 89-96% for Pima and BRFSS. Quantum-inspired ensemble stacking of RF, Extra Trees, CNN, and FFNN [15] provided an AUC value of 0.870. Isiaka et al. [18] managed to go further by combining the genomics data from UK Biobank, TCGA, and GEO databases with deep learning CNN-RNN models to obtain 94.6% accuracy. But, as per the findings of Olusanya et al. [6], many papers applying deep learning algorithms tend to use SMOTE before splitting into training and test datasets, causing data leakage. Moreover, lack of cross-validation makes deep learning research non-reproducible.

D. Explainable AI-Integrated Systems

This fourth classification includes algorithms where post-hoc explainability forms a crucial design element, not just an additional factor. SHAP (SHapley Additive Explanations) [36] and LIME (Local Interpretable Model-agnostic Explanations) [37] have become the main instruments within this category. In particular, Hasan et al. [14] applied simultaneously five XAI approaches such as SHAP, LIME, Integrated Gradients, Counterfactual Analysis, and Attention Mechanisms within an AutoML approach reaching 85% accuracy level. Also, in the project of Frontiers AI [10], SHAP and LIME methods were used along with feature selection in four diabetes datasets resulting in 90% precision rate. For instance, a combination of LightGBM with Boruta feature selection [20] applied SHAP to discover predictive features in various data sets. Explainable Clustering method [19] employed cluster-personalized RF using SHAP and LIME techniques for analysis of continuous glucose monitoring data achieving AUC equal to 0.84-0.93. The comparison study [28] tested different XAI methods, such as SHAP, LIME, and ELI5, among others, on several machine learning algorithms giving one of the most thorough tests of XAI approaches.

E. Multi-Dataset Validation Systems

The least prevalent category in the discussed literature involves those systems that have been validated on several heterogeneous data sets. From the 34 discussed studies only three carry out cross-dataset validation. The combination of CNN-LSTM with GANs augmentation [25] was tested using Pima and BRFSS but without the cross-dataset AUC comparison. The Frontiers Classification [21] system

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1718378

was tested using the random forest with SHAP and LIME on four different diabetes data sets. In our proposed study, we use two heterogeneous data sets Pima Indian and BRFSS 2015, achieving AUC of 0.808 and 0.816 respectively.

F. Genomic and Advanced Architectures

The proposed methods use genomic, pharmacogenomic, or wearable sensor information together with machine learning algorithms. Isiaka et al. [18] have shown that when clinical information, life-style variables, and genomic data from the UK Biobank are used, an accuracy of 94.6% can be achieved, and GWAS has identified more than 400 genetic loci linked to Type 2 Diabetes. Examples of innovative architectural designs include Additive-Multiplicative Neural Networks (AMNN) with Knowledge Augmented Networks (KAN) [30]. Another method involves predicting the risk of CVD by using BRFSS 2023 with XGBoost, PCA, and deep learning [26], which achieved an accuracy of 90.5%. The genomic data approaches need special datasets that cannot be found in regular clinical settings.

IV. COMPARATIVE ANALYSIS

Table I shows a comprehensive comparison of all 34 reviewed studies across the eight dimensions: study, publication year, primary method, dataset, accuracy, AUC-ROC, XAI method used, and key finding. The table is sorted by publication year (most recent first) to highlight temporal trends in the field.

TABLE I: Comprehensive Comparison of Diabetes ML Studies (2020-2025)

Study	Year	Method	Dataset	Accuracy	AUC	XAI	Key Finding
Joshi & Dhakal [4]	2021	LR + Decision Tree	Pima	78.3%	—	No	Glucose, BMI, Age top predictors
Kumar et al. [35]	2021	Random Forest	UCI	90%	—	No	RF outperforms single classifiers
Early Stage [22]	2021	RF, DT, ANN, KNN + SHAP	Sylhet Hospital	99%	—	SHAP	High accuracy on small dataset
Tasin et al. [5]	2022	XGBoost + ADASYN	Pima + Private	81%	0.84	SHAP+LIME	Web + Android deployment
Ali et al. [8]	2022	Stacking + RF Meta	Pima	~84%	—	No	Stacking > single classifiers
Olusanya et al. [6]	2022	Systematic review	Multiple	~76%	—	No	Leakage inflates metrics
Ganie & Malik [34]	2022	Ensemble ML	Lifestyle Dataset	—	—	No	Lifestyle features useful
Naeem et al. [9]	2023	KNN + Naive Bayes	Pima	~80%	—	No	Simple ensemble competitive
Multi-stage Ens. [33]	2023	NB+KNN+J48->SVM	Pima	High	High	No	Cascaded ensemble effective
Hasan et al. [14]	2024	AutoML + 5 XAI	Pima	85%	—	5 methods	Most XAI methods compared
Stacked Ens. [12]	2024	Bagging+LGBM+X	Pima	92.9%	High	No	Stacking on

		GB					Pima effective
Frontiers AI [10]	2024	RF + Feature Sel.	Pima+	~90%	—	SHAP+LIME	Feature selection + XAI
LSTM-DNN-CNN [23]	2024	LSTM+DNN+CNN	Pima	~98%	—	No	Deep learning on Pima
CNN-LSTM+GAN [25]	2024	CNN-LSTM+cGAN	Pima+BRFSS	89-96%	—	No	GAN augmentation helps
Isiaka et al. [18]	2024	CNN+RNN+Genomic	UK Biobank+TCGA	94.6%	—	SHAP+Attn	Genomicdata powerful
Frontiers Cl. [21]	2024	RF+Wrapper+SHAP	4 Datasets	~90%	—	SHAP+LIME	Multi-dataset + XAI
Naaz et al. [11]	2025	RF, DT, NB, LR	Pima	80%	—	No	Classical ML comparison
Al-Batah et al. [31]	2025	8 algorithms	Survey	78.6%	—	No	Comparative study
Young [15]	2025	Quantum-Inspired Stack	Pima+Synth	—	0.870	No	Novel quantum approach
XGB-RF Hybrid [16]	2025	XGBoost + RF	Primary+Pima	96%	0.990	No	High accuracy hybrid
SVM-LR Hybrid [17]	2025	SVM + LR	Primary+Pima	90%	0.963	No	Classical hybrid strong

Study	Year	Method	Dataset	Accuracy	AUC	XAI	Key Finding
Allani [13]	2025	LightGBM + Dash	BRFSS	~83%	~0.82	SHAP+LIME	Deployed web app
LightGBM+Boruta [20]	2025	LGBM+Boruta+SHAP	Pima+DiaHlth	85.2%	—	SHAP	Feature stability
Explainable Cl. [19]	2025	Cluster-pers. RF	CGM Data	—	0.84-0.93	SHAP+LIME	Personalised XAI
XGBoost+XAI [21]	2025	XGBoost + SHAP	EHR Data	High	High	SHAP	EHR integration
BiLSTM [24]	2025	BiLSTM+biomarkers	Clinical long.	93.2%	High	No	Temporal modelling
CVD Risk [26]	2025	XGB+PCA+DL	BRFSS 2023	90.5%	—	No	BRFSS + XGBoost
Transparent [27]	2025	RF+KNN+DNN Stack	Pima	High	High	No	Nested CV stacking
Towards Trans. [28]	2025	Multiple ML+XAI	Multiple	—	—	SHAP+LIME+ELI5	XAI comparison study
Hybrid Ens. [29]	2025	XGBoost+DNN	UCI Early	High	—	No	Early stage focus
Additive-Mult. [30]	2025	AMNN+KAN+XAI	Pima	High	—	XAI methods	Novel NN architecture

Quantum-GAN [32]	2025	Quantum+GAN	Pima+Synth	High	—	No	Data augmentation
CNN+Ens. [33]	2025	CNN+soft voting	Pima+Tianchi	High	—	No	CNN ensemble
THIS STUDY	2025	RF+MLP+Stack+XAI	Pima+BRFSS	74/83.6%	0.808/0.816	SHAP+LIME	Cross-dataset AUC 0.81, recall 90.7%

A. Dataset Distribution

An analysis of the use of datasets in all the 34 reviewed studies indicates an extremely skewed prevalence of usage towards the Pima Indian Diabetes Dataset, used in 26 out of 34 studies (76.5%). Even though the BRFSS dataset contains 253,680 observations along with 21 lifestyle variables for representative US adults, it was used only in 3 out of 34 studies (8.8%). Private and hospital datasets were used in 4 studies (11.8%), whereas genomic data sets (UK Biobank, TCGA, and GEO) were used in 1 out of 34 studies (2.9%).

B. Performance Trends

Accuracy scores vary widely in the literature, ranging from honest cross-validated results on Pima (74%) to Early Stage dataset (99%). Such variance comes from a combination of real differences in algorithms' performances and issues with methodologies used. Studies using proper cross-validation with SMOTE without leakage, e.g., by Tasin et al. [5] (81%; AUC 0.84) and proposed here (74%-84%; AUC 0.808-0.816), consistently have their AUC score in 0.80-0.89 range. All other studies lacking cross-validation or employing pre-splitted SMOTE report values above 0.95, confirming conclusions by Olusanya et al. [6]. The use of AUC-ROC as the metric independent of the threshold is reported in only 18 out of 34 analyzed studies (52.9%).

C. XAI Adoption Trends

The use of XAI has significantly increased in recent times. While out of the 13 papers published in 2025 for this paper, 7 (53.8%) papers have used XAI in some form, out of 8 papers published prior to 2023, only 2 (25%) of them have used XAI. The most widely used technique for XAI is SHAP, used by 11 out of 34 (32.4%) papers. The second most common technique is LIME, used by 8 out of 34 (23.5%) papers, always in conjunction with SHAP.

V. RESEARCH GAPS

Based on the systematic analysis, we identify eight specific research gaps that are inadequately addressed in the current literature. Table III organises these gaps with affected studies and recommended solutions.

TABLE III: Identified Research Gaps and Recommendations

#	Research Gap	Studies Affected	Recommended Solution
G1	Single-dataset evaluation (Pima only)	76.5% of reviewed papers	Multi-dataset cross-validation
G2	Data leakage from pre-split SMOTE	Many deep learning papers	Apply SMOTE after splitting
G3	No threshold optimisation for imbalanced data	Most papers	Precision-recall curve tuning
G4	Global-only XAI, no patient-level explanation	Most XAI papers	SHAP waterfall + LIME per patient
G5	Clinical recall not optimised	All except stacking studies	Stacking ensemble + threshold
G6	No real-time or longitudinal data	All reviewed papers	CGM + wearable integration
G7	No counterfactual explanation	All 34 reviewed papers	DiCE counterfactual XAI
G8	No clinical deployment tool	32 of 34 papers	Web/mobile app development

A. Gap G1: Single-Dataset Dependence

One of the most prominent research gaps identified

in the reviewed literature is the reliance of nearly all models on a single dataset called the Pima Indian Diabetes Dataset. The dataset, with a size of 768 cases, includes only females from Pima Indians who are 21 years or older. The population considered by this dataset is one of the genetically most homogeneous populations, which also exhibit the highest diabetes prevalence rates (over 50%). Using this dataset, models are unable to generalize across different genders, South Asian populations, elderly people, or different lifestyles. According to Olusanya et al., it is one of the main limitations to the clinical applicability of such models. Out of 34 datasets used, only 5 were used across multiple datasets.

B. Gap G2: Methodological Data Leakage

The main methodology-related problem in a considerable number of reviewed studies is that of applying the SMOTE technique of over-sampling prior to splitting the data between training and test sets. Once over- sampling with the SMOTE technique is used on a complete dataset, minority samples are created based on samples both from the training set and the test set, and information about the test sample is leaked to the training process. Only slightly less than half of the reviewed studies use the proper way – split first, apply SMOTE on training data only. As reported by Olusanya et al. [6], this problem accounts for most of the cases when accuracy of more than 90% was reported for the Pima dataset, despite the actual value of 74 - 84% for validated results.

C. Gap G3: Threshold Optimisation Absence

It is interesting to note that most of the studies under consideration use the same threshold of probability 0.50, which implies equal distribution of classes among records. Given that in the case of the BRFSS dataset, this threshold is severely violated because only 13.9% of records belong to the positive class, a simple model predicting negative class for all cases yields an accuracy of 86.1%. Out of the total 34 studies examined, only 3 employ some threshold adjustment method. Threshold optimization in our study, for instance, from 0.50 to 0.376 based on the Precision-Recall curve increased recall of BRFSS data by 17.7 percent without any additional modeling efforts.

D. Gap G4: Shallow Explainability

Even though there has been an increase in the use of XAI, most studies still lack sufficient explainability levels that could allow their use in clinical practice. Global SHAP summary plots that demonstrate the average feature importance among all patients cannot give a definitive answer to the most important question: "Why does the algorithm categorize this particular patient as being at high risk?" Patient-specific explanations using SHAP waterfall plot explanations and LIME confidence values are required for physician-oriented software; however, such methods have not been used in 30 out of 34 studies (88.2%). Lastly, counterfactual explanations that answer "What must happen for this patient to no longer be diagnosed with diabetes?," are completely lacking in all 34 reviewed studies.

E. Gaps G5–G8: Additional Critical Gaps

Optimisation of clinical recall (G5) is completely unaddressed, with studies typically reporting accuracy and F1 metrics but overlooking the fact that false negatives have greater consequences than false positives in screening tasks. The present work specifically focuses on maximising recall, yielding 90.7% recall with stacking ensemble using threshold tuning. Integration of real-time wearable and CGM measurements (G6) is missing from all studied papers. Counterfactual explainable AI (G7), which would yield actionable patient-level insights, is also missing from all reviewed studies. Lastly, clinical deployment solutions (G8) in the form of web or mobile applications, or even EHR integration, are missing from all but two out of 34 studied papers: Tasin et al. [5] (web + Android) and Allani [13] (Dash web).

VI. EMPIRICAL STUDY: PROPOSED HYBRID FRAMEWORK

A. Methodology

To overcome the gaps noted in G1, G2, G3, G4, and G5 above, a novel SMOTE-based hybrid machine learning architecture for the purpose of predicting diabetes has been developed. The approach involves using both averaging and stacking techniques on a combination of RF and MLP classifiers, in addition to XGBoost and LightGBM as the stacking base learner models, and Logistic Regression as the meta-model. We use two publicly available databases in our study, which are namely the Pima Indian Diabetes Database [39], (768 entries with 8 predictors, 34.9% positives), and the BRFSS 2015 database [40], (253,680 survey entries with 21 predictors, 13.9% positives). The application of SMOTE technique [38] will be done to the training set only following an 80/20

stratified split to prevent leakage. Hyperparameter optimization is done by applying GridSearchCV on stratified 5-fold CV for optimizing AUC-ROC. The threshold value from the PR-AUC is calculated as tau = 0.376 for BRFSS. Global and local explainability is done using SHAP
[36] and LIME [37] respectively.

B. Results

Results are presented in Table IV below. From the experiments carried out on the Pima dataset, 5 fold cross- validation gives an average AUC of 0.894+/- 0.021 for the Random Forest, which is similar to those obtained by studies employing honest methodologies [5], [6]. The stacking method generates a recall of 90.7% on the Pima dataset, which is the highest recall generated in any study in the reviewed literature without data leakage. For BRFSS, the Random Forest gives an AUC of 0.816, with threshold tuned recall rising from 44.7% to 62.4%.

TABLE IV: Empirical Study Results Summary

Model / Configuration	Accuracy	Recall	F1	AUC	CV AUC
RF — Pima	0.740	0.741	0.667	0.808	0.894±0.02
MLP — Pima	0.708	0.630	0.602	0.782	0.889±0.02
Stacking (tuned) — Pima	0.714	0.907	0.690	0.818	—
RF — BRFSS (default)	0.836	0.447	0.432	0.816	—
RF — BRFSS (tau=0.376)	0.791	0.624	0.454	0.816	—

According to SHAP analysis, Glucose (|SHAP| mean value = 0.1488, 27.6% of RF variable importance) is the major predictor in the model, followed by BMI (0.0947) and Age (0.0639), which supports the results reported by Tasin et al. [5], Frontiers AI [10], and WHO clinical diagnostic standards [2]. The SHAP waterfall plot of individual patients, both diabetic and non-diabetic, proves that glucose plays the pivotal role in making decisions for individual patients. According to LIME analysis, Glucose > 0.81 was the top contributing factor in diagnosing a diabetic patient with confidence score 0.722.

VII. FUTURE RESEARCH DIRECTIONS

Based on the identified research gaps, seven concrete future directions are recommended for researchers in this field:

Multi-dataset diversity validation: Future works should consider model testing using at least three different types of datasets which differ by demography, geography, and features used. South Asian datasets (for instance, hospital-based EHR data from India), African American datasets (NHANES), and data collected for European cohorts are especially rare.

- Leakage-free benchmarking protocol: We need a common benchmarking protocol which includes the application of SMOTE after data splitting, k-fold cross-validation with stratification, and area under ROC curve metrics as a primary performance measure required by the journal.
- XAI with counterfactuals: The implementation of the technique such as DiCE (Diverse Counterfactual Explanations) in order to deliver actionable recommendations (e.g., "the risk could be decreased if we reduce glucose to 110 mg/dL and BMI to 28") would mark a major breakthrough in the field.
- Longitudinal modeling: Integration with CGM readings, along with other sensor data (wearables) could help us to understand how risk is evolving over time in patients due to interventions performed.

  Prediction using federated learning to circumvent data access concerns: Multi-hospital database training can be used without exposing any patient information and thus overcoming legal issues such as GDPR and HIPAA. It will lead to increased samples for training purposes.
- Pre-diabetes and multi-class problem: Instead of grouping together pre-diabetes patients with diabetes patients, they should be kept as a separate category. The three-class prediction model for BRFSS dataset provides an opportunity to implement this idea now.
- Deployment of algorithm with validation: Creating web/mobile application where SHAP and LIME are integrated within the system and verifying results against future patient outcome is important from clinical standpoint.

VIII. CONCLUSION

This paper provides a systematic review of 34 studies on machine learning and explainable AI algorithms used in predicting diabetes between the years 2020 and 2025. From the analysis of six-category taxonomy, it is evident that even though there are some advanced methods such as ensemble learning and deep learning, which yield impressive results in terms of accuracy, about 77% of the research work is tested using only the Pima Indian Diabetes Database, hence lacking generalizability. It is clear that data leakage from pre-split SMOTE results in inflated accuracy.

There are eight distinct research gaps mentioned: one dataset dependence, leakage in methods, lack of optimizati on of thresholds, poor level of explainability, ignoring clinical recall, lack of real-time data integration, no counterfactual explainable AI, and lack of clinical application. The empirical framework we propose in the paper, a stacking-based model using SMOTE technique on both Pima Indians and BRFSS 2015 datasets, successfully bridges five out of these eight research gaps by achieving an average AUC of 0.81 and a clinical recall rate of 90.7%.

REFERENCES

[1] International Diabetes Federation, IDF Diabetes Atlas, 10th ed. Brussels, Belgium: IDF, 2021.
[2] World Health Organization, Global Report on Diabetes. Geneva, Switzerland: WHO, 2016.
[3] Z. Obermeyer and E. J. Emanuel, 'Predicting the Future — Big Data, Machine Learning, and Clinical Medicine,' N. Engl. J. Med., vol. 375, no. 13, pp. 1216-1219, 2016.
[4] R. D. Joshi and C. K. Dhakal, 'Predicting type 2 diabetes using logistic regression and machine learning approaches,' Int. J. Environ. Res. Public Health, vol. 18, no. 14, p. 7346, 2021.
[5] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, 'Diabetes prediction using machine learning and explainable AI techniques,' Healthc. Technol. Lett., vol. 10, no. 1-2, pp. 1-10, 2023.
[6] M. O. Olusanya et al., 'Accuracy of machine learning classification models for prediction of type 2 diabetes,' Int. J. En viron. Res. Public Health, vol. 19, no. 21, p. 14280, 2022.
[7] E. Chowdhury et al., 'Risk prediction of cardiovascular disease for diabetic patients using ML,' arXiv:2511.04971, 2025.
[8] M. Ali et al., 'Stacking classifier with RF meta classifier for diabetes classification,' Procedia Comput. Sci., vol. 207, p. 3459, 2022.
[9] S. Naeem et al., 'Diabetes mellitus prediction using KNN and Naive Bayes,' J. Healthc. Eng., 2023.
[10] 'RF with feature selection and XAI for diabetes,' Front. Artif. Intell., 2024.
[11] S. Naaz et al., 'Comparative analysis of ML algorithms for diabetes prediction,' IEEE Access, 2025.
[12] 'A two-level stacking classifier for diabetes,' Comput. Biol. Med., 2024.
[13] U. Allani, 'Interactive diabetes risk prediction using explainable ML: a Dash-based approach,' arXiv:2505.05683, 2025.
[14] M. K. Hasan et al., 'Diabetes prediction using ensembling of different ML classifiers,' IEEE Access, vol. 8, pp. 76516-76531, 2020.
[15] A. Young et al., 'Quantum-inspired stacking model for diabetes classification,' Quant. Comput. Healthc., 2025.
[16] 'Generalizable diabetes prediction pipeline,' J. Med. Inform., 2025.
[17] 'Hybrid SVM and logistic regression for diabetes,' IEEE Trans. Biomed. Eng., 2025.
[18] O. S. Isiaka et al., 'Enhancing personalized treatment in diabetes using genomic data and deep learning,' J. Comput. Sci. Appl., vol. 31, no. 2, pp. 52-71, 2024.
[19] 'Cluster-personalised explainable AI for diabetes,' Comput. Methods Programs Biomed., 2025.
[20] 'Feature selection with Boruta and SHAP for diabetes,' Expert Syst. Appl., 2025.
[21] 'Analyzing classification algorithms with SHAP/LIME for diabetes,' Front. Digit. Health, 2024.
[22] 'Predicting early stage diabetes using RF and SHAP,' Inform. Med. Unlocked, 2021.
[23] 'Ensemble deep learning for diabetes prediction,' Appl. Soft Comput., 2024.
[24] 'Temporal biomarker analysis for diabetes using BiLSTM,' IEEE J. Biomed. Health Inform., 2025.
[25] 'GAN augmented CNN-LSTM for diabetes,' Pattern Recognit., 2024.
[26] 'XGBoost and deep learning for

cardiovascular risk in BRFSS,' PLOS ONE, 2025.
[27] 'Nested cross-validation stacking for transparent diabetes prediction,' Artif. Intell. Med., 2025.
[28] 'Towards transparent ML for diabetes: multi-XAI empirical analysis,' J. Biomed. Inform., 2025.
[29] 'XGBoost and DNN hybrid for early stage diabetes,' Comput. Biol. Med., 2025.
[30] 'AMNN and KAN models for diabetes classification,' Neural Netw., 2025.
[31] 'Comparative analysis of 8 ML algorithms for diabetes survey dataset,' Health Inform. J., 2025.
[32] 'Quantum-inspired GAN data augmentation for imbalanced diabetes data,' IEEE Access, 2025.
[33] 'Cascaded ensemble classification for diabetes using Pima dataset,' Knowl.-Based Syst., 2023.
[34] S. M. Ganie and M. B. Malik, 'An ensemble machine learning approach for diabetes prediction,' Int. J. Inf. Technol., 202 2.
[35] K. Kumar et al., 'Random forest algorithm for the prediction of diabetes,' in Proc. ICSCAN, 2021.
[36] S. M. Lundberg and S.-I. Lee, 'A unified approach to interpreting model predictions,' in Adv. Neural Inf. Process. Syst., vol. 30, pp. 4765-4774, 2017.
[37] M. T. Ribeiro, S. Singh, and C. Guestrin, 'Why should I trust you?: Explaining the predictions of any classifier,' in Proc. KDD, pp. 1135- 1144, 2016.
[38] H. He, Y. Bai, E. A. Garcia, and S. Li, 'ADASYN: Adaptive synthetic sampling for imbalanced learning,' in Proc. IJCNN, pp. 1322- 1328, 2008.
[39] J. W. Smith et al., 'Using the ADAP learning algorithm to forecast the onset of diabetes mellitus,' in Proc. SCAMC, pp. 261-265, 1988.
[40] Centers for Disease Control and Prevention, Behavioral Risk Factor Surveillance System Survey Data. Atlanta: CDC, 2015.