COMPREHENSIVE LITERATURE REVIEW: EXPLAINABLE DIABETES RISK PREDICTION THROUGH MACHINE LEARNING, SHAP ANALYSIS, AND COUNTERFACTUAL EXPLANATIONS

[1]D Siva Raja Kumar, [2]Makkapati srikala, [3]Nakka Venkatesh
[1,2,3]Assistant Professor, Department of CSE (AI&ML), CMR Engineering College, Hyderabad.
[1]d.sivarajakumar@cmrec.ac.in, [2]yadlasrikalanath@cmrec.ac.in, [3]venkatesh.n@cmrec.ac.in

ABSTRACT:

Diabetes mellitus is a rapidly growing global health challenge that requires accurate, interpretable, and actionable risk prediction systems for early detection and prevention. This study presents a comprehensive review of explainable diabetes risk prediction approaches that integrate machine learning, SHAP (SHapley Additive exPlanations), and counterfactual explanations. Recent advances in machine learning algorithms, particularly ensemble models such as XGBoost, Random Forest, and Gradient Boosting, have significantly improved predictive performance in diabetes risk assessment. However, the black-box nature of these models limits their clinical adoption. SHAP provides transparent global and local explanations by quantifying feature contributions, while counterfactual explanations generate actionable "what-if" scenarios that support personalized intervention planning. The review synthesizes recent developments in interpretable diabetes prediction, highlights the complementary strengths of SHAP and counterfactual reasoning, and discusses their applications in clinical decision support. Furthermore, current challenges, research gaps, and future directions are identified to facilitate the development of trustworthy, patient-centric, and clinically deployable diabetes risk prediction systems.
Keywords: *Diabetes Risk Prediction, Machine Learning, Explainable Artificial Intelligence (XAI), SHAP Analysis, Counterfactual Explanations, XGBoost, Clinical Decision Support Systems, Personalized Healthcare*

INTRODUCTION

Machine learning models have the potential to transform healthcare by enabling the construction of decision support systems, yet a major challenge is the lack of transparency and accountability, as many models do not provide understandable explanations for their recommendations [1]. Diabetes mellitus, a chronic metabolic disorder affecting millions globally, represents an ideal application domain for machine learning-based risk prediction systems. However, the inherent opacity of complex predictive models presents significant barriers to clinical adoption. In domains such as medical and healthcare, the interpretability and explainability of machine learning and artificial intelligence systems are crucial for building trust in their results, where errors from these systems can have severe and even life-threatening consequences for patients [2].

This literature review synthesizes recent advances in explainable artificial intelligence (XAI) as applied to diabetes risk prediction, with particular emphasis on two powerful interpretability frameworks: SHAP (SHapley Additive exPlanations) analysis and counterfactual explanations. These complementary approaches address the critical need for transparent, clinically actionable machine learning models in diabetes screening and prevention.

THE DIABETES PREDICTION IMPERATIVE AND MACHINE LEARNING FOUNDATIONS

An estimated 425 million people globally have diabetes, accounting for 12% of the world's health expenditures, and the number continues to grow, placing a huge burden on the healthcare system [3]. Early detection of diabetes risk is essential for implementing preventive interventions before disease onset. Diabetes prediction and treatment are important to reduce and avoid associated risks, with Machine Learning (ML) and Deep Learning (DL) frameworks proven effective in enhancing prediction accuracy [4].

The integration of Machine Learning (ML) and Artificial Intelligence (AI) has revolutionized diabetes care, offering innovative approaches to prediction, monitoring, and personalized management, with supervised learning models such as Random Forest and Support Vector Machines consistently demonstrating high accuracy and reliability in predicting diabetes risk [5]. Ensemble learning methods, particularly Gradient Boosting, emerged as superior techniques for predictive performance, while deep learning models, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), proved effective in analyzing unstructured data such as medical images and time-series glucose data [5].

Multiple machine learning architectures have been evaluated for diabetes prediction. Extreme Gradient Boosting (XGBoost), support vector machine (SVM), logistic regression (LR), and random forest (RF) models have been trained to predict 5-year type 2 diabetes mellitus (T2DM) risk, with both LR and RF models demonstrating effectiveness in the Chinese population [6]. The XGBoost model has demonstrated superior performance with an AUC value of 0.939, surpassing RF, SVM, DT, ANNs, Naive Bayes, and LR models in pre-diabetes risk prediction [7].

EXPLAINABLE ARTIFICIAL INTELLIGENCE: ADDRESSING THE BLACK BOX CHALLENGE

Explainable Artificial Intelligence (XAI) is emerging as a critical field to address the "black box" nature of many machine learning (ML) models, where while these models achieve high predictive accuracy, their opacity undermines trust, adoption, and ethical compliance in critical domains such as healthcare, finance, and autonomous systems [8]. The challenge of interpretability is particularly acute in healthcare, where clinicians and patients require understanding of the factors driving individual predictions.

Advanced machine learning models have received wide attention in assisting medical decision making due to the greater accuracy they can achieve, however their limited interpretability imposes barriers for practitioners to adopt them [9]. This interpretability gap creates a false choice between accuracy and transparency. The main limitation of diabetes prediction is that accuracy and interpretability have often been addressed separately, with most studies focused on achieving high accuracy while ignoring the effect of demographic constraints and class imbalance in the datasets [10].

Recent advances in interpretable machine learning have begun bridging this divide. The application of Explainable Boosting Machine (EBM), Light Gradient Boosting Machine (LGBM), and other interpretable models has demonstrated that these approaches can maintain high performance while providing inherent interpretability without relying on external analytical tools, thus reducing interpretation biases [11]. An interpretable machine learning approach using nationwide cross-sectional data has proven effective in predicting pre-diabetic risk while quantifying the impact and significance of potential risks on pre-diabetes [7].

SHAP ANALYSIS: THEORETICALLY GROUNDED FEATURE ATTRIBUTION

Theoretical Foundations and Advantages

SHAP (SHapley Additive exPlanations) is an explainable artificial intelligence tool that extracts feature importance from datasets and explains the contribution of features towards detection/prediction output both locally and globally [12]. Built on game-theoretic principles, SHAP provides a unified framework for interpreting machine learning predictions by computing each feature's contribution to shifting the model's output from the base value to the actual prediction.

SHAP provides stable feature importance across all machine learning models, particularly excelling in global explanation where LIME works efficiently on local interpretation, demonstrating that SHAP-based attribution provides a stable and consistent approach to XAI [13]. SHAP merged with Local Interpretable Model-agnostic Explanations (LIME) forms a comprehensive interpretation layer for predictive outcomes, helping identify feature influence across the dataset as well as the reasoning behind individual predictions [14].

SHAP Applications in Diabetes Risk Prediction

SHAP has become the dominant explainability method in diabetes prediction research. Using the Pima Indian Diabetes dataset with an AutoML-based ensemble model achieving 85.01% accuracy, researchers applied XAI techniques including SHAP to identify critical risk factors such as glucose and BMI, providing both global and patient-specific insights into critical risk factors [15]. These XAI methods enable transparent and actionable predictions, supporting clinical decision-making and allowing clinicians to explore feature importance and test hypothetical scenarios through interactive applications [15].

SHAP analysis provided that the glucose, BMI, and insulin are the most important features for diabetes discrimination, which is in line with the clinical symptoms of diabetes [10]. Consistently across XAI methods including SHAP, Glucose emerged as the most influential predictor, followed by BMI and Age, aligning with established clinical risk factors [16].

Global feature importance analysis via SHAP consistently identified Glucose, BMI, and Age as key predictors in diabetes risk prediction using Random Forest classifiers on the Pima Indians Diabetes dataset [17]. Individual feature importance using SHAP value and the XGBoost model showed that previous GDM diagnosis, maternal age, body mass index, and gravidity play a vital role in gestational diabetes mellitus diagnosis [18].

Multi-Level SHAP Interpretation: Summary, Dependence, and Force Plots

SHAP provides multiple visualization techniques for different interpretability needs. SHAP summary analysis for XGBoost models showed that urine output on day 1, age, blood urea nitrogen and body mass index were the top four contributors to sepsis mortality prediction, while SHAP dependence analysis demonstrated insightful nonlinear interactive associations between factors and outcome, and SHAP force analysis provided individual samples for model prediction [19]. These hierarchical interpretation methods enable understanding at both population and individual levels.

The SHAP summary plot showed that all features predicted frailty in order of importance, with cognitive function being considered the most important predictive feature, where the poorer the cognitive function, nighttime sleep duration, body mass index (BMI), and grip strength, the higher the risk of frailty, while the SHAP individual force plot clearly shows the model's prediction in the individual decision-making process [20].

COUNTERFACTUAL EXPLANATIONS: ACTIONABLE DECISION SUPPORT FOR PATIENTS

Conceptual Framework and Applications

Counterfactual explanations provide a link between what could have happened had input to a model been changed in a particular way, with modern approaches to counterfactual explainability in machine learning drawing connections to the established legal doctrine in many countries, making them appealing to fielded systems in high-impact areas such as finance and healthcare [21].

A critical distinction sets counterfactual explanations apart from feature importance methods: they are inherently prescriptive and actionable. Counterfactual explanations aim to show how changing the feature values of an entity can affect its position within the ranking defined by the model, offering actionable insights that enable individuals to actively manage lifestyle choices and potentially reduce their diabetes risk [22]. This patient-centric focus aligns with precision medicine principles.

Proximity-Constrained Counterfactuals for Diabetes Risk Assessment

Recent work has directly addressed diabetes risk counseling through counterfactual explanations. A pragmatic approach to generate proximity-constrained counterfactuals for assessing diabetic risk synergistically combines the capabilities of Local Interpretable Model-Agnostic Explanations and Diverse Counterfactual Explanations, with methodology focusing on perturbing the most influential features identified by different AI explainers to reverse decisions while maintaining proximity to the decision boundary [22]. The resulting counterfactuals offer actionable insights, enabling individuals to actively manage lifestyle choices and potentially reduce their diabetes risk, addressing constraints and integrating human assessments for a more comprehensive evaluation [22].

SHAP and Counterfactuals: Complementary Perspectives

Investigation of the relationship between SHAP and Counterfactual Explanations for understanding machine learning interpretability from both descriptive and prescriptive perspectives found that counterfactuals can be used to effectively understand how sensitive the machine learning model is to feature variation, while SHAP provides stable feature importance across all machine learning models [13]. Additional findings demonstrated that combining SHAP-based attribution with counterfactual sensitivity analysis provides a complete and reliable approach to XAI, which can improve interpretability as well as decision support in real-world applications [13].

This complementarity is essential: SHAP explains "why" a prediction was made (through feature contributions), while counterfactuals explain "what would change" the prediction (through actionable scenarios). Interactive robust explanation systems generate personalized counterfactual region explanations, empowering users to explore and compare multiple counterfactual options and develop a personalized actionable plan for obtaining their desired outcome [23].

Calibration and Robustness in Counterfactual Generation

The rapid adoption of artificial intelligence methods in healthcare is coupled with the critical need for techniques to rigorously introspect models and ensure that they behave reliably, with counterfactual explanations that synthesize small, interpretable changes to a given query while producing desired changes in model predictions becoming popular [24]. However, this under-constrained, inverse problem is vulnerable to introducing irrelevant feature manipulations, particularly when the model's predictions are not well-calibrated [24].

Advanced calibration-based approaches address this challenge. The TraCE (training calibration-based explainers) technique utilizes a novel uncertainty-based interval calibration strategy for reliably synthesizing counterfactuals, demonstrating superiority over several state-of-the-art baseline approaches through rigorous empirical studies [24]. Robust counterfactual methods remain valid across different patient profiles and clinical scenarios.

INTEGRATED FRAMEWORKS: COMBINING MULTIPLE EXPLAINABILITY APPROACHES

XGBoost and SHAP: The Dominant Clinical Model Architecture

Using the XGBoost algorithm to build prediction models for diabetic retinopathy and then employing the Shapley Additive exPlanation (SHAP) method to visually explain the results has rendered the output of the XGBoost model clinically interpretable, with the model showing that an HbA1c value greater than 8%, nephropathy, serum creatinine value greater than 100 µmol/L, insulin treatment and diabetic lower extremity arterial disease were associated with increased diabetic retinopathy risk [25].

This XGBoost-SHAP combination has become the reference standard across diabetes prediction applications. Using XGBoost with Shapley Additive Explanations analysis revealed gender, poverty-to-income ratio, sleep duration, smoking status, educational levels, race, age, high cholesterol, hypertension, and insulin use as the most influential predictors of depression in adults with type 2 diabetes mellitus [26]. A streamlined XGBoost model incorporating only these 10 most influential variables achieved an area under the curve of 0.886, closely matching the predictive capability of the full model, with the deployed web-based tool enabling rapid and individualized estimation of risk using routinely available clinical and demographic information [26].

LIME and SHAP: Comparative Strengths and Complementarity

SHAP is powerful in global explanation where LIME works efficiently on local interpretation, with these two different tools performing similarly on detection accuracies while showing strength and weakness in explaining model's behaviour both locally and globally [12]. Using both SHAP and LIME, research has determined that both the fivefold cross-validation and independent testing show that the proposed method significantly outperforms state-of-the-art methods while maintaining excellent model interpretability, helping to uncover potential predictive patterns and enabling clinicians to better understand the model's decision-making process [27].

LIME showed superior fidelity whereas SHAP showed improved sparsity and consistency across models, facilitating a comprehensive understanding of trait importance in obesity prediction [28]. SHAP and LIME improved interpretability by analyzing feature contributions in obesity prediction models, with hybrid stacking and voting models outperforming other ensemble methods while preserving interpretability [29].

DIABETES RISK PREDICTION MODELS: METHODOLOGICAL COMPARISONS

Feature Identification and Selection

Robust diabetes prediction models systematize feature engineering and selection processes. Using logistic regression based on variables of physical measurement and questionnaire to identify risk factors for type II diabetes mellitus, combined with machine learning classifiers including decision tree, random forest, AdaBoost, and extreme gradient boosting decision tree, researchers found that BMI was the most significant feature, followed by age, waist circumference, systolic pressure, ethnicity, smoking amount, fatty liver, hypertension, physical activity, drinking status, dietary ratio, drink amount, smoking status, and diet habit [3].

LASSO regression algorithm was used to conduct feature selection, ultimately identifying nine non-zero coefficient features associated with pre-diabetes, including age, TG, TC, BMI, Apolipoprotein B, TP, leukocyte count, HDL-C, and hypertension [7]. These systematic selection procedures ensure models incorporate clinically relevant predictors while avoiding overfitting.

Model Performance Benchmarking Across Algorithms

When comparing extreme gradient boosting, support vector machine, logistic regression, and random forest models in predicting 5-year type 2 diabetes mellitus risk, the logistic regression model showed good accuracy with respective areas under the ROC of 0.914 and 0.913 in training and validation sets, while the random forest model exhibited favorable AUCs of 0.998 and 0.838, with logistic regression displaying good fit for low-risk patients and random forest exhibiting satisfactory fit for both low- and high-risk patients [6].

Deep learning approaches have also been systematically evaluated. An innovative Ensemble Deep Learning model using Deep Learning architectures such as Artificial Neural Networks (ANN), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN) integrated with an ensemble learning-based stacking model, when implemented based on a stack ensemble model that applies meta-level models including stack-ANN, stack-CNN, and stack-LSTM, achieved robust performance with accuracy scores of 99.51%, 98.81%, and 98.45% on three different diabetes datasets [30].

SPECIALIZED APPLICATIONS: DIABETES COMPLICATIONS AND COMORBIDITIES

Diabetic Retinopathy Risk Prediction

To develop a self-reportable risk assessment tool for elderly type 2 diabetes mellitus patients evaluating risks of diabetic nephropathy, retinopathy, peripheral neuropathy, and diabetic foot using machine learning, five machine learning algorithms (XGBoost, LightGBM, Random Forest, TabPFN, CatBoost) were applied with SHAP analysis for feature importance, with the analysis identifying 33 risk factors including 6 shared risk factors and 27 unique risk factors [31].

A diabetic retinopathy risk prediction model integrating machine learning models and SHAP was established using the CatBoost algorithm, with SHAP model outputs interpreted to show that ALBCR, HbA1c, UPR24, NEPHROPATHY and SCR were positively correlated with diabetic retinopathy, while CP, HB, ALB, DBILI and CRP were negatively correlated [32].

Diabetic Nephropathy and Kidney Disease

Machine learning algorithms were employed to establish forecasting models for diabetic kidney disease risk prediction in type 2 diabetes mellitus patients, with the Light Gradient Boosting Machine (LightGBM) model demonstrating the highest accuracy in predicting 3-year risk of diabetic kidney disease onset, and using the importance ranking in the random forest package, the variables of age, urinary albumin-to-creatinine ratio, serum cystatin C, estimated glomerular filtration rate, and neutrophil percentage were selected as the predictors [33].

Depression Comorbidity in Diabetes

Among 2,837 participants with type 2 diabetes mellitus, 449 (15.8%) were identified as having comorbid depression, with the XGBoost model demonstrating the highest discriminative ability with a validation area under the receiver operating characteristic curve of 0.888, surpassing the performance of other algorithms [26]. This demonstrates how interpretable ML extends beyond glucose control to mental health outcomes crucial for diabetes self-management.

METHODOLOGICAL CONSIDERATIONS AND CLINICAL IMPLEMENTATION

Data Preprocessing and Class Imbalance Handling

Type-2 diabetes prediction frameworks undergo preprocessing techniques such as normalization, missing value imputation, correlation-based feature selection, and class imbalance handling using SMOTE, with multiple machine learning algorithms—including Logistic Regression, Random Forest, Support Vector Machine, Gradient Boosting, and Extreme Gradient Boosting (XGBoost)—trained and evaluated to identify the best-performing model [34].

Cross-Validation and External Validation

Robust methodologies employ rigorous validation strategies. Stability testing across similar instances revealed robust interpretability with minimal variance when applying SHAP and LIME to diabetes prediction, demonstrating that beyond predictive accuracy, Random Forest models combined with SHAP and LIME provide reliable and interpretable insights for diabetes prediction [17].

Clinical Decision Support System Integration

An interactive Streamlit application was developed to allow clinicians to explore feature importance and test hypothetical scenarios, with cross-validation confirming the model's robust performance across diverse datasets, demonstrating the integration of AutoML with XAI as a pathway to achieving accurate, interpretable models that foster transparency and trust while supporting actionable clinical decisions [15].

RESEARCH GAPS AND FUTURE DIRECTIONS

Limited Generalization and External Validation

Despite numerous successes, significant gaps remain. Despite advancements in machine learning for diabetes care, challenges such as data imbalance, limited external validation, and the need for explainable AI frameworks persist, underscoring the necessity for methodological rigor and standardization [5]. Many diabetes prediction models remain validated only on single institutional datasets, limiting generalizability.

Integrating Multimodal and Real-Time Data

Type-2 diabetes prediction frameworks integrating multimodal data including clinical parameters such as glucose level, blood pressure, insulin, BMI, demographic attributes, and lifestyle indicators including dietary habits, physical activity, stress level, and sleep patterns may show associations with improved predictive performance compared to clinical data alone [34]. Yet systematic integration of wearable and continuous glucose monitoring data remains underexplored.

Counterfactual Explanations at Scale

While counterfactual explanations show promise for patient engagement, computational scalability and robustness across diverse populations require further research. FACET, a novel explanation analytics system which supports a user in interactively refining counterfactual explanations for decisions made by tree ensembles, designs a novel type of counterfactual explanation called the counterfactual region, where unlike traditional counterfactuals which aim to find feature value changes that would result in a different predicted class label, this approach aims to identify changes that would bring predictions in line with a dynamic threshold determined by other data items [35]. Such advances could personalize diabetes prevention counseling.

Fairness and Health Equity

Emerging research addresses disparities in model performance across demographic groups. An integrated framework uncovers latent, non-linear associations through network-based exploration of nonlinear interactions combined with machine learning, with subgroup analysis showing stable predictions across racial groups, supporting fairness in diabetes risk prediction and offering a robust, interpretable, and equitable tool for precision diabetes risk modeling [36].

SYNTHESIS AND CONCLUSIONS

The convergence of three elements—machine learning for predictive power, SHAP for interpretable feature attribution, and counterfactual explanations for actionable guidance—has created a new paradigm for clinically adoptable diabetes risk prediction systems. In domains such as medical and healthcare, explainable artificial intelligence algorithms such as LIME and SHAP can provide explanations for complex machine learning models, improving trust in their predictions by providing feature importance and increasing confidence in the systems [2].

This research explores methodologies and frameworks to enhance the interpretability of ML models, focusing on techniques like feature attribution, surrogate models, and counterfactual explanations, and by balancing model complexity and transparency, highlights strategies to bridge the gap between performance and explainability [8]. The integration of XAI into ML workflows not only fosters trust but also aligns with regulatory requirements, enabling actionable insights for stakeholders [8].

The main advantage of the proposed interpretable machine learning approach is that prediction accuracy and interpretability can be achieved simultaneously which addresses the long-standing problem of diabetes prediction [10]. This achievement marks a transition from purely investigational machine learning applications to clinically deployable decision support tools.

Future diabetes risk prediction research should prioritize: (1) external validation across geographically and demographically diverse populations; (2) integration of continuous real-time biomarker data with traditional clinical measurements; (3) systematic comparison of SHAP-based global and counterfactual-based prescriptive approaches for patient engagement; (4) deployment of fully interpretable models alongside post-hoc explanation methods to understand the interpretability-accuracy frontier; and (5) evaluation of health outcomes when models are deployed with clinical integration and patient counseling. The field has achieved technical maturity in explainability; the next imperative is demonstrating clinical value and equitable impact.

REFERENCES

[1] M. Salimiparsa, "Counterfactual explanations for rankings," *Canadian AI*, 2023, doi: 10.21428/594757db.15b61c8c.

[2] S. Ali, F. Akhlaq, A. S. Imran, Z. Kastrati, S. M. Daudpota, and M. Moosa, "The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review," *Computers in Biology and Medicine*, 2023, doi: 10.1016/j.compbiomed.2023.107555.

[3] M. Xue, Y. Su, C. Li, S. Wang, and H. Yao, "Identification of potential type II diabetes in a large-scale chinese population using a systematic machine learning framework," *Journal of Diabetes Research*, 2020, doi: 10.1155/2020/6873891.

[4] A. S and N. S, "Comprehensive analysis of diabetes prediction using machine learning and deep learning approach," *2025 5th International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, 2025, doi: 10.1109/ICMNWC66779.2025.11354230.

[5] M. A. Alam, A. Sohel, K. Hasan, and M. A. Islam, "Machine learning and artificial intelligence in diabetes prediction and management: A comprehensive review of models," *Innovatech Engineering Journal*, 2024, doi: 10.70937/jnes.v1i01.41.

[6] H. Liu *et al.*, "Comparing the accuracy of four machine learning models in predicting type 2 diabetes onset within the chinese population: A retrospective study," *Journal of International Medical Research*, 2024, doi: 10.1177/03000605241253786.

[7] X. Li *et al.*, "Interpretable machine learning method to predict the risk of pre-diabetes using a national-wide cross-sectional data: Evidence from CHNS," *BMC Public Health*, 2025, doi: 10.1186/s12889-025-22419-7.

[8] D. Kulaklıoğlu, "Explainable AI: Enhancing interpretability of machine learning models," *Human-Computer Interaction*, 2024, doi: 10.62802/z3pde490.

[9] X. Gao, S. Alam, P. Shi, F. Dexter, and N. Kong, "Interpretable machine learning models for hospital readmission prediction: A two-step extracted regression tree approach," *BMC Medical Informatics and Decision Making*, 2023, doi: 10.1186/s12911-023-02193-5.

[10] S. Jayaprakash, A. Nellyet, and C. Ganesh, "Interpretable diabetes risk classification using optimized deep learning and baseline machine learning models," *International Conference on Independent Component Analysis and Signal Separation*, 2026, doi: 10.1109/ICASS69550.2026.11547809.

[11] Y. Duan *et al.*, "Intelligent diagnosis of kawasaki disease from real-world data using interpretable machine learning models." *Hellenic Journal of Cardiology*, 2024, doi: 10.1016/j.hjc.2024.08.003.

[12] A. Paul, S. Kumari, N. R. Navadia, and S. Sinha, "Analyzing the role of LIME and SHAP in explainable DoS attack detection for IoT systems," *2025 4th International Conference on Automation, Computing and Renewable Systems (ICACRS)*, 2025, doi: 10.1109/ICACRS67045.2025.11324170.

[13] R. S. S. Paramkusam and P. Mohanty, "Bridging local interpretability: Analyzing SHAP importance and counterfactual sensitivity for model transparency," *2026 4th International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA)*, 2026, doi: 10.1109/AIMLA67915.2026.11522491.

[14] K. C. Yarlagadda, "A dual-interpretability AI approach for high-stakes applications using SHAP and LIME," *2026 IEEE International Conference on Emerging Computing and Intelligent Technologies (ICoECIT)*, 2026, doi: 10.1109/ICoECIT68303.2026.11497917.

[15] R. Hasan, V. Dattana, S. Mahmood, and S. Hussain, "Towards transparent diabetes prediction: Combining AutoML and explainable AI for improved clinical insights," *Inf.*, 2024, doi: 10.3390/info16010007.

[16] U. Mitra, P. Sarkar, J. Mondal, and J. Kundu, "Enhancing interpretability in diabetics prediction: A comparative study of SHAP, LIME and permutation feature importance," *2025 AI-Driven Smart Healthcare for Society 5.0*, 2025, doi: 10.1109/IEEECONF64992.2025.10962890.

[17] M. A. Attalah, Purnawansyah, L. Syafie, A. R. Manga', L. N. Hayati, and Irawati, "Evaluating SHAP and LIME for interpretable diabetes risk prediction with random forest," *2025 9th International Conference On Electrical, Electronics And Information Engineering (ICEEIE)*, 2025, doi: 10.1109/ICEEIE66203.2025.11251687.

[18] N. Ali, W. Khan, A. Y. Ahmad, M. Masud, H. Adam, and L. Ahmed, "Predictive modeling for the diagnosis of gestational diabetes mellitus using epidemiological data in the united arab emirates," *Inf.*, 2022, doi: 10.3390/info13100485.

[19] B. He and Z. Qiu, "Development and validation of an interpretable machine learning for mortality prediction in patients with sepsis," *Frontiers in Artificial Intelligence*, 2024, doi: 10.3389/frai.2024.1348907.

[20] L. Yu, S. Cao, B. Song, and Y. Hu, "Predicting grip strength-related frailty in middle-aged and older chinese adults using interpretable machine learning models: A prospective cohort study," *Frontiers in Public Health*, 2024, doi: 10.3389/fpubh.2024.1489848.

[21] Verma *et al.*, "Counterfactual explanations and algorithmic recourses for machine learning: A review," *ACM Computing Surveys*, 2024, doi: 10.1145/3677119.

[22] M. F. Kabir, A. M. Mahakal, Y. Wang, A. M. R. Alsobeh, B. Al-Ahmad, and R. S. Alkhawaldah, “Proximity-constrained counterfactuals for explainable diabetes risk assessment,” *Applied Computational Intelligence and Soft Computing*, 2025, doi: 10.1155/acis/3424976.

[23] P. M. VanNostrand, D. M. Hofmann, L. Ma, B. Genin, R. Huang, and E. A. Rundensteiner, “Counterfactual explanation analytics: Empowering lay users to take action against consequential automated decisions,” *Proceedings of the VLDB Endowment*, 2024, doi: 10.14778/3685800.3685872.

[24] J. J. Thiagarajan, K. Thopalli, D. Rajan, and P. Turaga, “Training calibration-based counterfactual explainers for deep learning models in medical image analysis,” *Scientific Reports*, 2021, doi: 10.1038/s41598-021-04529-5.

[25] W. Li *et al.*, “Predictive model and risk analysis for diabetic retinopathy using machine learning: A retrospective cohort study in china,” *BMJ Open*, 2021, doi: 10.1136/bmjopen-2021-050989.

[26] Y. Tang *et al.*, “Development of an explainable machine learning model for predicting depression in adults with type 2 diabetes mellitus: A cross-sectional SHAP-based analysis of NHANES 2009–2023,” *Medicine*, 2026, doi: 10.1097/MD.0000000000047522.

[27] Y. Fan, X. Lu, and G. Sun, “IHCP: Interpretable hepatitis c prediction system based on black-box machine learning models,” *BMC Bioinformatics*, 2023, doi: 10.1186/s12859-023-05456-0.

[28] Y. Görmez *et al.*, “Prediction of obesity levels based on physical activity and eating habits with a machine learning model integrated with explainable artificial intelligence,” *Frontiers in Physiology*, 2025, doi: 10.3389/fphys.2025.1549306.

[29] S. M. Ganie, P. K. Pramanik, and Z. Zhao, “Lifestyle data-based multiclass obesity prediction with interpretable ensemble models incorporating SHAP and LIME analysis,” *Scientific Reports*, 2025, doi: 10.1038/s41598-025-20936-4.

[30] M. S. A. Reshan *et al.*, “An innovative ensemble deep learning clinical decision support system for diabetes prediction,” *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3436641.

[31] H. Liu *et al.*, “Characterizing clinical risk profiles of major complications in type 2 diabetes mellitus using deep learning algorithms,” *Frontiers in Endocrinology*, 2025, doi: 10.3389/fendo.2025.1657366.

[32] X. Wang, W. Wang, H. Ren, X. Li, and Y. Wen, “Prediction and analysis of risk factors for diabetic retinopathy based on machine learning and interpretable models,” *Heliyon*, 2024, doi: 10.1016/j.heliyon.2024.e29497.

[33] L. Zou, X. Wang, Z.-L. Hou, L. Sun, and J. Lu, “Machine learning algorithms for diabetic kidney disease risk predictive model of chinese patients with type 2 diabetes mellitus,” *Renal Failure*, 2025, doi: 10.1080/0886022X.2025.2486558.

[34] R. Sumathi and E. Priya, “Hybrid XGBoost–random forest ensemble model for early prediction of type-2 diabetes using multimodal clinical and lifestyle data,” *International Journal of Scientific Research in Engineering and Management*, 2025, doi: 10.55041/ijsrem54844.

[35] P. M. VanNostrand, H. Zhang, D. M. Hofmann, and E. A. Rundensteiner, “FACET: Robust counterfactual explanation analytics,” *Proceedings of the ACM on Management of Data*, 2023, doi: 10.1145/3626729.

[36] R. Balamurugan, K. Taniya, A. Sekar, and C. Gershenson, “Charting the complexity of diabetes risk using network-based exploration of nonlinear interactions,” *2025 IEEE First International Conference on Innovations in Engineering and Next-Generation Technologies for Sustainability (ICINVENTS)*, 2025, doi: 10.1109/ICINVENTS64613.2025.11401397.