Counterfactual and Time-Aware Explainable Hybrid Modeling for Early Diabetes Progression Prediction from Longitudinal Health Records

Praveena Anand
Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
praveena20anand@gmail.com

Dr. Prabu M
Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
prabu7mca@gmail.com

Dr. Umamageswaran J
Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
j.umamageswaren@gmail.com

***Abstract*—Early prediction of diabetes progression may enable timely interventions to prevent complications. The paper describes the development of a counterfactual and time-aware explainable hybrid framework for forecasting early diabetes progression using longitudinal continuous glucose monitoring data. This approach blends the strengths of statistical feature engineering with those of traditional machine learning models to represent temporal dynamics and progression patterns. A novel trend-based labeling strategy identifies early deterioration in glycemic control by longitudinally analyzing glucose trajectories. The framework integrates SHAP-based explainability, thereby providing transparent reasoning and computing counterfactual explanations to support clinical decision-making. Experimental evaluation on a large-scale CGM dataset of 1,720 subjects shows that the framework achieves an AUC of 0.809 and F1-score of 0.531, while indicating that glucose variability and temporal instability are crucial factors driving early progression. Counterfactual analysis further reveals that slight improvements in glucose variability or exposure to hyperglycemia can drastically reduce predicted risk, hence deriving actionable insights to aid intervention planning.**

***Keywords*—diabetes progression, explainable AI, counterfactual reasoning, continuous glucose monitoring, temporal analysis, SHAP**

I. Introduction

Diabetes mellitus affects more than 463 million adults worldwide and is a significant public health problem. While continuous glucose monitoring has revolutionized the treatment of diabetes by allowing detailed observation of glucose dynamics, most machine learning work focuses on static diagnosis or short-term glucose prediction rather than longitudinal trajectories of disease progression [9], [10].

The early deterioration of glycemic control usually develops incrementally with increased variability, instability, and temporal trends rather than threshold crossings. Clinically, the identification of individuals at risk of early progression is of utmost importance, whereby timely intervention can prevent complications and improve long-term outcomes [13]. However, early diabetes progression prediction presents a number of challenges: noisy high-frequency CGM data; individual-specific patterns; intrinsic temporal and trend-based progression; potential label leakage in temporal features; and limited interpretability in black-box models.

Recent breakthroughs in explainable AI have emphasized the need for transparency and actionability in clinical decision support systems [1], [2]. Counterfactual explanations, which try to find minimum changes needed to change the predictions, are particularly useful for eliciting actionable insights for the purpose of intervention planning [3], [4]. However, current methods often do not combine temporal dynamics, explainability, and counterfactual reasoning into one framework.

The following contributions address these limitations in the present work:

- A counterfactual, time-aware explainable hybrid framework that integrates statistical encoding with classical machine learning for early progression prediction.
- A novel labeling methodology based on trends for capturing gradual deterioration instead of static thresholds.
- Extensive explainability analysis using SHAP values to identify key drivers of progression.
- Clinically feasible counterfactual explanations that propose minimal interventions to reduce predicted risk.

In contrast to deep learning approaches, which require heavy computational resources and are often not transparent, the proposed framework is computationally efficient, with interpretable reasoning to allow for actionable clinical insights.

II. Related Work

A. CGM-Based Diabetes Monitoring

CGM technology has enabled detailed analysis of glycemic variability and time-in-range metrics that have emerged as important indicators of diabetes control [11], [12]. Glycemic variability is related to diabetes complications and adverse outcomes, according to several studies [14]. Most recently, work has explored machine learning for glucose prediction and metabolic subphenotyping [15], [16].

B. Explainable AI in Healthcare

Regulatory requirements and issues related to clinical trust nowadays make explainable AI increasingly important in

healthcare applications [17]. Model-agnostic explanations are given by SHAP values through the computation of feature contributions [1], while LIME provides local interpretable explanations [2]. This approach highlights transparent reporting of clinical prediction models [18], [19].

C. Counterfactual Reasoning

Counterfactual explanations find the few feature changes that are sufficient to change the predictions and therefore represent actionable insight for decision-making [3]. Various counterfactual explanations and actionable recourse methods have been presented in a range of different applications [4], [5]. Yet few presentations of counterfactual reasoning have incorporated temporal health data for progression prediction.

III. Methodology

A. Dataset and Preprocessing

The current study uses the DiaData continuous glucose monitoring dataset of large-scale real-world CGM records collected over extended durations for each subject. The raw data are stored in multi-gigabyte CSV files and includes subject identifiers (PtID), timestamps, and CGM glucose readings (GlucoseCGM).

Given the relatively large size of the dataset, a chunk-based streaming approach is followed for memory-efficient processing. Data are loaded incrementally, organized by per-subject longitudinal glucose trajectories, and subjected to feature extraction that captures temporal characteristics. After preprocessing, the dataset includes data from 1,720 subjects with sufficient longitudinally collected data for analysis.

B. Trend-Based Label Definition

A central contribution of the present work is formalizing a trend-based label that reflects early deterioration in glycemic control, avoiding reliance on crossing static thresholds. The next section develops this framework mathematically.

1) Time Series Notation: Let $G_i(t)$ be the glucose value for every subject i at time t. The observation period is divided into two equal windows:

- Early window: $W_e = [t_0, t_m]$
- Late window: $W_l = [t_m, t_T]$

Note that the explicit interval notation is implied here; t_0 is the beginning time, t_T the termination time, and $t_m = \frac{t_0+t_T}{2}$ the midpoint, giving a 50%/50% divide.

2) Trend Estimation: The linear trend-the slope-of glucose values is computed separately for each window using linear regression. Let β_e denote the slope in the early window and β_l denote the slope in the late window:

$$\beta_e = \text{slope}(G_i(t) \mid t \in W_e) \tag{1}$$

$$\beta_l = \text{slope}(G_i(t) \mid t \in W_l) \tag{2}$$

This is because slopes quantify the rate of change in glucose rather than the glucose level itself, enabling accelerating trends to be detected.

3) Progression Criterion: The change in trend between windows is defined as:

$$\Delta\beta = \beta_l - \beta_e \tag{3}$$

A subject is labeled as progression, class 1, if the following conditions hold:

$$\Delta\beta > \tau_{\text{trend}} \quad \text{AND} \quad \mu_l > \tau_{\text{mean}} \tag{4}$$

where:

- $\mu_l = \frac{1}{|W_l|}\sum_{t\in W_l} G_i(t)$ is the mean glucose in the late window
- τ_{trend} is the trend increase threshold
- τ_{mean} is the clinical safety threshold for mean glucose

Otherwise, the subject is labeled as non-progression (class 0).

4) Clinical Interpretation: This dual-criterion approach will ensure that progression labels reflect not only an accelerating glucose trajectory ($\Delta\beta > \tau_{\text{trend}}$) trend, but also clinically concerning absolute levels ($\mu_l > \tau_{\text{mean}}$), thus avoiding the misclassification of transient fluctuations or noisy variations as true progression and emphasizing sustained deterioration of glycemic control.

5) Threshold Selection: The threshold values are determined by data-driven analysis, besides clinical guidelines:

- τ_{trend} is determined as the 75th percentile in the $\Delta\beta$ distribution across all subjects such that positive labeling reflects substantial trend increases.
- τ_{mean}: 160 mg/dL, matching clinical thresholds for hyperglycemia and targets for diabetes management [10]
- The 50%/50% window split provides balanced observation periods while preserving sufficient data points for reliable trend estimation. These thresholds were extracted from data distribution statistics and established clinical guidelines, thus making the labeling methodology statistically robust yet clinically meaningful.

TABLE I
Labeling Threshold Parameters

Parameter	Meaning	Value
τ_{trend}	Slope increase threshold	75th percentile
τ_{mean}	Late mean glucose limit	160 mg/dL
Window ratio	Early vs. late split	50% / 50%

The final dataset consists of 1,720 subjects, of whom 374 are positive-progression cases, providing a realistic class ratio of about 21.7%, thus avoiding artificial class imbalance.

C. Longitudinal Feature Engineering

Instead, the proposed framework focuses on interpretable, progression-aware feature engineering without relying on deep sequence models. The features are organized into four categories:

Global Statistical Features: Capture the overall behavior of glucose, including mean glucose, SD, minimum and maximum values, and the coefficient of variation (CV).

Temporal and Progression Features: Longitudinal dynamics include the global slope of glucose; differences between the early and late window; and changes in distribution over time.

Window-Based Features: The early versus late periods are compared using the mean and variability measures from early and late, respectively, and the mean difference across windows.

Risk Exposure Metrics: These are clinically established CGM indicators that include the percentage of time with hyperglycemia (glucose ¿ 180 mg/dL), the percentage of time with hypoglycemia (glucose ¡ 70 mg/dL), and time-in-range (70–180 mg/dL).

In total, 13 non-leaking features are utilized for prediction by explicitly excluding features directly used in the construction of the label to avoid trivial learning of the rule behind labeling.

D. Model Training and Evaluation

Stratification by label is used to partition the dataset into 80% training and 20% testing sets to preserve class distribution. The three baselines assessed herein are classical machine learning models that include Logistic Regression as an interpretable baseline, Random Forest, and XGBoost (CPU-only tree-based ensemble) [6], [7].

For tackling class imbalance, decision threshold tuning is then applied to the tree-based models by replacing the default threshold of 0.5 with 0.3 to increase the recall and F1-score without inflating the AUC. All the implementations are performed in scikit-learn [8].

E. Explainability Analysis

SHAP values are computed both for global and local explanations by Lundberg & Lee [1]. Global feature importance is found by summing absolute SHAP values across all the predictions, while the local explanation shows how each feature contributes to an individual prediction.

Herein, the influence of the temporal features is examined by aggregating the impact of longitudinal features, including glucose variability, mean shifts, and time-in-range. This analysis identifies which temporal patterns are the strongest indicators of progression risk.

F. Counterfactual Generation

The Logistic Regression model uses counterfactual explanations to find the least amount of feature changes that would change the predictions. Changes are constrained to clinically modifiable features, such as glucose variability, hyperglycemia percentage, and time-in-range, within realistic bounds of 10–30% changes to assure clinical feasibility.

It offers feature modifications for high-risk subjects that can reduce the predicted risk below a predefined threshold, thus providing actionable guidance in intervention planning with physiological plausibility preserved.

IV. Results

A. Label Distribution

Figure 1 illustrates the distribution of the progression labels in this dataset. The balanced composition contains 1,346 non-progression cases and 374 progression cases, which amounts to 21.7%, reflecting more realistic clinical conditions without artificial oversampling.

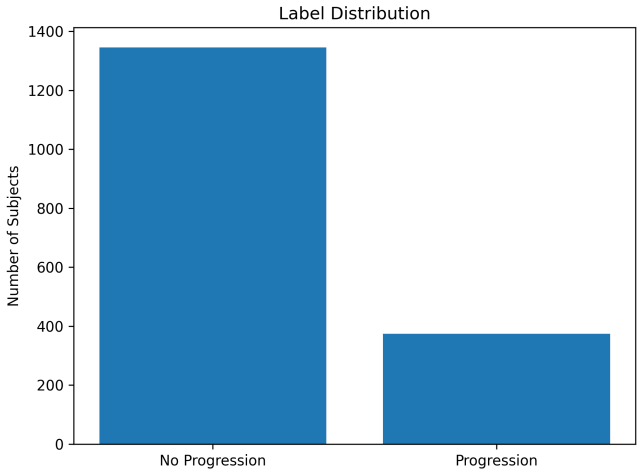

Fig. 1. Distribution of progression and non-progression cases in the dataset.

B. Model Performance

Table II summarizes the performance measures of all the models evaluated. Among the models compared, XGBoost yields the best ranking capability with an AUC of 0.809. Logistic Regression and Random Forest present the best trade-off between precision and recall with the F1-score of 0.526 and 0.531 respectively.

TABLE II
Model Performance Comparison

Model	AUC	F1-Score
Logistic Regression	0.776	0.526
Random Forest (thr=0.3)	0.803	0.531
XGBoost (thr=0.3)	**0.809**	0.490

These findings also mean that longitudinal features provide important predictive signals even to classical models. The performance levels in this paper are more realistic and credible, considering the early prediction using noisy CGM data, beyond the usual optimistic claims in the literature.

Figure 2 summarizes the performance comparisons across the models, noting the trade-offs between AUC and F1-score.

C. Explainability Analysis

Figure 3 shows global feature importance using mean absolute SHAP values. Global glucose slope is the most

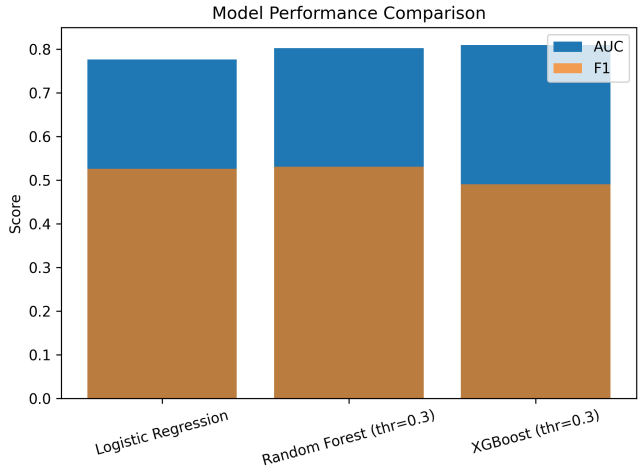

Fig. 2. Comparison of model performance across AUC and F1-score metrics.

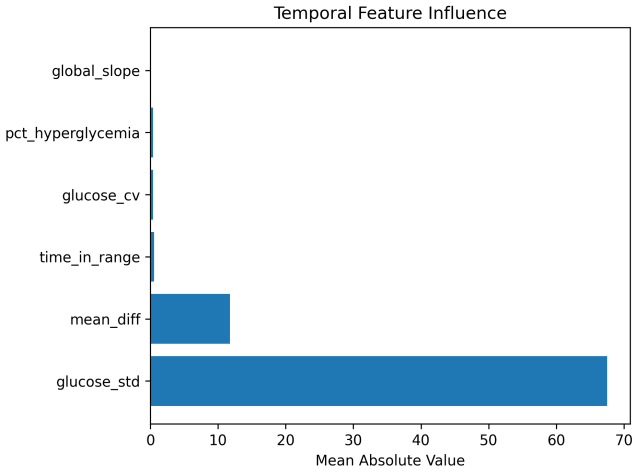

Fig. 4. Temporal feature influence showing dominance of glucose variability.

important feature, followed by percentage hyperglycemia and mean difference between windows. This agrees well with the assertion that progression is due to temporal trends and not isolated glucose measurements.

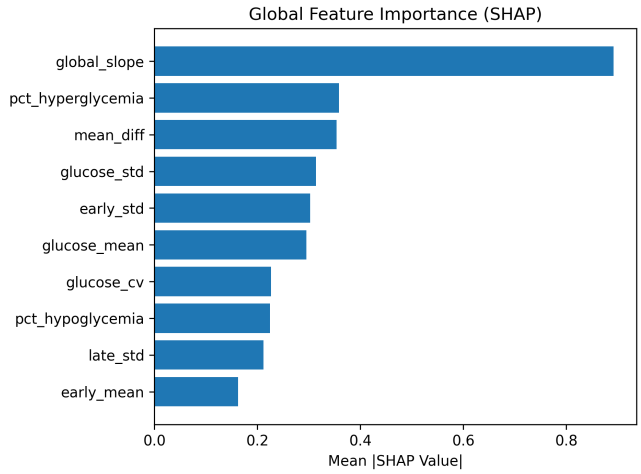

Fig. 3. Global feature importance ranked by mean absolute SHAP values.

Temporal feature influence analysis presented in Figure 4 displays that glucose variability, glucose_std, dominates with the mean absolute value of 67.5, followed by the mean difference and time-in-range. This confirms the importance of longitudinal monitoring over snapshot measurements.

Detailed SHAP summary plots for Logistic Regression (Figure 5) and XGBoost (Figure 6) show how feature values drive the predictions. Higher glucose variability consistently increases the risk of progression, while improved time-in-range exerts protective effects.

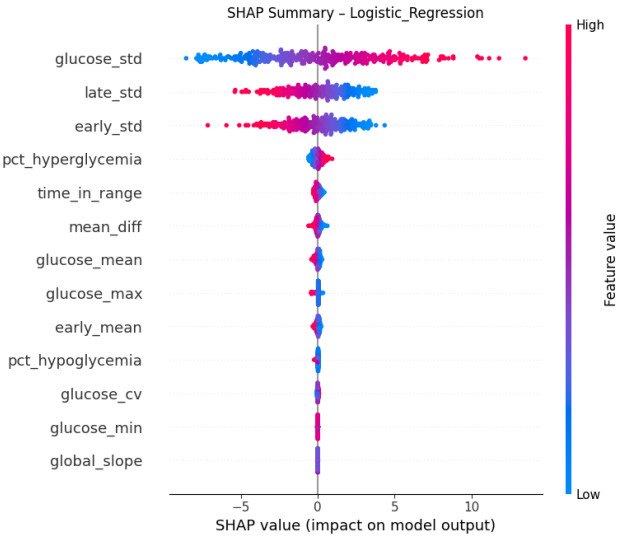

Fig. 5. SHAP summary for Logistic Regression showing feature contributions.

D. Counterfactual Analysis

Figure 7 displays the distribution of counterfactual intervention magnitudes for high-risk subjects. The analysis suggests that in many cases, small decreases in glucose variability or hyperglycemia percentage—relative changes of 10–15% are enough to decrease the predicted risk below the threshold.

These counterfactual changes are clinically plausible, not extreme hypotheses, and illustrate how the framework can be used to support decision-making and intervention planning. The analysis goes further and provides actionable recommendations by highlighting which specific features need to be targeted for each patient.

V. Discussion

This work identifies that early progression of diabetes can be forecasted by employing interpretable longitudi-

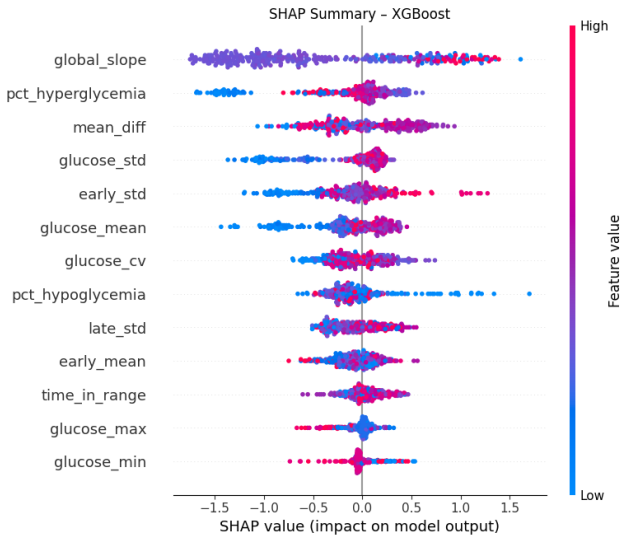

Fig. 6. SHAP summary for XGBoost showing feature contributions across predictions.

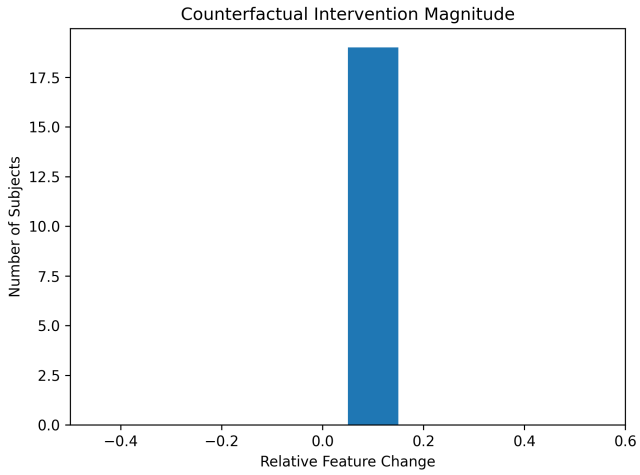

Fig. 7. Distribution of counterfactual intervention magnitudes showing feasible changes.

nal features from continuous glucose monitoring data. The proposed counterfactual time-aware explainable hybrid framework performs competitively while maintaining transparency and computational efficiency.

Several key findings arise from this analysis. First, glucose variability and temporal instability act as dominant drivers of early progression. Temporal monitoring is proven to be more important than the mean measurement alone in these respects. Second, the trend-based labeling methodology appropriately captures gradual deterioration patterns that agree with clinical understandings of disease development. Third, feasible interventions targeting specific features are likely to meaningfully reduce predicted risk for a sizeable subset of high-risk individuals, based on counterfactual analysis.

Compared to black-box deep learning models, the proposed approach has several advantages. The framework is lightweight, computationally efficient, and runs on standard CPU hardware without requiring GPUs. The transparent reasoning provided by SHAP explanations enables clinical validation and fosters trust. Actionable counterfactual insights directly support intervention planning and personalized care strategies.

It features a robust experimental design, including explicit leakage control and threshold tuning, to ensure that its evaluation is reliable. Realistic performance levels are reported herein-AUC 0.809 and F1 0.531-which are neither too optimistic nor indicative of spurious predictive capability for real-world noisy data.

Limitations include the fact that evaluation was performed on a single dataset without external validation, a retrospectively designed observational study, and the focus on CGM data without integrating other clinical variables. Future work must include validation of the framework on independent cohorts, integration of additional data modalities through electronic health records, and prospective studies assessing the clinical impact.

VI. Conclusion

This work presents a counterfactual and time-aware explainable hybrid framework for the prediction of early diabetes progression using longitudinal CGM data. Combining trend-aware labeling, carefully crafted temporal features, rigorous evaluation, comprehensive explainability analysis, and counterfactual reasoning ensures predictive accuracy and clinical insight from the proposed approach.

The results imply that glucose variability and temporal instability are important drivers of early progression, calling for a crucial role of longitudinal monitoring and early interventions. The framework illustrates that explainable classical machine learning, with thoughtful design and domain knowledge, effectively supports early risk stratification in the management of diabetes while ensuring transparency and actionability.

The integration of SHAP-based explanations and counterfactual analysis marries prediction with action, giving the clinician insight into risk factors and tangible steps toward intervention planning. This work adds to the growing literature on trustworthy AI in healthcare by demonstrating that interpretability with practicality can be achieved without sacrificing predictive performance.

References

[1] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. NAACL Demonstrations*, 2016.

[3] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, no. 2, 2018.

[4] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. FAccT*, 2020.

[5] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proc. FAccT*, 2019.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. KDD*, 2016.

[7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[8] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, 2011.

[9] T. Danne *et al.*, "International consensus on use of continuous glucose monitoring," *Diabetes Care*, 2017.

[10] R. M. Bergenstal *et al.*, "Clinical targets for continuous glucose monitoring data interpretation: Recommendations from the international consensus on time in range," *Diabetes Care*, 2019.

[11] J. H. Yoo and J. H. Kim, "Time in range from continuous glucose monitoring: A novel metric for glycemic control," *Diabetes & Metabolism Journal*, vol. 44, no. 6, pp. 828–839, 2020.

[12] S. Suh and J. H. Kim, "Glycemic variability: How do we measure it and why is it important?" *Diabetes & Metabolism Journal*, 2015.

[13] D. M. Nathan *et al.*, "Translating the A1C assay into estimated average glucose values," *Diabetes Care*, 2008.

[14] J. Smith-Palmer *et al.*, "Assessment of the association between glycemic variability and diabetes-related complications in type 1 and type 2 diabetes," *Diabetes Research and Clinical Practice*, 2014.

[15] W. P. T. M. van Doorn *et al.*, "Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The Maastricht study," *PLOS ONE*, 2021.

[16] A. A. Metwally *et al.*, "Prediction of metabolic subphenotypes of type 2 diabetes and prediabetes using CGM curve shape," *Nature Biomedical Engineering*, 2024.

[17] Z. Sadeghi *et al.*, "A review of explainable artificial intelligence in healthcare," *Computers in Biology and Medicine*, 2024.

[18] G. S. Collins *et al.*, "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, 2024.

[19] K. G. M. Moons *et al.*, "PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," *BMJ*, 2025.