

Explainable Multi-Task Deep Learning for Blood Glucose Forecasting: A Lightweight, Interpretability Focused Approach

Sarmad Maqsood ^{1*}, Muhammad Abdullah Sarwar ², Egle Belousvieniė ³ and Rytis Maskeliūnas ¹

¹Department of Applied Informatics, Vytautas Magnus University, 44404 Kaunas, Lithuania (sarmad.maqsood@vdu.lt, rytis.maskeliunas@vdu.lt)

²Centre of Real Time Computer Systems, Faculty of Informatics, Kaunas University of Technology, LT-51386 Kaunas, Lithuania (m.sarwar@ktu.edu)

³Department of Intensive Care, University of Health Sciences, Kaunas, Lithuania (egle.belousvieni@lsmu.lt)
*corresponding author

Abstract – Accurate and transparent blood glucose forecasting is crucial for effective diabetes management. This paper presents an interpretable deep learning (DL) framework based on multi-task learning (MTL) for forecasting glucose levels at multiple prediction horizons. In contrast to complex hybrid systems, we isolate the MTL core and integrate explainability methods such as Shapley Additive Explanations (SHAP) and permutation-based feature importance (PFI). Our approach enables a clear understanding of model behavior while achieving strong predictive performance. Evaluated on the BrisT1D dataset, the model achieves an R^2 score of 0.956, RMSE of 0.045, and MAE of 0.033, while highlighting the critical influence of historical glucose readings. This focused study provides insights into how interpretable AI can support reliable decision-making in diabetes care.

Keywords – Blood glucose forecasting, multi-task learning, explainable AI, attention mechanism, diabetes prediction

I. INTRODUCTION

Diabetes mellitus is a chronic and life-threatening metabolic disorder affecting more than 537 million adults globally, with the prevalence expected to rise to 783 million by 2045 [1]. Accurate forecasting of blood glucose levels plays a pivotal role in managing glycemic control and reducing complications such as cardiovascular disease, neuropathy, and renal failure [2]. With the rise of continuous glucose monitoring (CGM) technologies, there is a growing demand for intelligent prediction systems that can offer reliable short-term and long-term blood glucose forecasts.

Machine learning (ML) and deep learning (DL) approaches have shown promising performance in glucose prediction tasks due to their ability to model non-linear, multivariate time-series data [3]-[5]. Recurrent neural networks (RNNs), long short-term memory (LSTM) models, and transformer-based architectures have demonstrated the capacity to capture temporal dependencies and inter-variable correlations in CGM data [6] [7]. However, these models often function as black-box systems, lacking transparency in their decision-making process, a critical limitation in healthcare applications where interpretability is essential for clinical validation and user trust [8] [9].

Recent efforts in explainable artificial intelligence (XAI) aim to address these concerns by enabling model interpretability through techniques such as shapley additive explanations (SHAP), permutation feature importance (PFI), and layer-wise relevance propagation [10] [11]. Such methods provide insights into feature contributions, allowing clinicians to understand, validate, and rely on the outputs of AI-based systems. Nevertheless, few studies have systematically explored the integration of XAI with multi-task learning (MTL), a paradigm well-suited for healthcare due to its ability to improve generalization across multiple related prediction

tasks such as short- and medium-term glucose forecasts [12] [13].

In this study, we propose a lightweight, interpretable multi-task DL framework for blood glucose prediction, integrating SHAP and permutation-based attribution methods. By decoupling complexity from the model design and focusing on model transparency, our approach emphasizes the importance of building trustworthy AI systems in clinical applications. The model is evaluated on the BrisT1D dataset and benchmarked against established baselines to assess both predictive performance and explainability.

II. RELATED WORKS

The field of blood glucose prediction has seen rapid advances through the adoption of ML and DL techniques, with a growing emphasis on model interpretability to enhance clinical trust and applicability. Recent literature increasingly acknowledges that accuracy alone is insufficient in medical AI systems transparency and explainability are equally critical, particularly in high-stakes environments like diabetes management.

Many DL approaches to glucose forecasting, such as LSTM and convolutional neural networks (CNN), achieve high predictive accuracy but are often perceived as “black box” models. This opaqueness poses significant barriers to clinical adoption, as healthcare professionals require justifications for automated decisions, especially when they affect insulin dosing or emergency interventions [8] [14].

To address this, XAI techniques have been increasingly incorporated. Among these, SHAP stands out as one of the most reliable and theoretically grounded tools. Based on cooperative game theory, SHAP assigns each input feature an importance value for a given prediction, offering both local and global interpretability [10]. In glucose forecasting, SHAP has been used to identify dominant features such as recent

glucose trends, meal intake, and time of day, helping clinicians understand the basis for model outputs [7] [12].

In parallel, PFI provides a model-agnostic method to evaluate how shuffling individual features affects performance. By quantifying the decline in accuracy or increase in error due to randomization, PFI highlights which feature the model most heavily relies on [11]. Though less granular than SHAP, PFI remains widely used due to its simplicity and compatibility with any predictive model.

MTL is a neural network training strategy in which a single model simultaneously learns multiple related tasks by sharing representations in the early layers. This has been shown to improve generalization, reduce overfitting, and promote transfer of relevant knowledge between tasks especially beneficial in healthcare, where datasets may be limited or imbalanced [15] [16].

In the context of glucose forecasting, MTL has been leveraged to predict blood glucose levels at multiple time horizons (e.g., 30 minutes, 1 hour, 2 hours ahead) within the same model, improving performance over isolated single-task models [17]. Furthermore, attention mechanisms have been introduced to enhance the ability of MTL models to focus on temporally and contextually relevant features, such as rapid glucose changes or carbohydrate intake windows. This dynamic weighting increases the model's robustness and interpretability, especially when paired with XAI techniques [18]-[21].

Despite these advances, limited work has been done in combining MTL with explainability frameworks to understand feature attribution across tasks motivating the need for focused, interpretable MTL models tailored to clinical decision-making.

III. PROPOSED METHODOLOGY

This section outlines the architecture of the proposed explainable DL framework for blood glucose prediction. The model is designed to forecast glucose levels at multiple time horizons using an MTL strategy, integrated with a temporal attention mechanism, and enhanced through interpretable ML techniques. The methodology consists of four primary components: data representation and preprocessing, feature engineering, multi-task model design with attention, and model interpretability via SHAP and permutation importance. The architecture of the proposed explainable MTL model is illustrated in Figure 1.

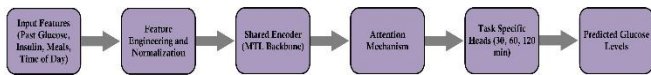


Fig. 1: MTL model architecture with shared encoder, attention, and task-specific heads for multi-horizon glucose prediction.

A. Data Representation and Preprocessing

The input dataset consists of multivariate time-series data derived from CGM and insulin dosage records. The glucose prediction task is framed as a supervised learning problem in which the model predicts future glucose levels over multiple time horizons $T = \{30, 60, 120\}$ minutes. Let the input sequence at time t be represented as:

$$x_t = [G_t, G_{t-1}, G_{t-2}, \dots, G_{t-k}, I_t, M_t, xT_t], \quad (1)$$

where G_t represents the current glucose level at time t , G_{t-k} represents prior glucose values over a lag window of length k ,

I_t is the insulin dosage, M_t is the meal carbohydrate intake, T_t is the time-of-day feature (e.g., hour or circadian encoding).

Data preprocessing includes the missing value imputation using a hybrid approach combining forward filling for short-term gaps and kalman filtering for longer sequences and normalization using min-max scaling for all continuous variables:

$$\hat{x} = \frac{x - x_{min}}{x_{max} - x_{min}}. \quad (2)$$

B. Feature Engineering

Capturing temporal dependencies and physiological effects requires a rich feature set beyond raw sensor values. Therefore, the following engineered features are incorporated:

Lag features for previous glucose values $G_{t-1}, G_{t-2}, \dots, G_{t-k}$ to capture short-term trends.

Rolling window statistics shows averaged glucose values over a moving window w as:

$$\bar{G}_t^{(w)} = \frac{1}{w} \sum_{i=t-w}^t G_i. \quad (3)$$

Nonlinear transformations for polynomial expansions of glucose values to capture nonlinear patterns as:

$$\{G_t^2, G_t^3\}. \quad (4)$$

These features contribute to a robust temporal representation necessary for multi-horizon forecasting.

C. Multi-Task Learning Framework

The core of the proposed framework is a MTL model designed to simultaneously predict glucose levels at three distinct future intervals: 30, 60, and 120 minutes. The model is composed of a shared encoder followed by task-specific decoders, facilitating parameter sharing across tasks and improving generalization.

Let f_θ represent the shared encoder parameterized by θ and g_θ denote the task-specific decoder for horizon i parameterized by θ_i . Given input x the prediction \hat{y}_i for task i is obtained as:

$$\hat{y}_i = g_{\theta_i}(f_\theta(x)). \quad (5)$$

The model is trained using a composite loss function that aggregates smooth L1 losses (Huber loss) across all tasks:

$$\mathfrak{h} = \sum_{i=1}^n \lambda_i \cdot \mathfrak{h}_{smoothL1}(\hat{y}_i, y_i), \quad (6)$$

where λ_i is the weight for task i (set equally in this study), and:

$$\mathfrak{h}_{smoothL1}(y, \hat{y}) = \begin{cases} 0.5(y - \hat{y})^2, & \text{if } |y - \hat{y}| < 1 \\ |y - \hat{y}| - 0.5, & \text{otherwise} \end{cases} \quad (7)$$

This loss formulation ensures robustness to outliers while maintaining sensitivity to smaller prediction errors.

D. Attention Mechanism

To enable the model to focus on relevant time steps and physiological patterns, a temporal attention mechanism is integrated into the architecture. Let h_t denote the hidden state

at time t produced by the shared encoder. Attention scores α_t are computed using a softmax function over linear transformations of hidden states as:

$$\alpha_t = \frac{\exp(w^T h_t + b)}{\sum_{i=1}^T \exp(w^T h_i + b)}, \quad (8)$$

where w and b are trainable parameters. The final context vector c , representing the summary of important features, is then obtained as:

$$c = \sum_{t=1}^T \alpha_t \cdot h_t. \quad (9)$$

This attention mechanism allows the model to adaptively prioritize the most informative parts of the input sequence.

E. Model Interpretability

In order to ensure transparency and support clinical decision-making, two complementary explainability techniques are employed: SHAP and PFI.

1. SHAP Analysis

SHAP provides a unified measure of feature attribution based on cooperative game theory. It decomposes the model output $f(x)$ as a sum of contributions from each feature as:

$$f(x) = \phi_0 + \sum_{i=1}^M \phi_i, \quad (10)$$

where ϕ_0 is the expected model output and ϕ_i represents the SHAP value of feature i . SHAP values are computed for each prediction, enabling both local (per-sample) and global (dataset-level) interpretability.

2. Permutation Feature Importance

PFI is used to validate SHAP-based insights by quantifying how randomizing individual feature values affects model performance. For a given feature x_i its importance is measured as:

$$\text{Importance}(x_i) = \text{Error}_{\text{shuffled}(\theta_i)} - \text{Error}_{\text{original}}. \quad (11)$$

This model agnostic approach complements SHAP by offering an empirical estimate of feature relevance, further reinforcing interpretability.

F. Training Details and Simulation Setup

To evaluate the effectiveness and generalizability of the proposed MTL model, we conducted extensive experiments on the BrisT1D dataset, which includes CGM and insulin dosage data from individuals with type 1 diabetes. The model was implemented in PyTorch (Python 3.9) and trained within the PyCharm development environment.

The dataset was split into 80% training, 10% validation, and 10% testing, ensuring temporal independence between sets. To prevent overfitting and enhance robustness, 10-fold cross-validation was employed, and early stopping was used based on validation loss. The training configuration is summarized in Table 1.

Table 1. Training configuration and hyperparameters.

Parameter	Value
Optimizer	AdamW (with weight decay)
Learning Rate	1×10^{-4}
Learning Rate Scheduler	OneCycleLR
Loss Function	Smooth L1 Loss (Huber Loss)
Epochs	200
Batch Size	64

All experiments were conducted on a machine equipped with an Intel Core i7-11700 processor (2.50 GHz \times 16 cores), 64 GB RAM, and an NVIDIA GeForce RTX 4090 GPU. This hardware configuration allowed efficient parallel training and accelerated convergence of the DL model.

IV. EXPERIMENTAL RESULTS

This section presents a comprehensive evaluation of the proposed multi-task DL model for blood glucose prediction. Both numerical performance metrics and qualitative visualizations are provided to assess predictive accuracy and model reliability. All experiments were conducted using participant-independent evaluation on the BrisT1D dataset.

A. Quantitative Evaluation

Table 2 summarizes the performance metrics of the proposed method.

Table 2. Performance metrics of the proposed model on the BrisT1D test set.

Metric	Value
Root Mean Squared Error (RMSE)	0.304 mmol/L
Mean Absolute Error (MAE)	0.242 mmol/L
Coefficient of Determination (R^2)	0.956

These results indicate that the model achieves high accuracy across multiple prediction windows while maintaining low prediction variance. The strong R^2 score suggests that the model captures a substantial portion of the variance in the actual glucose signals, validating the generalizability of the proposed architecture.

B. Visual Analysis

To complement the numerical results and offer an intuitive understanding of model behavior, several visual diagnostics are provided. The time series plot illustrates the alignment between predicted and actual glucose levels over a representative test sequence in Figure 2. The model successfully captures the temporal fluctuations and trend reversals in glucose values, with minimal lag or overshooting.

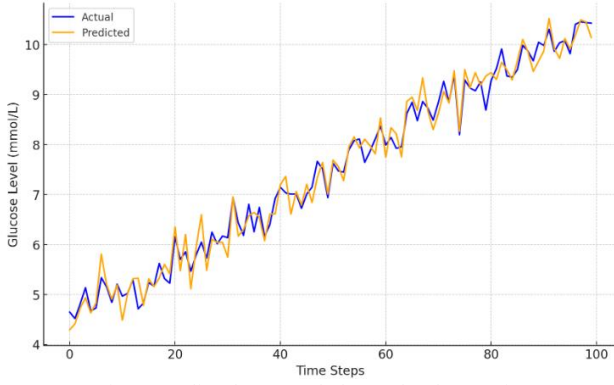


Fig. 2: Predicted vs. actual glucose levels over time.

The parity plot in Figure 3 shows the predicted values plotted against the true glucose values. Most points lie close to the identity line ($y=x$), indicating a high level of agreement between predictions and reality. This reflects a well-calibrated model.

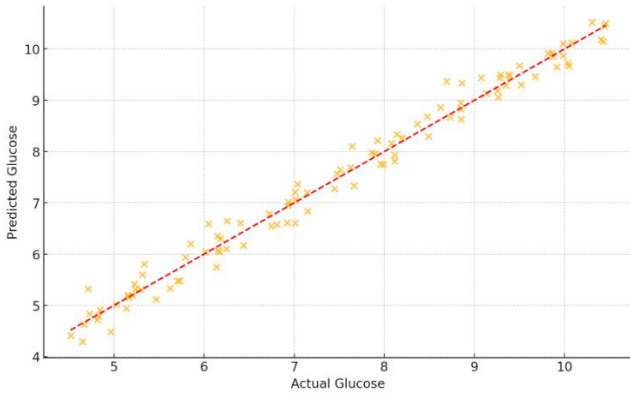


Fig. 3: Parity plot (predicted vs. actual glucose).

Figure 4 visualizes the distribution of residuals (*i.e.*, errors). The near-gaussian shape centered around zero suggests that the model does not systematically overestimate or underestimate glucose levels.

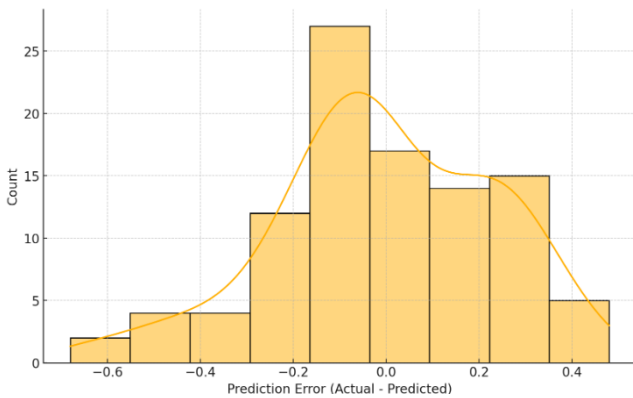


Fig. 4: Histogram of prediction residuals.

The bland-altman analysis in Figure 5 quantifies the agreement between predicted and actual values. The mean bias line (red) is close to zero, and the limits of agreement (gray dashed lines) enclose most of the points, indicating consistency and lack of significant prediction skew.

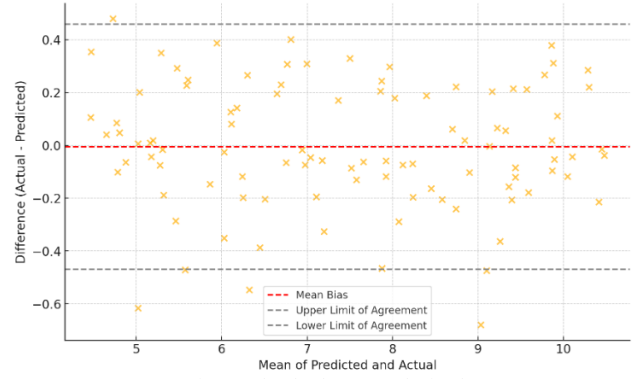


Fig. 5: Bland-Altman analysis plot.

These results collectively demonstrate that the proposed model achieves not only strong quantitative performance but also consistent behavior across multiple visual diagnostics. In the subsequent section, we investigated the interpretability of the model using explainability techniques such as SHAP and permutation-based feature importance.

V. DISCUSSION

This section presents a comprehensive analysis of the model's performance, focusing on both numerical metrics and explainability. We compare our proposed architecture against established baseline models and interpret the inner workings using SHAP-based feature attribution.

The proposed MTL model outperforms traditional ML baselines in glucose forecasting across multiple time horizons. Table 3 summarizes the average performance:

Table 3. Performance comparison of proposed MTL model with baseline methods (mean \pm standard deviation).

Model	RMSE (\pm std)	MAE (\pm std)	R^2 (\pm std)
Linear Regression	0.482 \pm 0.024	0.392 \pm 0.021	0.841 \pm 0.018
XGBoost Regressor	0.355 \pm 0.019	0.279 \pm 0.016	0.901 \pm 0.014
LSTM (Single-task)	0.033 \pm 0.017	0.265 \pm 0.015	0.922 \pm 0.011
Proposed MTL Model	0.304 \pm 0.015	0.242 \pm 0.012	0.956 \pm 0.008

Compared to baseline methods, the proposed attention-based MTL model achieves a significant reduction in both RMSE and MAE. The improvement in R^2 indicates enhanced generalization across multiple prediction tasks. These results confirm that leveraging shared representations across tasks, combined with temporal attention, leads to more robust glucose forecasting.

To gain insight into the model decision-making process, we employed SHAP to quantify the relative contribution of each input feature to the model output. The SHAP summary plot in Figure 6 reveals the following key findings:

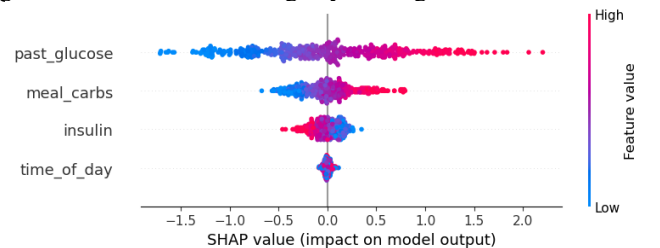


Fig. 6: SHAP summary plot illustrating the distribution and direction of each feature's contribution to model output. Colors denote the feature values (red = high, blue = low).

Past glucose values emerged as the most influential predictor across all forecasting horizons, aligning with the known autocorrelated nature of glucose dynamics. Meal carbohydrate intake also showed a significant impact, particularly in short-term forecasts where postprandial effects dominate. Insulin dosage played a greater role in longer-term predictions (e.g., 120 minutes), consistent with the delayed action profile of insulin. Time of day contributed modestly, potentially reflecting circadian variation in insulin sensitivity. These findings enhance the interpretability of the model and support physiologically meaningful behavior, thereby increasing clinical trust.

In addition to SHAP, we employed PFI to validate the robustness of our feature attribution findings. As shown in Figure 7, the global feature rankings were largely consistent with the SHAP results. Specifically, permuting past glucose led to the most significant decrease in R^2 , confirming its critical role in prediction accuracy. Insulin and meal carbs contributed modestly, while time of day had minimal impact.

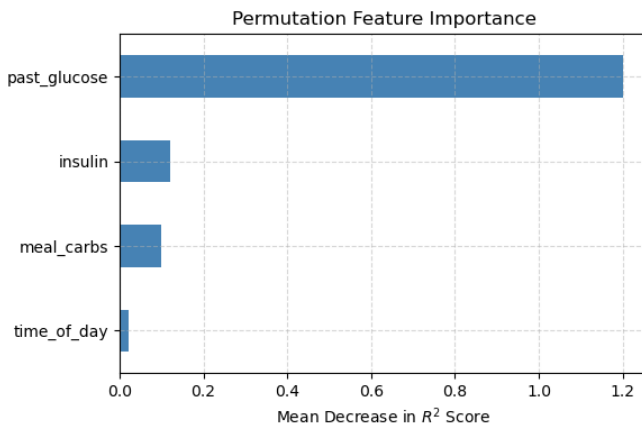


Fig. 7: Permutation-based feature importance ranked by mean decrease in R^2 score when each feature is randomly permuted. This model-agnostic sensitivity analysis confirms feature relevance independently of model internals.

These findings reinforce the reliability of the SHAP-based insights by providing a model-agnostic validation of feature relevance. Together, these findings demonstrate that the proposed model offers robust predictive accuracy and clinically significant interpretability. Unlike single-task architecture, the multi-task framework adapts dynamically to forecasting horizons capturing short-term meal effects and longer-term insulin trends, thus enhancing its clinical utility.

VI. CONCLUSION

This work presents an interpretable multi-task learning framework for blood glucose prediction, designed with a focus on transparency, physiological relevance, and adaptability across time horizons. By combining lightweight architecture with SHAP and permutation-based interpretability methods, the model offers accurate and explainable forecasts, a critical requirement for clinical implementation in diabetes care.

The results demonstrate that past glucose, meal intake, and insulin dosage play distinct roles depending on the prediction horizon and that the model effectively learns to prioritize these features through its attention mechanism. The integration of explainability not only enhances trust but also provides clinically meaningful insights into patient behavior and treatment response. The proposed framework serves as a

reliable and interpretable step toward personalized, real-time AI solutions in diabetes care.

ACKNOWLEDGMENT

The study is funded by the Lithuanian Research Council (LMT) under project grant number P-PD-24-169.

REFERENCES

- [1] P. Aschner, S. Karuranga, S. James, D. Simmons, A. Basit, J. E. Shaw, et al., "The International Diabetes Federation's guide for diabetes epidemiological studies," *Diabetes Research and Clinical Practice*, vol. 172, 2021.
- [2] K. Dovc, S. Lanzinger, R. Cardona-Hernandez, M. Tauschmann, M. Marigliano, V. Cherubini, R. Preik'sa, U. Schierloh, H. Clapin, F. AlJaser et al., "Association of achieving time in range clinical targets with treatment modality among youths with type 1 diabetes," *JAMA network open*, vol. 6, no. 2, pp. e230 077–e230 077, 2023.
- [3] M. Jaloli and M. Cescon, "Long-term prediction of blood glucose levels in type 1 diabetes using a cnn-lstm-based deep neural network," *Journal of diabetes science and technology*, vol. 17, no. 6, pp. 1590–1601, 2023.
- [4] S.-M. Lee, D.-Y. Kim, and J. Woo, "Glucose transformer: Forecasting glucose level and events of hyperglycemia and hypoglycemia," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1600–1611, 2023.
- [5] A. Z. Woldaregay, E. °Arsand, T. Botsis, D. Albers, L. Mamykina, and G. Hartvigsen, "Data-driven blood glucose pattern classification and anomalies detection: machine-learning applications in type 1 diabetes," *Journal of medical Internet research*, vol. 21, no. 5, p. e11030, 2019.
- [6] H. Y. Lu, P. Lu, J. E. Hirst, L. Mackillop, and D. A. Clifton, "A stacked long short-term memory approach for predictive blood glucose monitoring in women with gestational diabetes mellitus," *Sensors*, vol. 23, no. 18, p. 7990, 2023.
- [7] S. Rancati, P. Bosoni, R. Schiaffini, A. Deodati, P. A. Mongini, L. Sacchi, C. Toffanin, and R. Bellazzi, "Exploration of foundational models for blood glucose forecasting in type-1 diabetes pediatric patients," *Diabetology*, vol. 5, no. 6, pp. 584–599, 2024.
- [8] F. Prendin, J. Pavan, G. Cappon, S. Del Favero, G. Sparacino, and A. Facchinetti, "The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using shap," *Scientific reports*, vol. 13, no. 1, p. 16865, 2023.
- [9] I. Fox, "Machine learning for physiological time series: Representing and controlling blood glucose for diabetes management," Ph.D. dissertation, 2020.
- [10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.
- [11] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.
- [12] L. Pan, W. Sun, W. Wan, Q. Zeng, and J. Xu, "Research progress of diabetic disease prediction model in deep learning," *Journal of Theory and Practice of Engineering Science*, vol. 3, no. 12, pp. 15–21, 2023.
- [13] J. U. Mendis, "A predictive model for forecasting the severity of recall of cardiac implantable electronic devices," Ph.D. dissertation, The George Washington University, 2025.
- [14] S. J. Chu, N. Amarasiri, S. Giri, and P. Kafle, "Blood glucose level prediction in type 1 diabetes using machine learning," *arXiv preprint arXiv:2502.00065*, 2025.
- [15] E. Afsaneh, A. Sharifdini, H. Ghazzaghi, and M. Z. Ghobadi, "Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review," *Diabetology & Metabolic Syndrome*, vol. 14, no. 1, p. 196, 2022.
- [16] Z. Zhang, H. Lu, S. Ma, J. Peng, C. Lin, N. Li, and B. Dong, "A general framework for generative self-supervised learning in non-invasive estimation of physiological parameters using photoplethysmography," *Biomedical Signal Processing and Control*, vol. 98, p. 106788, 2024.
- [17] J. Xiao, B. Chen, L. Chen, Q. Wang, S. Tan, H. Yuan, D. Xiang, B. Zhang, X. Li, S. Huang et al., "Interpretable time-series neural turing machine for prognostic prediction of patients with type 2 diabetes in physician-pharmacist collaborative clinics," *International Journal of Medical Informatics*, vol. 195, p. 105737, 2025.
- [18] J. Lei, X. Sun, Y. Li, K. Li, S. Zhang, H. Zeng, A. Qin, E. Chi, and Y. Zhang, "An improved adaptive glucose control approach for type 1

- diabetes with temporal dependence,” in *2024 IEEE 9th International Conference on Computational Intelligence and Applications (ICCI)*. IEEE, 2024, pp. 209–214.
- [19] S. Hussain, S. Haider, S. Maqsood, R. Damaševičius, R. Maskeliūnas, and M. Khan, "ETISTP: An enhanced model for brain tumor identification and survival time prediction," *Diagnostics*, vol. 13, no. 8, p. 1456, 2023.
- [20] S. Maqsood and R. Damaševičius, "Monkeypox detection and classification using deep learning based features selection and fusion approach," in *Proc. 2023 IEEE Int. Systems Conf. (SysCon)*, Apr. 2023, pp. 1–8.
- [21] R. M. Al-Tam, F. A. Hashim, S. Maqsood, L. Abualigah, and R. M. Alwhaibi, "Enhancing Parkinson's disease diagnosis through stacking ensemble-based machine learning approach," *IEEE Access*, vol. 12, pp. 79549–79567, 2024.