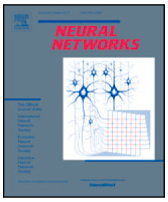
Neural Networks 185 (2025) 107229

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Full Length Article

# GARNN: An interpretable graph attentive recurrent neural network for predicting blood glucose levels via multivariate time series

Chengzhe Piao [a], Taiyu Zhu [b], Stephanie E. Baldeweg [c,d], Paul Taylor [a], Pantelis Georgiou [e], Jiahao Sun [f], Jun Wang [g], Kezhi Li [a,*]

[a] Institute of Health Informatics, University College London, London, NW1 2DA, UK
[b] Department of Psychiatry, University of Oxford, Oxford, OX3 7JX, UK
[c] Department of Diabetes & Endocrinology, University College London Hospitals, London, NW1 2PG, UK
[d] Centre for Obesity & Metabolism, Department of Experimental & Translational Medicine, University College London, London, WC1E 6JF, UK
[e] Centre for Bio-Inspired Technology, Department of Electrical and Electronic Engineering, Imperial College London, London, SW7 2AZ, UK
[f] FLock.io, London, WC2H 9JQ, UK
[g] Department of Computer Science, University College London, London, WC1E 6EA, UK

ARTICLE INFO

*Keywords:*
Blood glucose prediction
Interpretable
Graph
Recurrent neural networks
Diabetes

ABSTRACT

Accurate prediction of future blood glucose (BG) levels can effectively improve BG management for people living with type 1 or 2 diabetes, thereby reducing complications and improving quality of life. The state of the art of BG prediction has been achieved by leveraging advanced deep learning methods to model multimodal data, i.e., sensor data and self-reported event data, organized as multi-variate time series (MTS). However, these methods are mostly regarded as “black boxes” and not entirely trusted by clinicians and patients. In this paper, we propose interpretable graph attentive recurrent neural networks (GARNNs) to model MTS, explaining variable contributions via summarizing variable importance and generating feature maps by graph attention mechanisms instead of post-hoc analysis. We evaluate GARNNs on four datasets, representing diverse clinical scenarios. Upon comparison with fifteen well-established baseline methods, GARNNs not only achieve the best prediction accuracy but also provide high-quality temporal interpretability, in particular for postprandial glucose levels as a result of corresponding meal intake and insulin injection. These findings underline the potential of GARNN as a robust tool for improving diabetes care, bridging the gap between deep learning technology and real-world healthcare solutions.

## 1. Introduction

Diabetes is directly responsible for over a million deaths worldwide every year (WHO, 2023) due to complications arising from type 1 diabetes mellitus (T1DM) and type 2 diabetes mellitus (T2DM). The autoimmune reaction of people with T1DM destroys the cells in the pancreas which produce endogenous insulin, while people with T2DM predominantly have insulin resistance, which inhibits their ability to utilize insulin effectively. Difficulties to manage BG levels by endogenous insulin leads to hypoglycemia and hyperglycemia, causing serious health problems (Bloomgarden, 2004; Mora, Roche, & Rodríguez-Sánchez, 2023). Hence, effective self-management for BG levels is the key to the treatment (Woldaregay et al., 2019), because increased “Time in Range” has been shown to reduce the likelihood of complications (Bezerra, Neves, Neves, & Carvalho, 2023).

Continuous glucose monitoring (CGM) systems offer the ability to track BG levels every few minutes, generating continuous BG trajectories. BG level prediction (BGLP) based on CGM data (Cichosz, Kronborg, Jensen, & Hejlesen, 2021; Naumova, Pereverzyev, & Sivananthan, 2012; Plis, Bunescu, Marling, Shubrook, & Schwartz, 2014) allows people with diabetes to avoid hypo- and hyperglycemia by taking precautions in real-time. Recent work (Karim, Vassányi, & Kósa, 2020; Nemat, Khadem, Elliott, & Benaissa, 2023; Zhu et al., 2023; Zhu, Li, Herrero and Georgiou, 2023; Zhu et al., 2022) leverages multimodal data by organizing it as MTS in BGLP. In this case, apart from CGM data, the MTS input also includes sensor data, e.g., heart rate, and self-reported events, e.g., the amount of carbohydrate intake and bolus insulin injection. While these methods have the potential to further improve BGLP by levaraging the rich hidden information of MTS,

* Corresponding author.
*E-mail addresses:* chengzhe.piao.21@ucl.ac.uk (C. Piao), taiyu.zhu@psych.ox.ac.uk (T. Zhu), s.baldeweg@ucl.ac.uk (S.E. Baldeweg), p.taylor@ucl.ac.uk (P. Taylor), pantelis@imperial.ac.uk (P. Georgiou), sun@flock.io (J. Sun), junwang@cs.ucl.ac.uk (J. Wang), ken.li@ucl.ac.uk (K. Li).

https://doi.org/10.1016/j.neunet.2025.107229
Received 22 February 2024; Received in revised form 20 January 2025; Accepted 27 January 2025
Available online 5 February 2025

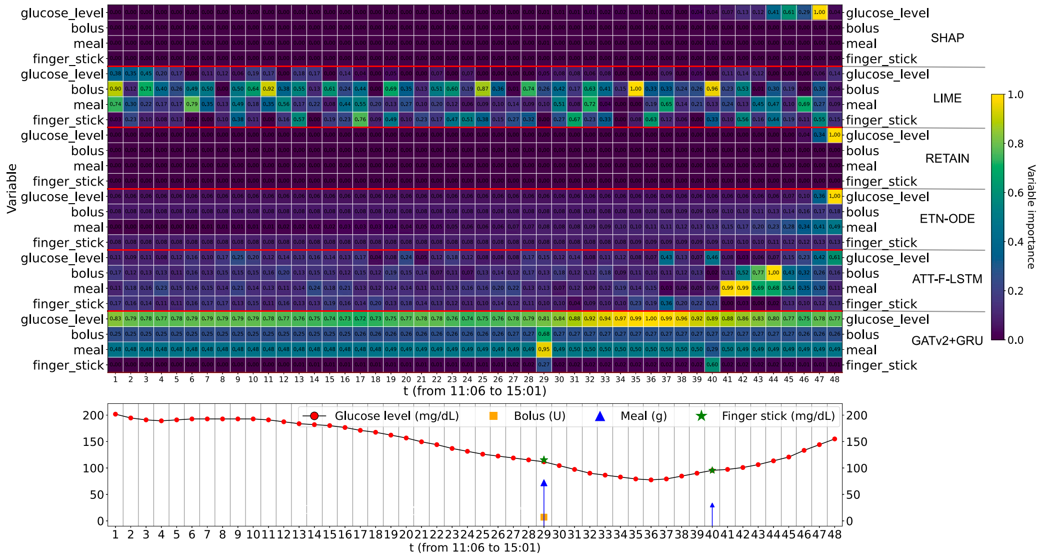


**Fig. 1.** Some feature maps of participant 1005 in ArisesT1DM. The bottom sub-figure is the visualization of a historical multivariate time serires, only showing “glucose_level (continuous glucose monitoring)”, “bolus”, “meal” and “finger_stick (capillary blood glucose test)”. The heat maps on the top are the feature maps of different methods. The x-axis/y-axis of heatmaps is the timestep/variable. The value in the cell is the variable importance, scaled to $[0, 1]$. Abbr.: SHAP (SHapley Additive exPlanations, Lundberg and Lee (2017)), LIME (Local Interpretable Model-agnostic Explanations, Ribeiro, Singh, and Guestrin (2016)), RETAIN (REverse Time AttentIoN, Choi et al. (2016)), ETN-ODE (Explainable Tensorized Neural Ordinary Differential Equations, Gao, Yang, Zhang, Huang, and Goulermas (2023)), ATT-F-LSTM (ATTention of Feature before Long Short-Term Memory (Gandin, Scagnetto, Romani, & Barbati, 2021)) and our proposed method “GATv2+GRU” (Graph Attention NeTworks version 2 by Brody, Alon, and Yahav (2022) and Gated Recurrent Unit by Cho, van Merrienboer, Bahdanau, and Bengio (2014)).

the lack of interpretability makes them less trustworthy. It is vital to understand how each variable contributes towards prediction rather than solely improving prediction accuracy.

However, post-hoc analysis methods, e.g., gradient-based attribution methods (Ancona, Ceolini, Öztireli, & Gross, 2018), are computationally inefficient (Guo, Lin, & Antulov-Fantulin, 2019) and difficult to be used by the researchers and clinicians without machine learning knowledge. Shapley Additive exPlanations (SHAP) values, as discussed by Lundberg and Lee (2017), are utilized alongside Long Short-Term Memory (LSTM, Hochreiter and Schmidhuber (1997)) models in BGLP (Cappon et al., 2020; Prendin et al., 2023). It should be noted that while they offer significant insights into the importance of a limited number of variables, their scope in temporal variable importance is somewhat restricted.

Comparably, attention-based recurrent neural networks (RNNs), e.g., RETAIN (Choi et al., 2016), ETN-ODE (Gao et al., 2023) and ATT-F-LSTM (Gandin et al., 2021), can inherently learn variable importance by the attention mechanisms during training. Nevertheless, the feature maps built on the variable importance are unhelpful to understand BGLP. Hence, we aim to propose a novel attention-based interpretable model that can rank variables in accordance with domain knowledge and generate understandable feature maps. For instance, Fig. 1 demonstrates the comparison of existing interpretable methods in interpreting the time-dependent importance of self-reported events for BG trajectories. Other interpretable methods fail to present their correct focus on self-reported events, as shown in their feature maps. On the contrary, our proposed method “GATv2+GRU” provides insights into the importance of variables when valid observations are available for those variables. For instance, in our feature map, the importance of “bolus” suddenly increases from 0.27 to 0.68 at $t = 29$, because this participant administrated bolus insulin at this timestep. When data is incomplete or invalid, like when missing points are filled with average values, the importance of variables tends to become stable at certain numbers. In such cases ($t \neq 29$), the importance of “bolus” typically remains near 0.26, reflecting an average importance due to this data padding. Besides, it can precisely capture the local maxima and minima. When $t = 36$, the “glucose_level” achieves its lowest local minima, and its importance increases to 1.0.

Both post-hoc analysis methods (SHAP and LIME) and attention-based methods (RETAIN, ETN-ODE, and ATT-F-LSTM) fail to generate significant feature maps for explaining future glucose predictions due to the complex temporal dependencies in multivariate time series data. SHAP and LIME, designed for independent and identically distributed (i.i.d.) data, struggle with the sequential nature of time series. ETN-ODE, which uses variable-wise temporal attention and variable attention to summarize outputs from parallel GRUs, fails to capture interactions among variables and introduces temporal biases. RETAIN, despite using an attention mechanism to combine embeddings of MTS, is influenced by RNN outputs. ATT-F-LSTM attempts to minimize RNN influence by applying attention mechanisms prior to RNN layers but fails to effectively capture useful information.

Comparatively, our proposed methods leverage graph attention networks to dynamically and explicitly model the correlations among variables at each time step. In this framework, multiple variables are treated as nodes, with their correlations represented by edges on the graph. The weights on these edges are learnable and changeable at different time steps. Variable importance, used for feature maps, is derived by aggregating these weights through a series of operations. Crucially, this importance is calculated before the data passes through the RNN structure, avoiding biases introduced by RNNs. By reducing temporal bias and explicitly modeling correlations among variables, our proposed model can generate significant feature maps.

When compared with both non-interpretable and interpretable approaches, our proposed methods outperform others in BGLP. Additionally, they offer effective explanations for MTS by inherently providing detailed insights. To summarize, the contributions of this paper are as follows:

- We propose Graph Attentive Recurrent Neural Networks, denoted as GARNNs, combining Graph Attention neTworks (GAT by Velickovic et al. (2018) or GATv2 by Brody et al. (2022)), with RNNs. We leverage GAT/GATv2 to explicitly model correlations among various variables, resulting in the inherent learning of temporal variable importance. Subsequently, RNNs are employed to aggregate temporal features for the prediction of future BG levels.

- We propose a novel temporal variable importance for MTS based on graph attention mechanisms. It can respectively summarize and generate significant variable importance ranking and feature maps in BGLP. We employ three properties and their corresponding explanations to demonstrate the creation, characteristics, and influence of temporal variable importance.
- Our methods, along with fifteen baseline approaches (thirteen for prediction comparisons and ten for interpretability comparisons), are assessed using four datasets. This evaluation focuses on the accuracy of predictions and the significance of variable ranking and feature maps. Not only do our methods exhibit outstanding performance, but they also attribute medically justifiable significance to crucial variables, including those with sparse signals like insulin injections and meal intake.

## 2. Related work

Interpretable methods can be generally divided into post-hoc analysis methods and attention-based methods.

Some temporal models (Shen, Wei, & Wang, 2023; Succetti, Rosato, & Panella, 2023) possess strong predictive abilities yet fall short in interpretability. However, it is still possible to derive explanations via post-hoc analysis methods, e.g., gradient-based analyzing (Bykov, Hedström, Nakajima, & Höhne, 2022; Shrikumar, Greenside, & Kundaje, 2017), variable contribution to the temporal shift of model (Tonekaboni, Joshi, Campbell, Duvenaud, & Goldenberg, 2020), Local Interpretable Model-agnostic Explanations (LIME, Ribeiro et al. (2016)). SHapley Additive exPlanations (SHAP, Lundberg and Lee (2017)), shapley based feature attributions (Kwon & Zou, 2022) and local rule based explanation (Rajapaksha & Bergmeir, 2022).

However, considering the efficiency and the target to explain the modeling of each MTS, the above methods are not considered in BGLP. Instead, we focus on attention-based methods, where variable importance based on attention is inherently learned during training.

Tensorized RNN methods (Aguiar, Santos, Watkinson, & Zhu, 2022; Chu et al., 2020; Gao et al., 2023; Guo et al., 2019; Shamout, Zhu, Sharma, Watkinson, & Clifton, 2020) leveraged parallel RNNs to model MTS, where each time series was modeled by a certain RNN. Then, they used attention to aggregate the outputs of parallel RNNs. In this regard, IMV-TENSOR (Guo et al., 2019) leveraged variable-wised temporal attention and variable attention to summarize all the outputs of parallel LSTMs, where the attention was directly mapped from the outputs of LSTMs by fully-connected layers. Similarly, instead of using parallel RNNs, Hsieh, Wang, Sun, and Honavar (2021) utilized parallel 1-dimensional convolution neural networks for modeling MTS aided by attention.

These parallel structures explicitly split the contribution of variables in the prediction, but they ignored the interaction among the variables. Besides, extracting variable importance from the temporal attention which worked on the outputs of RNNs could cause temporal biases. For example, as shown in Fig. 1, we cannot infer the variable importance of “meal” from the feature map of ETN-ODE, when $t = 29$. This is because some important content at $t = 29$ is passed to the last several timesteps by RNNs, and feature maps are built on the outputs of RNNs. While RETAIN (Choi et al., 2016) and RAIM (Xu, Biswal, Deshpande, Maher, & Sun, 2018) did not incorporate parallel structures and utilized the attention mechanism to combine embeddings of MTS instead of RNN outputs, their attention process was still influenced by the outputs from RNNs. For instance, RETAIN employed two separate RNNs to create attention at both the visit and variable levels. This RNN-derived attention was then applied to aggregate the MTS embeddings. Consequently, in RETAIN, the significance assigned to variables was somewhat biased by the RNNs, as illustrated in Fig. 1.

Even though certain researchers had attempted to minimize the influence of RNNs on the importance of temporal variables, the effectiveness of this importance remained uncertain. Gandin et al. (2021) and Kaji et al. (2019) had implemented attention mechanisms prior to the RNN layers, ensuring that the variable importance was not skewed by the RNNs. However, this approach, which was directed either by variable-wise temporal attention or time-wise variable attention, still led their models to erroneously focus on some irrelevant data points, particularly in the context of sparse signals. This issue is evident in models like ATT-F-LSTM (Gandin et al., 2021), as showcased in Fig. 1. Our approach leverages graph attention neural networks to explicitly model relationships among variables before applying RNNs, providing a strong foundation for calculating variable importance. This enables us to propose methods for generating high-quality variable rankings and feature maps, which we consider the key contribution of this work.

On the other hand, a mechanistic model (Man et al., 2014) uses known physical, biological, or chemical principles to describe and predict the behavior of a system through mathematical equations, such as ordinary differential equations (ODEs). These models capture essential mechanisms and offer interpretability and validation by adhering to established equations. In the context of glucose dynamics, a mechanistic model represents the physiological processes of insulin and glucose interaction within the body. Specifically, this model uses ODEs to simulate the dynamics of plasma glucose and insulin, endogenous glucose production, and glucose utilization, capturing the causal relationships and dependencies within these physiological systems. Recently, studies such as Miller, Foti, and Fox (2020), Wang, Levine, Shi, and Fox (2023) and Zou, Levine, Zaharieva, Johari, and Fox (2024) have proposed hybrid models that integrate the traditional UVA/Padova biomedical simulator (Man et al., 2014) with deep learning methods. This integration allows mechanistic models to leverage real patient data patterns, generating more accurate glucose responses to insulin delivery or meal consumption.

However, the interpretability provided by these mechanistic models differs from our objectives in this paper. While mechanistic models provide interpretability and validation by following established scientific principles and theories, our goal is to provide interpretability by calculating variable importance for each variable at each time step. Therefore, in our experiments, we can only compare the prediction accuracy of our methods with the latest mechanistic methods from Zou et al. (2024).

## 3. Preliminaries

Given a directed graph $\mathcal{G} \triangleq (\mathcal{N}, \mathcal{E}, \mathbf{E})$, we assume that the graph $\mathcal{G}$ has $N$ nodes, denoted as $\mathcal{N} \triangleq \{1, \dots, n, \dots, N\}$. Each node $n$ is connected with its neighborhood $\mathcal{N}^n \subseteq \mathcal{N}$, through the edges of $\mathcal{E}$. Besides, the presentations of all the nodes are $\mathbf{E} = [\mathbf{e}^1 \dots \mathbf{e}^n \dots \mathbf{e}^N] \in \mathbb{R}^{E \times N}$, where the presentation of the node $n$ is $\mathbf{e}^n$. Assuming that there are multiple layers in Graph Attention neTworks (GATs), defined as $\mathcal{L} \triangleq \{1, \dots, l, \dots, L\}$, each node $n$ in layer $l$ can receive and aggregate neural messages from $\mathcal{N}^n$. Then, one of the aggregation approaches is:

$$\mathbf{e}^{n,l+1} = \sum_{j \in \mathcal{N}^n} \tilde{s}^{n,j,l} (\mathbf{W}^l \mathbf{e}^{j,l} + \mathbf{b}^l), \tag{1}$$

$$\tilde{s}^{n,j,l} = \mathrm{Softmax}(s^{n,j,l}), \tag{2}$$

where the neural message from node $j \in \mathcal{N}^n$ is $\mathbf{W}^l \mathbf{e}^{j,l} + \mathbf{b}^l$; learnable parameters are $\mathbf{W}^l$ and $\mathbf{b}^l$; the attention weight from $j$ to $n$ is $\tilde{s}^{n,j,l} \in \mathbb{R}^1$.

We temporarily omit the superscript $l$ for simplicity. The attention weight $\tilde{s}^{n,j}$ is gotten by normalizing the score $s^{n,j}$. The score from node $j$ to node $n$ is $s^{n,j}$, as:

$$\text{GAT}: \quad s^{n,j} = \mathrm{LeakyReLU}(\mathbf{a}'^{\top} [\mathbf{W}\mathbf{e}^n + \mathbf{b}; \mathbf{W}\mathbf{e}^j + \mathbf{b}]), \tag{3}$$

$$\text{GATv2}: \quad s^{n,j} = \mathbf{a}^{\top} \mathrm{LeakyReLU}(\mathbf{W}' [\mathbf{e}^n; \mathbf{e}^j] + \mathbf{b}), \tag{4}$$

where $\mathbf{a} \in \mathbb{R}^A$ and $\mathbf{a}' \in \mathbb{R}^{2A}$ are learnable parameters; constant $A$ is a hyperparameter; the concatenation of vectors is $[;]$. LeakyReLU is the default activation function for both GAT and GATv2. According to Velickovic et al. (2018), LeakyReLU introduces nonlinearity in the attention mechanism, ensuring non-zero gradients for all inputs, efficient

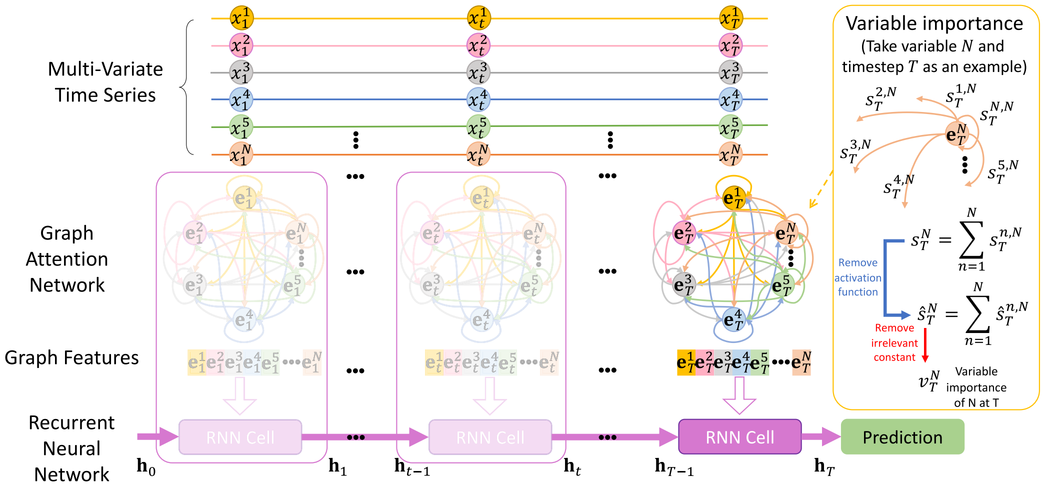

Multi-Variate Time Series
Graph Attention Network
Graph Features
Recurrent Neural Network
$x_1^1$ $x_1^2$ $x_1^3$ $x_1^4$ $x_1^5$ $x_1^N$
$x_t^1$ $x_t^2$ $x_t^3$ $x_t^4$ $x_t^5$ $x_t^N$
$x_T^1$ $x_T^2$ $x_T^3$ $x_T^4$ $x_T^5$ $x_T^N$
$\mathbf{e}_T^1$ $\mathbf{e}_T^2$ $\mathbf{e}_T^3$ $\mathbf{e}_T^4$ $\mathbf{e}_T^5$ $\mathbf{e}_T^N$
RNN Cell
Prediction
$\mathbf{h}_0$ $\mathbf{h}_1$ $\mathbf{h}_{t-1}$ $\mathbf{h}_t$ $\mathbf{h}_{T-1}$ $\mathbf{h}_T$
Variable importance
(Take variable $N$ and timestep $T$ as an example)
$s_T^{1,N}$ $s_T^{2,N}$ $s_T^{3,N}$ $s_T^{4,N}$ $s_T^{5,N}$ $s_T^{N,N}$
$s_T^N = \sum_{n=1}^N s_T^{n,N}$
Remove activation function
$\hat{s}_T^N = \sum_{n=1}^N \hat{s}_T^{n,N}$
Remove irrelevant constant
$v_T^N$ Variable importance of N at T

**Fig. 2.** Graph attentive recurrent neural networks (GARNNs). The observation of the variable $n \in \mathcal{N} \triangleq \{1,\dots,n,\dots,N\}$ at timestep $t \in \{1,\dots,t,\dots,T\}$ is $x_t^n$. The total length of the historical multivariate time series is $T$. The attention score from $j$ to $n$ is $s_t^{n,j}$ ($j \in \mathcal{N}$) for aggregating neural messages at node $n$. The variable importance of $j$ is $v_t^j$ which is gotten from $s_t^j$. The hidden state is $\mathbf{h}_t$.

gradient propagation, and stable training dynamics.

The above score function can be rewritten as:

$$\text{GAT}: \quad s^{n,j} = \mathrm{f}(\mathbf{q}^n, \mathbf{k}^j) = \mathrm{LeakyReLU}\left(\mathbf{a}_1^\top \mathbf{q}^n + \mathbf{a}_2^\top \mathbf{k}^j\right), \tag{5}$$

$$\text{GATv2}: \quad s^{n,j} = \mathrm{f}(\mathbf{q}^n, \mathbf{k}^j) = \mathbf{a}^\top \mathrm{LeakyReLU}(\mathbf{q}^n + \mathbf{k}^j), \tag{6}$$

where $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^A$. The query $\mathbf{q}^n$ is $\mathbf{W}_1\mathbf{e}^n + \mathbf{b}_1$, and the key $\mathbf{k}^j$ is $\mathbf{W}_2\mathbf{e}^j + \mathbf{b}_2$. As for Eq. (5), the learnable parameters $\{\mathbf{W}_1, \mathbf{b}_1\}$ should be the same as $\{\mathbf{W}_2, \mathbf{b}_2\}$.

Based on the findings presented by Brody et al. (2022), the scoring mechanism in GAT and GATv2 differs, with GAT having a static approach and GATv2 featuring a dynamic one. As a result, GATv2 demonstrates greater expressiveness in modeling intricate features of graphs, offering enhanced capabilities compared to its predecessor.

**Static scoring**: given queries $\{\mathbf{q}^n | n \in \mathcal{N}\}$ and keys $\{\mathbf{k}^m | m \in \mathcal{M}\}$, it holds $\mathrm{f}(\mathbf{q}^n, \mathbf{k}^m) \le \mathrm{f}(\mathbf{q}^n, \mathbf{k}^{m'=m^{max}})$, where $\forall n \in \mathcal{N}$, $\forall m \in \mathcal{M}$, $\exists m' \in \mathcal{M}$, $\mathrm{f}: \mathbb{R}^A \times \mathbb{R}^A \to \mathbb{R}^1$.

**Dynamic scoring**: given queries $\{\mathbf{q}^n | n \in \mathcal{N}\}$ and keys $\{\mathbf{k}^m | m \in \mathcal{M}\}$, it holds $\mathrm{f}(\mathbf{q}^n, \mathbf{k}^m) < \mathrm{f}(\mathbf{q}^n, \mathbf{k}^{m'=\phi(n)})$, where $\forall n \in \mathcal{N}$, $\exists m' \in \mathcal{M}$, $\forall m \in \mathcal{M}$ and $m \ne \phi(n)$. Meanwhile, function $\phi: \mathcal{N} \to \mathcal{M}$ and $\mathrm{f}: \mathbb{R}^A \times \mathbb{R}^A \to \mathbb{R}^1$.

## 4. Proposed model

### 4.1. Problem definition

Blood glucose (BG) levels represent the amount of glucose present in the blood. Glucose is a type of sugar and is the primary source of energy for the body's cells. BG levels are measured in milligrams of glucose per deciliter of blood (mg/dL) or in millimoles of glucose per liter of blood (mmol/L).

**Blood glucose level prediction based on multivariate time series (BGLP-MTS)**: given the values of variables $\mathcal{N}$ from historical timesteps $\mathcal{T} \triangleq \{1,\dots,t,\dots,T\}$, i.e., $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_t \dots \mathbf{x}_T] \in \mathbb{R}^{N \times T}$, predict the BG level $y_{T+H}$. The vector $\mathbf{x}_t = [x_t^1 \ \dots \ x_t^n \ \dots \ x_t^N]^\top$, and $H$ is a prediction horizon. We let $n = 1$ be the target variable, i.e., $x_t^1 = y_t$. The rest variables ($n > 1$) are exogenous variables when $N > 1$.

### 4.2. Overview

Our proposed models, GARNNs, build a graph at each timestep and use each node $n$ of the graph $\mathcal{G}$ to represent a variable $n$ of MTS, assuming each graph is initially a complete graph (see Fig. 2). Then, the input $\mathbf{e}_t^n$ of Eq. (1) is

$$\mathbf{e}_t^n = \mathrm{ReLU}(\mathbf{w}^n x_t^n + \mathbf{b}^n), \tag{7}$$

where learnable parameters $\mathbf{w}^n \in \mathbb{R}^E$. Then, we can use Eqs. (1)–(2) aided by Eq. (5) or Eq. (6) to update representations of $n$. Next, we collect the latest representations $\mathbf{e}_t^{1:N}$ and concatenate them as $\mathbf{e}_t = [\mathbf{e}_t^1; \dots; \mathbf{e}_t^n; \dots; \mathbf{e}_t^N]$.

After explicitly modeling correlations of these variables, we collect $\mathbf{e}_t$ of all timesteps, denoted as $\mathbf{e}_{1:T}$. Then, we leverage RNN to aggregate them, as:

$$\mathbf{h}_{1:T} = \mathrm{RNN}(\mathbf{e}_{1:T}, \mathbf{h}_{0:T-1}), \tag{8}$$

where we utilize gated recurrent unit (GRU, Cho et al. (2014)) as $\mathrm{RNN}(\cdot)$ to aggregate temporal features in this paper.

The GRU is a type of RNN architecture introduced to address the vanishing gradient problem in traditional RNNs. GRUs simplify the LSTM architecture by combining the forget and input gates into a single update gate, and merging the cell state and hidden state, making them computationally more efficient while maintaining performance. GRUs are particularly effective in sequence modeling tasks such as time series prediction and natural language processing. The structure of GRU is as follows:

$$\mathbf{r}_t = \sigma(\mathbf{W}_{re}\mathbf{e}_t + \mathbf{b}_{re} + \mathbf{W}_{rh}\mathbf{h}_{t-1} + \mathbf{b}_{rh}), \tag{9}$$

$$\mathbf{z}_t = \sigma(\mathbf{W}_{ze}\mathbf{e}_t + \mathbf{b}_{ze} + \mathbf{W}_{zh}\mathbf{h}_{t-1} + \mathbf{b}_{zh}), \tag{10}$$

$$\mathbf{n}_t = \tanh(\mathbf{W}_{ne}\mathbf{e}_t + \mathbf{b}_{ne} + \mathbf{r}_t * (\mathbf{W}_{nh}\mathbf{h}_{t-1} + \mathbf{b}_{nh})), \tag{11}$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) * \mathbf{n}_t + \mathbf{z}_t * \mathbf{h}_{t-1}, \tag{12}$$

where $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters, and $*$ is the element-wise product; $\sigma$ is the sigmoid activation function.

Finally, the prediction is $\hat{y}_{T+H} = \mathrm{MLP}(\mathbf{h}_T)$, where $\mathrm{MLP}(\cdot)$ consists of fully connected neural networks.

Given training examples $\mathcal{I} \triangleq \{1,\dots,i,\dots,I\}$, the objective function is:

$$J(\theta) = \frac{1}{I}\sum_i \left(\hat{y}_{T+H}(i) - y_{T+H}(i)\right)^2 + \frac{\lambda}{2}\|\theta\|_2^2, \tag{13}$$

where $\theta$ are all the learnable parameters of our proposed model; $\lambda$ is a hyperparameter.

### 4.3. Interpretability

In this section, we introduce three key properties that establish a robust framework for interpreting variable importance in GAT and its variant, GATv2. Property 1 hypothesis the existence of a universal interpretable variable importance for GAT and GATv2, inspired by the RETAIN model (Choi et al., 2016) which uses attention mechanisms to weigh the contributions of variables. This property provides a unified method for calculating variable importance based on learnable parameters and input, leveraging graph structures. Property 2 asserts that variable importance is derived from static scoring, a concept adapted from the work of Brody et al. (2022), highlighting that static scoring offers more stable and self-contained variable importance measures compared to dynamic scoring, which varies with different interactions. Lastly, Property 3 addresses the bounded differences between prediction modeling and variable importance calculation, with bounds dependent on the LeakyReLU parameter $\alpha$, ensuring that these differences remain minimal and controlled. Together, these properties underscore the reliability, consistency, and interpretability of variable importance in GAT and GATv2 models.

**Property 1.** *There exists a universal interpretable variable importance of GAT and GATv2.*

**Explanation.** In BGLP scenario, we assume each node $n$ connects with all of the nodes of $\mathcal{N}$. Then, as for the neighbor nodes $\mathcal{N}^n$ of the node $n$, we have $\mathcal{N}^n = \mathcal{N}$. Considering Eqs. (1)–(2) and Eqs. (5)–(6), for the variable $n$, the higher the score $s_t^{n,j}$, the more important $j$ to $n$. The mean of the scores from $j$ to all variables of $\mathcal{N}$ is denoted as:

$$s_t^j = \frac{1}{N} \sum_n s_t^{n,j}. \tag{14}$$

It is the impact of the variable $j$ on all the variables. Theoretically, a higher level of importance assigned to variable $j$ corresponds to an increased mean score $s_t^j$. As a result, an enhanced amount of information from $j$ is conveyed to and preserved within the embedding $\mathbf{e}_t$ collected from GAT/GATv2, thereby amplifying the contribution of $j$ in the prediction process.

However, we do not regard $s_t^j$ as the variable importance of $j$ at timestep $t$. Instead, we extract its variable importance from $s_t^j$ by removing irrelevant information. Specifically, considering that $\mathrm{LeakyReLU}(x)$ is monotonic with respect to $x$, we remove $\mathrm{LeakyReLU}(\cdot)$ from Eq. (14). GAT and GATv2 can be organized in the same format, as:

$$\hat{s}_t^j = \frac{1}{N} \sum_n \hat{s}_t^{n,j} = \frac{1}{N} \sum_n (\mathbf{a}_1^\top \mathbf{q}_t^n + \mathbf{a}_2^\top \mathbf{k}_t^j), \tag{15}$$

$$= \underbrace{\mathbf{a}_2^\top \mathbf{k}_t^j}_{\text{variable importance}} + \frac{1}{N} \sum_n \mathbf{a}_1^\top \mathbf{q}_t^n \tag{16}$$

where $\mathbf{a}_1$ is unequal and equal to $\mathbf{a}_2$ in GAT and GATv2 respectively. Then, we define the variable importance $\upsilon_t^j$ by removing the irrelevant item $\frac{1}{N} \sum_n \mathbf{a}_1^\top \mathbf{q}_t^n$ from $\hat{s}_t^j$:

$$\upsilon_t^j \triangleq \mathbf{a}_2^\top \mathbf{k}_t^j = \underbrace{\mathbf{a}_2^\top \mathbf{W}_2}_{\text{variable contribution}} \underbrace{\mathbf{e}_t^j}_{\text{variable embedding}} + \underbrace{\mathbf{a}_2^\top \mathbf{b}_2}_{\text{constant bias}}. \tag{17}$$

On the other hand, we can consider multiple layers of GAT or GATv2, the variable importance $\upsilon_t^j$ is defined as:

$$\upsilon_t^j = \frac{1}{L} \sum_l \upsilon_t^{j,l}, \quad \upsilon_t^{j,l} \triangleq \mathbf{a}_2^{l\,\top} \mathbf{k}_t^{j,l}. \tag{18}$$

Given $\mathcal{I}$, the variable importance of $\upsilon_j$ over $\mathcal{I}$ is:

$$\upsilon^j(\mathcal{I}) = \frac{1}{IT} \sum_{i,t} \upsilon_t^j(i), \tag{19}$$

where $\upsilon_t^j(i)$ is the variable importance of $j$ at timestep $t$ of the $i$th sample; the total number of training examples is $I$; the length of the MTS is $T$. ■

As in Eq. (17), our proposed variable importance $\upsilon_t^j$ is fully understandable. It consists of a variable contribution $\mathbf{a}_2^\top \mathbf{W}_2$, a variable embedding $\mathbf{e}_t^j$ and a constant bias $\mathbf{a}_2^\top \mathbf{b}_2$. The variable contribution and constant bias are learnable, so $\upsilon_t^j$ is directly guided by the variable embedding $\mathbf{e}_t^j$. When $L = 1$, the variable importance $\upsilon^j(\mathcal{I})$ is only affected by $j$. When $L > 1$, the variable importance $\upsilon^j(\mathcal{I})$ considers correlations of $j$ and other variables.

**Property 2.** *The variable importance $\upsilon_t^j$ is from static scoring.*

**Explanation.** Considering that $\hat{s}_t^{n,j} = \mathbf{a}_1^\top \mathbf{q}_t^n + \mathbf{a}_2^\top \mathbf{k}_t^j$, when $n$ and $t$ is fixed, $\mathbf{a}_1^\top \mathbf{q}_t^n$ can be seen as a constant. $\hat{s}_t^{n,j}$ is largely affected by $\mathbf{k}_t^j$. There can exist a $j' \in \mathcal{N}$, making $\hat{s}_t^{n,j} \le \hat{s}_t^{n,j'=j^{max}}$. This observation holds for any $n$ if $t$ is fixed.

The variable importance $\upsilon_t^j$ is extracted from $\hat{s}_t^j$ which is the mean of $\hat{s}_t^{n,j}$ over $n$, so $\upsilon_t^j$ is from static scoring. ■

The use of static scoring for determining variable importance aligns with our expectations, primarily because dynamic scoring does not guarantee consistent significance of variables. For instance, in the case of GATv2, if we arrange variable $j$ based on the value of $s_t^{n,j}$ among $\{s_t^{n,j} | j \in \mathcal{N}\}$, the dynamic scoring leads to considerable fluctuations in the ranking of $j$ within $\mathcal{N}$ as $n$ varies. Consequently, using $s_t^j$ from GATv2 as a measure of variable importance becomes unreliable.

In contrast, the ranking of $j$ in $\mathcal{N}$ based on the value of $\hat{s}_t^{n,j}$ in $\{\hat{s}_t^{n,j} | j \in \mathcal{N}\}$ remains constant despite changes in $n$. Averaging $\hat{s}_t^{n,j}$ over $n$ to reevaluate the ranking of $j$ does not alter its position, emphasizing the necessity for static scoring in assessing variable importance. Furthermore, eliminating $\frac{1}{N} \sum_n \mathbf{a}_1^\top \mathbf{q}_t^n$ from $\hat{s}_t^j$ also maintains the ranking of $j$, supporting this approach.

Additionally, the implementation of variable importance based on static scoring does not interfere with the dynamic scoring function of GATv2 in mapping out variable correlations. This ensures that GATv2 remains both effective in modeling and consistent in calculating variable importance.

**Property 3.** *The difference between $\hat{s}_t^j$ and $s_t^j$ is bounded by small values depending on the slope $\alpha \in [0, 1]$ of* $\mathrm{LeakyReLU}(\cdot)$.

**Explanation.** In terms of GATv2 (see Eq. (6)), we have:

$$s_t^j = \frac{1}{N} \sum_n \mathbf{a}^\top \tilde{\mathbf{I}}_t^{n,j} (\mathbf{q}_t^n + \mathbf{k}_t^j) = \frac{1}{N} \sum_n \mathbf{a}^\top \tilde{\mathbf{I}}_t^{n,j} \mathbf{m}_t^{n,j}, \tag{20}$$

where $\tilde{\mathbf{I}}_t^{n,j}$ is an indicate diagonal matrix, as:

$$\tilde{\mathbf{I}}_t^{n,j} = \mathrm{Diag}(i_{t,1}^{n,j}, \dots, i_{t,a}^{n,j}, \dots, i_{t,A}^{n,j}),$$

$$i_{t,a}^{n,j} = \begin{cases} 1, & m_{t,a}^{n,j} \ge 0 \\ \alpha, & m_{t,a}^{n,j} < 0, \alpha \in [0, 1], \end{cases}$$

and $m_{t,a}^{n,j}$ is the $a$th value of the vector $\mathbf{m}_t^{n,j}$.

$$s_t^j - \hat{s}_t^j = \frac{1}{N} \sum_n \mathbf{a}^\top \tilde{\mathbf{I}}_t^{nj} \mathbf{m}_t^{n,j} - \frac{1}{N} \sum_n \mathbf{a}^\top \mathbf{m}_t^{n,j}, \tag{21}$$

$$= \|\mathbf{a}\|_2 \left\| \frac{1}{N} \sum_n (\tilde{\mathbf{I}}_t^{n,j} - \mathbf{I}) \mathbf{m}_t^{n,j} \right\|_2 \cos\beta, \tag{22}$$

where vetorial angle is $\beta$. The boundary of the difference between $\hat{s}_t^j$ and $s_t^j$ is:

$$0 \le |s_t^j - \hat{s}_t^j| \le \|\mathbf{a}\|_2 \left\| \frac{1}{N} \sum_n (\tilde{\mathbf{I}}_t^{n,j} - \mathbf{I}) \mathbf{m}_t^{n,j} \right\|_2. \tag{23}$$

Eq. (13) has a soft constraint for $\|\theta\|_2 < \epsilon$. Based on the theory of Lagrange multipliers, there exists a $\lambda$ value that is equivalent to the hard constraint for $\|\theta\|_2 < \epsilon$, where $\epsilon$ is a small positive value. Then, for a learnable parameter $w$, we can have $|w| < \epsilon, \forall w \in \theta$. We assume each value of the input vector $\mathbf{e}_t^{j,l}$ of each GATv2 layer belongs to $[-c, c]$. Given that $\|\mathbf{a}\|_2 < \sqrt{A}\epsilon$ and $\left\| \frac{1}{N} \sum_n (\tilde{\mathbf{I}}_t^{n,j} - \mathbf{I}) \mathbf{m}_t^{n,j} \right\|_2 \le (1-\alpha) 2\sqrt{A}(Ec+1)\epsilon$, we have:

$$0 \le |s_t^j - \hat{s}_t^j| \le (1 - \alpha) 2A(Ec + 1)\epsilon^2. \tag{24}$$

Similarly, in terms of GAT, we have:

$$0 \le s_t^j - \hat{s}_t^j \le (1 - \alpha) 2A(Ec + 1)\epsilon^2. \quad ■ \tag{25}$$

The bounds can be treated as the gaps between the modeling for prediction and the calculation of variable importance. The gaps are affected by $\alpha$. Nevertheless, changing $\alpha$ has a slight impact on variable importance (see Fig. 4). Hence, we hold the view that the small gaps between the prediction and the variable importance can be ignored.

**Table 1**
The statistics of four datasets.

| Demographic | OhioT1DM | AT1DM | ShanghaiT1DM | ShanghaiT2DM |
|---|---|---|---|---|
| No. of participants (male/female) | 12 (7/5) | 12 (6/6) | 12 (7/5) | 100 (44/56) |
| Age (years) | 20–60 | 30–49 | 37–73 | 22–97 |
| Diabetes type | T1DM | T1DM | T1DM | T2DM |
| CGM | Medtronic Enlite | Empatica E4 | FreeStyle Libre H | FreeStyle Libre H |
| CGM sampling frequency ($\delta t$) | Every 5 min | Every 5 min | Every 15 min | Every 15 min |
| No. of days of CGM data per patient | 54 ± 2 | 49 ± 4 | 15 ± 9 | 12 ± 6 |
| No. of CGM records per patient | 13871 ± 1015 | 13324 ± 1081 | 1307 ± 849 | 1122 ± 580 |
| Mean of CGM data (mg/dL) | 159.35 ± 16.34 | 161.25 ± 26.02 | 166.51 ± 27.81 | 141.05 ± 29.70 |
| Standard deviation of CGM data (mg/dL) | 58.11 ± 6.15 | 57.06 ± 13.50 | 62.75 ± 11.95 | 40.89 ± 12.86 |
| Time in range (%) | 63.54 ± 9.70 | 63.29 ± 16.00 | 53.84 ± 12.26 | 77.74 ± 17.41 |
| Time below range (%) | 3.30 ± 2.25 | 2.92 ± 1.91 | 6.65 ± 6.45 | 2.54 ± 7.44 |
| Time above range (%) | 33.15 ± 10.71 | 33.78 ± 17.01 | 39.51 ± 16.40 | 19.72 ± 17.78 |
| Coefficient of variation (%) | 36.63 ± 3.70 | 35.14 ± 4.47 | 38.30 ± 7.16 | 28.89 ± 6.22 |
| Low blood glucose index | 0.88 ± 0.48 | 0.78 ± 0.46 | 1.63 ± 1.55 | 0.96 ± 1.97 |
| High blood glucose index | 7.15 ± 2.45 | 7.59 ± 4.17 | 8.87 ± 3.55 | 4.42 ± 3.72 |
| Other variables | Self-reported data; Sensor band (Empatica or basis peak) data | Self-reported data; Sensor band (Empatica E4) data | Self-reported data | Self-reported data; |

## 5. Experiments

Our experiments can be divided into two parts: evaluation of prediction performance and evaluation of variable importance. Before evaluating the methods, we first introduce the four datasets used in our experiments in Section 5.1. Then, we describe the baseline methods employed in our experiments in Section 5.2, which include traditional interpretable machine learning methods, post-hoc analysis methods, attention-based methods, uninterpretable deep learning methods, and neural network-based mechanism methods.

For the evaluation of prediction performance, excluding post-hoc analysis methods, we assess the prediction performance of all methods in Section 5.4 using the metrics defined in Section 5.3. We also vary key hyperparameters of our proposed methods, such as the number of GAT/GATv2 layers, to analyze their impact on prediction accuracy. Different variations of our proposed methods are evaluated alongside the selected baseline methods on four datasets.

Next, we evaluate the quality of variable importance by examining both instant variable importance ($v_t^j$) and summarized variable importance ($v^j(\mathcal{I})$) for all interpretable methods, including variations of our proposed methods and interpretable baseline methods. Summarized variable importance, which can be used for ranking variables, is evaluated based on medical knowledge and statistical data analysis in Section 5.5. We also consider the impact of the $\alpha$ parameter from our proposed methods on variable importance.

Finally, the quality of instant variable importance, represented as feature maps, is evaluated using three criteria: data validity, signal importance, and sparse signal detection in Section 5.6.

### 5.1. Datasets

We select four datasets containing multimodal data with high quality, where T1DM and T2DM groups in Shanghai dataset (Zhao et al., 2023) are regarded as two datasets. All datasets used in the study, except for ArisesT1DM (NCT ID: NCT03643692), are publicly accessible. The ArisesT1DM dataset was gathered in full compliance with relevant legal regulations. Consequently, ethical approval is not further required for this research. Details of these four dataset are presented in Table 1.

**OhioT1DM** (Marling & Bunescu, 2020): it is developed to promote and facilitate research in BGLP by providing eight weeks of CGM, insulin, physiological sensor, and self-reported life-event data for 12 adults with T1DM. Participants reported insulin doses, meal times, exercise, sleep, stress, and illness via a smartphone app. Data collection involved Medtronic insulin pumps, CGM sensors, and fitness bands (Basis Peak and Empatica Embrace). The dataset, de-identified according to HIPAA guidelines, requires a Data Use Agreement (DUA) for access, ensuring ethical use and protecting participant privacy.

**ShanghaiT1DM and ShanghaiT2DM** (Zhao et al., 2023):
it includes data from T1DM (n = 12) and T2DM (n = 100) patients in Shanghai, China. Data collection was conducted under real-life conditions, capturing clinical characteristics, laboratory measurements, medications, CGM readings, and daily dietary information. The study was ethically approved by the Ethics Committee of Shanghai Fourth People's Hospital and Shanghai East Hospital, affiliated with Tongji University, and informed consent was obtained from all participants. This ensures the protection of sensitive patient information, and data use is restricted to research purposes with appropriate ethical considerations.

**ArisesT1DM** (Zhu et al., 2022): The study utilized a clinically validated sensor wristband and CGM device under free-living conditions, collecting data from 12 adult participants with T1D over a six-week period. Participants were asked to log daily events such as insulin doses, meal macronutrient composition, alcohol intake, stress, illness, and exercise in a smartphone app. The study received ethical approval from the London - Fulham Research Ethics Committee (trial protocol 18/LO/1096), and all participants provided informed consent. The dataset included glucose and wristband data with specific focus on the impact of non-invasive physiological data on predicting glycemic events. In order to simplify the exhibition of experiment results, we leverage abbreviations of variable names (Zhu et al., 2022), e.g., electrodermal activity (EDA).

In this study, different from Zeevi et al. (2015), we did not use an external cohort to evaluate our proposed methods, because our focus is on developing personalized models. Our approach involves fine-tuning each model using individual patient data, thereby generating personalized models tailored to each patient's unique behavior. For evaluation, we divided each patient's data sequentially into training, validation, and test sets from different time periods. This approach aligns with other mainstream glucose prediction methods, such as those described by Cappon et al. (2020) and Prendin et al. (2023). OhioT1DM has been originally divided into training data and testing data by time for each participant. We further split the training data into the training part (80%) and validation part (20%). In terms of the rest three datasets, we respectively split the data into training data (60%), validation data (20%) and testing data (20%) by time for each participant. We use sliding windows to generate examples $(\mathbf{X}, y_{T+H})$, i.e., $T = 48$ and $H = 6$ for $\delta t = 5$ min, or $T = 16$ and $H = 2$ for $\delta t = 15$ min. All the examples are normalized by standard normalization and padded by zeros. We consider almost all the variables (see Fig. 3) in each dataset. However, certain variables with poor data quality, such as "stressors" in the OhioT1DM dataset, were excluded from our analysis.

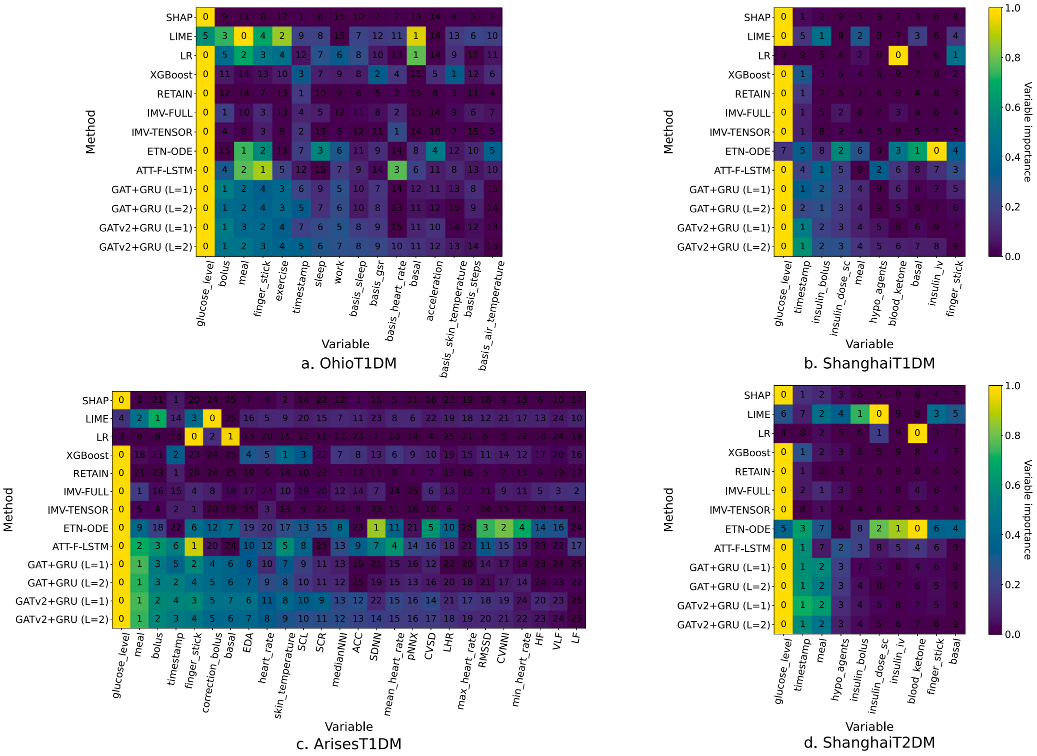


**Fig. 3.** The variable importance $v^j(\mathcal{I})$, scaled to $[0,1]$, of different methods, where $j$ is a variable from the variable set $\mathcal{N}$, and $\mathcal{I}$ contains all training examples. The number of graph layers is $L$. The x-axis/y-axis is variable/method. Brighter cell means higher $v^j(\mathcal{I})$. The number in a cell is the ranking place of the variable ranked by $v^j(\mathcal{I})$. Explanation of methods can refer to Section 5.2. EDA: electrodermal activity; SCL: skin conductance level; SCR: skin conductance response; medianNNI: median value of NN intervals; ACC: average 3D acceleration; SDNN: standard deviation of normal to normal (NN) intervals; pNNX: percentage of successive NN intervals greater than 50 ms; CVSD: coefficient of variation of successive differences; LHR: low-/high-frequency power ratio; RMSSD: root mean square of successive differences between adjacent NNs; CVNNI: coefficient of variation of NN intervals; HF: high frequency of heart rate in frequency domain; VLF: very high frequency of heart rate in frequency domain; LF: low frequency of heart rate in frequency domain.

We selected a 30-min prediction horizon (PH) because, following carbohydrate ingestion, BG levels typically begin to rise after 10 to 15 min (Tena, Garnica, Lanchares, & Hidalgo, 2021). Thus, a 30-min PH is the minimum duration necessary to implement corrective actions effectively (Balasooriya & Nanayakkara, 2020; Contador, Colmenar, Garnica, Velasco, & Hidalgo, 2022).

### *5.2. Baselines*

**SHapley Additive exPlanations (SHAP**, Lundberg and Lee (2017)) **and Local Interpretable Model-agnostic Explanations (LIME**, Ribeiro et al. (2016)): they are model-agnostic methods, providing variable importance for any methods. Given that our proposed methods got the best predicting performance compared with all of baseline methods, we leverage SHAP and LIME to explain our proposed methods for comparisons. We directly treat the absolute value of the variable importance provided by SHAP/LIME as $v^j$, and $v^j(\mathcal{I})$ is gotten by Eq. (19).

**Linear Regression (LR)**: it is a linear method. We flatten the input **X** and use scikit-learn (Pedregosa et al., 2011) to fit models. We aggregate the coefficients of the model as $v^j(\mathcal{I})$.

**eXtreme Gradient Boosting (XGBoost**, Chen and Guestrin (2016)): it is an optimized distributed gradient boosting approach. We flatten the input **X** to fit models. We regard the average gain as $v^j(\mathcal{I})$, where the gain is collected across all splits when using variables.

**REverse Time AttentIoN (RETAIN**, Choi et al. (2016)): it is an interpretable RNN model. The variable importance is calculated by the outputs of two RNNs, some learnable parameters and the input value of a variable. We use the absolute value of the variable importance of RETAIN as $v^j_t$, and $v^j(\mathcal{I})$ is gotten by Eq. (19).

**Interpretable Multi-Variable Long Short-Term Memories (IMV-LSTMs,** Guo et al. (2019)): both **IMV-TENSOR** and **IMV-FULL** are interpretable LSTMs by generating variable importance and variable-wise temporal importance. We directly average the variable importance of IMV-LSTMs with $\mathcal{I}$ examples and regard the mean variable importance as $v^j(\mathcal{I})$. Meanwhile, we treat the variable-wise temporal importance as $v^j_t$.

**Explainable Tensorized Neural Ordinary Differential Equations (ETN-ODE,** Gao et al. (2023)): it consists of: (1) Tensorized GRU; (2) tandem attention; (3) ordinary differential equation network. Part 1 and 2 are for the interpretation, which is similar as IMV-TENSOR, and part 3 is for the arbitrary-step prediction.

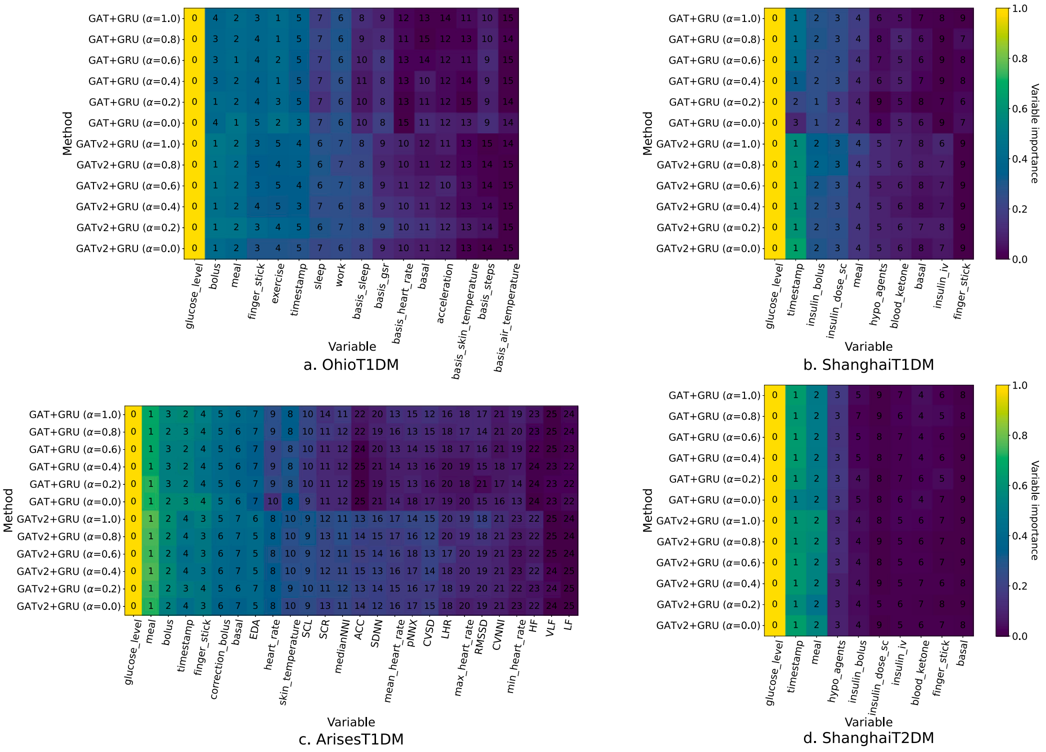
**ATTention of Time series before Long Short-Term Memory (ATT-T-LSTM,** Kaji et al. (2019)): it separately adds temporal attention for each variable before passing through LSTM. We regard the temporal attention as $v^j_t$, while $v^j(\mathcal{I})$ cannot be calculated by this method.

**ATTention of Features before Long Short-Term Memory (ATT-F-LSTM**, Gandin et al. (2021)): it adds variable attention at each timestep before LSTM. Hence, the attention weight of a variable $j$ at timestep $t$ is $v^j_t$, and $v^j(\mathcal{I})$ is calculated via Eq. (19).

**Neural Basis Expansion Analysis for Interpretable Time Series forecasting (N-BEATS**, Oreshkin, Carpov, Chapados, and Bengio (2020)): it is a non-interpretable deep learning model for univariate time series modeling. It utilizes a stack of fully connected neural network layers to model time series data, offering remarkable accuracy and flexibility.

**Neural Hierarchical Interpolation for Time Series forecasting (NHiTS**, Challu et al. (2023)): it is a non-interpretable deep learning model for MTS modeling, extending the N-BEATS and performing better in long-horizon prediction.

**Latent Parameter dynamics (LP), LP + State Closure (LPSC) and Mechanistic Neural ODE (MNODE)** (Zou et al., 2024): LP augments a



a. OhioT1DM

b. ShanghaiT1DM

c. ArisesT1DM

d. ShanghaiT2DM

**Fig. 4.** The variable importance $v^j(\mathcal{I})$, scaled to $[0,1]$, of "GAT+GRU" and "GATv2+GRU" changes with different $\alpha$, where $\alpha$ is a hyperparameter of $\mathrm{LeakyReLU}(\cdot)$. Two graph layers are considered, and $j$ is a variable from the variable set $\mathcal{N}$, and $\mathcal{I}$ contains all training examples. The x-axis/y-axis is variable/method. Brighter cell means higher $v^j(\mathcal{I})$. The number in a cell is the ranking place of the variable ranked by $v^j(\mathcal{I})$. Explanation of methods can refer to Section 5.2. EDA: electrodermal activity; SCL: skin conductance level; SCR: skin conductance response; medianNNI: median value of NN intervals; ACC: average 3D acceleration; SDNN: standard deviation of normal to normal (NN) intervals; pNNX: percentage of successive NN intervals greater than 50 ms; CVSD: coefficient of variation of successive differences; LHR: low-/high-frequency power ratio; RMSSD: root mean square of successive differences between adjacent NNs; CVNNI: coefficient of variation of NN intervals; HF: high frequency of heart rate in frequency domain; VLF: very high frequency of heart rate in frequency domain; LF: low frequency of heart rate in frequency domain.

mechanistic model, UVA/Padova (Man et al., 2014), by incorporating time-varying parameters governed by latent dynamics. This approach enhances the model's flexibility while retaining some of the underlying mechanistic structure, allowing it to adapt to changing conditions more effectively. LPSC extends the LP model by adding a state closure mechanism, where state include variables such as BG levels and insulin concentrations. This combines the flexibility of latent parameters with a correction term that adjusts the state dynamics based on observed residuals, thereby improving the model's accuracy and robustness. MNODE integrates mechanistic models with neural networks by using adjacency matrices to maintain dependencies between states. This approach allows the model to learn state dynamics, i.e., insulin-glucose dynamics, flexibly while preserving the causal relationships encoded in the mechanistic framework.

Our proposed method Graph Attentive Recurrent Neural Networks (**GARNNs**) are represented by Graph Attention neTworks (GAT by Velickovic et al. (2018)) or GATv2 by Brody et al. (2022) and Gated Recurrent Unit (GRU, Cho et al. (2014)), i.e., **"GAT+GRU"** and **"GATv2+GRU"**.

All the deep methods are implemented by PyTorch 1.11.0 following their original codes on github and run with NVIDIA RTX 3090 Ti. All the methods except LR are trained four times by changing the random seed. In terms of XGBoost, we search for the learning rate in $\{0.01, 0.1, 1.0\}$, n_estimators in $\{50, 100, 200\}$, max_depth in $\{3, 4, 5, 6, 7\}$, gamma in $\{0.5, 1, 1.5, 2, 5\}$ and min_child_weight in $\{1, 5, 10\}$. For all the deep methods, we search for the learning rate in $\{10^{-3}, 10^{-4}, 10^{-5}\}$. For IMV-LSTM and ETN-ODE, we find the variable-wise hidden state size in $\{8, 16, \ldots, 512/N\}$. In terms of the rest deep methods, we search for hidden state size in $\{128, 256, 512\}$. Besides, we also find $\lambda$ in $\{10^{-4}, 10^{-5}, 10^{-6}\}$. We choose hyperparameters by the performance of the validation data based on the metrics in the following subsection. Please refer to https://github.com/ChengzhePiao/garnn_public for more details.

### 5.3. Metrics

Considering the root mean square error (RMSE), mean absolute percentage error (MAPE) and mean absolute error (MAE), we also leverage the glucose-specific RMSE, denoted as gRMSE (Favero, Facchinetti, & Cobelli, 2012; Zhu et al., 2022), to evaluate all the methods. The gRMSE penalizes overestimation in hypoglycemia and underestimation in hyperglycemia. These two conditions can cause severe consequences to patients' health. Besides, time lag is determined by analyzing the correlation between the forecasted BG levels and the actual readings from CGM.

On the other hand, we also introduce Clarke Error Grids (CEG, Clarke, Cox, Gonder-Frederick, Carter, and Pohl (1987)) and Parkes Error Grids (PEG, Parkes, Slatin, Pardo, and Ginsberg (2000)) to evaluate the prediction performance. The CEG categorizes the clinical significance of differences between reference glucose values and predicted values. Developed in the 1980s by clinicians, the CEG consists of five zones. Zone A indicates clinically accurate results, while Zone B represents acceptable errors that lead to correct treatment decisions. Zones C, D, and E represent increasing levels of severity in treatment decisions, with Zone E indicating severely dangerous treatment decisions. The PEG, developed later, incorporates a broader consensus, including feedback from regulatory bodies, patients, and industry representatives, and also consists of five zones with similar clinical implications. Hence, for these two metrics, we aim for a higher percentage of predicted results in Zone A.

**Table 2**
Prediction of blood glucose levels in OhioT1DM.

| Method | RMSE (mg/dL) | MAPE (%) | MAE (mg/dL) | gRMSE (mg/dL) | Time lag (min) |
|---|---|---|---|---|---|
| LR | 22.19 ± 0.00(2.79)‡ | 10.90 ± 0.00(2.13)‡ | 15.92 ± 0.00(1.95)‡ | 27.69 ± 0.00(3.72)‡ | 8.34 ± 0.00(6.06)‡ |
| XGBoost | 22.51 ± 0.04(3.32)‡ | 10.92 ± 0.04(2.24)‡ | 16.08 ± 0.05(2.26)‡ | 28.89 ± 0.08(4.77)‡ | 9.36 ± 0.12(6.47)‡ |
| RETAIN | 20.30 ± 0.08(2.64)‡ | 9.78 ± 0.04(1.81)‡ | 14.41 ± 0.03(1.75)‡ | 25.48 ± 0.16(3.47)‡ | 7.39 ± 0.09(4.95)‡ |
| IMV-FULL | 21.61 ± 0.32(2.99)‡ | 10.21 ± 0.18(1.79)‡ | 15.18 ± 0.23(1.89)‡ | 27.16 ± 0.40(4.16)‡ | 6.42 ± 0.28(4.50) |
| IMV-TENSOR | 20.15 ± 0.03(2.77)‡ | 9.54 ± 0.02(1.82)‡ | 14.00 ± 0.02(1.75)‡ | 25.42 ± 0.07(3.79)‡ | 7.57 ± 0.11(4.86)‡ |
| ETN-ODE | 21.00 ± 0.22(2.92)‡ | 10.11 ± 0.13(1.95)‡ | 14.78 ± 0.18(1.93)‡ | 26.41 ± 0.32(3.94)‡ | 8.64 ± 0.35(5.22)‡ |
| ATT-T-LSTM | 21.62 ± 0.45(2.98)‡ | 10.46 ± 0.25(1.99)‡ | 15.32 ± 0.39(1.96)‡ | 27.08 ± 0.54(3.95)‡ | 8.12 ± 0.36(5.36)‡ |
| ATT-F-LSTM | 20.31 ± 0.06(2.69)‡ | 9.78 ± 0.04(1.81)‡ | 14.29 ± 0.04(1.72)‡ | 25.52 ± 0.09(3.67)‡ | 7.48 ± 0.15(5.24)‡ |
| N-BEATS | 20.15 ± 0.05(2.56)‡ | 9.62 ± 0.03(1.77)‡ | 14.11 ± 0.04(1.68)‡ | 25.31 ± 0.07(3.37)‡ | 7.98 ± 0.12(5.23)‡ |
| NHiTS | 20.14 ± 0.03(2.47)‡ | 9.60 ± 0.02(1.74)‡ | 14.07 ± 0.02(1.61)‡ | 25.24 ± 0.07(3.20)‡ | 7.55 ± 0.23(4.61)‡ |
| LP | 20.45 ± 0.38(2.88)‡ | 9.76 ± 0.24(1.89)‡ | 14.35 ± 0.34(1.88)‡ | 25.78 ± 0.49(3.89)‡ | 7.63 ± 0.83(5.33)‡ |
| LPSC | 20.79 ± 0.48(2.87)‡ | 10.01 ± 0.39(1.87)‡ | 14.67 ± 0.49(1.87)‡ | 26.03 ± 0.67(3.75)‡ | 8.76 ± 0.87(5.88)‡ |
| MNODE | 20.01 ± 0.04(2.76)‡ | 9.60 ± 0.03(1.82)‡ | 14.04 ± 0.04(1.72)‡ | 25.18 ± 0.07(3.71)‡ | 7.62 ± 0.22(5.18)‡ |
| GAT+GRU (L = 1) | 19.03 ± 0.07(2.40) | 9.10 ± 0.03(1.77) | 13.37 ± 0.03(1.65) | 23.75 ± 0.09(3.18) | 6.24 ± 0.14(4.45) |
| GAT+GRU (L = 2) | 19.08 ± 0.04(2.38) | 9.08 ± 0.02(1.76) | 13.37 ± 0.02(1.64) | 23.82 ± 0.08(3.15)∗ | 6.19 ± 0.25(4.51) |
| GATv2+GRU (L = 1) | **18.97 ± 0.06(2.43)** | **9.07 ± 0.01(1.78)** | **13.34 ± 0.02(1.68)** | **23.65 ± 0.10(3.21)** | **6.19 ± 0.14(4.47)** |
| GATv2+GRU (L = 2) | 19.11 ± 0.15(2.45) | 9.08 ± 0.03(1.78) | 13.38 ± 0.06(1.68) | 23.89 ± 0.22(3.29) | 6.30 ± 0.14(4.75) |

∗ $p \leq 0.05$; †$p \leq 0.01$; ‡$p \leq 0.005$;
Total historical timetamps is $T = 48$; Prediction horizon is $H = 6$; BG levels are sampled every $\delta t = 5$ min;
RMSE: root mean square error; MAPE: mean absolute percentage error;
MAE: mean absolute error; gRMSE: glucose-specific RMSE;
The result is formatted as "$mean \pm sd_1(sd_2)$";
$sd_1$ is the standard deviation after running the experiments four times by changing random seed;
$sd_2$ is the standard deviation of the metric results across the participants.
Explanation of methods can refer to Section 5.2.

**Table 3**
Prediction of blood glucose levels in ShanghaiT1DM.

| Method | RMSE (mg/dL) | MAPE (%) | MAE (mg/dL) | gRMSE (mg/dL) | Time lag (min) |
|---|---|---|---|---|---|
| LR | 22.57 ± 0.00(23.53)‡ | 11.47 ± 0.00(6.38)† | 15.81 ± 0.00(13.32)‡ | 26.33 ± 0.00(26.77)‡ | 2.50 ± 0.00(5.59) |
| XGBoost | 22.68 ± 0.14(12.65)‡ | 17.19 ± 0.18(14.52)‡ | 17.67 ± 0.07(10.86)‡ | 29.01 ± 0.27(17.30)‡ | 2.50 ± 0.00(5.59) |
| RETAIN | 16.25 ± 0.94(5.93)‡ | 10.77 ± 0.71(6.73)‡ | 12.28 ± 0.57(5.06)‡ | 20.00 ± 1.41(8.42)‡ | 3.12 ± 0.62(6.04)∗ |
| IMV-FULL | 13.63 ± 0.10(2.48) | 9.01 ± 0.18(3.45)∗ | 10.57 ± 0.12(2.22) | 16.65 ± 0.13(3.96) | 2.50 ± 0.00(5.59) |
| IMV-TENSOR | 13.88 ± 0.18(2.69) | 9.26 ± 0.23(3.64)‡ | 10.80 ± 0.17(2.45)‡ | 16.91 ± 0.23(4.10)∗ | 3.44 ± 0.54(6.27)∗ |
| ETN-ODE | 15.38 ± 0.23(3.35)‡ | 10.19 ± 0.29(4.18)‡ | 11.91 ± 0.21(3.04)‡ | 19.15 ± 0.33(5.14)‡ | 3.75 ± 0.00(6.50)∗ |
| ATT-T-LSTM | 14.03 ± 0.15(3.03) | 9.15 ± 0.15(3.80)∗ | 10.82 ± 0.14(2.77)∗ | 16.94 ± 0.22(4.29)∗ | 2.50 ± 0.00(5.59) |
| ATT-F-LSTM | 14.31 ± 0.21(2.97)‡ | 9.35 ± 0.23(3.78)‡ | 10.95 ± 0.14(2.62)‡ | 17.30 ± 0.28(4.17)‡ | 1.56 ± 0.54(4.51) |
| N-BEATS | 14.60 ± 0.27(3.01)‡ | 9.53 ± 0.31(3.43)‡ | 11.36 ± 0.26(2.61)‡ | 17.59 ± 0.40(4.54)‡ | 2.81 ± 0.54(5.82) |
| NHiTS | 14.86 ± 0.53(3.23)‡ | 9.56 ± 0.85(3.31)‡ | 11.41 ± 0.55(2.68)‡ | 18.00 ± 1.00(4.76)‡ | 3.44 ± 0.54(6.27)∗ |
| LP | 17.49 ± 0.30(4.34)‡ | 11.84 ± 0.43(5.09)‡ | 13.59 ± 0.31(3.85)‡ | 21.65 ± 0.46(6.38)‡ | 3.75 ± 0.00(6.50)∗ |
| LPSC | 16.85 ± 0.46(3.96)‡ | 11.10 ± 0.37(4.34)‡ | 13.00 ± 0.39(3.37)‡ | 20.66 ± 0.68(5.76)‡ | 3.75 ± 0.00(6.50)∗ |
| MNODE | 13.92 ± 0.08(2.62)∗ | 9.12 ± 0.17(3.42)‡ | 10.76 ± 0.11(2.30)‡ | 16.93 ± 0.18(4.01)∗ | 2.81 ± 0.54(5.82) |
| GAT+GRU (L = 1) | 13.80 ± 0.12(2.82) | 8.93 ± 0.07(3.70) | 10.53 ± 0.06(2.51) | 16.69 ± 0.14(4.49) | 1.25 ± 0.00(4.15) |
| GAT+GRU (L = 2) | 14.01 ± 0.34(2.85) | 9.11 ± 0.27(3.75)∗ | 10.66 ± 0.22(2.47) | 17.07 ± 0.47(4.56) | **0.62 ± 0.62(2.07)**∗ |
| GATv2+GRU (L = 1) | **13.62 ± 0.22(2.78)** | **8.74 ± 0.33(3.54)** | **10.38 ± 0.21(2.43)** | **16.44 ± 0.37(4.48)** | 1.88 ± 0.62(4.87) |
| GATv2+GRU (L = 2) | 13.98 ± 0.34(2.95)∗ | 8.97 ± 0.35(3.72) | 10.60 ± 0.26(2.54) | 16.90 ± 0.50(4.68) | 1.25 ± 0.00(4.15) |

∗ $p \leq 0.05$; †$p \leq 0.01$; ‡$p \leq 0.005$;
Total historical timetamps is $T = 16$; Prediction horizon is $H = 2$; BG levels are sampled every $\delta t = 15$ min;
RMSE: root mean square error; MAPE: mean absolute percentage error;
MAE: mean absolute error; gRMSE: glucose-specific RMSE;
The result is formatted as "$mean \pm sd_1(sd_2)$";
$sd_1$ is the standard deviation after running the experiments four times by changing random seed;
$sd_2$ is the standard deviation of the metric results across the participants.
Explanation of methods can refer to Section 5.2.

### 5.4. *Comparison of prediction performance*

In terms of RMSE, MAPE, MAE, gRMSE and time lag, the evaluation of prediction is shown in Tables 2–5. We also leveraged Wilcoxon test to evaluate the significance between "GATv2+GRU ($L$ = 1)" and other methods, where $p \leq 0.05$ means statistically significant. We have some observations based on these tables.

(1) Firstly, our proposed method, "GATv2+GRU ($L$ = 1)", outperforms all the baselines, while LR and XGboost perform worst.

(2) Deep methods perform better than non-deep methods. Non-interpretable methods (N-Beats and NHiTS) cannot promise better predicting performance compared with interpretable methods.

(3) Compared with "GAT+GRU", the dynamic scoring in "GATv2+GRU" can slightly improve the prediction performance in this scenario. Hence, the modeling via dynamic scoring can bring limited advantages to BGLP.

(4) Compared with neural network-based mechanistic methods (LP, LPSC, and MNODE), our method still performs better. Among these

**Table 4**
Prediction of blood glucose levels in ArisesT1DM.

| Method | RMSE (mg/dL) | MAPE (%) | MAE (mg/dL) | gRMSE (mg/dL) | Time lag (min) |
|---|---|---|---|---|---|
| LR | 25.17 ± 0.00(7.54)‡ | 12.14 ± 0.00(3.45)‡ | 18.20 ± 0.00(5.01)‡ | 31.93 ± 0.00(10.19)‡ | 11.67 ± 0.00(6.32)‡ |
| XGBoost | 24.84 ± 0.02(5.63)‡ | 12.07 ± 0.02(3.18)‡ | 17.99 ± 0.02(3.85)‡ | 32.32 ± 0.03(7.76)‡ | 13.37 ± 0.32(7.24)‡ |
| RETAIN | 21.46 ± 0.11(4.35)‡ | 10.41 ± 0.05(2.55)‡ | 15.49 ± 0.07(2.97)‡ | 27.27 ± 0.17(5.98)‡ | 11.01 ± 0.13(6.37)‡ |
| IMV-FULL | 24.44 ± 0.34(5.58)‡ | 11.93 ± 0.16(2.83)‡ | 17.89 ± 0.28(3.89)‡ | 30.89 ± 0.46(7.37)‡ | 11.26 ± 0.38(6.68)‡ |
| IMV-TENSOR | 21.48 ± 0.23(4.57)‡ | 10.31 ± 0.04(2.52)‡ | 15.35 ± 0.09(2.95)‡ | 27.41 ± 0.33(6.27)‡ | 10.78 ± 0.09(6.35)‡ |
| ETN-ODE | 23.18 ± 0.26(5.33)‡ | 10.99 ± 0.15(2.77)‡ | 16.38 ± 0.23(3.35)‡ | 29.75 ± 0.35(7.35)‡ | 12.20 ± 0.66(6.74)‡ |
| ATT-T-LSTM | 25.60 ± 0.64(6.04)‡ | 12.29 ± 0.26(2.85)‡ | 18.56 ± 0.44(4.02)‡ | 32.68 ± 0.82(8.31)‡ | 13.08 ± 0.48(6.49)‡ |
| ATT-F-LSTM | 21.85 ± 0.18(5.06)‡ | 10.49 ± 0.08(2.55)‡ | 15.72 ± 0.12(3.27)‡ | 27.75 ± 0.26(6.94)‡ | 10.62 ± 0.18(6.18)‡ |
| N-BEATS | 21.76 ± 0.03(4.48)‡ | 10.52 ± 0.02(2.57)‡ | 15.64 ± 0.02(3.03)‡ | 27.54 ± 0.04(6.11)‡ | 11.28 ± 0.21(6.25)‡ |
| NHiTS | 21.85 ± 0.04(4.54)‡ | 10.55 ± 0.02(2.61)‡ | 15.66 ± 0.03(3.07)‡ | 27.61 ± 0.05(6.17)‡ | 11.41 ± 0.25(6.25)‡ |
| LP | 22.32 ± 0.70(4.80)‡ | 10.68 ± 0.33(2.66)‡ | 15.96 ± 0.49(3.12)‡ | 28.52 ± 1.00(6.55)‡ | 11.61 ± 0.43(6.41)‡ |
| LPSC | 22.83 ± 0.71(4.90)‡ | 11.03 ± 0.47(2.66)‡ | 16.38 ± 0.61(3.18)‡ | 29.18 ± 1.26(6.76)‡ | 12.16 ± 0.49(7.08)‡ |
| MNODE | 21.53 ± 0.05(4.80)‡ | 10.30 ± 0.03(2.54)‡ | 15.33 ± 0.03(3.01)‡ | 27.39 ± 0.07(6.64)‡ | 10.89 ± 0.17(6.43)‡ |
| GAT+GRU (L = 1) | 20.02 ± 0.12(3.94) | 9.70 ± 0.05(2.30) | 14.50 ± 0.07(2.70) | 25.18 ± 0.15(5.35) | 9.57 ± 0.45(5.22) |
| GAT+GRU (L = 2) | 20.00 ± 0.11(3.91) | **9.66 ± 0.05(2.29)** | 14.48 ± 0.07(2.67) | 25.15 ± 0.14(5.28) | **9.38 ± 0.48(5.61)** |
| GATv2+GRU (L = 1) | **19.97 ± 0.07(3.93)** | 9.68 ± 0.05(2.26) | 14.47 ± 0.05(2.69) | **25.11 ± 0.13(5.31)** | 9.53 ± 0.28(5.26) |
| GATv2+GRU (L = 2) | 20.02 ± 0.11(3.81) | 9.67 ± 0.05(2.28) | **14.46 ± 0.08(2.61)** | 25.20 ± 0.13(5.18) | 9.81 ± 0.08(5.95) |

∗ $p \leq 0.05$; †$p \leq 0.01$; ‡$p \leq 0.005$;
Total historical timetamps is $T = 48$; Prediction horizon is $H = 6$; BG levels are sampled every $\delta t = 5$ min;
RMSE: root mean square error; MAPE: mean absolute percentage error;
MAE: mean absolute error; gRMSE: glucose-specific RMSE;
The result is formatted as "$mean \pm sd_1(sd_2)$";
$sd_1$ is the standard deviation after running the experiments four times by changing random seed;
$sd_2$ is the standard deviation of the metric results across the participants.
Explanation of methods can refer to Section 5.2.

**Table 5**
Prediction of blood glucose levels in ShanghaiT2DM.

| Method | RMSE (mg/dL) | MAPE (%) | MAE (mg/dL) | gRMSE (mg/dL) | Time lag (min) |
|---|---|---|---|---|---|
| LR | 17.10 ± 0.00(13.04)‡ | 9.34 ± 0.00(4.27)‡ | 12.03 ± 0.00(7.64)‡ | 19.62 ± 0.00(16.17)‡ | 1.50 ± 0.00(4.50)‡ |
| XGBoost | 16.75 ± 0.02(5.64)‡ | 11.09 ± 0.04(7.58)‡ | 12.79 ± 0.03(5.22)‡ | 20.12 ± 0.04(7.41)‡ | 1.09 ± 0.06(4.43) |
| RETAIN | 14.82 ± 0.42(13.45)‡ | 7.79 ± 0.05(3.52)‡ | 9.77 ± 0.11(4.44)‡ | 17.10 ± 0.42(15.29)‡ | 0.53 ± 0.48(8.11) |
| IMV-FULL | 11.84 ± 0.05(3.04)† | 6.97 ± 0.07(2.32) | 8.66 ± 0.06(2.32)∗ | 13.67 ± 0.07(3.91) | 0.82 ± 0.13(3.41) |
| IMV-TENSOR | 12.14 ± 0.13(3.15)‡ | 7.29 ± 0.12(2.79)‡ | 8.96 ± 0.08(2.48)‡ | 14.11 ± 0.12(4.08)‡ | 1.50 ± 0.00(4.50)‡ |
| ETN-ODE | 12.90 ± 0.30(3.58)‡ | 7.66 ± 0.17(3.01)‡ | 9.44 ± 0.20(2.71)‡ | 15.03 ± 0.34(4.63)‡ | 1.05 ± 0.18(3.81)† |
| ATT-T-LSTM | 13.92 ± 0.58(11.99)‡ | 7.41 ± 0.04(2.73)‡ | 9.34 ± 0.11(4.09)‡ | 15.91 ± 0.61(13.07)‡ | 0.90 ± 0.00(3.85) |
| ATT-F-LSTM | 12.76 ± 0.56(5.53)‡ | 7.29 ± 0.01(2.59)‡ | 9.08 ± 0.08(2.84)‡ | 14.72 ± 0.53(6.37)‡ | 1.05 ± 0.00(3.83)† |
| N-BEATS | 12.15 ± 0.03(3.19)‡ | 7.12 ± 0.05(2.09)‡ | 8.90 ± 0.03(2.44)‡ | 14.08 ± 0.05(4.06)‡ | 1.35 ± 0.00(4.29)‡ |
| NHiTS | 12.12 ± 0.05(3.17)‡ | 7.08 ± 0.08(2.09)‡ | 8.85 ± 0.06(2.39)‡ | 14.04 ± 0.09(4.03)‡ | 1.35 ± 0.11(4.29)‡ |
| LP | 14.65 ± 0.11(3.87)‡ | 8.64 ± 0.18(2.75)‡ | 10.73 ± 0.16(2.90)‡ | 17.27 ± 0.15(4.81)‡ | 2.70 ± 0.00(6.14)‡ |
| LPSC | 14.73 ± 0.07(3.83)‡ | 8.72 ± 0.15(2.74)‡ | 10.85 ± 0.16(2.92)‡ | 17.33 ± 0.11(4.79)‡ | 2.78 ± 0.13(6.20)‡ |
| MNODE | 12.59 ± 0.13(3.34)‡ | 7.27 ± 0.04(2.22)‡ | 9.08 ± 0.02(2.41)‡ | 14.51 ± 0.11(4.13)‡ | 1.54 ± 0.06(4.55)‡ |
| GAT+GRU (L = 1) | 11.78 ± 0.05(3.10) | 6.97 ± 0.05(2.36)‡ | 8.63 ± 0.05(2.37)† | 13.62 ± 0.06(3.96) | 0.86 ± 0.06(3.49) |
| GAT+GRU (L = 2) | 11.75 ± 0.04(3.07) | 6.98 ± 0.04(2.47) | 8.63 ± 0.02(2.38) | 13.60 ± 0.06(3.93) | 0.79 ± 0.12(3.33) |
| GATv2+GRU (L = 1) | **11.72 ± 0.02(3.03)** | **6.93 ± 0.05(2.29)** | **8.59 ± 0.03(2.34)** | **13.55 ± 0.05(3.88)** | **0.79 ± 0.06(3.34)** |
| GATv2+GRU (L = 2) | 11.74 ± 0.04(3.03) | 7.00 ± 0.05(2.60) | 8.62 ± 0.04(2.36) | 13.58 ± 0.04(3.87) | 0.79 ± 0.06(3.34) |

∗ $p \leq 0.05$; †$p \leq 0.01$; ‡$p \leq 0.005$;
Total historical timetamps is $T = 16$; Prediction horizon is $H = 2$; BG levels are sampled every $\delta t = 15$ min;
RMSE: root mean square error; MAPE: mean absolute percentage error;
MAE: mean absolute error; gRMSE: glucose-specific RMSE;
The result is formatted as "$mean \pm sd_1(sd_2)$";
$sd_1$ is the standard deviation after running the experiments four times by changing random seed;
$sd_2$ is the standard deviation of the metric results across the participants.
Explanation of methods can refer to Section 5.2.

three methods, MNODE primarily relies on neural networks, effectively removing traditional ODE functions. Instead, it introduces causality in its loss function and uses adjacency matrices to encode the dependencies between states. This further confirms that primarily neural network-based models perform the best in BGLP.

Similarly, as shown in Tables 6–7, our proposed methods mostly achieve the best performance. In terms of the percentage of BG predictions in Zone A, representing accurate predictions, our methods achieve over 88% and 89% in CEG and PEG, respectively. For the combined percentage of predictions in Zones A and B, indicating acceptable predictions, our methods achieve over 98% and 99% in CEG and PEG, respectively. Figs. 5–6 visualize the CEG and PEG results for our method, "GATv2+GRU", showing that predictions are concentrated within Zones A and B, consistent with the tables.

Furthermore, we separately recorded the time consumption of evaluation with all testing examples on four datasets (see Table 8). Regarding running efficiency, traditional machine learning methods (LR and XGBoost) are the fastest, consuming the least amount of time. Uninterpretable methods (N-BEATS and NHiTS) follow, running faster than neural network-based interpretable methods (RETAIN to ATT-F-LSTM and our proposed methods) and mechanistic methods (LP, LPSC, MNODE).

**Table 6**
Percentage of blood glucose predictions in Zone A of the Clarke Error Grid (%).

| Method | OhioT1DM | AriseskT1DM | ShanghaiT1DM | ShanghaiT2DM |
|---|---|---|---|---|
| LR | 85.40 ± 0.00(5.33)‡ | 82.71 ± 0.00(8.43)‡ | 87.60 ± 0.00(8.29)† | 91.25 ± 0.00(8.68)‡ |
| XGBoost | 85.59 ± 0.06(5.48)‡ | 82.65 ± 0.09(7.64)‡ | 79.01 ± 0.28(17.16)‡ | 85.62 ± 0.28(16.58)‡ |
| RETAIN | 88.03 ± 0.13(4.35)‡ | 86.07 ± 0.04(6.42)‡ | 88.20 ± 0.51(12.95)‡ | 94.30 ± 0.23(5.03)‡ |
| IMV-FULL | 87.26 ± 0.43(4.15)‡ | 83.13 ± 0.53(6.82)‡ | 90.24 ± 0.47(8.65) | 95.46 ± 0.08(3.77) |
| IMV-TENSOR | 88.63 ± 0.02(4.26)‡ | 86.30 ± 0.09(6.58)‡ | 89.11 ± 0.64(10.36)‡ | 95.20 ± 0.31(4.60)∗ |
| ETN-ODE | 87.23 ± 0.28(4.72)‡ | 84.68 ± 0.32(7.00)‡ | 87.28 ± 0.62(10.70)‡ | 94.12 ± 0.37(5.23)‡ |
| ATT-T-LSTM | 86.41 ± 0.63(4.84)‡ | 81.82 ± 0.59(7.33)‡ | 89.65 ± 0.59(10.49)† | 95.01 ± 0.08(4.19)‡ |
| ATT-F-LSTM | 88.18 ± 0.11(4.23)‡ | 86.26 ± 0.12(6.26)‡ | 89.38 ± 0.43(10.15)‡ | 95.05 ± 0.07(4.06)‡ |
| N-BEATS | 88.47 ± 0.05(4.21)‡ | 85.67 ± 0.05(6.50)‡ | 89.54 ± 1.02(7.76)‡ | 95.12 ± 0.09(4.21)‡ |
| NHiTS | 88.59 ± 0.11(4.00)‡ | 85.73 ± 0.18(6.41)‡ | 89.76 ± 2.14(6.33)‡ | 95.24 ± 0.05(4.14)‡ |
| LP | 88.33 ± 0.59(4.42)‡ | 85.56 ± 0.87(6.72)‡ | 84.26 ± 0.64(10.98)‡ | 92.04 ± 0.35(5.93)‡ |
| LPSC | 87.74 ± 0.96(4.42)‡ | 84.79 ± 1.12(6.66)‡ | 85.22 ± 0.49(10.01)‡ | 91.96 ± 0.25(5.89)‡ |
| MNODE | 88.76 ± 0.09(4.14)‡ | 86.29 ± 0.11(6.57)‡ | 89.84 ± 0.64(8.70)† | 95.00 ± 0.06(4.10)‡ |
| GAT+GRU (L = 1) | 89.72 ± 0.09(3.98) | 88.00 ± 0.23(5.62) | 90.70 ± 0.35(8.54) | 95.39 ± 0.12(3.96) |
| GAT+GRU (L = 2) | 89.69 ± 0.02(3.91) | **88.16 ± 0.13(5.49)** | 90.36 ± 0.65(8.43) | 95.43 ± 0.07(3.99) |
| GATv2+GRU (L = 1) | 89.74 ± 0.07(3.97) | 88.06 ± 0.13(5.52) | **91.24 ± 1.32(7.48)** | **95.51 ± 0.04(3.84)** |
| GATv2+GRU (L = 2) | **89.82 ± 0.12(3.97)** | 88.13 ± 0.10(5.49) | 90.54 ± 0.87(8.48) | 95.45 ± 0.14(3.94) |

∗ $p \leq 0.05$; † $p \leq 0.01$; ‡ $p \leq 0.005$;
The result is formatted as "$mean \pm sd_1(sd_2)$";
$sd_1$ is the standard deviation after running the experiments four times by changing random seed;
$sd_2$ is the standard deviation of the metric results across the participants.
Explanation of methods can refer to Section 5.2.

**Table 7**
Percentage of blood glucose predictions in Zone A of the Parkes Error Grid (%).

| Method | OhioT1DM | ArisesT1DM | ShanghaiT1DM | ShanghaiT2DM |
|---|---|---|---|---|
| LR | 87.48 ± 0.00(4.31)‡ | 85.04 ± 0.00(6.82)‡ | 90.83 ± 0.00(7.29)‡ | 95.56 ± 0.00(5.33)‡ |
| XGBoost | 87.62 ± 0.09(4.47)‡ | 84.89 ± 0.06(6.12)‡ | 82.94 ± 0.14(17.10)‡ | 96.13 ± 0.03(7.64)‡ |
| RETAIN | 89.58 ± 0.06(3.46)‡ | 88.21 ± 0.12(4.86)‡ | 90.81 ± 0.87(11.30)‡ | 98.06 ± 0.05(2.11)‡ |
| IMV-FULL | 88.79 ± 0.34(3.42)‡ | 85.14 ± 0.43(5.38)‡ | **95.50 ± 0.24(3.37)** | 98.32 ± 0.05(1.79)∗ |
| IMV-TENSOR | 90.41 ± 0.05(3.19)‡ | 88.52 ± 0.09(4.71)‡ | 94.64 ± 0.91(4.51) | 98.25 ± 0.05(1.88)‡ |
| ETN-ODE | 89.41 ± 0.23(3.54)‡ | 87.09 ± 0.29(5.24)‡ | 91.83 ± 0.29(7.03)‡ | 98.03 ± 0.07(1.96)‡ |
| ATT-T-LSTM | 88.25 ± 0.54(3.91)‡ | 83.95 ± 0.67(5.77)‡ | 94.20 ± 0.75(5.52) | 98.19 ± 0.05(1.83)‡ |
| ATT-F-LSTM | 90.03 ± 0.10(3.33)‡ | 88.55 ± 0.14(4.72)‡ | 93.95 ± 0.72(4.90)‡ | 98.21 ± 0.05(1.91)‡ |
| N-BEATS | 90.02 ± 0.06(3.31)‡ | 87.92 ± 0.08(4.92)‡ | 94.33 ± 0.45(4.03)‡ | 97.82 ± 0.03(2.22)‡ |
| NHiTS | 90.22 ± 0.06(3.20)‡ | 87.82 ± 0.15(4.96)‡ | 93.75 ± 1.08(4.12)‡ | 97.86 ± 0.04(2.09)‡ |
| LP | 89.93 ± 0.42(3.40)‡ | 87.59 ± 0.78(4.99)‡ | 88.56 ± 0.92(9.16)‡ | 97.31 ± 0.11(2.59)‡ |
| LPSC | 89.46 ± 0.71(3.53)‡ | 87.06 ± 0.92(4.94)‡ | 90.04 ± 1.35(6.79)‡ | 97.22 ± 0.12(2.67)‡ |
| MNODE | 90.50 ± 0.05(3.31)‡ | 88.65 ± 0.08(4.57)‡ | 94.92 ± 0.33(3.24) | 97.79 ± 0.15(2.05)‡ |
| GAT+GRU (L = 1) | 91.12 ± 0.09(3.08) | 89.77 ± 0.12(4.44) | 94.54 ± 0.55(4.30) | 98.40 ± 0.03(1.68) |
| GAT+GRU (L = 2) | 91.04 ± 0.10(3.06) | 89.74 ± 0.17(4.38) | 94.29 ± 0.82(4.38) | 98.44 ± 0.04(1.66) |
| GATv2+GRU (L = 1) | **91.13 ± 0.12(3.13)** | 89.82 ± 0.12(4.36) | 95.07 ± 0.87(4.26) | **98.42 ± 0.03(1.67)** |
| GATv2+GRU (L = 2) | 91.09 ± 0.03(3.11) | **89.85 ± 0.09(4.32)** | 94.27 ± 1.08(4.76)∗ | 98.41 ± 0.02(1.65) |

∗ $p \leq 0.05$; † $p \leq 0.01$; ‡ $p \leq 0.005$;
The result is formatted as "$mean \pm sd_1(sd_2)$";
$sd_1$ is the standard deviation after running the experiments four times by changing random seed;
$sd_2$ is the standard deviation of the metric results across the participants.
Explanation of methods can refer to Section 5.2.

When comparing among the interpretable methods, taking OhioT1DM as an example, testing time on the OhioT1DM dataset varies from 2.105 s to 28.121 s, with most interpretable methods taking between 2 to 7 s. ETN-ODE has the longest testing time. Among the variations of our proposed methods, one layer of GAT/GATv2 consumes less than 7 s, while two layers of GAT/GATv2 double the testing time consumption.

Additionally, we provide the theoretical time complexity in Table 9. The results in Table 9 are generally consistent with those in Table 8, although differences in implementation can impact actual time consumption. For instance, according to Table 9, IMV-TENSOR should outperform IMV-FULL in terms of speed. However, Table 8 shows that IMV-TENSOR is slightly slower. This discrepancy arises because IMV-TENSOR was implemented serially in our experiments, while a parallel implementation would likely make it faster. As shown in Table 9, excluding ETN-ODE, RNN-based methods (RETAIN, IMV-FULL, IMV-TENSOR, ATT-T-LSTM, ATT-F-LSTM, LP, LPSC, MNODE and our proposed methods) are expected to have similar testing times, which is confirmed by the results in Table 8. While ETN-ODE is also RNN-based, the time consumption of the ODE solver makes it slower, as indicated in Table 8. Additionally, LP, LPSC, and MNODE are RNN-based as well, since they use LSTM to encode historical time series. Additionally, since the results in Table 8 represent the total testing time for all examples across each of the four datasets, as for our proposed methods, the time required to test a single example is significantly less than one second, making it negligible for daily use.

### 5.5. Interpretation of variable importance

After seeing the variable ranking lists w.r.t. $v^j(\mathcal{I})$ in four datasets in Fig. 3, we have the observations as follows:

(1) The target variable, "glucose_level", should be the most important in BGLP (Bevan & Coenen, 2020; Zhu et al., 2022). GARNNs, ATT-F-LSTM, IMV-LSTMs, RETAIN, XGBoost and SHAP can consistently focus on the target variable, while the rest methods often lose focus.

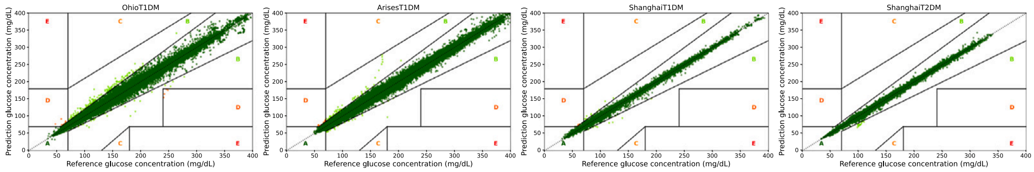


**Fig. 5.** Visualization of Clarke Error Grid by GATv2+GRU ($L$ = 1).

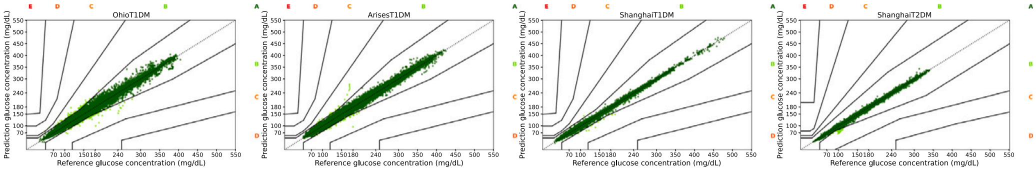


**Fig. 6.** Visualization of Parkes Error Grid by GATv2+GRU ($L$ = 1).

**Table 8**
Time consumption of evaluation with all testing examples (seconds)

| Method | OhioT1DM | AriseT1DM | ShanghaiT1DM | ShanghaiT2DM |
|---|---|---|---|---|
| LR | 0.017 | 0.032 | 0.002 | 0.002 |
| XGBoost | 0.028 | 0.027 | 0.002 | 0.002 |
| RETAIN | 2.567 | 2.708 | 0.166 | 1.372 |
| IMV-FULL | 4.978 | 5.187 | 0.261 | 1.909 |
| IMV-TENSOR | 6.225 | 6.411 | 0.290 | 2.277 |
| ETN-ODE | 28.121 | 27.033 | 1.062 | 7.478 |
| ATT-T-LSTM | 2.105 | 2.345 | 0.150 | 1.234 |
| ATT-F-LSTM | 2.612 | 2.520 | 0.175 | 1.406 |
| N-BEATS | 0.901 | 1.011 | 0.112 | 0.771 |
| NHiTS | 1.425 | 1.447 | 0.138 | 1.219 |
| LP | 2.297 | 2.279 | 0.158 | 1.323 |
| LPSC | 4.487 | 4.633 | 0.221 | 1.702 |
| MNODE | 3.341 | 2.984 | 0.188 | 1.483 |
| GAT+GRU (L = 1) | 6.986 | 7.734 | 0.315 | 2.642 |
| GAT+GRU (L = 2) | 11.190 | 12.990 | 0.496 | 3.738 |
| GATv2+GRU (L = 1) | 5.691 | 6.348 | 0.316 | 2.267 |
| GATv2+GRU (L = 2) | 8.528 | 10.233 | 0.410 | 3.003 |

(2) GARNNs typically outperform baseline models by assigning significant importance to “timestamp” in the rankings, particularly for Shanghai T1DM and T2DM where the sample frequency of CGM is relatively low. This characteristic aligns with clinical observations that BG fluctuations are linked to personal lifestyles, exhibiting distinct temporal patterns. It is significant that useful exogenous variables are highlighted when lacking enough values from the target variable in Shanghai datasets.

(3) Given that we predict the future BG of CGM instead of “finger_stick”, i.e., capillary BG test, more difference between them may reduce the importance of “finger_stick”. The MAE between “glucose_level” and “finger_stick” in OhioT1DM and AriseT1DM is lower, i.e., 20.92 and 9.97 mg/dL, respectively. The MAE between them in ShanghaiT1DM and ShanghaiT2DM are respectively 29.03 and 22.31 mg/dL. Hence, GARNNs give special highlights to “finger_stick” in OhioT1DM and AriseT1DM but less importance in ShanghaiT1DM and ShanghaiT2DM because of the larger MAE, while the variable importance ranking of other methods does not have similar observations.

(4) Apart from “glucose_level”, “finger_stick” and “timestamp”, both bolus insulin (“bolus”, “correction_bolus”, “insulin_dose_sc” and “insulin_bolus”) and carbohydrate intake (“meal”) should be important variables as well, because they can rapidly affect the BG levels within short periods. Basal insulin (“basal”), also called “background insulin”, steadily controls BG levels for long periods. Besides, “insulin_iv” is an intravenous insulin infusion for super serious hyperglycemia under some extreme circumstances.

Therefore, in terms of daily cases, bolus insulin tends to be more important than basal insulin in BGLP. GARNNs follow this knowledge in four datasets, while other methods fail to do that. Besides, in AriseT1DM, GARNNs give “bolus” more importance than “correction_bolus”, because the latter is an extra insulin taken during hyperglycemia. The amount of the latter is much less than the former.

On the other hand, a predominant reason for T2DM is that cells respond inactively to insulin. Hence, compared with the variable ranking in ShanghaiT1DM, GARNNs reduce the importance of bolus insulin and give more importance to “meal” and non-insulin hypoglycemic agents, “hypo_agents”, in ShanghaiT2DM.

(5) In OhioT1DM, the self-reported events, i.e., “exercise” and “sleep”, should be more important than the sensor data, e.g., “basis_heart_rate”, “basis_gsr”, etc. This is because both exercising and sleeping can indirectly cause changes in BG levels. Unlike the other methods, “GATv2+GRU” reflects this knowledge. Besides, “sleep” and “basis_sleep” are self-reported and sensor-detected, respectively, but “basis_sleep” misses 56.89% of the sleeping intensity data, so GARNNs give more importance to “sleep”.

(6) In the analysis of the number of layers $L$ in GAT or GATv2, it is observed that this can slightly alter the ranking of variable importance $v^j(\mathcal{I})$. For instance, different configurations like “GATv2+GRU($L$ = 2)” show varying importance rankings for variables like “sleep” and “work”. This was further substantiated by a feature ablation study in OhioT1DM (Piao & Li, 2023), indicating that “sleep” improves prediction accuracy more than “work”. The preference for “sleep” by “GATv2+GRU($L$ = 2)” seems more aligned with these findings, suggesting that multiple layers in GATv2 might capture more detailed interactions among variables. Similarly, there are subtle differences in rankings between GAT and GATv2, with GAT-based methods favoring “work” over “sleep”. It is hypothesized that GAT’s slightly lesser interpretability might be linked to its relatively weaker prediction performance, as indicated in Tables 2–5.

Despite these variations, both models consistently classify variables into categories like most important, very important, somewhat important, and less important. For example, both models consistently identify “glucose_level” as the most important variable across all datasets. This is followed by other closely related and significant variables that have a direct impact on BG levels, such as carbohydrate intake, classified as very important. Subsequently, both models also pay attention to variables that indirectly affect BG levels, like exercise in the OhioT1DM dataset. These are considered somewhat important variables. Finally, both GAT and GATv2 place other less significant variables at the lower end of the ranking.

(7) Based on Property 3, the gaps between the prediction and the calculation of variable importance are affected by $\alpha$. Fig. 4 shows the

**Table 9**
Testing time complexity of methods.

| Method | Testing time complexity | Comments |
|---|---|---|
| LR | $O(INT)$ | $I$: no. of data points; $N$: no. of input variables; $T$: sequence length. |
| XGBoost | $O(rd)$ | $r$: no. of trees; $d$: maximum depth of the trees. |
| RETAIN | $O(TND + THD + TH^2)$ | $D$: input size of RNN; $H$: hidden size. |
| IMV-FULL | $O(THN + TH^2)$ | |
| IMV-TENSOR | $O(TH + TH^2/N)$ | |
| ETN-ODE | $O(TH + TH^2/N)$ + ODE Solver Time | |
| ATT-T-LSTM | $O(NT^2 + THD + TH^2)$ | |
| ATT-F-LSTM | $O(TN^2 + THD + TH^2)$ | |
| N-BEATS | $O(BTH)$ | $B$: no. of blocks. |
| NHiTS | $O(TH(1-b^B)/(1-b))$ | $b$: expressivity ratios of NHiTS for reducing the amount of parameters for each layer. |
| LP | $O(THN + TH^2)$ | |
| LPSC | $O(THN + TH^2)$ | |
| MNODE | $O(THN + TH^2)$ | |
| GAT+GRU | $O(TNAE + TN^2A + THNE + TH^2)$ | $A$: output size of embedding transformation; $E$: embedding size of each variable. |
| GATv2+GRU | $O(TNAE + TN^2A + THNE + TH^2)$ | |

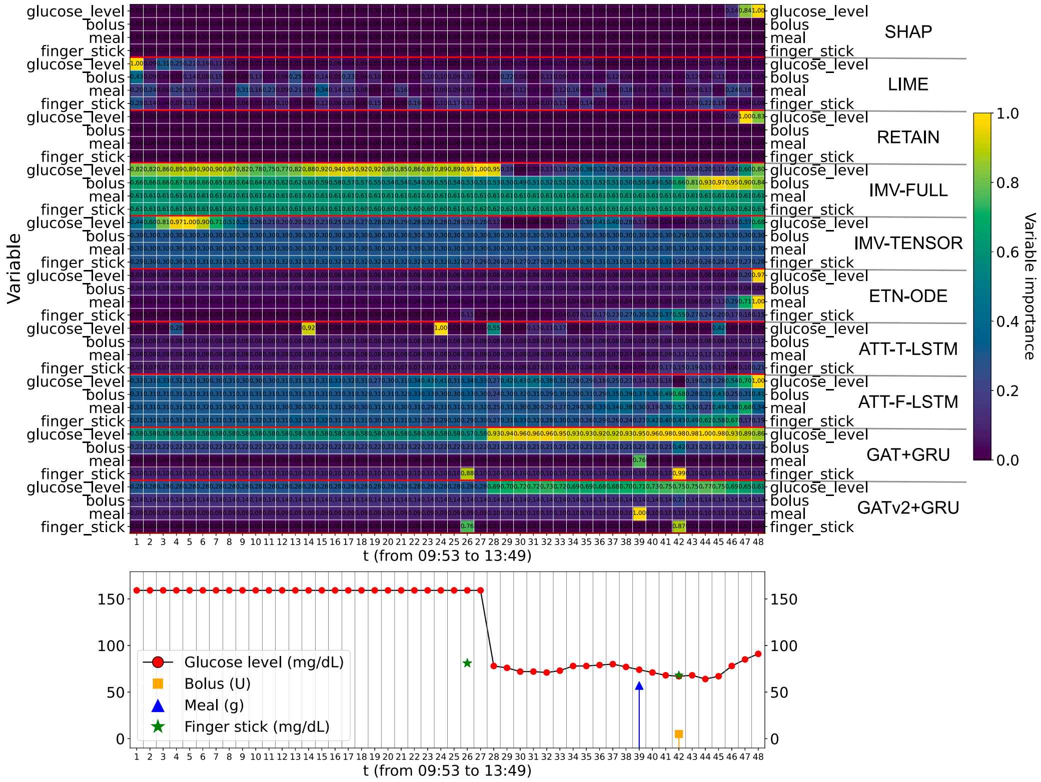


**Fig. 7.** The bottom sub-figure is the visualization of a historical multi-variate time series of the patient 591 in OhioT1DM, only showing “glucose_level”, “bolus”, “meal” and “finger_stick”. The heatmaps on the top are the feature maps of interpretable baseline methods and our proposed methods (“GAT+GRU” and “GATv2+GRU” with two graph layers). The x-axis/y-axis is the timestep $t$ or variable name. The value in the cell is the variable importance $v_t^j$, scaled to $[0,1]$. Explanation of methods can refer to Section 5.2.

variable importance ranking of GARNNs when changing $\alpha$. We find that the gaps cannot significantly alter the variable importance, especially for “GATv2+RNN”. Even with different gap sizes, the changes in variable importance remain small but still acceptable.

### 5.6. Interpretation of feature maps

CGM might be inaccurate after being worn for a period, so calibration by “finger_stick” can make it back to normal. According to the bottom sub-figure of Fig. 7, participant 591 finds the CGM is unreliable, so this patient takes the first “finger_stick” at $t = 26$. Then, this patient calibrates “glucose_level” of CGM with “finger_stick”, and “glucose_level” goes down and back to accurate measurements after $t = 27$. Next, this patient has a meal at $t = 39$; this patient takes “bolus” and another “finger_stick” at $t = 42$.

Hence, “glucose_level” ($t > 27$) should be more important than “glucose_level” ($t \leq 27$). Based on the feature maps of the methods, GARNNs accurately highlight the importance of “glucose_level” ($t > 27$). Furthermore, GARNNs exactly capture the sparse signals, i.e., “bolus”, “meal” and “finger_stick”, considering the related cells are markedly brighter when $t = 26, 39$ and 42. Besides, GARNNs precisely focus on the lowest BG level, when $t = 44$. However, the feature maps of other methods are not quite informative. All of them cannot explicitly show the usage of sparse signals. Meanwhile, they fail to focus “glucose_level” time series reasonably.

More feature maps are shown in Fig. 8. Note that the “meal” in ShanghaiT1DM or ShanghaiT2DM means the amount of food intake rather than carbohydrate intake. The variable importance is scaled to $[0,1]$ as well. The observations still hold that GARNNs ($L = 2$) can correctly focus the important values of “glucose_level”, i.e., $t = 4$ and

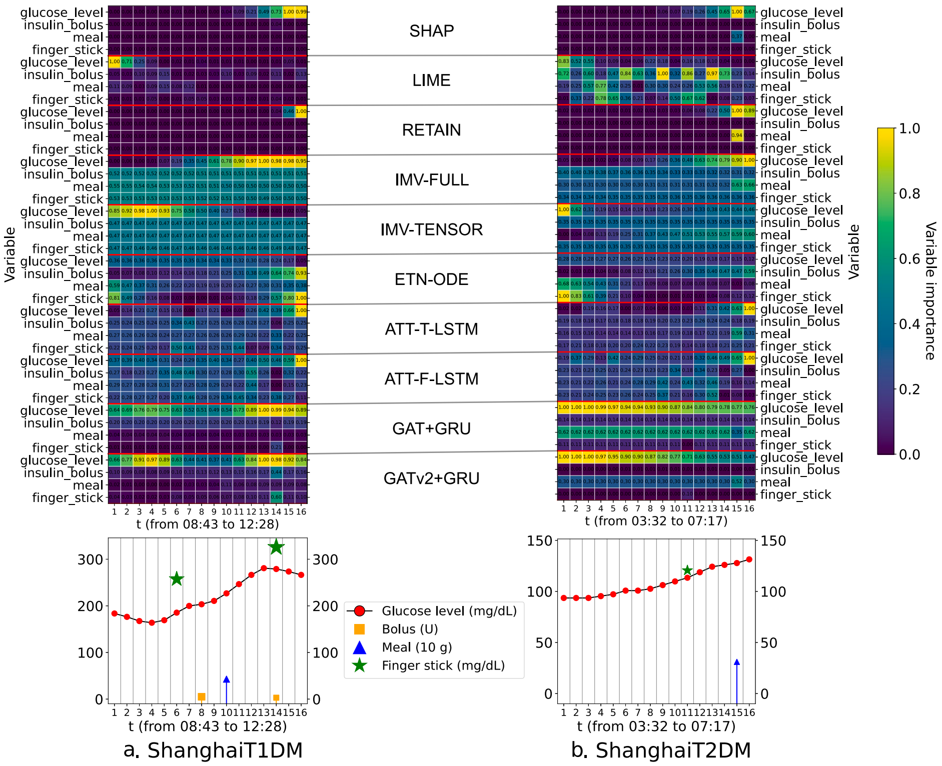


**Fig. 8.** The bottom sub-figure is the visualization of a historical multi-variate time series of the patient 1001/2030 in ShanghaiT1DM/ShanghaiT2DM. The heatmaps on the top are the feature maps of interpretable baseline methods and our proposed methods ("GAT+GRU" and "GATv2+GRU" with two graph layers). The x-axis/y-axis is the timestep $t$ or variable name. The value in the cell is the variable importance $v_t^j$, scaled to $[0, 1]$. Explanation of methods can refer to Section 5.2.

$t = 13$ in Fig. 8a, and $t = 1$ in Fig. 8b. Important sparse signals gain more attention as well, i.e., $t = 14$ in Fig. 8a, and $t = 15$ in Fig. 8b.

## 6. Discussion

### 6.1. Research scope

Our study specifically focuses on blood glucose level prediction (BGLP). This problem definition is stated in our work (see Section 4.1) and aligns with the BGLP 2018 and 2020 challenge, as outlined in Bach, Bunescu, Farri, Guo, Hasan, Ibrahim, Marling, Raffa, Rubin, and Wu (2018) and Bach, Bunescu, Marling, and Wiratunga (2020). In this challenge, deep learning-based methods achieved the best performance, such as those by Bhimireddy, Sinha, Oluwalade, Gichoya, and Purkayastha (2020), Daniels, Herrero, and Georgiou (2020) and Rubin-Falcone, Fox, and Wiens (2020). This challenge focuses on achieving high prediction accuracy rather than the interpretability of BG predictions. Nevertheless, we have included N-BEATS (Oreshkin et al., 2020) and NHiTS (Challu et al., 2023) in our experiments, as the championship model (Rubin-Falcone et al., 2020) was based on N-BEATS, and NHiTS serves as its upgraded version.

The target of personalized BGLP is to enhance the precision and effectiveness of predictive models tailored to individual characteristics by utilizing detailed personal data. This approach aims to create highly accurate and reliable models for daily diabetes management for both T1DM and T2DM patients, using data collected from wearable devices and self-reports. These models are specifically customized to individual needs and variations. This differs from models designed for specific scenarios, such as predicting postprandial glucose for personalized nutrition (Zeevi et al., 2015). Such models sometimes rely on data that cannot be easily collected by patients, making them less suitable for daily diabetes management.

Despite the excellent performance of deep learning methods in BGLP, their main limitation is the lack of interpretability. Our aim is to introduce interpretability to deep learning models while maintaining high prediction accuracy. While our proposed interpretable method is not specifically designed to explain the impact of individual meals on postprandial glucose levels like (Howard, Guo, & Hall, 2020), it offers broader applications for various glucose prediction tasks. Specifically, our method determines the relative importance of each variable (scaled between 0 and 1). This variable importance can help clinicians infer the contribution of a variable in glucose predictions compared to other factors, potentially offering a versatile tool for personalized diabetes management in various scenarios, such as night glucose dynamics and the impact of exercise on glucose levels.

Meanwhile, even though we compared with mechanistic models (LP, LPSC and MNODE), we do not extend this problem to artificial pancreas or artificial beta cell projects, such as those by Albers et al. (2017), Ha and Sherman (2020) and Sirlanci, Levine, Low Wang, Albers, and Stuart (2023), or simulator models like (Miller et al., 2020; Wang et al., 2023), but concentrate purely on BGLP. Mechanistic models differ from the interpretable methods in this paper, which demonstrate the contribution of each variable to the prediction by providing variable importance. Traditional equation-based models are interpretable by adhering to established scientific laws and theories, emphasizing the glucose-insulin response, but they often have limited performance due to higher modeling errors. These errors arise because mathematical models primarily adjust parameters, resulting in lower accuracy and flexibility compared to deep learning models. Traditional models often struggle with the high variability in BG levels across different individuals, influenced by factors such as diet, exercise, stress, and other health conditions. Simply adjusting parameters in traditional models is insufficient to address significant individual BG variances.

In contrast, deep learning models achieve higher accuracy and personalization by incorporating various variables flexibly. They can integrate a wide range of data inputs, such as CGM readings, insulin dosages, meal information, physical activity, and even sleep patterns. This holistic approach allows deep learning models to quickly and effectively capture the nuances of individual BG variations, providing more personalized predictions of glucose levels.

### 6.2. Multimodal data for glucose prediction

Multimodal data refers to data collected from multiple sources or modalities, each capturing different aspects of the same phenomenon (Mitri, Schneider, Specht, & Drachsler, 2018). In the context of BGLP, this involves integrating self-reported data (e.g., insulin and meal intake) with sensor-based data (e.g., Continuous Glucose Monitoring or CGM). The inclusion of additional sensor data from wearable devices (e.g., heart rate) and more self-reported information (e.g., exercise and sleep) further exemplifies the multimodal nature of this data.

The advanced state of BG prediction using multimodal data is demonstrated by the BGLP 2020 challenge championship (Rubin-Falcone et al., 2020). Recent studies highlight the potential of AI-enabled wearable devices for noninvasive BG monitoring and forecasting, indicating promising avenues for future development and widespread adoption (Ahmed et al., 2023; Ahmed, Aziz, Qidwai, Abd-Alrazaq & Sheikh, 2023). Research has also shown the effectiveness of leveraging multimodal data, including physiological data from mobile devices, for managing both T1DM (Kuang et al., 2021; Zhu, Kuang et al., 2023) and T2DM (Pai et al., 2024; Tsai, Li, Lam, Li, & Ho, 2019).

Recognizing the significant potential of multimodal data in BGLP, our methods provide explanations for each prediction by assessing variable importance, explaining the contribution of each variable to the prediction. Additionally, variable importance can be averaged over the training data to rank variables, indicating the most influential ones in BGLP. Our approaches bridge the gap between leveraging multimodal data and providing explanations in deep learning models.

### 6.3. Deep learning for clinical usage

Deep learning has been leveraged for medical applications. Esteva et al. (2021) emphasizes the potential of AI techniques, particularly deep learning, to extract valuable insights from medical data. Their work summarizes the progress in convolutional neural networks and their applications in medical imaging, medical video analysis, and clinical deployment. Pati et al. (2023) presents GaNDLF, an open-source deep learning framework designed to facilitate scalable end-to-end clinical workflows. GaNDLF aims to lower the barriers to developing, training, and deploying deep learning algorithms in the clinical and scientific communities. It focuses on enhancing the stability, reproducibility, interpretability, and scalability of these processes. Lee et al. (2024) discusses the integration of a deep learning algorithm with point-of-care testing platforms, significantly reducing diagnostic assay times while maintaining high accuracy.

In the context of deep learning-based BGLP for clinical use, Kim et al. (2020) utilized RNNs to predict glucose levels in hospital patients. The study, led by clinician Dae-Yeon Kim, aimed to aid medical personnel in monitoring and controlling BG levels in hospitalized patients with T2DM. Data was collected using a CGM device over a week from 20 patients, and three types of RNNs (simple RNN, GRU, and LSTM) were tested. The GRU model outperformed the other variants, supporting our decision to incorporate GRU into our proposed method.

Meanwhile, Zale and Mathioudakis (2022) highlights the growing use of machine learning approaches to predict glucose trends in hospitalized patients. Notably, most of the approaches reviewed are interpretable machine learning algorithms such as XGBoost and Logistic Regression, which predict discrete glucose levels. The preference for these interpretable methods likely stems from their transparency and ease of understanding, which are crucial in clinical settings. Apart from Kim et al. (2020), there are fewer deep learning-based models successfully leveraged in actual clinical applications for diabetes care. This might be attributed to the opacity of most deep learning methods. In contrast, interpretable machine learning algorithms such as XGBoost and Logistic Regression are popular for clinical usage. This also confirms that our proposed interpretable methods could enable more deep learning methods to be actually leveraged for real clinical applications. Our experiments have shown that our proposed methods can provide more accurate predictions and significant explanations for glucose prediction compared to traditional machine learning methods such as XGBoost and Linear Regression.

### 6.4. Variable importance of BGLP models

In this paper, we present two types of variable importance in this paper: instant variable importance and summarized variable importance.

- **Instant Variable Importance** ($v_t^j$): This can be leveraged for generating feature maps for each prediction, which provide variable importance at each historical time point $t$ (see Figs. 1, 7, 8). Given that instant variable importance is provided for each variable $j$ at each time point $t$ and scaled between [0, 1], it is also significant to compare the variable importance for a specific prediction by averaging the instant variable importance over time.
- **Summarized Variable Importance** ($v^j(\mathcal{I})$): The summarized variable importance is generated by aggregating all instant variable importance values of the variable $j$ from the training set $\mathcal{I}$. It can be leveraged for ranking variables effectively. (see Figs. 3–4)

Although existing research has explored variable importance, its application to glucose prediction often falls short. Our proposed approach, which includes both instant and summarized variable importance, offers significant benefits.

The benefits of instant variable importance are numerous. For diet optimization, analyzing feature maps for each prediction can reveal the variable importance of each meal, aiding in optimizing dietary habits. For insulin dosing, understanding the importance of each insulin dose through feature maps can help in fine-tuning dosing schedules and types, thereby improving glycemic control. In terms of exercise planning, recognizing the impact of exercise intensity through feature maps can guide patients in scheduling physical activities to minimize glucose fluctuations. Feature maps are also especially useful when input data is inaccurate or padded with imputed values, as they help focus on valid readings, making predictions more reliable. Lastly, effective self-management is enhanced by advising patients to monitor the most impactful variables closely. For example, patients with T2DM who cannot leverage insulin effectively might find non-insulin medications and meal intake more crucial. Feature maps can help these patients choose suitable medicines and diets to stabilize their glucose levels.

Summarized variable importance provides a ranking list of variables, which is especially helpful when dealing with extensive physiological data, such as heart rate. By focusing on the most influential variables, healthcare providers can develop highly personalized treatment plans, such as offering tailored nutritional advice to patients whose glucose levels are highly sensitive to diet. Additionally, understanding variable importance aids in feature selection, potentially reducing model complexity by eliminating less important variables without sacrificing accuracy. In resource-constrained settings, such as mobile applications, this focus ensures that efforts are concentrated on collecting and ensuring the accuracy of the most critical variables, thus optimizing resource use.

It is challenging to quantitatively evaluate variable importance because explanations vary with different inputs, making it impractical to invite experts to label each prediction. Therefore, we evaluated variable importance from two perspectives. Firstly, for instant variable importance (feature maps) in Figs. 1, 7 and 8, we compared the quality of interpretability based on data validity, signal importance, and sparse signal detection. For example, in Fig. 7, our methods excluded invalid CGM readings, highlighted severe hypoglycemia points, and captured sparse signals like bolus and meal data, outperforming baseline methods. Secondly, for summarized variable importance (Figs. 3–4), we

**Table 10**
Root Mean Square Error (RMSE) of methods on four datasets.

| Method | OhioT1DM | AristesT1DM | ShanghaiT1DM | ShanghaiT2DM |
|---|---|---|---|---|
| LSTM | 21.20 ± 0.13(2.84)‡ | 23.46 ± 0.25(5.00)‡ | 15.78 ± 1.10(3.97)‡ | 11.86 ± 0.04(3.21)‡ |
| GRU | 19.88 ± 0.04(2.63)‡ | 22.16 ± 0.02(4.43)‡ | 14.42 ± 0.08(2.87)‡ | 12.00 ± 0.14(3.57)‡ |
| GaAN | 19.89 ± 0.11(2.93)‡ | 21.17 ± 0.09(4.65)‡ | 14.42 ± 0.22(3.19)‡ | 15.95 ± 2.13(25.18)‡ |
| GATv2 | 19.72 ± 0.03(2.88)‡ | 21.18 ± 0.10(4.66)‡ | 14.28 ± 0.29(2.79)‡ | 11.79 ± 0.03(3.10)‡ |
| GaAN+LSTM | 19.84 ± 0.22(2.71)‡ | 20.27 ± 0.13(3.98)‡ | 14.24 ± 0.82(3.35) | 11.95 ± 0.16(3.34)‡ |
| GaAN+GRU | 19.50 ± 0.21(2.79)‡ | 20.15 ± 0.15(3.87)‡ | 13.63 ± 0.40(2.73) | 11.96 ± 0.11(3.28)‡ |
| GATv2+LSTM | 19.25 ± 0.13(2.49)‡ | 20.16 ± 0.09(4.10)‡ | 14.39 ± 0.47(3.41)‡ | 11.81 ± 0.03(3.10) |
| **GATv2+GRU** | **18.97 ± 0.06(2.43)** | **19.97 ± 0.07(3.93)** | **13.62 ± 0.22(2.78)** | **11.72 ± 0.02(3.03)** |

$*\ p \leq 0.05$; $\dagger p \leq 0.01$; $\ddagger p \leq 0.005$;
The result is formatted as "$mean \pm sd_1(sd_2)$";
$sd_1$ is the standard deviation after running the experiments four times by changing random seed;
$sd_2$ is the standard deviation of the metric results across the participants.
Explanation of methods can refer to Sections 5.2 and 6.6.

used medical knowledge and statistical analysis. For instance, our methods in ShanghaiT2DM reduce the importance of bolus insulin and increase the importance of meals and non-insulin hypoglycemic agents compared to ShanghaiT1DM. In OhioT1DM, our methods prioritize self-reported sleep over sensor-detected sleep due to significant missing data. The OhioT1DM dataset, widely used in research, supports our variable rankings. Butt, Khosa, and Iftikhar (2023) summarized input features from 624 studies, confirming that CGM data is the most important, followed by insulin and meal, with exercise being less important, and other features being even less so. It aligns with our findings, further validating our methods.

### 6.5. Graph attention mechanisms for BGLP

Graph attention focuses on learning attention coefficients for each node's neighbors, allowing it to selectively emphasize important connections and ignore less relevant ones. This approach is especially useful in graph-based data where the relevance of connections can vary significantly. Existing attention mechanisms, such as self-attention (Vaswani et al., 2017) and temporal attention (Luong, Pham, & Manning, 2015), have also been applied to blood glucose level prediction (BGLP) (Bevan & Coenen, 2020; Zhu et al., 2024).

Temporal attention is specifically designed for sequence data, focusing on temporal relationships, and may not fully exploit the underlying graph structure when used in a graph context. Self-attention, while also designed for sequence data, can model the correlations of nodes in a completely connected graph, where each node connects to all other nodes. Both self-attention and graph attention can be utilized in such scenarios.

We choose graph attention over self-attention because graph attention does not require generating key, query, and value vectors from node embeddings to calculate node weights, as self-attention does. Instead, node embeddings are directly leveraged with learnable parameters to generate edge weights. This simplified node weight generation can potentially highlight variable importance by removing irrelevant information, as demonstrated in this paper.

Regarding the specialties of graph attention, a multi-layer graph attention mechanism can extract more complex features. However, based on the prediction results in Tables 2–7, increasing the number of graph attention layers does not further improve prediction performance. In fact, GATv2+GRU (L = 1) generally performs the best. As shown in Section 5.5(6), there is no significant difference in the quality of variable importance between using one and two graph attention layers. Therefore, we believe that more detailed features do not enhance model performance in terms of prediction capability and variable importance quality. General features are sufficient in this scenario.

Regarding time consumption, a comparison between GARNNs and existing methods shows that one layer of GAT/GATv2 consumes less than 7 s when tested on all examples across four datasets, similar to most interpretable methods. However, two layers of GAT/GATv2 double the testing time consumption, affecting the model's efficiency. Fortunately, for the BGLP problem, multiple layers of GAT/GATv2 are unnecessary. One layer of graph attention is sufficient, with time consumption comparable to other interpretable methods.

### 6.6. Ablation study and module selection

The main objective of this work is to interpret predictions on BGLP using MTS models. Graph-based structures can explicitly model correlations among different nodes (variables). This can provide possibility to interpret correlations among variables. Therefore, we aimed to combine graph neural networks (GNNs) with RNNs, with the goal of making the model interpretable for MTS.

Since MTS data is inherently temporal, we believed it was important that the graph structure should also vary over time. For this reason, we did not focus on static graph models like Graph Convolutional Networks (GCNs, Kipf and Welling (2017)), where the edge weights need to be predefined by domain knowledge and remain fixed throughout training. In contrast, the edge weights in models like GAT/GATv2 are learnable, allowing the graph weights to vary over time. This time-varying property is crucial, as it enables us to leverage these varying weights to extract the variable importance of each node, as discussed in Section 4.3.

There are other attention-based GNNs, such as Gated Attention Networks (GaAN, Zhang et al. (2018)), which do not require predefined static graphs. However, we did not consider these models because they are less suitable for extracting variable importance. For example, GaAN relies on dot-product attention, where a query from one node (variable $n$) performs a dot product with the key from another node $j$, generating attention scores. Extracting variable importance based on these attention scores is challenging. In contrast, GAT/GATv2 introduces a learnable vector $\mathbf{a}$, which is used to perform a dot product with the concatenation of a query and a key. This mechanism provides a potential pathway for summarizing and extracting learnable variable importance, as explained in Section 4.3.

Interpretability is our primary concern, which is why we selected GAT/GATv2 for this work. After addressing interpretability, we then explored how to improve prediction performance. As shown in Table 10, the combination of GATv2 and GRU, i.e., GATv2+GRU, achieves the lowest RMSE on BGLP. Each component alone performs worse, highlighting the success of this combination.

When we replaced GRU with other RNNs (e.g., GATv2+LSTM) or replaced GATv2 with other GNNs (e.g., GaAN+GRU), these combinations did not outperform GATv2+GRU. Other combinations, such as GaAN+LSTM, also showed similar trends, reinforcing that GATv2+GRU achieved the best prediction performance on BGLP in our experiments.

However, the methods we propose for extracting variable importance from GAT/GATv2 are not limited to this specific combination with GRU. The RNN component in our framework can be replaced with other types of RNNs, and variable importance can still be calculated based on GAT/GATv2. Exploring other RNNs or time series models could be an avenue for future work, though this is outside the scope of this paper, where interpretability remains the central focus.

## 7. Conclusion and limitations

In this work, we propose GARNNs, novel interpretable models, by incorporating graph attention networks and RNNs for BGLP. One notable advantage of GARNNs is their ability to provide significant variable importance for ranking variables and generating feature maps. In experiments across four datasets, GARNNs demonstrated superior performance compared to fifteen baseline models. GARNNs outperform the baselines in both the accuracy of predictions and the quality of explanations regarding the contribution of variables.

Then, GARNN can potentially be applied to both T1DM and T2DM. As discussed in Section 5.5 (4), GARNNs can learn different patterns for these two groups. However, we need to further validate this observation with additional T2DM datasets in the future.

Once well-trained, GARNNs are capable of performing real-time predictions. However, a limitation of this work is the absence of online training capabilities in our algorithms. We plan to address this in future work.

Another limitation of this work is its narrow focus, primarily concentrating on specific aspects of BG management without extensively exploring other relevant scenarios, such as the development of interpretable models for insulin advice. Consequently, future work will aim to expand the applicability of the proposed interpretable models across a broader range of BG management scenarios.

## CRediT authorship contribution statement

**Chengzhe Piao:** Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Taiyu Zhu:** Writing – review & editing, Validation, Formal analysis, Data curation. **Stephanie E. Baldeweg:** Writing – review & editing, Conceptualization. **Paul Taylor:** Writing – review & editing, Resources. **Pantelis Georgiou:** Writing – review & editing, Resources, Data curation. **Jiahao Sun:** Writing – review & editing, Conceptualization. **Jun Wang:** Writing – review & editing, Supervision, Conceptualization. **Kezhi Li:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Chengzhe Piao reports financial support was provided by UK Research and Innovation. Kezhi Li reports financial support was provided by Rosetrees Trust. Kezhi Li reports financial support was provided by Great Ormond Street Hospital Children's Charity. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

We thank the support from UKRI Center for Doctoral Training in AI-enabled healthcare systems [EP/S021612/1] and University College London Overseas Research Scholarships, United Kingdom. We also appreciate the support from Rosetrees Trust, United Kingdom (Grant number: UCL-IHE-2020\102) and Great Ormond Street Hospital, United Kingdom (Charity ref. X12018). The study sponsors had no study involvement.

## Data availability

The OhioT1DM, ShanghaiT1DM and ShanghaiT2DM datasets are publicly accessible online via their respective websites. However, the ArisesT1DM dataset is confidential and needs authorization for access.

## References

Aguiar, H., Santos, M. D., Watkinson, P. J., & Zhu, T. (2022). Learning of cluster-based feature importance for electronic health record time-series. In *ICML'22* (pp. 161–179).

Ahmed, A., Aziz, S., Abd-Alrazaq, A., Farooq, F., Househ, M., & Sheikh, J. (2023). The effectiveness of wearable devices using artificial intelligence for blood glucose level forecasting or prediction: Systematic review. *Journal of Medical Internet Research*, *25*, Article e40259.

Ahmed, A., Aziz, S., Qidwai, U., Abd-Alrazaq, A., & Sheikh, J. (2023). Performance of artificial intelligence models in estimating blood glucose level among diabetic patients using non-invasive wearable device data. *Computer Methods and Programs in Biomedicine Update*, *3*, Article 100094.

Albers, D. J., Levine, M., Gluckman, B., Ginsberg, H., Hripcsak, G., & Mamykina, L. (2017). Personalized glucose forecasting for type 2 diabetes using data assimilation. *PLoS Computational Biology*, *13*(4), Article e1005232.

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2018). Towards better understanding of gradient-based attribution methods for deep neural networks. In *ICLR'18*.

Bach, K., Bunescu, R. C., Farri, O., Guo, A., Hasan, S. A., Ibrahim, Z. M., Marling, C., Raffa, J., Rubin, J., & Wu, H. (Eds.), (2018). *CEUR workshop proceedings*: *vol. 2148*, *Proceedings of the 3rd international workshop on knowledge discovery in healthcare data co-located with the 27th international joint conference on artificial intelligence and the 23rd European conference on artificial intelligence (IJCAI-ECAI 2018), Stockholm, Schweden, July 13, 2018*. CEUR-WS.org, URL: https://ceur-ws.org/Vol-2148.

Bach, K., Bunescu, R. C., Marling, C., & Wiratunga, N. (Eds.), (2020). *CEUR workshop proceedings*: *vol. 2675*, *Proceedings of the 5th international workshop on knowledge discovery in healthcare data co-located with 24th European conference on artificial intelligence, KDH@ECAI 2020, Santiago de Compostela, Spain & Virtually, August 29-30, 2020*. CEUR-WS.org, URL: https://ceur-ws.org/Vol-2675.

Balasooriya, K., & Nanayakkara, N. D. (2020). Predicting short-term changing blood glucose level of diabetes patients using noninvasive data. In *IEEE TENCON'20* (pp. 31–36). http://dx.doi.org/10.1109/TENCON50793.2020.9293823.

Bevan, R., & Coenen, F. (2020). Experiments in non-personalized future blood glucose level prediction. In *KDH@ECAI'20* (pp. 100–104).

Bezerra, M. F., Neves, C., Neves, J. S., & Carvalho, D. (2023). Time in range and complications of diabetes: a cross-sectional analysis of patients with Type 1 diabetes. *Diabetology & Metabolic Syndrome*, *15*(1), 244. http://dx.doi.org/10.1186/s13098-023-01219-2.

Bhimireddy, A. R., Sinha, P., Oluwalade, B., Gichoya, J. W., & Purkayastha, S. (2020). Blood glucose level prediction as time-series modeling using sequence-to-sequence neural networks. In *CEUR workshop proceedings*: *vol. 2675*, *KDH@ECAI'20* (pp. 125–130).

Bloomgarden, Z. T. (2004). Diabetes complications . *Diabetes Care*, *27*(6), 1506–1514. http://dx.doi.org/10.2337/diacare.27.6.1506.

Brody, S., Alon, U., & Yahav, E. (2022). How attentive are graph attention networks? In *ICLR'22*.

Butt, H., Khosa, I., & Iftikhar, M. A. (2023). Feature transformation for efficient blood glucose prediction in type 1 diabetes mellitus patients. *Diagnostics*, *13*(3), 340.

Bykov, K., Hedström, A., Nakajima, S., & Höhne, M. M. (2022). NoiseGrad - enhancing explanations by introducing stochasticity to model weights. In *AAAI'22* (pp. 6132–6140). http://dx.doi.org/10.1609/AAAI.V36I6.20561.

Cappon, G., Meneghetti, L., Prendin, F., Pavan, J., Sparacino, G., Favero, S. D., et al. (2020). A personalized and interpretable deep learning based approach to predict blood glucose concentration in type 1 diabetes. In *KDH@ECAI'20* (pp. 75–79).

Challu, C., Olivares, K. G., Oreshkin, B. N., Ramírez, F. G., Canseco, M. M., & Dubrawski, A. (2023). NHITS: neural hierarchical interpolation for time series forecasting. In *AAAI'23* (pp. 6989–6997). http://dx.doi.org/10.1145/2939672.2939778.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *ACM KDD'16* (pp. 785–794). http://dx.doi.org/10.1145/2939672.2939785.

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In *SSST@EMNLP'14* (pp. 103–111). http://dx.doi.org/10.3115/V1/W14-4012.

Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., & Stewart, W. F. (2016). RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. In *NIPS'16* (pp. 3504–3512).

Chu, Y., Wang, X., Ma, J., Jia, K., Zhou, J., & Yang, H. (2020). Inductive granger causal modeling for multivariate time series. In *ICDM'20* (pp. 972–977).

Cichosz, S. L., Kronborg, T., Jensen, M. H., & Hejlesen, O. K. (2021). Penalty weighted glucose prediction models could lead to better clinically usage. *Computers in Biology and Medicine*, *138*, Article 104865. http://dx.doi.org/10.1016/J.COMPBIOMED.2021.104865.

Clarke, W. L., Cox, D., Gonder-Frederick, L. A., Carter, W., & Pohl, S. L. (1987). Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care*, *10*(5), 622–628.

Contador, S., Colmenar, J. M., Garnica, O., Velasco, J. M., & Hidalgo, J. I. (2022). Blood glucose prediction using multi-objective grammatical evolution: analysis of the "agnostic" and "what-if" scenarios. *Genetic Programming and Evolvable Machines*, 1–32.

Daniels, J., Herrero, P., & Georgiou, P. (2020). Personalised glucose prediction via deep multitask networks. In *CEUR workshop proceedings*: *vol. 2675*, *KDH@ECAI'20* (pp. 110–114).

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., et al. (2021). Deep learning-enabled medical computer vision. *Npj Digital Medicine*, *4*(1), 5.

Favero, S. D., Facchinetti, A., & Cobelli, C. (2012). A glucose-specific metric to assess predictors and identify models. *IEEE Transactions on Biomedical Engineering*, *59*(5), 1281–1290. http://dx.doi.org/10.1109/TBME.2012.2185234.

Gandin, I., Scagnetto, A., Romani, S., & Barbati, G. (2021). Interpretability of time-series deep learning models: A study in cardiovascular patients admitted to intensive care unit. *Journal of Biomedical Informatics*, *121*, Article 103876. http://dx.doi.org/10.1016/j.jbi.2021.103876.

Gao, P., Yang, X., Zhang, R., Huang, K., & Goulermas, J. Y. (2023). Explainable tensorized neural ordinary differential equations for arbitrary-step time series prediction. *IEEE Transactions on Knowledge and Data Engineering*, *35*(6), 5837–5850. http://dx.doi.org/10.1109/TKDE.2022.3167536.

Guo, T., Lin, T., & Antulov-Fantulin, N. (2019). Exploring interpretable LSTM neural networks over multi-variable data. In *ICML'19* (pp. 2494–2504).

Ha, J., & Sherman, A. (2020). Type 2 diabetes: one disease, many pathways. *American Journal of Physiology-Endocrinology and Metabolism*, *319*(2), E410–E426.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. http://dx.doi.org/10.1162/NECO.1997.9.8.1735.

Howard, R., Guo, J., & Hall, K. D. (2020). Imprecision nutrition? Different simultaneous continuous glucose monitors provide discordant meal rankings for incremental postprandial glucose in subjects without diabetes. *The American Journal of Clinical Nutrition*, *112*(4), 1114–1119.

Hsieh, T., Wang, S., Sun, Y., & Honavar, V. G. (2021). Explainable multivariate time series classification: A deep neural network which learns to attend to important variables as well as time intervals. In *ACM WSDM'21* (pp. 607–615). http://dx.doi.org/10.1145/3437963.3441815.

Kaji, D. A., Zech, J. R., Kim, J. S., Cho, S. K., Dangayach, N. S., Costa, A. B., et al. (2019). An attention based deep learning model of clinical events in the intensive care unit. *PLoS One*, *14*(2), Article e0211057. http://dx.doi.org/10.1371/journal.pone.0211057.

Karim, R. A. H., Vassányi, I., & Kósa, I. (2020). After-meal blood glucose level prediction using an absorption model for neural network training. *Computers in Biology and Medicine*, *125*, Article 103956. http://dx.doi.org/10.1016/J.COMPBIOMED.2020.103956.

Kim, D.-Y., Choi, D.-S., Kim, J., Chun, S. W., Gil, H.-W., Cho, N.-J., et al. (2020). Developing an individual glucose prediction model using recurrent neural network. *Sensors*, *20*(22), 6460.

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *ICLR'17*.

Kuang, L., Zhu, T., Li, K., Daniels, J., Herrero, P., & Georgiou, P. (2021). Live demonstration: an IoT wearable device for real-time blood glucose prediction with edge AI. In *IEEE BioCAS'21* (pp. 01–01).

Kwon, Y., & Zou, J. Y. (2022). WeightedSHAP: Analyzing and improving Shapley based feature attributions. In *NIPS'22* (pp. 34363–34376).

Lee, S., Park, J. S., Woo, H., Yoo, Y. K., Lee, D., Chung, S., et al. (2024). Rapid deep learning-assisted predictive diagnostics for point-of-care testing. *Nature Communications*, *15*(1), 1695.

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. In *NIPS'17* (pp. 4765–4774).

Luong, T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *EMNLP'15* (pp. 1412–1421). http://dx.doi.org/10.18653/V1/D15-1166.

Man, C. D., Micheletto, F., Lv, D., Breton, M., Kovatchev, B., & Cobelli, C. (2014). The UVA/PADOVA type 1 diabetes simulator: new features. *Journal of Diabetes Science and Technology*, *8*(1), 26–34.

Marling, C., & Bunescu, R. C. (2020). The OhioT1DM dataset for blood glucose level prediction: Update 2020. In *KDH@ECAI'20* (pp. 71–74).

Miller, A. C., Foti, N. J., & Fox, E. (2020). Learning insulin-glucose dynamics in the wild. In *Machine learning for healthcare conference* (pp. 172–197). PMLR.

Mitri, D. D., Schneider, J., Specht, M., & Drachsler, H. (2018). From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer-Assisted Learning*, *34*(4), 338–349. http://dx.doi.org/10.1111/JCAL.12288.

Mora, T., Roche, D., & Rodríguez-Sánchez, B. (2023). Predicting the onset of diabetes-related complications after a diabetes diagnosis with machine learning algorithms. *Diabetes Research and Clinical Practice*, *204*, Article 110910. http://dx.doi.org/10.1016/j.diabres.2023.110910.

Naumova, V., Pereverzyev, S. V., & Sivananthan, S. (2012). A meta-learning approach to the regularized learning - Case study: Blood glucose prediction. *Neural Networks*, *33*, 181–193. http://dx.doi.org/10.1016/J.NEUNET.2012.05.004.

Nemat, H., Khadem, H., Elliott, J., & Benaissa, M. (2023). Causality analysis in type 1 diabetes mellitus with application to blood glucose level prediction. *Computers in Biology and Medicine*, *153*, Article 106535. http://dx.doi.org/10.1016/J.COMPBIOMED.2022.106535.

Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2020). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *ICLR'20*.

Pai, A., Santiago, R., Glantz, N., Bevier, W., Barua, S., Sabharwal, A., et al. (2024). Multimodal digital phenotyping of diet, physical activity, and glycemia in Hispanic/Latino adults with or at risk of type 2 diabetes. *Npj Digital Medicine*, *7*(1), 7.

Parkes, J. L., Slatin, S. L., Pardo, S., & Ginsberg, B. H. (2000). A new consensus error grid to evaluate the clinical significance of inaccuracies in the measurement of blood glucose. *Diabetes Care*, *23*(8), 1143–1148.

Pati, S., Thakur, S. P., Hamamcı, İ. E., Baid, U., Baheti, B., Bhalerao, M., et al. (2023). GaNDLF: the generally nuanced deep learning framework for scalable end-to-end clinical workflows. *Communications Engineering*, *2*(1), 23.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://dx.doi.org/10.5555/1953048.2078195.

Piao, C., & Li, K. (2023). Blood glucose level prediction: A graph-based explainable method with federated learning. http://dx.doi.org/10.48550/ARXIV.2312.12541, CoRR abs/2312.12541. arXiv:2312.12541.

Plis, K., Bunescu, R. C., Marling, C., Shubrook, J., & Schwartz, F. (2014). A machine learning approach to predicting blood glucose levels for diabetes management. In *AAAI workshop'14*.

Prendin, F., Pavan, J., Cappon, G., Del Favero, S., Sparacino, G., & Facchinetti, A. (2023). The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using SHAP. *Scientific Reports*, *13*(1), 16865. http://dx.doi.org/10.1038/s41598-023-44155-x.

Rajapaksha, D., & Bergmeir, C. (2022). LIMREF: local interpretable model agnostic rule-based explanations for forecasting, with an application to electricity smart meter data. In *AAAI'22* (pp. 12098–12107). http://dx.doi.org/10.1609/AAAI.V36I11.21469.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *ACM KDD'16* (pp. 1135–1144). http://dx.doi.org/10.1145/2939672.2939778.

Rubin-Falcone, H., Fox, I., & Wiens, J. (2020). Deep residual time-series forecasting: Application to blood glucose prediction. In *CEUR workshop proceedings*: *vol. 2675*, *KDH@ECAI'20* (pp. 105–109).

Shamout, F. E., Zhu, T., Sharma, P., Watkinson, P. J., & Clifton, D. A. (2020). Deep interpretable early warning system for the detection of clinical deterioration. *IEEE Journal of Biomedical and Health Informatics*, *24*(2), 437–446. http://dx.doi.org/10.1109/JBHI.2019.2937803.

Shen, L., Wei, Y., & Wang, Y. (2023). GBT: Two-stage transformer framework for non-stationary time series forecasting. *Neural Networks*, *165*, 953–970. http://dx.doi.org/10.1016/J.NEUNET.2023.06.044.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *ICML'17* (pp. 3145–3153).

Sirlanci, M., Levine, M. E., Low Wang, C. C., Albers, D. J., & Stuart, A. M. (2023). A simple modeling framework for prediction in the human glucose–insulin system. *Chaos. An Interdisciplinary Journal of Nonlinear Science*, *33*(7).

Succetti, F., Rosato, A., & Panella, M. (2023). An adaptive embedding procedure for time series forecasting with deep neural networks. *Neural Networks*, *167*, 715–729. http://dx.doi.org/10.1016/J.NEUNET.2023.08.051.

Tena, F., Garnica, O., Lanchares, J., & Hidalgo, J. I. (2021). Ensemble models of cutting-edge deep neural networks for blood glucose prediction in patients with diabetes. *Sensors*, *21*(21), 7090.

Tonekaboni, S., Joshi, S., Campbell, K., Duvenaud, D. K., & Goldenberg, A. (2020). What went wrong and when? Instance-wise feature importance for time-series black-box models. In *NIPS'20* (pp. 799–809).

Tsai, C.-W., Li, C.-H., Lam, R. W.-K., Li, C.-K., & Ho, S. (2019). Diabetes care in motion: Blood glucose estimation using wearable devices. *IEEE Consumer Electronics Magazine*, *9*(1), 30–34.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *NIPS'17* (pp. 5998–6008).

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In *ICLR'18*.

Wang, K. A., Levine, M. E., Shi, J., & Fox, E. B. (2023). Learning absorption rates in glucose-insulin dynamics from meal covariates. http://dx.doi.org/10.48550/ARXIV.2304.14300, CoRR abs/2304.14300. arXiv:2304.14300.

WHO (2023). Diabetes. https://www.who.int/health-topics/diabetes. (Accessed 19 December 2023).

Woldaregay, A. Z., Årsand, E., Walderhaug, S., Albers, D. J., Mamykina, L., Botsis, T., et al. (2019). Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes. *Artificial Intelligence in Medicine*, *98*, 109–134. http://dx.doi.org/10.1016/J.ARTMED.2019.07.007.

Xu, Y., Biswal, S., Deshpande, S. R., Maher, K. O., & Sun, J. (2018). RAIM: recurrent attentive and intensive model of multimodal patient monitoring data. In *ACM KDD'18* (pp. 2565–2573). http://dx.doi.org/10.1145/3219819.3220051.

Zale, A., & Mathioudakis, N. (2022). Machine learning models for inpatient glucose prediction. *Current Diabetes Reports*, *22*(8), 353–364.

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., et al. (2015). Personalized nutrition by prediction of glycemic responses. *Cell*, *163*(5), 1079–1094.

Zhang, J., Shi, X., Xie, J., Ma, H., King, I., & Yeung, D. (2018). GaAN: Gated attention networks for learning on large and spatiotemporal graphs. In *UAI'18* (pp. 339–349).

Zhao, Q., Zhu, J., Shen, X., Lin, C., Zhang, Y., Liang, Y., et al. (2023). Chinese diabetes datasets for data-driven machine learning. *Scientific Data*, *10*(1), 35. http://dx.doi.org/10.1038/s41597-023-01940-7.

Zhu, T., Kuang, L., Daniels, J., Herrero, P., Li, K., & Georgiou, P. (2023). IoMT-enabled real-time blood glucose prediction with deep learning and edge computing. *IEEE Internet of Things Journal*, *10*(5), 3706–3719. http://dx.doi.org/10.1109/JIOT.2022.3143375.

Zhu, T., Kuang, L., Piao, C., Zeng, J., Li, K., & Georgiou, P. (2024). Population-specific glucose prediction in diabetes care with transformer-based deep learning on the edge. *IEEE Transactions on Biomedical Circuits and Systems*, *18*(2), 236–246. http://dx.doi.org/10.1109/TBCAS.2023.3348844.

Zhu, T., Li, K., Herrero, P., & Georgiou, P. (2023). Personalized blood glucose prediction for type 1 diabetes using evidential deep learning and meta-learning. *IEEE Transactions on Biomedical Engineering*, *70*(1), 193–204. http://dx.doi.org/10.1109/TBME.2022.3187703.

Zhu, T., Uduku, C., Li, K., Herrero, P., Oliver, N., & Georgiou, P. (2022). Enhancing self-management in type 1 diabetes with wearables and deep learning. *Npj Digital Medicine*, *5*(1), 78. http://dx.doi.org/10.1038/S41746-022-00626-5.

Zou, B. J., Levine, M. E., Zaharieva, D. P., Johari, R., & Fox, E. B. (2024). Hybrid square neural ODE causal modeling. http://dx.doi.org/10.48550/ARXIV.2402.17233, CoRR abs/2402.17233. arXiv:2402.17233.