AttenGluco: Multimodal Transformer-Based Blood Glucose Forecasting on AI-READI Dataset

Ebrahim Farahmand*, Reza Rahimi Azghan*, Nooshin Taheri Chatrudi*, Eric Kim*, Gautham Krishna Gudur†,
Edison Thomaz†, Giulia Pedrielli*, Pavan Turaga*, Hassan Ghasemzadeh*

Abstract*— Diabetes is a chronic metabolic disorder characterized by persistently high blood glucose levels (BGLs), leading to severe complications such as cardiovascular disease, neuropathy, and retinopathy. Predicting BGLs enables patients to maintain glucose levels within a safe range and allows caregivers to take proactive measures through lifestyle modifications. Continuous Glucose Monitoring (CGM) systems provide real-time tracking, offering a valuable tool for monitoring BGLs. However, accurately forecasting BGLs remains challenging due to fluctuations due to physical activity, diet, and other factors. Recent deep learning models show promise in improving BGL prediction. Nonetheless, forecasting BGLs accurately from multimodal, irregularly sampled data over long prediction horizons remains a challenging research problem. In this paper, we propose AttenGluco[1], a multimodal Transformer-based framework for long-term blood glucose prediction. AttenGluco employs cross-attention to effectively integrate CGM and activity data, addressing challenges in fusing data with different sampling rates. Moreover, it employs multi-scale attention to capture long-term dependencies in temporal data, enhancing forecasting accuracy. To evaluate the performance of AttenGluco, we conduct forecasting experiments on the recently released AIREADI dataset, analyzing its predictive accuracy across different subject cohorts including healthy individuals, people with prediabetes, and those with type 2 diabetes. Furthermore, we investigate its performance improvements and forgetting behavior as new cohorts are introduced. Our evaluations show that AttenGluco improves all error metrics, such as root mean square error (RMSE), mean absolute error (MAE), and correlation, compared to the multimodal LSTM model, which is widely used in state-of-the-art blood glucose prediction. AttenGluco outperforms this baseline model by about 10% and 15% in terms of RMSE and MAE, respectively.

I. Introduction

According to the World Health Organization [1], the prevalence of type 2 diabetes has increased significantly over the last decades. In 2022, 14% of adults aged 18 years and older were living with diabetes, double the 7% reported in 1990 [2]. This increase is attributed to various factors such as sedentary lifestyles, stress, poor diet, and an aging population [3]. As a result, type 2 diabetes poses a significant public health challenge that requires urgent attention and intervention. Poor management of type 2 diabetes can lead to the progression of chronic health complications and an increased risk of both hyperglycemic and hypoglycemic events. Effectively managing blood glucose levels through consistent monitoring and accurate forecasting is crucial as early intervention measures to prevent hyperglycemic and hypoglycemic events. Accurate glucose prediction is essential for optimizing insulin dosages, meal planning, and exercise habits to maintain blood glucose levels within a safe range.

CGM devices have been developed as an advanced technology to support diabetes management. CGM devices provide valuable insights into blood glucose fluctuations by collecting continuous glucose signals. The CGM data allows patients to monitor fluctuations and trends in their blood glucose levels more effectively by providing real-time blood glucose level measurements. Thus, CGM devices have grown significantly in recent years, making them a widely adopted tool for diabetes prevention. Furthermore, physiological and behavioral variables, such as physical activity levels (e.g., walking or running) and stress levels, affect blood glucose fluctuation [4]. Therefore, the accurate forecasting of blood glucose levels can be evaluated by combining BGL signals with other physiological and behavioral variables. This data integration enables a more comprehensive and personalized approach to managing diabetes, especially for individuals with type 2 diabetes.

Recently, artificial intelligence (AI) and machine learning algorithms have played a critical role in the control and prediction of blood glucose levels. These advanced technologies leverage data from CGM devices and integrate with physiological signals, such as stress levels, heart rate, and physical activity signals. By analyzing these complex datasets, the algorithms can identify trends and patterns in blood glucose fluctuations with high accuracy.

Sequential machine learning models, notably Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) [5], are extensively employed in forecasting time-series signals due to their ability to capture temporal dependencies. GRU-based models outperform traditional methods in univariate time-series classification tasks [6]. Moreover, these models are highly effective for time-series forecasting when optimized with suitable algorithms [7]. These models have also been extensively applied in predicting Type 1 diabetes outcomes [8]. However, they often struggle to capture long-term dependencies inherent in time-series data, which in turn limits their effectiveness in long-term forecasting [9]. Research indicates that while LSTMs are designed to manage longer sequential correlations compared to traditional RNNs, they still encounter challenges in memorizing extended sequences [10].

Recently, transformers have emerged as a powerful model

*Arizona State University, Phoenix, AZ, USA,

† The University of Texas at Austin, Austin, TX, USA

[1] Code base is available at: https://github.com/rzarhmi/AttenGluco

for capturing long-term dependencies in time-series data, primarily through the use of attention mechanisms. Unlike traditional models, which have limited memory and struggle with long-term dependencies, transformers leverage attention mechanisms to effectively capture these dependencies [10]. The attention mechanism within transformers allows the model to weigh the importance of different time steps, which enables them to focus on relevant parts of the sequence when making predictions [11]. This capability is particularly beneficial in applications such as forecasting BGL in type 2 diabetes, where understanding long-range temporal relationships is crucial. Moreover, transformers are well-suited for handling time-series data collected at varying sampling rates. Traditional models often face challenges when dealing with such data due to inconsistencies in temporal resolution. Transformers, however, can manage these variations effectively [12], [13].

Key Limitations and Associated Challenges: We highlight the key limitations of state-of-the-art work and present the associated challenges of blood glucose prediction here.

- Difficulty in achieving accurate long-term blood glucose forecasting.
- Mismatched temporal resolutions in data sources (e.g., CGM readings, physiological, and behavioral variables).
- Limited clinical datasets, especially for populations with type 2 diabetes.

To address these challenges, *a novel accurate forecasting algorithm for long-term prediction is required for individuals with type 2 diabetes.* In this paper, an accurate forecasting model using a Transformer architecture is proposed. The Transformer model is made up of a combination of two attention mechanisms (e.g., cross-attention and multi-scale attention). Cross-attention captures long-term dependencies in temporal data and handles the various temporal resolutions in data sources. The multi-scale attention captures the influence of external time series variables (e.g., physiological signals, behavioral data) on blood glucose levels. Furthermore, to the best of our knowledge, our work is the first to investigate the problem of blood glucose forecasting on AI-READI dataset [14], [15]. The following list summarizes the novel contributions of our work.

- **We propose a Transformer-based architecture with the new attention layers** to forecast blood glucose levels accurately, especially for long-term forecasting.
- **We developed a hybrid attention mechanism of cross-attention and multi-scale attention** to forecast blood glucose levels.
- **We used various body variables**, such as activity along with BGL to enhance BGL prediction precision.
- **We applied our proposed forecasting blood glucose model on the Flagship AI-READI dataset** for patients with type 2 diabetes.
- **We performed multiple experiments** to evaluate the model's accuracy across different subject cohorts and analyze both its performance gains and forgetting behavior as new cohorts were introduced.

II. Related Work

Recent years have seen a surge in blood glucose management technologies. CGM systems, wearable health monitoring devices, and automated insulin delivery systems (AIDS) collectively provide real-time data and partial automation for diabetes care [2], [16]. CGMs offer continuous monitoring of glucose levels, synchronizing with mobile applications for timely alerts on hyperglycemia and hypoglycemia [1]. Wearable sensors further extend coverage to physiological and behavioral metrics, such as heart rate variability and physical activity levels [3], [17]. By combining CGM outputs with additional signals, AIDS can regulate insulin dosage more precisely [18]. However, limitations persist in terms of sensor calibration, missing data, and user non-adherence [16].

Early prediction efforts relied on statistical and time-series models, notably Autoregressive Integrated Moving Average (ARIMA) [19]. Although ARIMA and similar approaches are straightforward, they often fail to capture the complex, nonlinear patterns of glycemic fluctuation. Machine learning (ML) techniques, such as support vector regression and random forests, typically reduce forecasting error by 5–15% compared to ARIMA [19], [20]. However, they still struggle with deeper temporal dependencies over longer prediction windows [20].

Deep learning techniques, widely applied across various domains such as healthcare [21], [22], [23], and classification tasks [24], offer improved forecasting accuracy by effectively modeling intricate temporal dependencies and nonlinear patterns in time series data. Furthermore, integrating causal knowledge into learning frameworks [25], [26] can enhance adaptability and facilitate knowledge transfer across different environments. LSTM architectures are proposed to mitigate vanishing and exploding gradients in recurrent neural networks [4]. By gating internal states, LSTMs retain long-term context for extended horizons, outperforming classical ML methods in certain datasets [4]. Despite these improvements, LSTM-based models often demand significant computational resources and meticulous tuning, making them less flexible for large-scale or highly variable glucose data [10]. GRUs streamline the gating structure of LSTMs, converging 15–25% faster for some time-series tasks [5], [6]. Nevertheless, GRUs still encounter challenges related to sensor inaccuracies, incomplete user logs, and irregular sampling rates [10]. Hybrid methods that integrate convolutional layers with recurrent modules reduce some errors by 2–4% [6], yet extensive clinical validation for blood glucose forecasting remains limited.

Transformers adopt self-attention instead of recurrent loops which facilitates parallel learning over extensive sequences [11]. This approach outperforms RNNs by 5–10% in mean squared error (MSE) for long-horizon predictions [9], [27]. However, many existing implementations assume large, consistent datasets with minimal missing points. Glucose monitoring, conversely, often faces sensor dropouts and user non-adherence, limiting straightforward application [28]. Gluformer [29] developed a transformer-driven blood glucose forecasting model by providing uncertainty intervals rather than single-point estimates. Although a 1–2 mg/dL improvement in short-horizon RMSE has been observed, the absence of multi-scale/cross-attention hinders the model's ability to integrate additional clinical or activity data [29].

In summary, current blood glucose prediction models have notable limitations. ARIMA struggles with nonlinearities [19], [20], while ML models such as support vector regression improve RMSE but fail in long-range forecasting [19], [20]. LSTMs and GRUs improve but suffer from irregular sample rates of various sensors [10], [4]. Recurrent models still encounter inefficiencies for long-horizons forecasting [10]. Transformers, including Gluformer, introduce multi-head attention but lack effective cross-attention for integrating multimodal data [30], [28]. A more robust approach combining multi-scale and cross-attention is needed for accurate, real-world glucose forecasting.

III. Proposed Method

In this section, we introduce our proposed framework for blood glucose prediction. An overview of the AttenGluco framework is shown in Fig. 1. The framework comprises three main components: (a) a sensing module that gathers physiological and behavioral signals from wearable sensors, (b) a preprocessing module for time-series data preparation, and (c) a machine learning forecasting model utilizing the Transformer architecture for blood glucose prediction. Our transformer-based model predicts blood glucose levels (BGL) in individuals with type 2 diabetes by incorporating CGM data alongside activity information. The attention mechanism within the Transformer facilitates the effective integration of multi-time series signals recorded at different sampling rates. Additionally, it is well-suited for predicting highly fluctuating signals such as BGLs. To validate the effectiveness of our proposed model, we conduct experiments using the publicly available AI-READI (Flagship) dataset. The following sections provide a detailed explanation of the forecasting problem and key components of AttenGluco.

A. Forecasting Problem

The problem of blood glucose forecasting with multimodal input data can be formulated as a time series prediction task. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k]$ represent a set of k sensor-derived measurements in the sensing data component. The observation from the ith sensor is denoted as $\mathbf{x}_i = [x_{i,1}, \ldots, x_{i,t}]^\top$, where t is the sampling duration. Our proposed framework, AttenGluco, leverages CGM data ($\mathbf{x}_\text{g}$) and activity data such as walking steps ($\mathbf{x}_\text{ws}$) and walking time intervals ($\mathbf{x}_\text{wi}$), which represent the duration between consecutive walking events. The multi-step forecasting output is expressed as $\hat{\mathbf{x}}_g = [x_{g,t+1}, \ldots, x_{g,t+m}]^\top$, where m represents the number of predicted time steps, commonly referred to as the prediction horizon (PH). Mathematically, the forecasting task can be formulated as $\hat{\mathbf{x}}_g = f(\mathbf{X}; \boldsymbol{\Theta})$, where f represents the forecasting model, parameterized by Θ, which is learned during the training process.

B. AttenGluco

AttenGluco is composed of two primary stages. The first stage, data preparation, focuses on collecting and processing physiological and behavioral data to serve as input for the forecasting model. This stage also includes data interpolation to handle missing values and normalization for consistency. During this phase, BGLs are recorded using a CGM device, while additional behavioral metrics, such as physical activity, are gathered from wearable sensors such as smartwatches, as depicted in Fig. 1. The second stage is the multimodal forecasting model, which utilizes these preprocessed inputs for blood glucose prediction.

The forecasting model is developed based on the Transformer architecture. This architecture leverages an attention mechanism to extract time-dependent patterns from fused irregular time-series data while also capturing long-term dependencies. This approach enables the model to effectively process complex temporal relationships. The structure of our proposed Transformer-based forecasting model is shown in Fig. 2.

The standard transformer architecture is typically made up of an encoder-decoder for data reconstruction. However, we modified this design for our forecasting model and framed it as a supervised learning task. Specifically, we eliminated the decoder and only utilized the encoder for data representation learning. Our customized transformer architecture incorporates two attention mechanisms: cross-attention and multi-scale attention. The cross-attention mechanism integrates various time series data with variant sample rates, while the multi-scale attention captures temporal dependencies within the signals to reduce the effect of random noise [31]. By incorporating these attention mechanisms, our Transformer-based approach enhances the accuracy of BGL forecasting.

Our Transformer architecture consists of embedding and positional encoding layers, followed by cross-attention, feed-forward, Add & Norm layers, and a multi-scale attention block. The input variables $\mathbf{x}_\text{g}$, $\mathbf{x}_\text{ws}$, and $\mathbf{x}_\text{wi}$ are initially processed through an embedding layer $f_\text{embed}(\cdot)$, then passed through a positional encoding function $f_\text{pos}(\cdot)$, producing the transformed representations $\mathbf{X}_\text{G}$, $\mathbf{X}_\text{WS}$, and $\mathbf{X}_\text{WI}$, respectively. Each resulting matrix resides in $\mathbb{R}^{t \times d_\text{model}}$, where t represents the sampling duration and d_model is a hyperparameter. The multi-head attention mechanism in Transformer architectures [11] functions by scaling values ($\mathbf{V} \in \mathbb{R}^{t \times d_\text{model}}$) based on the relationships between keys ($\mathbf{K} \in \mathbb{R}^{t \times d_\text{model}}$) and queries ($\mathbf{Q} \in \mathbb{R}^{t \times d_\text{model}}$). The mathematical formulation of the attention mechanism is presented in Eq. 1.

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_\text{model}}}\right)\mathbf{V} \qquad (1)$$

We designed a two-branch cross-attention layer, where both branches receive $\mathbf{X}_\text{G}$ as the query. In one branch, the keys and values correspond to $\mathbf{X}_\text{WS}$, while in the other, they correspond to $\mathbf{X}_\text{WI}$. The cross-attention (CA) of the first branch is computed using Eqs. 2 and 3.

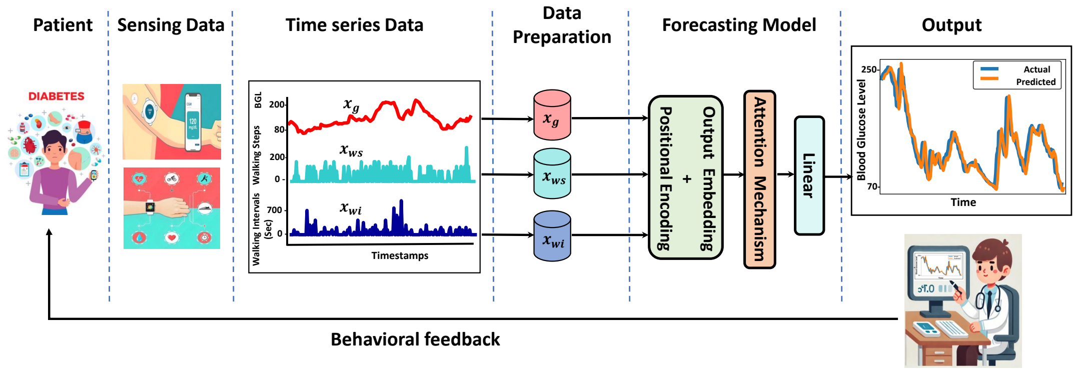

Fig. 1: Overview of the AttenGluco framework including sensing module, data preparation, and forecasting model.

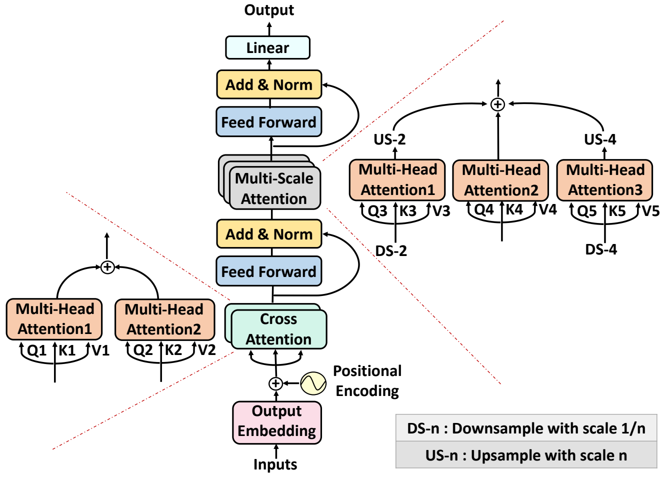

Fig. 2: AttenGluco model architecture consists of cross-attention and multi-scale attention to forecast BGL.

$$\mathrm{CA}\left(\mathbf{X}_{\mathrm{G}}, \mathbf{X}_{\mathrm{WS}}, \mathbf{X}_{\mathrm{WS}}\right) = [\mathbf{H}_1, \ldots, \mathbf{H}_{m_H}]\mathbf{W}_H^{\mathrm{CA}} \tag{2}$$

$$\mathbf{H}_h = \mathrm{Attention}(\mathbf{X}_{\mathrm{G}}\mathbf{W}_{\mathbf{Q}}^{\mathrm{CA}}, \mathbf{X}_{\mathrm{WS}}\mathbf{W}_{\mathbf{K}}^{\mathrm{CA}}, \mathbf{X}_{\mathrm{WS}}\mathbf{W}_{\mathbf{V}}^{\mathrm{CA}}) \tag{3}$$

where $\mathbf{W}_{\mathbf{Q}}^{\mathrm{CA}}$, $\mathbf{W}_{\mathbf{K}}^{\mathrm{CA}}$, and $\mathbf{W}_{\mathbf{V}}^{\mathrm{CA}}$ are weight matrices specific to the attention head and belong to $\mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$. Moreover, $\mathbf{W}_H^{\mathrm{CA}} \in \mathbb{R}^{(m_H \cdot d_{\text{model}}) \times d_{\text{model}}}$ is the final weight matrix that projects the concatenated attention head outputs into the original model dimension. The attention mechanism for the second branch follows the same computation, with the $\mathbf{X}_{\mathrm{WI}}$ as both the key and the query.

Then, the attention outputs from both branches are combined to incorporate cross-attention information. The resulting data is passed through a linear feedforward network followed by an Add & Norm module. The processed output, $\mathbf{X}_{\mathrm{CA}}$, is then fed into a multi-scale attention mechanism comprising three multi-head attention branches, each designed for different downsampling (DS) rates. These branches apply downsampling factors of 1, 2, and 4, where a factor of 1 indicates no downsampling, as illustrated in Fig. 2.

For the first branch, the multi-scale attention mechanism (MA) on $\mathbf{X}_{\mathrm{CA}}$ is computed by using Eqs. 4 and 5.

$$\mathrm{MA}(\mathbf{X}_{\mathrm{CA}}, \mathbf{X}_{\mathrm{CA}}, \mathbf{X}_{\mathrm{CA}}) = [\mathbf{H}_1, \ldots, \mathbf{H}_{m_H}]\mathbf{W}_H^{\mathrm{MA}} \tag{4}$$

$$\mathbf{H}_h = \mathrm{Attention}(\mathbf{X}_{\mathrm{CA}}\mathbf{W}_{\mathbf{Q}}^{\mathrm{MA}}, \mathbf{X}_{\mathrm{CA}}\mathbf{W}_{\mathbf{K}}^{\mathrm{MA}}, \mathbf{X}_{\mathrm{CA}}\mathbf{W}_{\mathbf{V}}^{\mathrm{MA}}) \tag{5}$$

Each attention branch utilizes query, key, and value weight matrices, $\mathbf{W}_{\mathbf{Q}}^{\mathrm{MA}}$, $\mathbf{W}_{\mathbf{K}}^{\mathrm{MA}}$, and $\mathbf{W}_{\mathbf{V}}^{\mathrm{MA}}$, all belonging to $\mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$. The outputs from all attention heads are concatenated and projected back into the original model dimension using the final weight matrix $\mathbf{W}_H^{\mathrm{MA}} \in \mathbb{R}^{(m_H \cdot d_{\text{model}}) \times d_{\text{model}}}$. The remaining two branches follow the same computational process but operate on downsampled input data. This approach improves the model's capability to capture both fine-grained details and long-term temporal dependencies within the input signals.

The outputs from the three multi-scale attention branches are summed and passed through a feed forward network, an Add & Norm block, and a fully connected layer. This final configuration generates m predicted CGM values. Each prediction corresponds to a measurement taken every 5 minutes, meaning that m samples collectively provide forecasts for $m\times$ 5 minutes into the future. In summary, Algorithm 1 describes the data processing pipeline in AttenGluco.

IV. Results & Discussion

In this section, we first introduce the AI-READI dataset used to train AttenGluco. We then compare its performance against a baseline model consisting of a 1D-CNN and LSTM for blood glucose forecasting to highlight the significance of our model for providing accurate forecasting. The baseline model, a multimodal LSTM, is commonly employed in state-of-the-art blood glucose prediction. The comparison is conducted using error metrics, including Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), as well as correlation analysis. We investigate various training and testing scenarios to comprehensively evaluate the performance of AttenGluco.

A. Dataset Description

The dataset used in this study is the publicly available AI-READI Flagship Dataset. This dataset is designed to advance

Algorithm 1 AttenGluco model

Input: Preprocessed and normalized data, including CGM signal ($\mathbf{x}_{\text{g}}$), walking steps ($\mathbf{x}_{\text{ws}}$), and walking time intervals ($\mathbf{x}_{\text{wi}}$), Cross-attention block (CA), Multi-scale atention block (MA), Embedding function (f_{embed}), Positional encoding function (f_{pos}), Two Add & Norm block ($f_{\text{AN}}^{(1)}$,$f_{\text{AN}}^{(2)}$), Two Feedforward model ($f_{\text{FF}}^{(1)}$,$f_{\text{FF}}^{(2)}$), Linear model (f_{lin})
Output: Predicted BGL $\hat{\mathbf{x}}_g$

```
Begin
    $[\mathbf{X}_{\text{G}}, \mathbf{X}_{\text{WS}}, \mathbf{X}_{\text{WI}}] \leftarrow f_{\text{pos}}\left(f_{\text{embed}}\left([\mathbf{x}_{\text{g}}, \mathbf{x}_{\text{ws}}, \mathbf{x}_{\text{wi}}]\right)\right)$
    $\mathbf{X}_{\text{CA1}} \leftarrow \text{CA}(\mathbf{X}_{\text{G}}, \mathbf{X}_{\text{WS}}, \mathbf{X}_{\text{WS}})$
    $\mathbf{X}_{\text{CA2}} \leftarrow \text{CA}(\mathbf{X}_{\text{G}}, \mathbf{X}_{\text{WI}}, \mathbf{X}_{\text{WI}})$
    $\mathbf{X}_{\text{CA}} \leftarrow f_{\text{AN}}^{(1)}\left(f_{\text{FF}}^{(1)}\left(\mathbf{X}_{\text{CA1}} + \mathbf{X}_{\text{CA2}}\right)\right)$
    $\mathbf{X}_{\text{CA}}^{(2)}, X_{\text{CA}}^{(4)} \leftarrow \text{Downsample}(\mathbf{X}_{\text{CA}}, 2), \text{Downsample}(\mathbf{X}_{\text{CA}}, 4)$
    $\mathbf{X}_{\text{MA1}} \leftarrow \text{MA}(\mathbf{X}_{\text{CA}}, \mathbf{X}_{\text{CA}}, \mathbf{X}_{\text{CA}})$
    $\mathbf{X}_{\text{MA2}} \leftarrow \text{Upsample(MA(}\mathbf{X}_{\text{CA}}^{(2)}, \mathbf{X}_{\text{CA}}^{(2)}, \mathbf{X}_{\text{CA}}^{(2)}\text{), 2)}$
    $\mathbf{X}_{\text{MA3}} \leftarrow \text{Upsample(MA(}\mathbf{X}_{\text{CA}}^{(4)}, \mathbf{X}_{\text{CA}}^{(4)}, \mathbf{X}_{\text{CA}}^{(4)}\text{), 4)}$
    $\mathbf{X}_{\text{MA}} \leftarrow f_{\text{AN}}^{(2)}\left(f_{\text{FF}}^{(2)}\left(\mathbf{X}_{\text{MA1}} + \mathbf{X}_{\text{MA2}} + \mathbf{X}_{\text{MA3}}\right)\right)$
    $\hat{\mathbf{x}}_g \leftarrow f_{\text{lin}}(\mathbf{X}_{\text{MA}})$
    return $\hat{\mathbf{x}}_g$
End
```

AI and machine learning research on Type 2 Diabetes Mellitus (T2DM). Collected from 1,067 participants across three U.S. sites. It includes individuals with and without T2DM, balanced across sex, race, and diabetes severity. The dataset consists of four categories: healthy individuals, individuals with prediabetes, individuals with T2DM on oral medication, and individuals with T2DM on insulin.

A key feature of the dataset is its multi-modal structure, where participants were monitored over ten days using a Dexcom G6 CGM for real-time blood glucose, a Garmin Vivosmart 5 for physical activity and heart rate variability, and a LeeLab Anura sensor for environmental factors such as air quality and temperature. The dataset also includes survey data, clinical assessments, and retinal imaging. Daily step counts are recorded via an accelerometer, with occasional gaps due to device recharging. The heart rate sensor also computed a stress index (0-100) based on heart rate variability.

For this study, CGM data and walking activity (steps and intervals) are extracted as key features. After filtering out subjects with missing data, 896 participants are included in the final analysis, distributed as follows: 323 healthy individuals, 207 pre T2DM, 258 with T2DM on oral medication, and 108 with T2DM on insulin.

B. Experimental Setup

The baseline model follows a 1D-CNN architecture coupled with an LSTM. The 1D-CNN consists of two convolutional layers with 64 and 128 filters, each using a kernel size of 3. This is followed by a two-layer LSTM with 128 and 64 output features. The LSTM output is then passed through an MLP composed of three fully connected layers.

Both AttenGluco and the baseline model receive a sliding window of historical data covering 6.66 hours (400 minutes) as input. Training is conducted for 300 epochs with a learning rate of 0.001, optimizing the Mean Squared Error (MSE) using the Adam optimizer. Forecasting performance is assessed across three prediction horizons (PHs): 5 minutes, 30 minutes, and 60 minutes. To ensure consistency, each model undergoes five independent training runs. Model performance is assessed across all subjects, with comparisons based on RMSE [32], MAE [32], and Correlation [33].

As mentioned in section IV-A, the AI-READI dataset categorizes subjects into four cohorts (healthy, pre-T2DM, oral, and insulin) based on diabetes severity. To evaluate AttenGluco's performance across these cohorts, we conducted three distinct experiments under different scenarios (subject training, cohort-wise fine-tuning, and forgetting analysis) and compared the results with the baseline model. The details of each scenario will be discussed in the following sections.

1) Isolated Subject Training: In this scenario, the CGM and activity data of AI-READI participants are first grouped according to their respective cohorts. The proposed model is then applied to each subject individually, with 85% of their data used for training and the remaining 15% reserved for testing. After evaluating one subject, the model is reinitialized before being trained and tested on the next. Table I presents the average error metrics for AI-READI participants across each cohort separately.

TABLE I: Comparison of baseline and AttenGluco performance across different cohorts in the isolated subject scenario. The best results are highlighted in bold.

Cohort	RMSE		MAE		Correlation	
	Baseline	AttenGluco	Baseline	AttenGluco	Baseline	AttenGluco
Healthy	18.04	**16.05**	13.02	**11.12**	0.38	**0.49**
Pre-T2DM	19.95	**18.27**	15.12	**13.65**	0.49	**0.57**
Oral	25.01	**22.56**	17.9	**15.74**	0.55	**0.64**
Insulin	29.9	**27.18**	22.28	**19.93**	0.59	**0.67**

Table I depicts that our proposed method surpassed the baseline model in all performance metrics. For instance, compared to the baseline model, AttenGluco improves the RMSE metric by 11.03%, 8.42%, 9.79%, and 9.09% for the healthy, pre-T2DM, oral, and insulin cohorts, respectively.

2) Cohort-Wise Fine-Tuning: In the cohort-wise fine-tuning scenario, the model is trained progressively within each participant category, unlike the isolated subject scenario where it is reset for each subject. Here, the model is first trained on one subject and then fine-tuned sequentially across the other subjects in the same category, with each subject serving as both training and testing data. This process continues until all subjects in a category have been used. Once a category is completed, the model is reinitialized before moving on to the next cohort. The average performance metrics for each category are presented in Table II. This approach enables the model to gradually adapt to variations within each cohort; therefore, it achieves better performance than the previous scenario.

TABLE II: Performance comparison between the baseline model and AttenGluco across different cohorts in the cohort-wise fine-tuning scenario. The best scores are highlighted in bold.

Cohort	**RMSE**		**MAE**		**Correlation**	
	Baseline	AttenGluco	Baseline	AttenGluco	Baseline	AttenGluco
Healthy	17.79	**15.45**	12.79	**10.96**	0.44	**0.53**
Pre-T2DM	19.77	**17.47**	14.41	**12.46**	0.51	**0.6**
Oral	23.37	**20.45**	16.93	**14.71**	0.57	**0.67**
Insulin	28.22	**25.04**	20.51	**18.03**	0.68	**0.75**

Referring to Table II, we conclude that AttenGluco outperforms the baseline model across all performance metrics. It improves RMSE by 13.15%, 11.63%, 12.49%, and 11.27% for the healthy, pre-T2DM, oral, and insulin cohorts, respectively.

Fig. 3 demonstrates that as more subjects are added into each cohort, the model's performance progressively improves in this scenario. This results in lower errors for newer subjects when used for testing. Notably, the reduction in test error is more significant in AttenGluco, indicating that its performance could further improve with a larger training dataset. For improved clarity and better visibility, we illustrate only 80 subjects of each cohort in Fig. 3 while maintaining the overall distribution and trends of the complete dataset.

Moreover, we evaluate the AttenGluco's forecasting RMSE at different PH values of 5, 30, and 60 minutes. Since CGM data is recorded at 5-minute intervals, a PH of 5 minutes corresponds to $m = 1$ sample, a PH of 30 minutes corresponds to $m = 6$ samples, and a PH of 60 minutes corresponds to $m = 12$ samples. Table III presents a comparison of AttenGluco and the baseline model across different PHs. As shown, increasing the PH leads to a higher RMSE for both models. However, while the baseline model experiences a significant drop in performance, AttenGluco maintains a relatively stable RMSE. This demonstrates AttenGluco's robustness in long-term forecasting.

TABLE III: RMSE Comparison of Baseline and AttenGluco Models Across Different PHs

Cohort	**Baseline RMSE**			**AttenGluco RMSE**		
	5 min	**30 min**	**60 min**	**5 min**	**30 min**	**60 min**
Healthy	7.35	14.37	17.79	7.63	12.38	15.45
Pre-T2DM	7.94	15.43	19.77	8.70	13.50	17.47
Oral	9.15	17.73	23.37	9.33	15.21	20.45
Insulin	12.11	21.00	28.22	11.94	18.55	25.04

3) Continual Learning and Forgetting Analysis: Even though transferring the model to and fine-tuning it on new subjects enhances the model's performance on new data, it simultaneously leads to the loss of previously learned knowledge. This phenomenon, known as catastrophic forgetting [34], is a well-known issue that happens with model retraining. The problem becomes more pronounced when there is a significant distribution shift between the old and new data, which causes the model to prioritize recent patterns while disregarding past ones.

We hypothesize that a distribution shift exists among the four cohorts, which potentially causes the model to forget previously learned information as new cohorts are introduced. To measure forgetting in both models, we evaluate their performance on prior cohorts after completing training on new ones. In this scenario, the model continues training on all subjects of the cohorts without reinitialization. The results are presented in Fig. 4, where the x-axis represents the training cohorts, and each grouped bar chart illustrates the model's performance after training on the respective cohort. As shown, the introduction of new cohorts degrades both model's retention of previous knowledge.

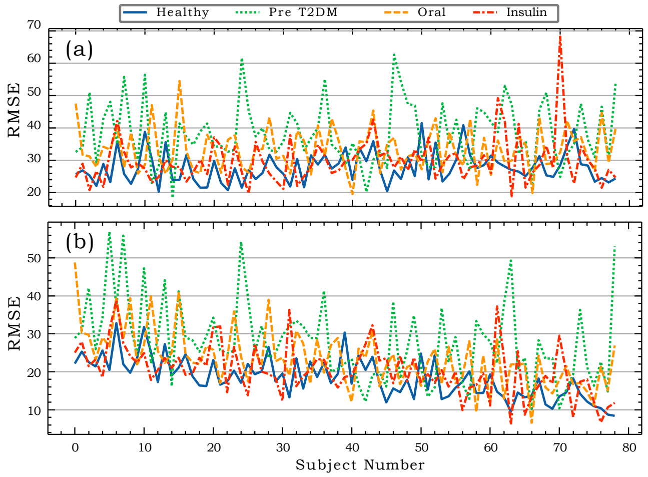

Fig. 3: The RMSE of each subject in (a) the baseline model and (b) AttenGluco under the cohort-wise scenario. As observed, AttenGluco's RMSE decreases significantly within each cohort as more subjects are incorporated into the fine-tuning process.

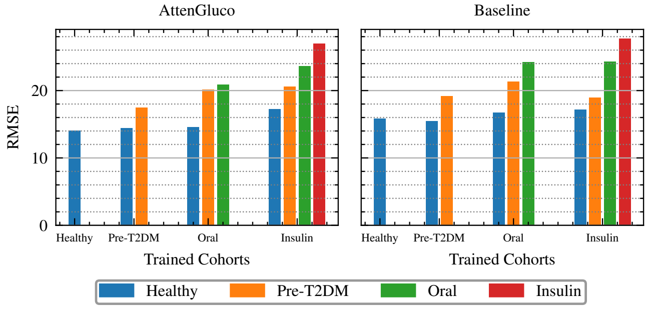

Fig. 4: Fine-tuning the model on new cohorts leads to the loss of knowledge from previous ones.

V. Conclusion

In this study, we proposed AttenGluco, a multimodal Transformer-based framework for long-term blood glucose forecasting using CGM and activity data. By integrating cross-attention and multi-scale attention, our model effectively fuses heterogeneous time-series data and captures long-term dependencies. Our evaluation on the AI-READI dataset demonstrated that AttenGluco outperforms baseline models under different test and train scenarios across various subject cohorts. AttenGluco improved RMSE by about 10% in the isolated subject training scenario. In the cohort-wise fine-tuning scenario, RMSE improvements are even more pronounced, with reductions of about 12%. Additionally, AttenGluco achieved higher correlation scores across all groups, further validating its enhanced predictive capability. Our analysis of forecasting accuracy at different prediction horizons (5, 30, and 60 minutes) shows that AttenGluco consistently outperformed the baseline model, with the most notable gains observed at longer horizons, where it reduced RMSE by up to 3.18 compared to the baseline. Furthermore, our forgetting analysis revealed that AttenGluco maintains lower error rates when fine-tuned on new cohorts. By improving long-term blood glucose forecasting, AttenGluco has the potential to advance precision medicine for diabetes care, enabling more proactive and individualized interventions to maintain optimal glucose levels.

VI. Acknowledgment

This work was supported in part by the National Science Foundation under grant IIS-2402650. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations.

References

[1] M. Abdul Basith Khan, M. J. Hashim, J. K. King, R. D. Govender, H. Mustafa, and J. Al Kaabi, "Epidemiology of type 2 diabetes—global burden of disease and forecasted trends," *Journal of epidemiology and global health*, vol. 10, no. 1, pp. 107–111, 2020.

[2] World Health Organization, "Urgent action needed as global diabetes cases increase four-fold over past decades," World Health Organization, News release, November 2024. [Online]. Available: https://www.who.int/

[3] J. Raffin, P. de Souto Barreto, A. P. Le Traon, B. Vellas, M. Aubertin-Leheudre, and Y. Rolland, "Sedentary behavior and the biological hallmarks of aging," *Ageing Research Reviews*, vol. 83, p. 101807, 2023.

[4] M. M. H. Shuvo and S. K. Islam, "Deep multitask learning by stacked long short-term memory for predicting personalized blood glucose concentration," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1612–1623, 2023.

[5] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[6] N. Elsayed, A. S, and M. Bayoumi, "Deep gated recurrent and convolutional network hybrid model for univariate time series classification," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, 2019. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2019.0100582

[7] A. Makinde, "Optimizing time series forecasting: A comparative study of adam and nesterov accelerated gradient on lstm and gru networks using stock market data," 2024. [Online]. Available: https://arxiv.org/abs/2410.01843

[8] A. R. Patil, J. Schug, C. Liu, D. Lahori, H. C. Descamps, A. Naji, K. H. Kaestner, R. B. Faryabi, and G. Vahedi, "Modeling type 1 diabetes progression using machine learning and single-cell transcriptomic measurements in human islets," *Cell Reports Medicine*, vol. 5, no. 5, 2024.

[9] J. Kim, H. Kim, H. Kim, D. Lee, and S. Yoon, "A comprehensive survey of time series forecasting: Architectural diversity and open challenges," *arXiv preprint arXiv:2411.05793*, 2024.

[10] Y. Kong, Z. Wang, Y. Nie, T. Zhou, S. Zohren, Y. Liang, P. Sun, and Q. Wen, "Unlocking the power of lstm for long term time series forecasting," *arXiv preprint arXiv:2408.10006*, 2024.

[11] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[12] P. Chen, Y. Zhang, Y. Cheng, Y. Shu, Y. Wang, Q. Wen, B. Yang, and C. Guo, "Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting," *arXiv preprint arXiv:2402.05956*, 2024.

[13] Y. Zhang, L. Ma, S. Pal, Y. Zhang, and M. Coates, "Multi-resolution time-series transformer for long-term forecasting," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024, pp. 4222–4230.

[14] AI-READI Consortium (2024), "AI-READI: rethinking data collection, preparation and sharing for propelling AI-based discoveries in diabetes research and beyond," *Nature Metabolism*, 2024. [Online]. Available: https://doi.org/10.1038/s42255-024-01165-x

[15] AI-READI Consortium, "Flagship Dataset of Type 2 Diabetes from the AI-READI Project (2.0.0) [Data set]," *FAIRhub.*, 2024. [Online]. Available: https://doi.org/10.60775/fairhub.2

[16] F. J. Pasquel, M. C. Lansang, K. Dhatariya, and G. E. Umpierrez, "Management of diabetes and hyperglycaemia in the hospital," *The lancet Diabetes & endocrinology*, vol. 9, no. 3, pp. 174–188, 2021.

[17] Z. Guan, H. Li, R. Liu, C. Cai, Y. Liu, J. Li, X. Wang, S. Huang, L. Wu, D. Liu *et al.*, "Artificial intelligence in diabetes management: advancements, opportunities, and challenges," *Cell Reports Medicine*, 2023.

[18] D. Lovic, A. Piperidou, I. Zografou, H. Grassos, A. Pittaras, and A. Manolis, "The growing epidemic of diabetes mellitus," *Current vascular pharmacology*, vol. 18, no. 2, pp. 104–109, 2020.

[19] J. Xie and Q. Wang, "Benchmarking machine learning algorithms on blood glucose prediction for type i diabetes in comparison with classical time-series models," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 11, pp. 3101–3124, 2020.

[20] O. Mujahid, I. Contreras, and J. Vehi, "Machine learning techniques for hypoglycemia prediction: trends and challenges," *Sensors*, vol. 21, no. 2, p. 546, 2021.

[21] R. R. Azghan, N. C. Glodosky, R. K. Sah, C. Cuttler, R. McLaughlin, M. J. Cleveland, and H. Ghasemzadeh, "Personalized modeling and detection of moments of cannabis use in free-living environments," in *2023 IEEE 19th International Conference on Body Sensor Networks (BSN)*. IEEE, 2023, pp. 1–4.

[22] A. Mamun, K. S. Leonard, M. P. Buman, and H. Ghasemzadeh, "Multimodal time-series activity forecasting for adaptive lifestyle intervention design," in *2022 IEEE-EMBS International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 2022, pp. 1–4.

[23] R. R. Azghan, N. C. Glodosky, R. K. Sah, C. Cuttler, R. McLaughlin, M. J. Cleveland, and H. Ghasemzadeh, "Cudle: Learning under label scarcity to detect cannabis use in uncontrolled environments using wearables," *IEEE Sensors Journal*, 2025.

[24] N. T. Chatrudi, W. Clegern, R. Hager, L. Nelson, and H. Ghasemzadeh, "Wavelet-augmented self-supervised learning for accurate classification of cognitive workload," in *2024 IEEE 20th International Conference on Body Sensor Networks (BSN)*, 2024, pp. 1–4.

[25] J. Corazza, H. P. Aria, D. Neider, and Z. Xu, "Expediting reinforcement learning by incorporating temporal causal information," in *Causal Representation Learning Workshop at NeurIPS 2023*, 2023.

[26] S. M. Alsadat, N. Baharisangari, Y. Paliwal, and Z. Xu, "Distributed reinforcement learning for swarm systems with reward machines," in *2024 American Control Conference (ACC)*, 2024, pp. 33–38.

[27] E. Farahmand, S. B. Soumma, N. T. Chatrudi, and H. Ghasemzadeh, "Hybrid attention model using feature decomposition and knowledge distillation for glucose forecasting," *arXiv preprint arXiv:2411.10703*, 2024.

[28] E. Acuna, R. Aparicio, and V. Palomino, "Analyzing the performance of transformers for the prediction of the blood glucose level considering imputation and smoothing," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 41, 2023.

[29] R. Sergazinov, M. Armandpour, and I. Gaynanova, "Gluformer: Transformer-based personalized glucose forecasting with uncertainty quantification," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[30] T. Zhu, T. Chen, L. Kuangt, J. Zeng, K. Li, and P. Georgiou, "Edge-based temporal fusion transformer for multi-horizon blood glucose prediction," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2023, pp. 1–5.

[31] A. Shabani, A. Abdi, L. Meng, and T. Sylvain, "Scaleformer: Iterative multi-scale refining transformers for time series forecasting," *arXiv preprint arXiv:2206.04038*, 2022.

[32] A. Arefeen and H. Ghasemzadeh, "Glysim: Modeling and simulating glycemic response for behavioral lifestyle interventions," in *2023 IEEE*

EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE, 2023, pp. 1–5.

[33] L. Zhang, S. Huang, A. Das, E. Do, N. Glantz, W. Bevier, R. Santiago, D. Kerr, R. Gutierrez-Osuna, and B. J. Mortazavi, "Joint embedding of food photographs and blood glucose for improved calorie estimation," in *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2023, pp. 1–4.

[34] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.