Computer Methods and Programs in Biomedicine 265 (2025) 108737

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

Personalized glucose forecasting for people with type 1 diabetes using large language models

Francisco J. Lara-Abelenda [a], David Chushig-Muzo [a], Pablo Peiro-Corbacho [a], Ana M. Wägner [b], Conceição Granja [c], Cristina Soguero-Ruiz [a,*]

[a] Department of Signal Theory and Communications, Telematics and Computing Systems, Rey Juan Carlos University, Madrid, Spain
[b] Instituto Universitario de Investigaciones Biomédicas y Sanitarias, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain
[c] Norwegian Centre for E-health Research, University Hospital of North, Norway, Tromsø, Norway

ARTICLE INFO

Keywords:
Large language models
Glucose forecasting
Transformers
GPT
Time series forecasting
Continuous glucose monitor
Type 1 diabetes

ABSTRACT

Background and objective: Type 1 Diabetes (T1D) is an autoimmune disease that requires exogenous insulin via Multiple Daily Injections (MDIs) or subcutaneous pumps to maintain targeted glucose levels. Despite the advances in Continuous Glucose Monitoring (CGM), controlling glucose levels remains challenging. Large Language Models (LLMs) have produced impressive results in text processing, but their performance with other data modalities remains unexplored. The aim of this study is three-fold. First, to evaluate the effectiveness of LLM-based models for glucose forecasting. Second, to compare the performance of different models for predicting glucose in T1D individuals treated with MDIs and pumps. Lastly, to create a personalized approach based on patient-specific training and adaptive model selection.

Methods: CGM data from the T1DEXI study were used for forecasting glucose levels. Different predictive models were evaluated using the mean absolute error (MAE) and the root mean squared error and considering the Prediction Horizons (PHs) of 60, 90, and 120 min.

Results: For short-term PHs (60 and 90 min), the personalized approach achieved the best results, with an average MAE of 15.7 and 20.2 for MDIs, and a MAE of 15.2 and 17.2 for pumps. For long-term PH (120 min), TIDE obtained an MAE of 19.8 for MDIs, whereas Patch-TST obtained a MAE of 18.5.

Conclusion: LLM-based models provided similar MAE values to state-of-the-art models but presented a reduced variability. The proposed personalized approach obtained the best results for short-term periods. Our work contributes to developing personalized glucose prediction models for enhancing glycemic control, reducing diabetes-related complications.

1. Introduction

Type 1 Diabetes (T1D) is an autoimmune disease that destroys the insulin-producing cells in the pancreas, producing insulin deficiency, which is the hormone that stimulates glucose transport from the bloodstream to cells [1]. To maintain glucose levels within a target range ($[70, 180]$ mg/dl), the main therapy in T1D patients is the administration of exogenous insulin [2] either by Multiple Daily Injections (MDIs) or by continuous subcutaneous insulin infusions with a pump [2].

MDIs are administered using pens or syringes, involving both basal (long-acting) and bolus (rapid-acting) insulin [3]. By contrast, pumps are wearable devices containing rapid-acting insulin that can supply both basal and bolus insulin continuously to the subcutaneous tissue [4]. Although both treatments aim to maintain glucose concentrations within a target range, pumps have proven to be more efficient in insulin delivery compared to MDIs [4]. These therapies combined with Continuous Glucose Monitoring (CGM) devices have led to better glycemic control, thus improving T1D patient's quality of life. CGM devices have also facilitated data collection, which is promising in developing personalized and data-driven models for glucose prediction, supporting clinical decision-making and glucose self-management [5]. Accurate glucose predictions may help people to make informed decisions about insulin dosing, diet, and exercise, leading to more effective glucose control [6].

Artificial Intelligence (AI) models based on Artificial Neural Networks (ANNs) architectures have proven excellent results for forecasting glucose in previous studies [7–9]. In particular, Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs)

* Corresponding author.
E-mail addresses: francisco.lara@urjc.es (F.J. Lara-Abelenda), david.chushig@urjc.es (D. Chushig-Muzo), pablo.peiro@urjc.es (P. Peiro-Corbacho), ana.wagner@ulpgc.es (A.M. Wägner), conceicao.granja@ehealthresearch.no (C. Granja), cristina.soguero@urjc.es (C. Soguero-Ruiz).

https://doi.org/10.1016/j.cmpb.2025.108737
Received 7 November 2024; Received in revised form 19 February 2025; Accepted 22 March 2025
Available online 2 April 2025

have been extensively used owing to their capability to capture short-term and long-term temporal dependencies in CGM data and are considered as baseline models in previous studies [10,11]. Despite their extensive usage, these methods often face limitations in capturing complex patterns inherent to time series. To address this, novel and sophisticated models based on convolutional neural networks [12], transformers [13,14] and Multiple Layer Perceptron (MLP)-based models (such as the TSMixer [15], TiDE [16], N-HiTS [17], NBEATS [18]) have shown excellent results for long-term forecasting [16,19]. However, for glucose forecasting, limited research has been conducted using these models. NBEATS and N-HiTS have obtained considerably better results compared with more traditional models (*e.g.,* ARIMA, linear regression, RNNs, and LSTMs) for glucose forecasting [20]. In contrast, TS-Mixer and TiDE have not yet been evaluated for glucose forecasting in the literature.

In recent years, Large Language Models (LLMs), originally designed for Natural Language Processing (NLP) tasks, have revolutionized the digital era and the way how humans interact with technology [21]. These models leverage vast amounts of text data to achieve unprecedented performance in NLP tasks and generative AI [21]. Thanks to their ability to capture complex and long-range dependencies, these models are promising to be used in other data modalities [22]. Several studies have explored the capability and limitations of LLMs for time series forecasting, including Time-LLM [23], TimeGPT [24], aLLM4TS [25], Lag-llama [26] among others. Among these models, TimeGPT is the only foundational model tested in short-term glucose forecasting for pediatric patients [27], reaching state-of-the-art performance. However, Time-LLM allows us to train with a small set of time series and requires fewer resources compared to other fine-tuned models, becoming promising for applications with scarce data.

In this study, we investigate the effectiveness of the LLM-based models for predicting glucose in patients diagnosed with T1D. Towards that end, we used glucose values belonging to people involved in the public dataset named *Type 1 Diabetes EXercise Initiative* (T1DEXI) [28]. The aim of this study is two-fold. First, we evaluated the effectiveness and performance of the LLM-based model named Time-LLM for forecasting glucose using CGM data. Second, we performed a comparative study of Time-LLM with other ANN-based models for predicting glucose, distinguishing between two insulin treatment modalities: MDIs and insulin pumps. Three different Prediction Horizons (PHs) were considered, including 60, 90, and 120 min. We quantitatively evaluated the performance of the models using the average values for Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) and Clarke Error Grid (CEG) analysis [29]. To create personalized models adapted to the CGM characteristics of participants, we follow two approaches: *(i)* patient-specific training, where each model is exclusively trained with CGM data from the same participant; and *(ii)* adaptive model selection, where eight different models are evaluated for each participant aiming to determine and select the best-fit model (measured by MAE). To our knowledge, this is one of the first studies proposing a personalized approach using LLM-based models to predict CGM time series data, comparing the performance of various models in T1D patients with different insulin administration methods (MDIs and pumps). In summary, our contributions are presented below. We introduce the application of Time-LLM for glucose forecasting, demonstrating its effectiveness for clinical applications. We also evaluate the feasibility of using LLMs with small datasets and without additional re-training, reducing computational time while achieving reasonable forecasting results. The source code and results for reproducibility can be accessed at the following link: github.com/ai4healthurjc/TimeLLM-CGM.

The rest of this paper is organized as follows: Section 2 presents the foundations of the models used for forecasting, the methodology followed and the dataset used, and the experimental setup; Section 3 presents the forecasting results in different PHs and considering individuals treated with MDIs and insulin pump. The discussion and conclusions are presented in Sections 4 and 5, respectively.

2. Methods

This section provides a comprehensive overview of the forecasting models used in this research, details the datasets employed, describes the proposed methodology for training the AI-based models (as illustrated in Fig. 1), and outlines the experimental setup in which the corresponding experiments were conducted.

2.1. Large language models for predicting time series

LLMs represent one of the most significant advancements in the field of AI in recent years. These models are based on transformer architectures [30], which are composed of encoder and decoder blocks as well as self-attention mechanisms. The encoder transforms input data into fixed-sized feature maps, whereas the decoder transforms these maps back into the input. The self-attention mechanism relates each input word to each other by establishing links between related words, identifying which previous tokens influence each generated token, thus capturing intricate dependencies and relationships within the text [30]. Generally, LLMs are trained on large text data, encompassing diverse domains and languages, to enhance their ability to generate human-like text, answer questions, and complete other language-related tasks. Among the most remarkable LLMs, we find the Generative Pre-trained Transformer (GPT) [31] created by OpenAI, Gemini (Google), Llama [32] (Meta), Mistral [33] (Mistral AI).

Although LLMs often produce impressive outputs, it remains to be explored how they perform in real-world scenarios with different data modalities. The application of LLMs for forecasting time series gathered considerable attention last year, as evidenced by recent studies [34–36]. Most LLM-based models were trained on time series characterized by periodicity and tendency [35,36], but a few studies have explored their application to non-periodic, real-world time series, particularly in the context of clinical data. As a result, the development of LLM-based applications for diabetes-related data has become a relevant research topic in recent months. In the literature, LLM-models have been explored for different aspects of diabetes management, such as providing personalized support for users [37,38], making an early detection of T2DM [39] or forecasting glucose levels in pediatric patients [27]. In this study, we used the model called Time-LLM [23] due to its novelty, performance, and reduced computational time, which make it highly efficient compared to other methods. Traditional LLMs require large volumes of data and fine-tuning to reach competitive and reasonable results, which can lead to prolonged periods for training models and higher computational complexity. Time-LLM uses a reprogramming layer framework to address the need for large data, reducing time and computational costs, thus allowing to obtaining reasonable results with datasets with small and medium sizes.

Time-LLM introduces a *reprogramming framework* that combines the language knowledge (from pre-trained word embedding) and time series information via linear projection and multi-head cross-attention layers. It reprogrammed an existing LLM (*e.g.,* GPT, Llama, BERT), called backbone, into a time series forecaster without requiring fine-tuning, thus reducing computational costs for training. Specifically, given a sequence of observations $\mathbf{X} \in \mathbb{R}^{N \times T}$ consisting of N different 1-dimensional variables across T time steps, a LLM is reprogrammed $f(\cdot)$ to understand the input time series and accurately forecast H future time steps, denoted by $\hat{\mathbf{Y}} \in \mathbb{R}^{N \times H}$. The aim is to minimize the mean squared error between the ground truth values $\mathbf{Y}$ and the predictions $\hat{\mathbf{Y}}$.

Before training Time-LLM, time series are normalized, segmented into patches, and embedded. These embedded patches are then reprogrammed using learned text prototypes to align the source (time series) and the target (language model). Next, we enhance the LLM ability to reason over time series data by prompting it with these reprogrammed patches. The LLM processes this input to generate output representations, which are subsequently projected to produce the final forecast. The Time-LLM framework comprises four primary components:

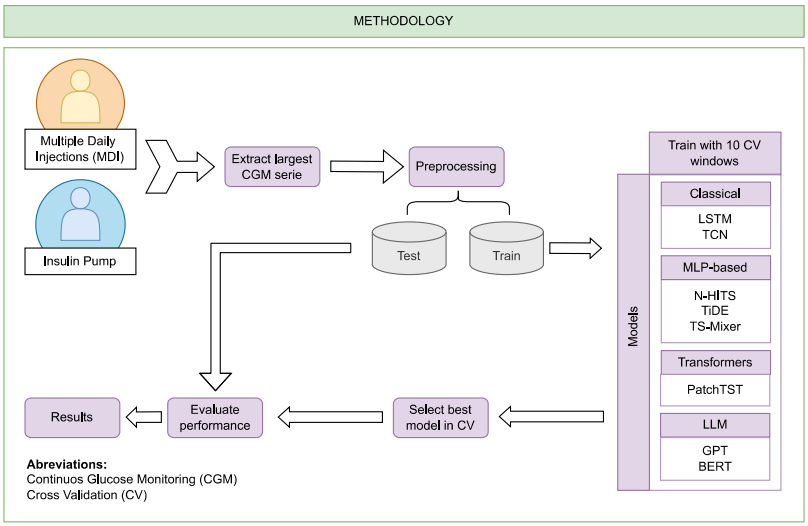

Fig. 1. A schematic diagram of workflow followed in the current work for predicting glucose values.

- *Input patching*: Each input sequence $\mathbf{X}^{(i)}$ is normalized via reversible instance normalization and segmented into consecutive overlapping or non-overlapping sequences of fixed length called *patches* named $\mathbf{X}_P^{(i)}$. This process aims to preserve local semantic information within each patch to construct a compact sequence of input tokens. Furthermore, it also works as a tokenization method, creating a compact sequence of input tokens, which helps reduce computational complexity. With these patches $\mathbf{X}_P^{(i)}$, an embedding was built using a simple linear layer as the patch embedder, generating $\hat{\mathbf{X}}_P^{(i)}$.
- *Patch reprogramming*: *Patch embeddings* are reconfigured into the representation space of the source data to align the modalities of time series and natural language, thereby activating the backbone's capabilities (*e.g.*, GPT, Llama, BERT) for processing time series. To address this challenge, $\hat{\mathbf{X}}_P^{(i)}$ is reprogrammed using pre-trained word embeddings $\mathbf{E} \in \mathbb{R}^{V \times D}$ from the backbone model, where V denotes the vocabulary size and D is the hidden dimension of the backbone model. However, without prior knowledge of which source tokens are most relevant, using the full embedding matrix $\mathbf{E}$ would create a large and potentially dense reprogramming space. To simplify this, authors in [23] used a linear approach to extract a smaller subset of text prototypes, represented as $\mathbf{E}' \in \mathbb{R}^{V' \times D}$, where $V' \ll V$. The text prototypes are designed to learn associations between language cues, and then these are combined to represent local patch information. Importantly, this process occurs entirely within the pre-trained space of the LLM, ensuring compatibility and efficiency. This method also enables the adaptive selection of relevant information using a multi-head cross-attention layer, and then this result is linearly projected to align the hidden dimensions with the backbone model, producing $\mathbf{O}^{(i)}$.
- *Prompt-as-prefix*: Prompting is an effective method for task-specific activation of LLMs. The prompts are used as prefixes to enrich the input context and guide the transformation of reprogrammed time series patches. It enhances the LLM's adaptability to downstream tasks while complementing patch reprogramming. In the Patch-as-Prefix approach, language models are prompted to predict time series values using natural language, but this method struggles with processing high-precision numerals and requires complex post-processing. On the other hand, the Prompt-as-Prefix approach addresses these challenges by integrating three crucial components: *(i)* dataset context; *(ii)* task instructions; and *(iii)* input statistics. The dataset context provides the LLM with critical background information about the input time series, which often varies significantly across different domains. Task instructions play a key role in guiding the LLM to transform patch embeddings for specific tasks.
- *Output projection*: Following the packing and feed-forwarding of prompt and patch embeddings through the frozen LLM, the prefixal part is discarded, and the output representation is obtained $O^{(i)}$. Subsequently, the output representation is flattened and linearly projected to obtain the final forecast $\hat{\mathbf{Y}}^{(i)}$.

2.2. Forecasting models based on neural networks

RNNs have been extensively utilized for sequence modeling tasks, including sequence forecasting and sequence labeling [40]. RNNs are characterized by cycles that feed activations from previous time steps back into the network to inform decisions about the current input. These activations, stored in the network's internal state, provide temporal contextual information. However, training conventional RNNs with gradient-based back-propagation often faces vanishing and exploding gradient problems [41]. LSTM networks seek to address these problems by including units called memory blocks in the recurrent hidden layer, which can control the flow of information [42]. LSTMs have been previously used in forecasting applications [42,43], and they have demonstrated reasonable results in glucose forecasting [42,44].

Despite the extended usage of RNNs and LSTMs, several studies have shown that convolutional architectures can outperform these models (in some applications) [40]. By skipping temporal connections, causal convolution filters can be applied to larger time spans while remaining computationally efficient. The temporal convolutional network (TCN) uses convolutional layers with dilated convolutions to capture temporal dependencies across a sequence. By stacking multiple convolutional layers, TCNs can handle long input sequences and produce accurate forecasts by mitigating vanishing gradient problems. TCN has been used to forecast series in multiple applications [45,46], including glucose forecasting where it has outperformed the results of LSTMs [47,48].

The emergence of transformer-based models, initially proposed for NLP tasks, has been gaining importance for the analysis of multiple data modalities and domains [49,50], including glucose classification and forecasting [51–53]. Transformers replace RNN cells with self-attention layers, point-wise fully connected layers, and positional encoding, seeking to capture long-range dependencies in sequences. PatchTST [14] is a transformer-based model that enhances locality and captures semantic information not available at the point level by aggregating time steps into subseries-level called patches. Although PatchTST has not yet been

used in glucose forecasting, it has outperformed RNNs in time series forecasting in other studies [54].

Despite recent progress in transformer-based models, they face difficulties in training due to their excessive computational complexity. For example, both attention and fully connected layers scale quadratically in memory and computational cost with the forecasting horizon length. To overcome these limitations, novel MLP-based models have been developed [15,17]. MLP-based models are models without any self-attention mechanisms, recurrent, or convolutional mechanisms, but they use statistical information for forecasting. In this paper, we evaluated three MLP-based models: TSMixer [15], NHiTS [17], and TiDE [16]. We selected these models based on their novelty and performance in time series forecasting in other domains [40,55].

TSMixer [15] is an MLP-based model that jointly learns temporal and cross-sectional representations of the time series by iteratively combining time and feature information of stacked mixing layers. TiDE [16] starts by encoding the past time series along with any associated covariates using dense MLPs. This encoding process creates a dense hidden representation filled with learned features that best describe the data characteristics. Then, another set of dense MLPs uses this hidden representation to generate future forecastings. The temporal decoder refines these forecastings by adapting them to future covariates. Additionally, residual connections are introduced to provide model regularization, preventing overfitting and reducing the risk of vanishing gradient problems [16]. NHITS [17], based on NBEATS [18], performs local nonlinear projections onto basis functions across multiple blocks. Each block consists of an MLP with ReLU non-linearities, which learns to produce coefficients for the backcast and forecast outputs of its basis. The backcast output is used to clean the inputs of subsequent blocks, while the forecasts are combined to form the final forecasting. Moreover, the blocks are grouped in stacks, each specialized in learning different characteristics of the data using a different set of basis functions.

2.3. Dataset description and preprocessing

In this work, we employed data belonging to a total of 497 adults with T1D from the dataset named T1DEXI [28], which was a randomized controlled trial study that analyzed the impact of different types of physical exercise on glucose levels. The participants were residents from the United States aged over 18 years with T1D for at least 2 years. Data were collected during the years 2019 and 2020 in compliance with the ethical principles that have their origin in the Declaration of Helsinki and with the standards of Good Clinical Practice. Adult participants were randomly assigned to six structured aerobic, interval, or resistance exercise sessions during four weeks. CGM data were obtained using either personal Dexcom G6 CGM or blinded Dexcom G6 CGM, with measurements taken at 5-minute intervals. Participants followed an intensive insulin regimen using either pumps or MDIs. Although several types of insulin pumps were utilized in the T1DEXI [28], we only considered the T-slim X2 with Control-IQ pump, a hybrid closed-loop system with automated insulin delivery, because it was the most widely used in the original study. The selected pump regimen included 190 participants, while the MDI regimen had 79 participants.

Regarding the preprocessing stage, each participant's CGM sequence consists of 28 days of data. However, we only used for training and validation a subsequence of the CGM data where the total amount of consecutive missing values did not reach one hour (60 min). This approach aimed to avoid having a large number of consecutive missing values, which can adversely affect the performance of the models to be evaluated. Then, we dealt with any remaining missing values in this subsequence using linear interpolation [56]. It is worth mentioning that our goal was not to validate the robustness of imputation techniques; on the contrary, we selected CGM sequences with the most available information to reduce the impact of any interpolation method. Moreover, this approach has been implemented in other forecasting models that used LLM, such as in [27].

After the interpolation, we continued to preprocess the time series by changing the sampling frequency from 5 min to 15 min. This change in frequency is based on improving the generalization of this implementation. Since one of the goals is to implement our methodology in a real-world application, the possibility of using CGM data extracted from any CGM device is very important. While most CGM devices can record glucose values at a 5-minute sampling frequency, some devices, like the Abbott FreeStyle Libre, resample the data every 15 min by averaging the 1-minute measurements [57]. The Abbott FreeStyle Libre is one of the most popular CGM devices [58]. Moreover, although CGM devices that record data at 15-minute intervals can be resampled to 5-minute intervals, the reverse conversion is more accurate. This is because downsampling a time series is generally more reliable than upsampling. Lastly, computational time and resources are critical factors for real-world implementation. By reducing the sampling rate, we can decrease computational time and resource usage while preserving the period analyzed by the model and the prediction horizon. Therefore, we chose to resample our glucose measurements and train our models using 15-minute interval data. This resampling approach has been used in several studies across the literature [59,60].

2.4. Methodology

This study aimed to design a methodology for training personalized models by selecting the most suitable model for a person from a set of eight data-driven models. Consequently, instead of selecting the same model (*e.g.,* LSTM or TCN) for all participants, we select the best option for each user from a set of eight different models, including LSTM, TCN, N-HITS, TiDE, TS-Mixer, PatchTST, Time-LLM-GPT, and Time-LLM-BERT. To achieve this, the CGM data of each participant is divided into training and testing splits. These splits are generated from different periods of the same participant's CGM data. A detailed description of how the training and testing division is performed is shown in Fig. 2. The amount of data used in the training and testing time series is determined based on five factors: the model input size ($input\_size$), the number of cross-validation windows (w), the number of test iterations (n), the step size (s), and the PH.

The length of the training set is calculated as follows:

$$\text{Train set length} = input\_size + w \cdot s_{train} + PH.$$

Similarly, the length of the test set is determined by:

$$\text{Test set length} = input\_size + n \cdot s_{test} + PH.$$

The selected $input\_size$ was determined through experimentation. We tested several intervals, including 48, 36, 24, and 12 h. Ultimately, the parameter $input\_size$ was fixed at 24 h as it provided the optimal balance between prediction accuracy and computational efficiency. Extending the input time frame beyond 24 h resulted in an exponential increase in computation time, which was deemed impractical for real-world applications.

The parameter w determines the number of cross-validation windows in which the models are evaluated to accurately identify the model (between the 8 options) best suited for each user, along with its optimal set of hyperparameters. To ensure that the models perform well with a specific set of hyperparameters across different time intervals, a relatively large w was selected, specifically $w = 50$. In contrast, the parameter n represents the number of test iterations where the model performance is assessed. We decided to evaluate the model over 5 time intervals. This choice reflects our aim to analyze the model's performance across distinct time spans while minimizing computational resource requirements.

Then, s determines the displacement of each one of the w cross-validation windows for the training and the n test iterations. Since the training set contains a high number of w, we aimed to ensure the methodology remains feasible for real-world implementation. To achieve this, we selected a training sample duration (s_{train}) of 15 min.

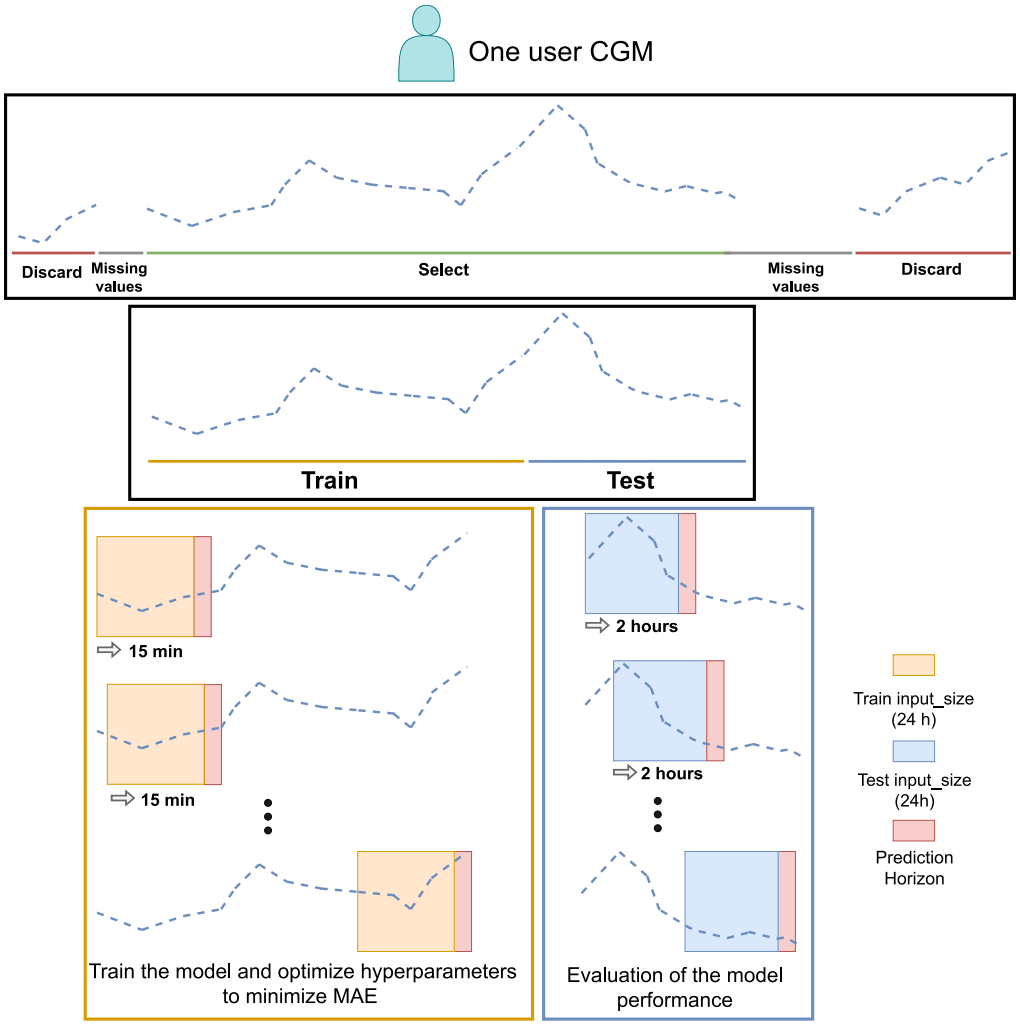

Fig. 2. Schematic diagram of the workflow followed to train the artificial intelligence models for each user. The process involves selecting the subsequence of the CGM data with the least missing values, dividing the data into training and testing sets, and training/evaluating models using various windows of the same time series.

On the other hand, in the test set, the value w is smaller because it is preferable to evaluate the mode in fewer iterations, but more separated from each other. As a result, to evaluate the performance of the models across multiple timespans, we increased the s_{test} value, selecting 2 h. Therefore, we evaluated the model during five time iterations, each separated by 2 h.

Finally, we evaluated the performance of all the models, including the personalized approach, across three different PHs: 60 min, 90 min, and 120 min. The choice of PH directly impacts the size of both the training and testing datasets, as these are adjusted according to the specific PH being evaluated. To evaluate the models' robustness and adaptability, we assessed their performance across multiple PHs, reflecting the varying requirements of real-world scenarios.

In conclusion, this pipeline processes a time series of glucose data, which is divided into training and testing partitions based on the previously described parameters. Therefore, the training set consists of 37.5, 38, and 38.5 h for the three PHs, respectively, while the test set includes 35, 35.5, and 36 h for the same three PH. Using the training partition, we trained eight models through cross-validation with 50 windows. For the personalized approach, we selected a different model for each user, choosing the one that achieved the best performance across the 50 cross-validation windows computed from the training set. Finally, these models, along with the personalized approach, were evaluated across the 5 iterations of the test set. Following this procedure, the computational time is significant only during the initial training phase, as the models do not need to be retrained. Once a model is trained and assigned to each user, which demands substantial computational resources and time, subsequent predictions are computed in just a few seconds since the model is not retrained. Moreover, the use of Time-LLM results in a significant reduction in computational time compared to other approaches, such as Time-GPT. This methodology was consistently applied to all participants, from both datasets (MDIs or insulin pumps), ensuring that the approach remained adaptable and effective for varying treatment modalities.

2.5. Experimental setup

In this study, eight models were used to forecast glucose levels using CGM data from MDI-treated patients and insulin pumps. For the six state-of-the-art models, the time series data were fed into the models without performing explicit feature engineering. Instead, a 24-hour window of data was selected as the input_size, and Min-Max normalization was applied before the training. In contrast, for the LLM-based models, the preprocessing involved segmenting the time series data into normalized patches and embedding the sequences, as detailed in Section 2.1.

For each one of the eight models, we explored several hyperparameter values (detailed in Table 1) for training and selected those that provided superior performance (measured by MAE) in the validation subset. The validation subset is composed of 50 cross-validation

Table 1
Summary of hyperparameters evaluated for different forecasting models.

Model	Hyperparameters	Values/options
LSTM	hidden layer size (encoder)	50, 100, 200, 300
	number of layers (encoder)	1, 2, 3, 4
	context size	5, 10, 50
	hidden layer size (decoder)	50, 100, 200, 300
	learning rate	0.0001, 0.001, 0.01, 0.1
	batch size	3, 6, 10
TCN	hidden layer size (encoder)	50, 100, 200, 300
	context size	5, 10, 50
	hidden layer size (decoder)	64, 128, 256
	learning rate	0.0001, 0.001, 0.01, 0.1
	batch size	3, 6, 10
N-HiTS	size of the windows	[2, 2, 1], [1, 1, 1], [2, 2, 2], [4, 4, 4], [8, 4, 1], [16, 8, 1]
	stack's coefficients	[168, 24, 1], [24, 12, 1], [180, 60, 1], [60, 8, 1], [40, 20, 1], [1, 1, 1]
	learning rate	0.0001, 0.001, 0.01, 0.1
	batch size	3, 6, 10
TiDE	hidden size	256, 512, 1024
	decoder output dimension	8, 16, 32
	temporal decoder dimension	32, 64, 128
	number encoder layers	1, 2, 3
	number decoder layers	1, 2, 3
	lower temporal projected dim	4, 8, 16
	dropout	0.0, 0.1, 0.2, 0.3, 0.5
	layernorm	True, False
	learning rate	0.0001, 0.001, 0.01, 0.1
	batch size	3, 6, 10
TS-Mixer	number of blocks	1, 2, 4, 6, 8
	dropout	0.3, 0.6, 0.9
	feed-forward layers in MLP	32, 64, 128
	batch size	3, 6, 10
	learning rate	0.0001, 0.001, 0.01, 0.1
Patch-TST	encoder layers	2, 4, 8
	hidden layer size	16, 128, 256
	number heads	4, 16, 32
	patch length	16, 24, 36
	learning rate	0.0001, 0.001, 0.01, 0.1
	batch size	3, 6, 10
Time-LLM-BERT	patch length	16, 24, 36
	learning rate	0.0001, 0.001, 0.01, 0.1
	batch size	3, 6, 10
Time-LLM-GPT	patch length	16, 24, 36
	learning rate	0.0001, 0.001, 0.01, 0.1
	batch size	3, 6, 10

windows with a stride of 15 min. To determine the best hyperparameter values in the validation subset, we used the tool `raytune` [61] that is used for distributed hyperparameter tuning and includes efficient search algorithms. For each model, we conducted 100 different combinations using a Bayesian optimizer. Due to computational constraints, we were unable to evaluate as many hyperparameters as desired for Time-LLM. To maintain consistency in time series length, we selected the 50 participants with the largest CGM from each cohort (*i.e.,* participants who used MDIs or pumps), resulting in a total of 100 participants.

To quantitatively evaluate the performance of forecasting models, we used the figures of merit MAE and RMSE, which are defined as follows.

$$MAE = \frac{1}{T_N}\sum_{t=1}^{T_N}|y_i - \hat{y}_i| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{T_N}\sum_{t=1}^{T_N}(y_i - \hat{y}_i)^2} \tag{2}$$

where T_N is the number of measurements in a PH, and $\hat{y}_i$ and y_i denotes the predicted and real value at time t. Both MAE and RMSE measure the errors between predicted and real values, being always positive values, and with 0 indicating accurate forecastings with no error.

To evaluate the performance of the models under critical conditions, such as episodes of hypoglycemia or hyperglycemia, we used the CEG plot. It is a tool used to visualize and quantify the discrepancies between predicted values by a model (y-axis) and reference values (x-axis). It is divided into five zones (A–E), where values within zones A and B are considered clinically acceptable, while those in zones C–E are potentially dangerous due to the risk of clinically significant errors. Therefore, for a model to perform well, almost all the points must be placed within the A and B zones. As a result, we used the CEG to analyze the number of points that fall in A–B zones, which is a key indicator of prediction errors and the performance of forecasting models.

In addition, we used boxplots to analyze the distributions of the errors for each model, including the personalized approach. Furthermore, we performed the Wilcoxon test [62] to compare the personalized model with the others, assessing whether there were statistically significant differences in performance. The Wilcoxon test is a non-parametric statistical test used to compare two paired groups without assuming a normal distribution. It evaluates whether the median differences between paired observations are significantly different from zero. As a consequence, it is perfect to assess differences in performance metrics, such as MAE, between two models in machine learning studies [63]. This combination of visualization and statistical testing provided a comprehensive evaluation of the models' results.

3. Results

This section presents the forecasting results based on CGM data collected from participants using either MDI or insulin pumps for their treatment. These results are presented for three different PHs: 60, 90, and 120 min. Finally, we analyze the errors generated by the models during CGM forecasting and highlight key insights gained from comparing the performance of all models across both cohorts (MDI and pump users).

3.1. Forecasting results in participants using multiple daily injections

In this subsection, we present the forecasting results of models trained using CGM data of participants who used MDIs as their insulin administration method. Table 2 presents the results of different models and evaluated in different PHs, including 60, 90, and 120 min. All models exhibit reasonable performance for the three PHs in average terms, with 15.7 ≤MAE≤ 24.3 (60 min), 20.2 ≤MAE≤ 25.9 (90 min), and 19.8 ≤MAE≤ 29.6 (120 min). It is important to indicate that when the STD is higher than the mean MAE or RMSE, it does not imply that the MAE or RMSE has negative values. Instead, it indicates the inter-patient variability of predicted values of glucose. For some individuals, the forecasting is accurate, resulting in lower MAE and RMSE values, whereas for others, the forecasting models perform worse, leading to higher MAE and RMSE values. While intra-patient variability reflects the glucose dynamics for a specific patient, inter-patient variability indicates how glucose concentrations of T1D patients vary significantly from patient to patient within a given population [64]. The inter-subject variability and multiple factors such as exogenous insulin, sleep disturbances, and physical exercise make glucose predictions complex and challenging, causing less accurate values in some scenarios.

Note that the personalized approach achieved the lowest MAE in mean and standard deviation (STD) for the 60-minute and 90-minute PHs, with 15.7 ± 12.5 and 20.2 ± 14.2, respectively. For the PH of 120 min, TiDE achieved the best performance, with an MAE of 19.8 ± 14.2. By comparing the LLM-based models, Time-LLM-GPT consistently outperformed Time-LLM-BERT in all PHs. It also achieved similar results compared to other models, exhibiting lower STDs in the

Table 2
Forecasting results (mean±standard-deviation) of the different models in participants using MDI as the insulin administration method. The results with the lowest mean for each PH are in bold.

Model	PH=60 minutes		PH=90 minutes		PH=120 minutes	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
LSTM	17.5 ± 18.4	19.8 ± 20.4	21.7 ± 17.9	25.8 ± 21.0	22.0 ± 15.9	26.3 ± 18.8
TCN	18.7 ± 20.2	21.0 ± 22.0	22.6 ± 18.8	26.4 ± 21.96	21.6 ± 14.5	26.2 ± 17.9
N-HiTS	19.0 ± 18.5	21.2 ± 20.2	21.8 ± 17.7	25.6 ± 21.0	23.4 ± 14.8	28.1 ± 18.5
TiDE	19.5 ± 19.4	21.7 ± 21.4	22.6 ± 18.7	26.5 ± 21.9	**19.8 ± 12.2**	**24.2 ± 15.3**
TSMixer	24.3 ± 20.9	26.4 ± 22.7	25.9 ± 19.9	30.3 ± 22.3	24.8 ± 14.7	29.5 ± 17.9
PatchTST	18.7 ± 19.9	20.9 ± 21.9	22.2 ± 17.8	26.1 ± 21.0	22.5 ± 15.3	27.6 ± 19.4
Time-LLM-GPT	19.4 ± 15.1	21.6 ± 15.6	21.4 ± 14.1	24.3 ± 15.2	23.8 ± 14.1	27.4 ± 15.8
Time-LLM-BERT	20.7 ± 16.5	22.8 ± 17.0	22.1 ± 14.5	25.0 ± 15.7	29.6 ± 18.2	33.0 ± 19.0
Personalized	**15.7 ± 12.5**	**17.5 ± 13.4**	**20.2 ± 14.2**	**23.5 ± 15.9**	21.9 ± 14.0	25.7 ± 15.8

60-minute and 90-minute PHs. For instance, in the 60-minute PH, Time-LLM-GPT showed an MAE of 19.4 ± 15.1, which is considerably lower than the best individual model LSTM (MAE 17.5 ± 18.4). Regarding the 90-minute PH, Time-LLM-GPT achieved the best performance (MAE 21.4 ± 14.1) compared to individual models and even demonstrated a lower STD compared to the personalized approach (MAE 20.2 ± 14.2). However, LLMs exhibited worse performance in the PH of 120 min, achieving the two worst MAE and RMSE values compared to the rest of the models.

To ensure an accurate evaluation of intra-patient variability and to compare it with inter-patient variability, we calculate intra-patient variability as the mean of the individual STDs obtained by measuring the absolute error for each patient. Appendix C presents a detailed analysis of the intra-patient and inter-patient variability for users employing MDI as their insulin administration method. Inter-patient variability is calculated as the STD of the MAE across all users. Table C.6 presents a reduced intra-patient variability compared with the inter-patient variability for users using MDI. For a PH of 60 min, all models exhibit an intra-patient variability ranging between 9.2 and 11.6, highlighting the consistency of predictions within individual users. In contrast, for inter-patient variability and in the PH of 60 min, the STD of the MAE ranges from 12.5 to 20.9, nearly doubling the variability observed in the intra-patient case. For longer PHs of 90 and 120 min, intra-patient variability stabilizes between 12.7 and 16.3, depending on the model. The TiDE model is the only one where intra-patient variability (14.4) exceeds inter-patient variability (12.2), likely due to its reduced overall variability for a long-predicted horizon. In conclusion, while the gap between intra-patient and inter-patient variability decreases as the PH increases for users on MDI, intra-patient variability remains more controlled and consistent than inter-patient variability.

In Fig. 3, we analyzed the distribution of the MAE for each model across different PHs using boxplots. For 60-minute PH, the personalized approach demonstrated the lowest median MAE (11.7) and the smallest interquartile range (IQR) among all models, with 75% of users achieving an MAE below 20.4. The small IQR is indicative of a strong and consistent performance. For 90-minute PH, the personalized approach achieved the lowest median MAE (15.3), with 75% of users achieving an MAE below 28.5. Both the median MAE and the IQR increased with the growth of the PH, as expected. For 120-minute PH, the personalized approach no longer shows the best performance, as the TiDE and LSTM models achieve lower medians with 16.6 and 17.0, respectively, compared to 19.8 for the personalized approach. However, the personalized approach still performed well, with 75% of users obtaining an MAE below or equal to 29.9, similar to the LSTM model. In this case, the TiDE achieved the best performance with the smallest IQR because 75% of the participants achieved an MAE lower than 26.5. Despite the personalized approach achieving lower MAE values and STD overall, the Wilcoxon test indicated that the differences were not statistically significant across all PHs. It is important to note that the differences may not be statistically significant due to the limited sample size of just 50 samples. In addition, to complement the previous results, we represented the RMSE distributions using boxplots in Appendix A for participants using MDI, in particular, in Fig. A.9.

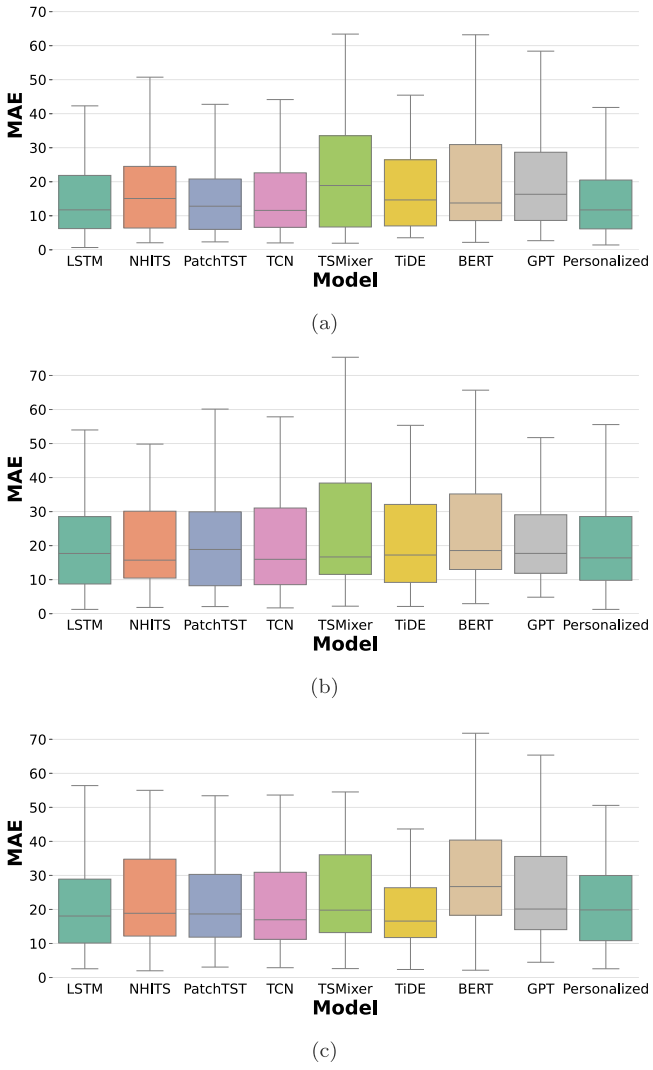

Fig. 3. Distribution of MAE values for different models trained with CGM data from participants using MDIs as the insulin administration method. Each boxplot was generated using a different PH: (a) 60 min; (b) 90 min; (c) 120 min.

Fig. 4 shows the CEG plots that compare the real and predicted CGM values obtained by Time-LLM-GPT and the personalized approach. We analyzed Time-LLM-GPT. It demonstrated stable performance across the CGM data of different participants, and the personalized approach because it achieved the best overall results. As shown in Fig. 4, both models reliably predict hyperglycemic events without errors. However, some inaccuracies occur when forecasting hypoglycemic events in both

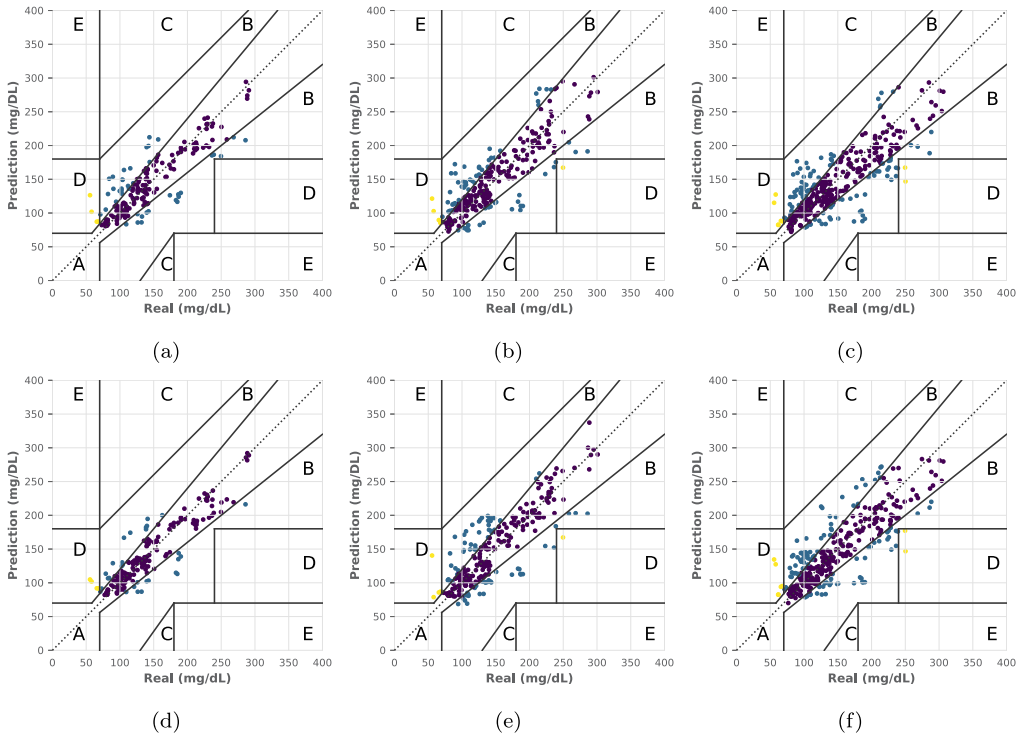

Fig. 4. Clarke error grid plots that compare predicted and real CGM values for all participants who used MDIs as the insulin administration method. To compare the performance of LLM-based models, we present only the CEG plots for: (a–c) Time-LLM using GPT as backbone at PH-60 min, 90 min, and 120 min, respectively ; and the (d–f) personalized method (model with the best MAE and RMSE values) at PH-60 min, 90 min, and 120 min, respectively.

Table 3
Percentages of errors in each zone of the CEG using Time-LLM-GPT and personalized approach for all participants who used MDIs.

Model	CEG Zone	PH=60 min	PH=90 min	PH=120 min
Time-LLM GPT	Zone A	76.0%	72.3%	70.0%
	Zone B	22.0%	26.0%	28.0%
	Zone C	0%	0%	0%
	Zone D	2.0%	1.7%	2.0%
	Zone E	0%	0%	0%
Personalized	Zone A	83.0%	71.3%	70.5%
	Zone B	15.0%	27.0%	27.5%
	Zone C	0%	0%	0%
	Zone D	2.0%	1.7%	2.0%
	Zone E	0%	0%	0%

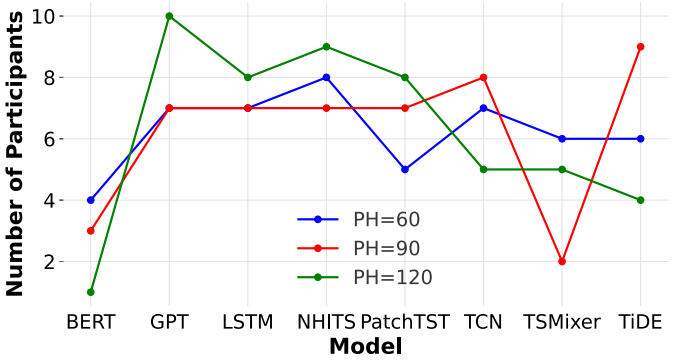

Fig. 5. Number of MDI participants in which each model reached the best performance in the validation set across different PHs.

approaches. Regarding Time-LLM-GPT, Table 3 presents the percentage of total values in each zone of the CEG plot. Zone A decreases as the prediction PH increases: 76.0%, 72.3%, and 70.0% for PHs of 60 min, 90 min, and 120 min, respectively. Conversely, the percentage of values in Zone B increases with larger PHs: 22%, 26.0%, and 28.0% for the same intervals. The proportions in Zones C and E remain constant at 0% across all PHs. Lastly, the percentage of values in Zone D remains relatively stable: 2.0%, 1.7%, and 2.0% for PHs of 60 min, 90 min, and 120 min, respectively. The percentage of total values in each zone of the CEG plot for the personalized approach are shown in Table 3, showing a higher percentage of total values within Zone A: 83%, 71.3%, and 70%, respectively for 60, 90 and 120 min. The percentages in Zone B are 15%, 27%, and 28%, while the percentages in Zone D remain the same as those observed for the Time-LLM-GPT model (2%, 1.7%, and 2%) across all PHs. As with the Time-LLM-GPT approach, the percentages in Zones C and E remain constant at 0% across all PHs.

Fig. 5 presents the number of MDI participants in which each model reached the best performance in the validation subset for different PHs. As shown, all models are employed by at least one participant in each PH, indicating that the selection of an accurate model is highly dependent on individual CGM variability. Time-LLM-GPT is the most frequently used, maintaining stable selection across the three PHs: 7 out of 50 participants at both the 60-minute and 90-minute PHs, and 10 participants at the 120-minute PH. NHiTS is the second most frequent model selected, being used by 7, 8, and 9 participants across the three PHs, respectively. By contrast, Time-LLM-BERT is the model less used, with its usage decreasing as the PH increases: 4 participants at the 60-minute PH, 3 participants at the 90-minute PH, and only 1 participant at the 120-minute PH.

3.2. Forecasting results in participants using insulin pump

In this subsection, we present the forecasting results of models trained using CGM data of participants who used insulin pumps. Table 4 presents the forecasting results of various models for three different PHs, including 60, 90, and 120 min. All models exhibit reasonable performance for the three PHs, with 15.2 $\leq$MAE$\leq$ 19.3 (60 min),

Table 4
Forecasting results (mean±standard-deviation) of different models trained with CGM data of participants treated with pumps as the insulin administration method. The results with the lowest mean for each PH are in bold.

Model	PH=60 minutes		PH=90 minutes		PH=120 minutes	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
LSTM	16.4 ± 12.3	18.4 ± 13.2	18.5 ± 13.6	21.1 ± 14.8	19.6 ± 13.9	23.1 ± 15.4
TCN	17.1 ± 12.6	18.8 ± 13.3	18.6 ± 13.3	21.3 ± 14.3	20.1 ± 12.7	23.6 ± 14.4
N-HiTS	15.8 ± 11.5	17.6 ± 12.3	17.6 ± 12.4	20.1 ± 13.6	18.7 ± 13.6	21.8 ± 15.0
TiDE	15.5 ± 10.8	17.3 ± 11.7	18.0 ± 12.7	20.6 ± 13.6	20.5 ± 13.0	23.5 ± 14.0
TSMixer	16.1 ± 12.0	17.9 ± 12.9	17.7 ± 12.6	20.5 ± 13.1	19.7 ± 13.4	23.0 ± 14.9
PatchTST	16.0 ± 12.1	17.7 ± 12.9	17.6 ± 11.8	20.1 ± 12.7	**18.5 ± 13.6**	**21.6 ± 14.8**
Time-LLM-GPT	16.9 ± 9.6	18.7 ± 10.1	18.5 ± 11.7	21.0 ± 12.6	19.6 ± 10.2	22.9 ± 10.8
Time-LLM-BERT	19.3 ± 12.6	21.1 ± 12.8	20.3 ± 12.7	22.5 ± 13.1	21.9 ± 13.0	24.5 ± 13.3
Personalized	**15.2 ± 10.1**	**17.0 ± 10.7**	**17.2 ± 11.0**	**19.0 ± 11.5**	19.0 ± 11.5	22.2 ± 12.4

17.2 ≤MAE≤ 20.3 (90 min), and 18.5 ≤MAE≤ 21.9 (120 min). It is worth mentioning that these models outperformed the results obtained of models trained with participants using MDI, and the STDs of all models are lower compared to the values for the MDI participants (see Table 2). The personalized approach achieved the lowest MAE and STD in the 60-minute and 90-minute PHs, with an MAE of 15.2 ± 10.1 and 17.2 ± 11.0 for 60 min and 90 min, respectively. For PH of 120 min, N-HiTS and Patch-TST outperformed the personalized approach (19.0 ± 11.5), reaching 18.7 ± 13.6 and 18.5 ± 13.6, respectively. The best individual model for PH=60 was TiDE (15.5 ± 10.8), and for PH=90 was Patch-TST (17.6 ± 11.8). Regarding the LLM models, Time-LLM-GPT consistently outperformed Time-LLM-BERT for all PHs, achieving reasonable forecasting results compared to other models, and exhibiting lower STDs for all three PHs. For instance, in the 60-minute PH, Time-LLM-GPT achieved an MAE of 16.9 ± 9.6, even lower than the personalized approach. For 120-minute PH, Time-LLM-GPT (MAE 19.6 ± 10.2) achieved the lowest STD among all models.

Table C.7 presents the inter-patient and intra-patient variability for patients using insulin pumps as their insulin administration method. Similar to the previous results, the findings indicate reduced intra-patient variability compared to inter-patient variability. For a PH of 60 min, all models exhibit a value below 9.6, demonstrating consistent predictions within individual users. In contrast, inter-patient variability shows values ranging between 9.5 and 12.6, highlighting a noticeable increase compared to the intra-patient case. For larger PHs of 90 and 120 min, the intra-patient variability stabilizes within the range of 10.6 to 11.5 for 90 min and 11.4 to 13.0 for 120 min. This intra-patient variability remains lower than inter-patient variability, which ranges from 11.0 to 13.6 for 90 min and 10.2 to 13.9 for 120 min. Notably, the Time-LLM-GPT model is the only one to exhibit higher intra-patient variability (13.0) compared to inter-patient variability (10.2) for PH = 120 min. These results indicate that increasing the PH decreases the gap between inter-patient and intra-patient variability, although intra-patient variability remains consistently more controlled than inter-patient variability.

Fig. 6 shows the distribution of the MAE for each model across different PHs using boxplots, specifically for participants utilizing pumps as their insulin administration method. For a 60-minute PH, the personalized approach demonstrates the second smallest IQR, with 75% of users achieving an MAE below 21.5. The TiDE model slightly outperformed the personalized approach, with 75% of the users achieving an MAE below 19.9, although both models shared a similar median MAE of 12.2. At a 90-minute PH, the personalized approach achieved the lowest median MAE (13.0) and the smallest IQR, with 75% of users obtaining an MAE below 20.8. Finally, for a 120-minute PH, the personalized approach maintained strong performance, achieving the third-lowest median MAE (15.1), surpassed by the PatchTST (12.8) and N-HiTS (14.3) models. Notably, the personalized approach exhibited a more compact IQR, with its upper bound reaching 23.9, compared with the PatchTST model which reached a value of 25.6. The median and upper bound of the IQR exhibited minimal increases as the PH extended, highlighting the consistent performance of the personalized approach across all PHs. Although the personalized approach achieved lower MAE values and standard deviations overall, the Wilcoxon test revealed that the differences were not statistically significant across all PHs, likely due to the limited sample size of only 50 participants. Additionally, we represented the RMSE distributions using boxplots in

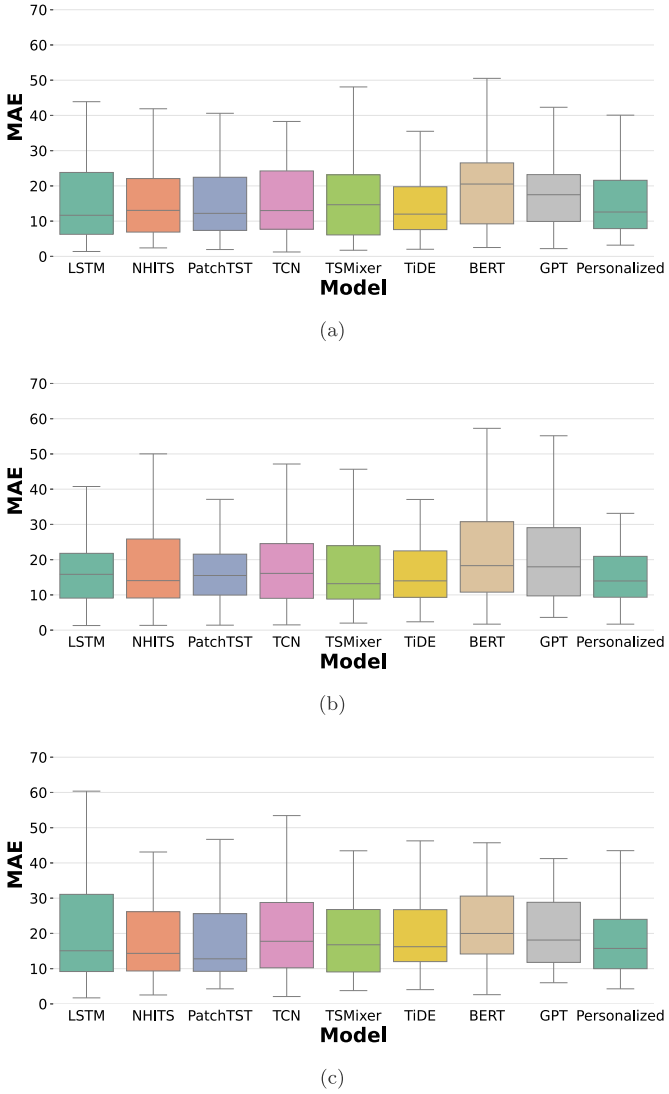

Fig. 6. Distribution of MAE values of different models trained with CGM data of participants treated with pumps as the insulin administration method. Each boxplot was generated using a different PH: (a) 60 min; (b) 90 min; (c) 120 min.

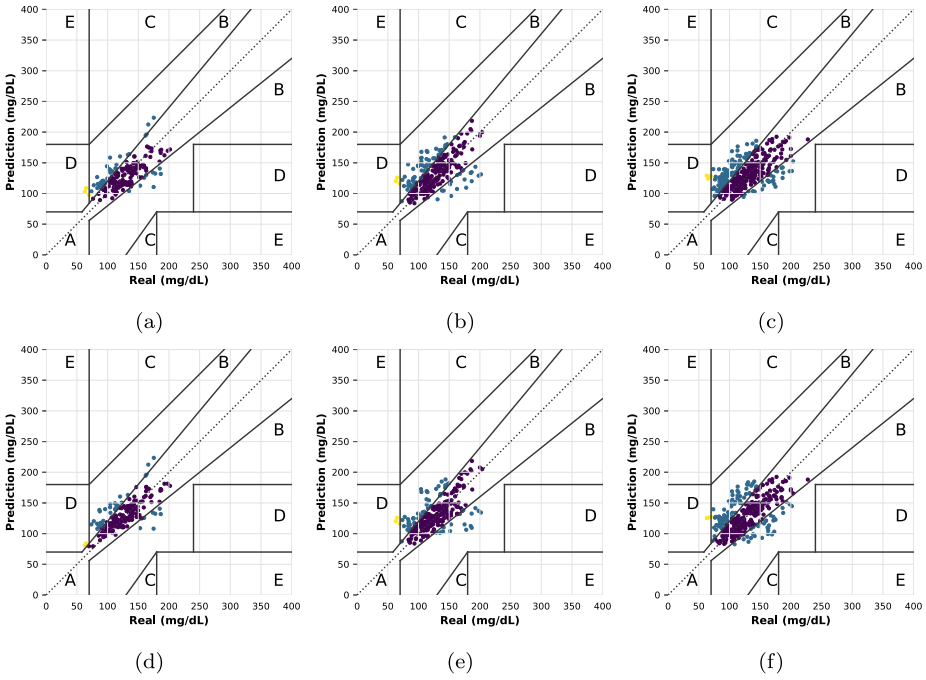

Fig. 7. Clarke Error Grid plots that compare predicted and real CGM values for all participants who used insulin pump as the insulin administration method. To compare performance of LLM-based models, we only present the CEG plots for: (a–c) Time-LLM using GPT as backbone at PH-60 min, 90 min, and 120 min, respectively ; and the (d–f) personalized method (model with the best MAE and RMSE values) at PH-60 min, 90 min, and 120 min, respectively.

Appendix A for participants using pumps as the insulin administration method, in particular, in Fig. A.10.

Fig. 7 shows the CEG plots that compare the real and estimated glucose values in participants using insulin pumps. Similar to the previous section, where Fig. 4 represents the Time-LLM-GPT and personalized approaches. Time-LLM-GPT was selected because of its stable performance across different participants, while the personalized approach achieved the best overall results. Fig. 7 demonstrates that all critical errors arise when forecasting hypoglycemic events, while this approach consistently succeeds in predicting hyperglycemic events without failure.

Table 5 presents the percentages of values located in each zone of the CEG plot for participants using pumps and the Time-LLM-GPT approach. The percentage of total values within Zone A decreases with larger PHs: 72.2%, 70.7%, and 65.7% for 60 min, 90 min, and 120 min, respectively. Conversely, the percentage of values in Zone B increases: 26.1%, 28.3%, and 33.6% for the same PHs. The percentages in Zones C and E remain constant at 0% across all PHs. Lastly, the percentage in Zone D remains relatively stable: 1.7%, 1%, and 0.8%. As a result, although the general error increases with larger PHs, the percentage of critical errors is maintained or even reduced as PH increases. In contrast, the personalized approach consistently outperformed Time-LLM-GPT as it is shown in Table 5. For PHs of 60 min, 90 min, and 120 min, the personalized approach achieved a higher percentage of total values within Zone A: 77.3%, 77.3%, and 70.1%, respectively. The percentages in Zone B are 21.7%, 21.7%, and 29.1%, while Zone D is relatively stable across all PHs, with values of 1.0%, 1.0%, and 0.8%. As with the Time-LLM-GPT approach, the percentages in Zones C and E consistently remained at 0% across all PHs.

Fig. 8 presents the number of insulin pump participants using each model across different PHs in the personalized approach. All models were employed by at least one participant in each PH, indicating that, similar to MDI participants, the selection of an accurate model is highly dependent on individual CGM variability. For the 60-minute PH, the two most common models are Time-LLM-GPT (used by 11 participants) and TCN (used by 13 participants). As the PH increases, the number of participants using Time-LLM-GPT remains stable, with 15 participants at the 90-minute PH and 12 at the 120-minute PH. In contrast, the number of participants using TCN decreases significantly, dropping to 9 at the 90-minute PH and only 2 at the 120-minute PH. The transformer-based model PatchTST also presented stable usage across the different forecasting horizons, being employed by 6 participants at the 60-minute PH, 9 participants at the 90-minute PH, and 9 participants at the 120-minute PH. Similar to the MDI cohort, Time-LLM-BERT is the least commonly used model among insulin pump participants, indicating its relatively lower suitability for this specific forecasting task.

Table 5
Percentages of errors in each zone of the CEG using Time-LLM-GPT and personalized approach for all participants who used pumps.

	CEG Zone	PH=60 min	PH=90 min	PH=120 min
Time-LLM GPT	Zone A	72.2%	70.7%	65.7%
	Zone B	26.1%	28.3%	33.5%
	Zone C	0%	0%	0%
	Zone D	1.7%	1.0%	0.8%
	Zone E	0%	0%	0%
Personalized	Zone A	77.3%	77.3%	70.1%
	Zone B	21.7%	21.7%	29.1%
	Zone C	0%	0%	0%
	Zone D	1.0%	1.0%	0.8%
	Zone E	0%	0%	0%

4. Discussion

In this paper, we evaluated the effectiveness and performance of the LLM-based model named Time-LLM for forecasting glucose values collected from CGM devices of patients diagnosed with T1D. Specifically, we compared the performance of these models against various ANN-based models by considering two cohorts, including participants using MDI and insulin pumps. This comparison is relevant because in the literature most predictive models have been primarily developed using CGM data of patients treated with insulin pumps, and limited research has been conducted using MDIs. The ANN-based models

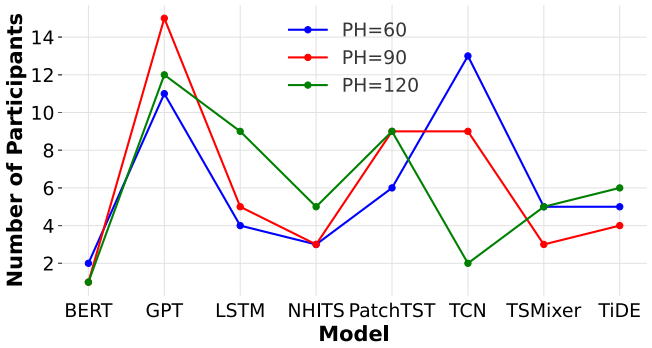

Fig. 8. Number of insulin pump participants in which each model reached the best performance in the validation set across different PHs.

included LSTM, TCN, N-HiTS, TS-Mixer, TiDE, and the transformer-based model named PatchTST. It is important to note that LLM-based models (with backbone BERT and GPT) along with TiDE, TS-Mixer, and Patch-TST, have not been explored for glucose forecasting. Lastly, we designed an adaptive personalized approach where each participant is assigned the model that best fits their data, rather than using the same model for all participants.

Regarding the insulin pump cohort, the models that achieved the lowest MAE for each PH were: personalized (15.2 ± 10.1) for PH=60 min, personalized (17.2 ± 11.0) for PH=90 min, and Patch-TST (18.5 ± 13.6) for PH=120 min. In contrast, for the MDI cohort, the best results across the three PHs were: personalized (15.7 ± 12.5) for PH=60 min, personalized (20.2 ± 14.2) for PH=90 min, and TiDE (19.8 ± 12.2) for PH=120 min. The MAE values in the insulin pump cohort are notably lower, indicating better overall performance of the models compared to participants using MDIs. Bear in mind that the closed loops already have a glucose prediction model embedded, corrections in the insulin infusion rate are performed, and less variability in glucose is expected. Furthermore, participants treated with insulin pumps consistently exhibited significantly lower STD across all models and PHs. This difference in variability between the two cohorts can be attributed to a higher short-term variability associated with the usage of MDIs, as it has been recognized in previous studies [65]. The difference in STD between the two cohorts decreases as the PH increases. This short-term variability affects the general performance of the models in the MDI cohort reducing the MAE of the models and increasing the STD values.

Reviewing the best results, for short-term forecasting, the best option was the proposed personalized approach, obtaining the best performance in both datasets for the 60 and 90-minute PHs. The RNN-based models perform considerably well, with LSTM consistently outperforming TCN models in both datasets. Notably, in the MDI cohort for PH=60, LSTM emerged as the best individual model, achieving an MAE of 17.5 ± 18.4. However, both LSTM and TCN models exhibited a considerably high STD, with a mean STD of 15.3 across all PHs in both datasets, compared to the overall mean STD of 13.7 for all models. MLP-based models generally outperformed RNN models. This is evident in the MDI cohort in PH of 120 min, where the TiDE model achieved the best overall performance with an MAE of 19.8 ± 12.29. This aligns with previous findings that TiDE can capture information for long-term forecasting [16]. Additionally, both N-HiTS and TiDE models performed well in the pump cohort, with TiDE being the best individual model in PH of 60 min (15.5 ± 10.8) and PH of 90 min (17.6 ± 12.4). Conversely, the TS-Mixer model did not perform well in glucose forecasting. It consistently ranks among the worst models across all PHs in both datasets. Lastly, PatchTST also provided reasonable results, particularly in the pump dataset and for larger forecasting horizons. It achieved the best performance in PH=120 and PH=90 in the pump dataset. However, PatchTST faced challenges in the CGM data of the MDI cohort, failing to capture the variability encountered, resulting in sub-optimal performance.

For the first time in the literature, we evaluated LLM-based models using as backbone both GPT and BERT to forecast glucose levels. We observed that the forecasting results of these models were similar to those values of RNN, MLP-based models, and transformer-based models. Moreover, the LLM approaches captured variability better and consistently achieved the lowest STD, surpassing the personalized approach in some cases. Time-LLM-GPT was the best individual model in the PH of 60 min for the MDI cohort. Finally, it is worth noting that Time-LLM-GPT outperformed the results of the Time-LLM-BERT in all cases. Reviewing the CEG plots, Time-LLM-GPT and the personalized approach performed similarly. The CEG results were better in the pump cohort than in the MDI cohort, as the percentage of values falling in the dangerous zone was lower. Considering only the personalized approach, in the pump dataset, only 1% of the values fell into the danger zone D, consistently across all PHs. In contrast, for the MDI dataset, this percentage varies between 1.8% and 2.1%. Both results are good, however, the minimum acceptable number of values located in zones A and B is 99% [29]. In both datasets, the values were only located in zones A, B, and D, so only the personalized approach in the pump cohort reaches the minimum 99% value and complies with the regulations.

Upon reviewing the figures representing the selected models in the personalized approach (Figs. 5 and 8), we observed that all models were selected for at least one participant across both cohorts. This distribution highlights the importance of accurately selecting the model to fit the unique characteristics of each participant's glucose variation levels. In both datasets, the number of participants selecting the Time-LLM-GPT model has maintained stable and high across all PHs. The proposed personalized approach outperformed the individual models in patients with MDI and insulin pumps in short-term forecasting (60 min and 90 min). However, for long-term forecasting, TiDE performed better for MDI, and Patch-TST works better for insulin pumps. Additionally, the models trained on the pump cohort outperformed those trained on the MDI cohort, especially in terms of STD, which was notably high in the MDI dataset.

By analyzing the MAE distributions for participants using MDI and pumps as their insulin delivery methods (Figs. 3 and 6), we extracted several insights. These plots reveal that, although the personalized approach does not always achieve the best scores, it consistently demonstrates strong and stable performance across the different PHs. In contrast, some models exhibit better performance at specific PHs but perform poorly at others. Furthermore, the personalized approach shows robust performance in both participant groups. The upper bounds of the IQR for the MDI group were 20.4, 28.5, and 29.9 for the 60-minute, 90-minute, and 120-minute PHs, respectively. For pump users, the corresponding upper bounds were 21.5, 20.8, and 23.9. This indicates that, while both results mean good performance, the results for participants using insulin pumps are slightly better compared to those using MDI, as evidenced by their lower mean, median, and upper IQR bounds for MAE. Despite these results, the Wilcoxon test indicated that the differences in performance between the personalized approach and the other models were not statistically significant. This outcome may be attributed to the limited sample size of only 50 participants. In statistical analysis, it is essential to consider statistical power, which is the probability that a study of a given size will detect a statistically significant difference when one truly exists. The commonly accepted threshold for adequate statistical power is 0.8 [66]. In this case, with a sample size of only 50, the Wilcoxon test fails to meet the required power of 0.8, using the formula from [66], highlighting the need for a larger sample size to achieve reliable statistical comparisons.

The methodology developed in this study will be integrated into a real-world mobile-based application as part of the WARIFA project [67]. WARIFA [68] is an international and multidisciplinary project involving researchers from multiple European Union countries,

whose aim is to create a comprehensive AI-based system for personalized early risk prediction of Non-Communicable Diseases (NCDs). The mobile application will be accessible via smartphones and the users will be able to add data from various data sources to assess NCD risks and provide personalized lifestyle recommendations. Within the application, users will initially upload three days of glucose measurements from a CGM device. Data will be used to train and select the model best suited to the user's characteristics. Once the model is calibrated, users will only need to provide 24 h of glucose measurements for the application to forecast glucose levels based on these inputs in the PH indicated by the user. The main goal is to empower individuals to take proactive steps in preventing risk episodes, promoting overall health, and reducing the burden on healthcare systems.

One of the key requirements for deploying an AI model in a real-world application is accurately assessing its resource consumption and environmental impact. This is particularly important because these tools often demand substantial energy, contributing significantly to a notable adverse impact on the environment. Consequently, it is essential to implement sustainability measures that enhance transparency, not only in terms of model performance and accuracy but also regarding their carbon footprint, which reflects energy usage (e.g., through tools like the ML emissions calculator) [69]. To address this, Appendix B provides a comprehensive analysis of the resource consumption of each model during the training phase for 50 users. This evaluation includes metrics such as time consumed, energy consumed, and CO2 emissions. Resource usage was evaluated by computing the mean and STD across the three PHs for both datasets. The environmental impact was assessed using an NVIDIA GPU ADA 4000 for training. As shown in Figs. B.11 and B.12, LLM-based models exhibit significantly higher time and energy consumption, resulting in greater carbon emissions compared to the other models across both datasets. Most models require less than 15 min to train for 50 users, except for the LLM-based models, which take approximately 30 min. For the personalized approach, where all models are trained for each user, the total resource consumption is calculated as the sum of resources used by all models. Training for 50 users takes around 2 h. Despite the increased computational time, the environmental impact remains limited due to relatively low CO2 emissions and energy consumption [70]. Furthermore, the training time for the personalized approach per user is under 3 min. While a large-scale implementation may not be feasible with a single GPU, leveraging cloud services can significantly reduce computational time, making such implementations viable.

Our research highlights the utility of using LLM models to forecast glucose values and emphasizes the necessity of a personalized approach to enhance model performance. However, this methodology has three significant limitations. First, the experiments were conducted on a dataset with a limited sample size of older adults with T1D, predominantly from a white ethnic group. This limitation underscores the need for further evaluation and extension of the model to more diverse populations. To partially address this limitation, we have applied this methodology to the OhioT1DM dataset [71]. The personalized model achieved good performance, with an MAE of 14.2 $\pm$ 5.21, 18.3 $\pm$ 7.1, and 22.1 $\pm$ 7.6 for the 60-minute, 90-minute, and 120-minute prediction horizons, respectively. Additionally, the model's training demands substantial computational power and resources, which poses challenges for implementing this pipeline on a large scale. Specifically, the Time-LLM models require considerable time and resources to train, making it impractical to test all possible hyperparameter combinations for each user. Lastly, the model requires a minimum amount of data without more than one hour of consecutive missing values to forecast glucose values. If the required data is not provided, the model will not produce any forecasts.

Since we evaluated the performance of LLM-based models in this work it is worth highlighting the importance of prompting. Prompting refers to the process of providing input (named 'prompt') to an LLM to generate a desired output. Broadly speaking, it is how users interact with LLMs, guiding them to produce accurate, coherent, and contextually appropriate responses. As stated, Time-LLM transforms time-series data into a text representation before processing it through LLM-based models, and prompting can impact results. Prior studies on prompting have demonstrated its great impact on performance in different tasks [72,73]. However, identifying the adequate prompt is not straightforward because it often takes a significant amount of time for *word tuning*, and it is sensitive to slight changes, impacting the performance of LLMs [74]. By summarizing, prompting is a key component because: *(i)* LLMs interpret structural, contextual, and semantic variations in prompts differently; *(ii)* changes in the prompt format, wording, or tokenization can influence model outputs; and *(iii)* different LLM versions (e.g., GPT, GPT-4, Llama, Llama2) might introduce architectural changes or additional pre-training knowledge, modifying the model's response to the same input prompt. This is a current limitation of most LLM-based models, and it should be considered when these models are used for glucose forecasting.

Additionally, although other LLM-based models have been proposed for the prognosis of type 2 diabetes [75], forecasting patient health trajectories [76], and predicting the onset of type 2 diabetes [39], limited research has been performed for glucose. In contrast to other LLM-based models for time series forecasting such as LLM4TS, LLMAD, or TimeCMA that require fine-tuning [77], Time-LLM does not require fine-tuning to reach reasonable forecasting results. Thanks to the reprogramming layer in Time-LLM, which serves as an interface to transform time series data into a format that LLMs can effectively process, any LLM-based model (e.g., GPT, Llama) can be used for glucose forecasting, being highly adaptive in contexts and applications where the computational complexity of an LLM can be a restriction. In our study, we evaluate the feasibility of using LLMs without additional re-training, reducing computational time and reaching reasonable forecasting results. Additionally, our approach serves as the groundwork for future advancements in time series modeling using LLMs, and more particularly in healthcare applications. Since the text generated by Time-LLM is promising to be able to predict glucose, we can leverage these embeddings for other complex tasks. For instance, this is a first step toward leveraging LLM embeddings for glucose forecasting, with the long-term goal of integrating them into fusion models that combine domain-specific features extracted from Type 1 Diabetes (T1D) patients (e.g., insulin dosage, physical exercise) with embeddings representations learned by LLMs. In the previous author's paper [78], fusion methods were used to predict severe hypoglycemia using multimodal data, and a SAX-based model was used to obtain a vector representation of the time series of glucose. Due to reasonable results of Time-LLM, in future work, we can obtain embeddings of glucose, and then use these representations for more complex fusion methods. This allows us to have a more overall representation of a patient's metabolic state, thus improving glucose forecasting paving the way for more effective personalized healthcare applications, and supporting clinical-decision decision-making.

In future work, we plan to focus on three key areas. First, due to limited computational resources, we aim to test a broader range of hyperparameters in the LLM-based models to improve their performance. Additionally, we will explore more advanced GPT models (e.g., GPT-3, GPT-4) and different versions of LLAMA to evaluate their effectiveness as the backbone for Time-LLM. Second, we intend to incorporate additional information, such as demographic data (e.g., age and weight) and lifestyle information (e.g., carbohydrates consumed during the period or exercises performed), to further enhance forecasting accuracy. Lastly, we aim to extend this analysis to a larger group of individuals, including populations using different types of insulin pumps.

5. Conclusion

In this paper, we investigated the effectiveness, performance, and limitations of LLM-based models for predicting glucose time series in patients diagnosed with T1D. We used real-world CGM data collected by the T1DEXI study, including two cohorts: patients treated with MDIs and insulin pumps. We considered three different PHs (60, 90, and 120 min) and quantitatively evaluated the performance of eight AI-based models using MAE, RMSE, and the CEG plot. The best forecasting results for the MDI cohort were reached using the personalized approach in short-term PHs (60 and 90 min), with an MAE of 15.7 ± 12.5, 20.2 ± 14.2, whereas for long-term PH (120 min), TIDE obtained an MAE of 19.8 ± 12.2. Regarding the pump cohort, we obtained an average MAE of 15.2 ± 10.1 and 17.2 ± 11.0 using the personalized approach, and 18.5 ± 13.6 with Patch-TST in the PH=90 min. The CEG analysis showed that a large proportion of the model's forecastings fell within zones A and B for 60, 90, and 120 min. It is worth noting that LLM-based models provided reasonable forecasting results, with similar MAE and RMSE values compared to novel and sophisticated ANN-based models, and with the lowest STD in both metrics. To the best of our knowledge, this is one of the first studies that investigates LLM-based models for predicting glucose time series and distinguishing between individuals with different insulin therapies (MDIs and pumps). Our work contributes to the development of personalized models for glucose forecasting, supporting clinical decision-making and improving the control of glucose levels of individuals with T1D.

CRediT authorship contribution statement

Francisco J. Lara-Abelenda: Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **David Chushig-Muzo:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Pablo Peiro-Corbacho:** Writing – review & editing, Software. **Ana M. Wägner:** Writing – review & editing, Validation. **Conceição Granja:** Writing – review & editing, Validation, Project administration. **Cristina Soguero-Ruiz:** Writing – review & editing, Supervision, Project administration, Methodology.

Declaration of competing interest

The authors 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

This work was supported by the European Commission through the H2020-EU.3.1.4.2, European Project WARIFA (Watching the risk factors: Artificial intelligence and the prevention of chronic conditions) under Grant Agreement 101017385; and by the Spanish Government by the Grant AAVis-BMR PID2019-107768RA-I00/AEI/10.13039/50110 0011033. The study sponsors have not been involved in any stage of the study. This publication is based on research using data from Jaeb Center for Health Research Foundation that has been made available through Vivli, Inc. Vivli has not contributed to or approved, and is not in any way responsible for, the contents of this publication.

Appendix A. Representation of the RMSE using boxplots.

See (Figs. A.9 and A.10).

Appendix B. Comprehensive analysis of resource consumption

See (Figs. B.11 and B.12).

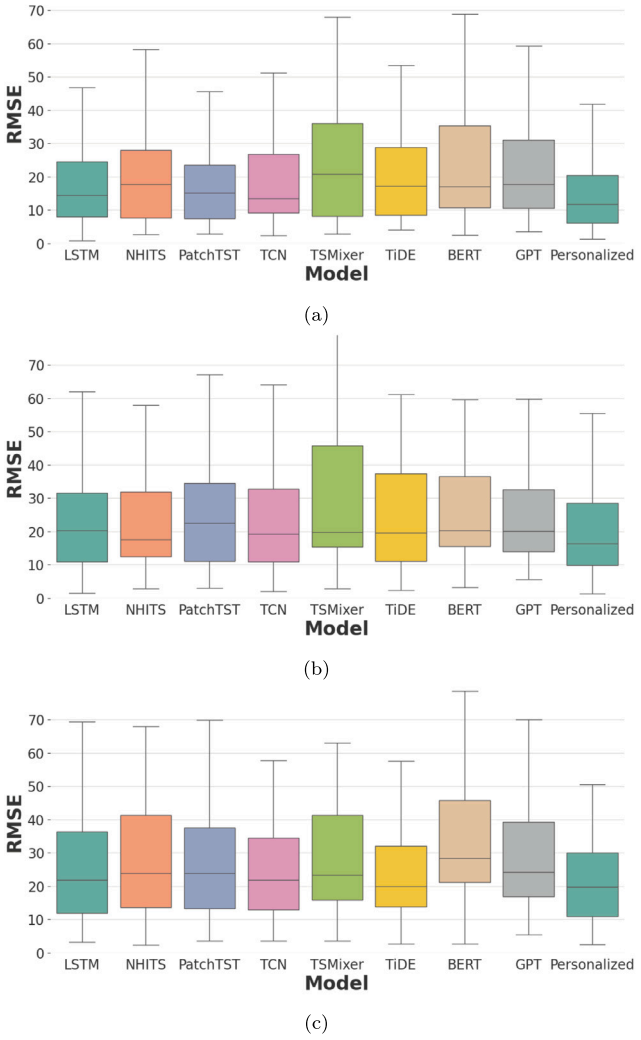

Fig. A.9. Distribution of the RMSE of the eight different models trained with CGM data from participants using MDI as the insulin administration method. Each boxplot was generated using a different PH: (a) 60 min; (b) 90 min, (c) 120 min.

Table C.6
Analysis of intra-patient and inter-patient variability among participants using MDI as their insulin administration method. The inter-patient variability is computed as the STD of the MAE using all users, while the intra-patient variability is computed as the mean of the different values of STD obtained by measuring the absolute error for each individual patient.

Model	PH = 60 minutes		PH = 90 minutes		PH = 120 minutes	
	Inter-Var.	Intra-Var.	Inter-Var.	Intra-Var.	Inter-Var.	Intra-Var.
LSTM	18.4	10.2	17.9	14.7	15.9	15.3
TCN	20.2	10.3	18.8	14.4	14.5	14.5
N-HiTS	18.5	10.2	17.7	13.8	14.8	14.7
TiDE	19.4	10.4	18.7	14.7	12.2	14.4
TSMixer	20.9	10.8	19.9	16.2	14.7	16.3
PatchTST	19.9	9.8	17.8	14.3	15.3	15.2
Time-LLM-GPT	15.1	11.6	14.1	13.0	14.1	13.9
Time-LLM-BERT	16.5	11.5	14.5	12.9	18.2	14.7
Personalized	12.5	9.2	14.2	12.7	14.0	13.5

Appendix C. Intra-patient and inter-patient variability analysis

See (Tables C.6 and C.7).

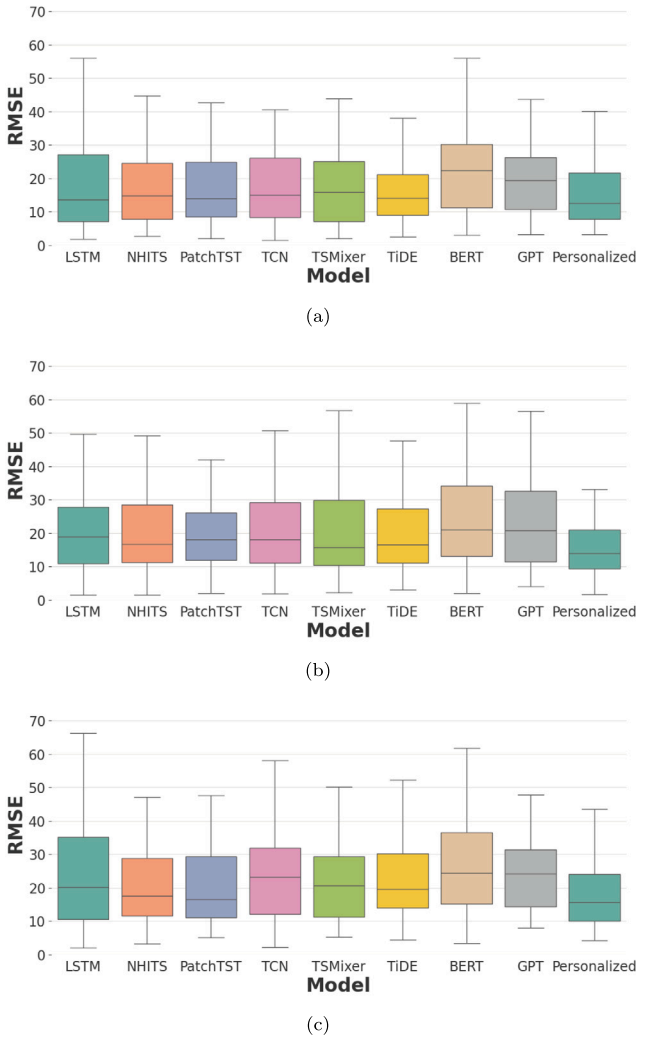

Fig. A.10. Distribution of the RMSE of the eight different models trained with CGM data from participants using pumps as the insulin administration method. Each boxplot was generated using a different PH: (a) 60 min; (b) 90 min, (c) 120 min.

Table C.7
Analysis of intra-patient and inter-patient variability among participants using insulin pumps as their insulin administration method. The inter-patient variability is computed as the STD of the MAE forecast for all users, while the intra-patient variability is computed as the mean of the different values of STD obtained by measuring the absolute error for each individual patient.

Model	PH = 60 minutes		PH = 90 minutes		PH = 120 minutes	
	Inter-Var.	Intra-Var.	Inter-Var.	Intra-Var.	Inter-Var.	Intra-Var.
LSTM	12.3	9.6	13.6	11.5	13.9	12.5
TCN	12.6	9.2	13.3	11.4	12.7	12.7
N-HiTS	11.5	9.1	12.4	10.6	13.6	11.7
TiDE	10.8	8.5	12.7	11.3	13.0	12.0
TSMixer	12.0	9.5	12.6	11.3	13.4	12.4
PatchTST	12.1	9.1	11.8	11.1	13.6	11.7
Time-LLM-GPT	9.6	9.5	11.7	11.2	10.2	13.0
Time-LLM-BERT	12.6	9.6	12.7	10.9	13.0	12.6
Personalized	10.1	9.5	11.0	11.0	11.5	11.4

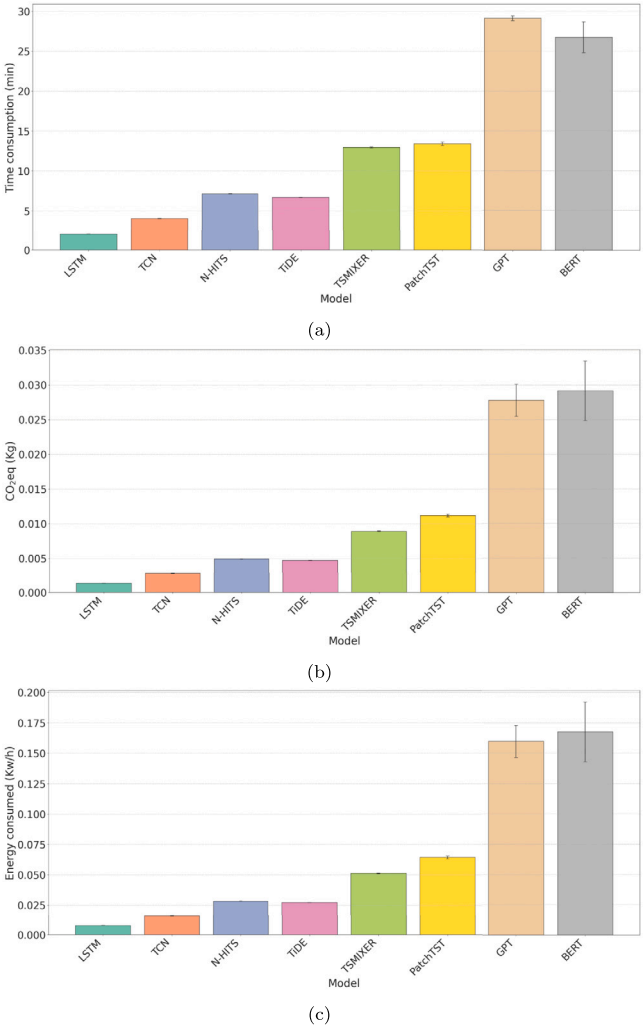

Fig. B.11. Comparison of resource consumption for each model across three prediction horizons, considering users employing multiple injection methods for insulin administration. The analysis evaluates (a) time consumption, (b) carbon emissions, and (c) energy consumption.

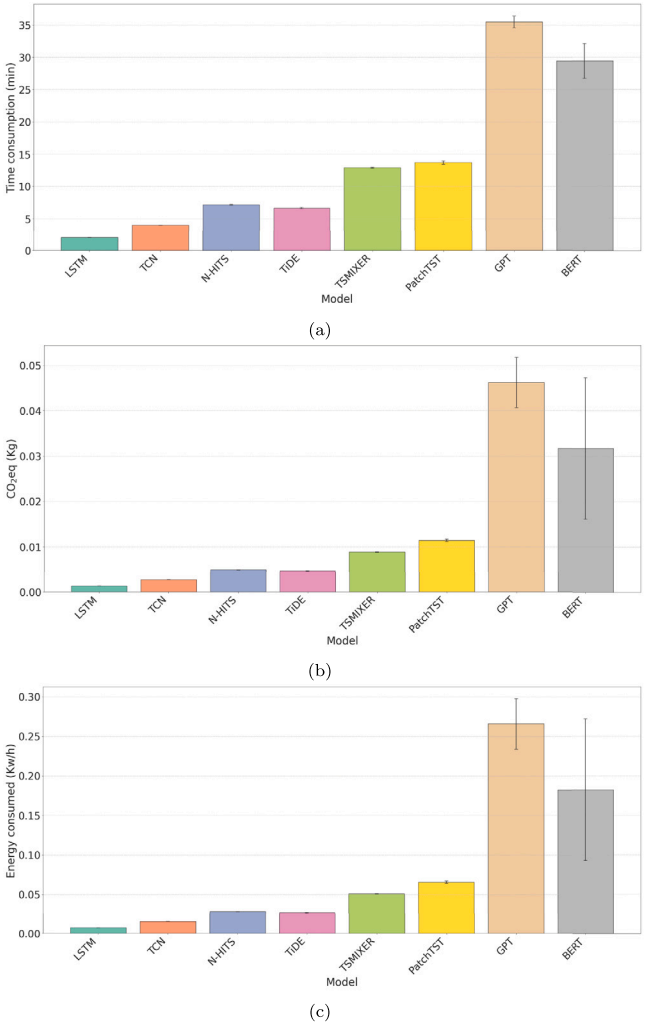

Fig. B.12. Comparison of resource consumption for each model across three prediction horizons, considering users with insulin pumps. The analysis evaluates (a) time consumption, (b) carbon emissions, and (c) energy consumption.

References

[1] P.G. Jacobs, N. Resalat, W. Hilts, G.M. Young, J. Leitschuh, J. Pinsonault, J. El Youssef, D. Branigan, V. Gabo, J. Eom, et al., Integrating metabolic expenditure information from wearable fitness sensors into an AI-augmented automated insulin delivery system: a randomised clinical trial, Lancet Digit. Heal. 5 (9) (2023) e607–e617.

[2] M. Vettoretti, A. Facchinetti, Combining continuous glucose monitoring and insulin pumps to automatically tune the basal insulin infusion in diabetes therapy: a review, Biomed. Eng. Online 18 (1) (2019) 37.

[3] A. Janež, C. Guja, A. Mitrakou, N. Lalic, T. Tankova, L. Czupryniak, A.G. Tabák, M. Prazny, E. Martinka, L. Smircic-Duvnjak, Insulin therapy in adults with type 1 diabetes mellitus: a narrative review, Diabetes Ther. 11 (2020) 387–409.

[4] R.W. Beck, R.M. Bergenstal, L.M. Laffel, J.C. Pickup, Advances in technology for management of type 1 diabetes, Lancet 394 (10205) (2019) 1265–1273.

[5] A.Z. Woldaregay, E. Årsand, S. Walderhaug, D. Albers, L. Mamykina, T. Botsis, G. Hartvigsen, Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes, Artif. Intell. Med. 98 (2019) 109–134.

[6] A. Della Cioppa, I. De Falco, T. Koutny, U. Scafuri, M. Ubl, E. Tarantino, Reducing high-risk glucose forecasting errors by evolving interpretable models for type 1 diabetes, Appl. Soft Comput. 134 (2023) 110012.

[7] K. Li, C. Liu, T. Zhu, P. Herrero, P. Georgiou, GluNet: A deep learning framework for accurate glucose forecasting, IEEE J. Biomed. Heal. Informat. 24 (2) (2019) 414–423.

[8] V. Felizardo, N.M. Garcia, N. Pombo, I. Megdiche, Data-based algorithms and models using diabetics real data for blood glucose and hypoglycaemia prediction–a systematic literature review, Artif. Intell. Med. 118 (2021) 102120.

[9] G. Yang, S. Liu, Y. Li, L. He, Short-term prediction method of blood glucose based on temporal multi-head attention mechanism for diabetic patients, Biomed. Signal Process. Control. 82 (2023) 104552.

[10] S.L. Cichosz, T. Kronborg, M.H. Jensen, O. Hejlesen, Penalty weighted glucose prediction models could lead to better clinically usage, Comput. Biol. Med. 138 (2021) 104865.

[11] S.L. Cichosz, M.H. Jensen, O. Hejlesen, Short-term prediction of future continuous glucose monitoring readings in type 1 diabetes: Development and validation of a neural network regression model, Int. J. Med. Informatics 151 (2021) 104472.

[12] Y.A. Andrade-Ambriz, S. Ledesma, M.-A. Ibarra-Manzano, M.I. Oros-Flores, D.-L. Almanza-Ojeda, Human activity recognition using temporal convolutional neural network architecture, Expert Syst. Appl. 191 (2022) 116287.

[13] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, L. Sun, Transformers in time series: A survey, in: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China, ijcai.org, 2023, pp. 6778–6786.

[14] Y. Nie, N.H. Nguyen, P. Sinthong, J. Kalagnanam, A time series is worth 64 words: Long-term forecasting with transformers, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023.

[15] V. Ekambaram, A. Jati, N. Nguyen, P. Sinthong, J. Kalagnanam, Tsmixer: Lightweight MLP-mixer model for multivariate time series forecasting, in: A.K. Singh, Y. Sun, L. Akoglu, D. Gunopulos, X. Yan, R. Kumar, F. Ozcan, J. Ye (Eds.), Proceedings of the 29th Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023, ACM, 2023, pp. 459–469.

[16] A. Das, W. Kong, A. Leach, S. Mathur, R. Sen, R. Yu, Long-term forecasting with tide: Time-series dense encoder, Trans. Mach. Learn. Res. (2023).

[17] C. Challu, K.G. Olivares, B.N. Oreshkin, F.G. Ramirez, M.M. Canseco, A. Dubrawski, Nhits: Neural hierarchical interpolation for time series forecasting, in: Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, Washington, DC, USA, ol. 37, No. 6, 2023, pp. 6989–6997.

[18] B.N. Oreshkin, D. Carpov, N. Chapados, Y. Bengio, N-BEATS: neural basis expansion analysis for interpretable time series forecasting, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020.

[19] P.H. Borghi, O. Zakordonets, J.P. Teixeira, A COVID-19 time series forecasting model based on MLP ANN, Procedia Comput. Sci. 181 (2021) 940–947.

[20] R. Sergazinov, E. Chun, V. Rogovchenko, N.J. Fernandes, N. Kasman, I. Gaynanova, GlucoBench: Curated list of continuous glucose monitoring datasets with prediction benchmarks, in: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024, OpenReview.net, 2024.

[21] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM Trans. Intell. Syst. Technol. 15 (3) (2024) 1–45.

[22] D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, M. Cornia, R. Cucchiara, The revolution of multimodal large language models: A survey, in: Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 13590–13618.

[23] M. Jin, S. Wang, L. Ma, Z. Chu, J.Y. Zhang, X. Shi, P. Chen, Y. Liang, Y. Li, S. Pan, Q. Wen, Time-LLM: Time series forecasting by reprogramming large language models, in: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024, 2024.

[24] A. Garza, M. Mergenthaler-Canseco, TimeGPT-1, 2023, arXiv preprint arXiv:2310.03589.

[25] Y. Bian, X. Ju, J. Li, Z. Xu, D. Cheng, Q. Xu, Multi-patch prediction: Adapting language models for time series representation learning, in: Forty-First International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024, 2024.

[26] K. Rasul, A. Ashok, A.R. Williams, A. Khorasani, G. Adamopoulos, R. Bhagwatkar, M. Biloš, H. Ghonia, N.V. Hassen, A. Schneider, et al., Lag-llama: Towards foundation models for time series forecasting, 2023, arXiv preprint arXiv:2310.08278.

[27] S. Rancati, P. Bosoni, R. Schiaffini, A. Deodati, P.A. Mongini, L. Sacchi, C. Toffanin, R. Bellazzi, Exploration of foundational models for blood glucose forecasting in type-1 diabetes pediatric patients, Diabetology 5 (6) (2024) 584–599.

[28] M.C. Riddell, Z. Li, R.L. Gal, P. Calhoun, P.G. Jacobs, M.A. Clements, C.K. Martin, F.J. Doyle III, S.R. Patton, J.R. Castle, et al., Examining the acute glycemic effects of different types of structured exercise sessions in type 1 diabetes in a real-world setting: the type 1 diabetes and exercise initiative (T1DEXI), Diabetes Care 46 (4) (2023) 704–713.

[29] N. Jendrike, A. Baumstark, U. Kamecke, C. Haug, G. Freckmann, ISO 15197: 2013 evaluation of a blood glucose monitoring system's measurement accuracy, J. Diabetes Sci. Technol. 11 (6) (2017) 1275–1276.

[30] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Vol. 1, No. 2, Association for Computational Linguistics, 2019.

[31] M. Schaefer, S. Reichl, R. ter Horst, A.M. Nicolas, T. Krausgruber, F. Piras, P. Stepper, C. Bock, M. Samwald, GPT-4 as a biomedical simulator, Comput. Biol. Med. 178 (2024) 108796.

[32] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, 2023, arXiv preprint arXiv:2302.13971.

[33] A.Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D.S. Chaplot, D.d.l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7B, 2023, arXiv preprint arXiv:2310.06825.

[34] M. Jin, H. Tang, C. Zhang, Q. Yu, C. Liu, S. Zhu, Y. Zhang, M. Du, Time series forecasting with LLMs: Understanding and enhancing model capabilities, 2024, arXiv preprint arXiv:2402.10835.

[35] C. Liu, Q. Xu, H. Miao, S. Yang, L. Zhang, C. Long, Z. Li, R. Zhao, Timecma: Towards LLM-empowered time series forecasting via cross-modality alignment, 2024, arXiv preprint arXiv:2406.01638.

[36] N. Gruver, M. Finzi, S. Qiu, A.G. Wilson, Large language models are zero-shot time series forecasters, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, la, USA, December 10 - 16, 2023, 2023.

[37] E. Healey, I. Kohane, LLM-CGM: A benchmark for large language model-enabled querying of continuous glucose monitoring data for conversational diabetes management, in: Biocomputing 2025: Proceedings of the Pacific Symposium, World Scientific, 2024, pp. 82–93.

[38] E. Healey, A. Tan, K. Flint, J. Ruiz, I. Kohane, Leveraging large language models to analyze continuous glucose monitoring data: A case study, 2024, MedRxiv.

[39] J.-E. Ding, P.N.M. Thao, W.-C. Peng, J.-Z. Wang, C.-C. Chug, M.-C. Hsieh, Y.-C. Tseng, L. Chen, D. Luo, C. Wu, et al., Large language multimodal models for new-onset type 2 diabetes prediction using five-year cohort electronic health records, Sci. Rep. 14 (1) (2024) 20774.

[40] Z. Chen, M. Ma, T. Li, H. Wang, C. Li, Long sequence time-series forecasting with deep learning: A survey, Inf. Fusion 97 (2023) 101819.

[41] R. Pascanu, T. Mikolov, Y. Bengio, On the difficulty of training recurrent neural networks, in: Proceedings of the 30th International Conference on Machine Learning,ICML 2013, Atlanta, GA, USA, 16-21 June 2013, Vol. 28, JMLR.org, 2013, pp. 1310–1318.

[42] M.F. Rabby, Y. Tu, M.I. Hossen, I. Lee, A.S. Maida, X. Hei, Stacked LSTM based deep recurrent neural network with kalman smoothing for blood glucose prediction, BMC Med. Inform. Decis. Mak. 21 (2021) 1–15.

[43] A.B. Amendolara, D. Sant, H.G. Rotstein, E. Fortune, LSTM-based recurrent neural network provides effective short term flu forecasting, BMC Public Health 23 (1) (2023) 1788.

[44] T.E. Idrissi, A. Idri, I. Abnane, Z. Bakkoury, Predicting blood glucose using an LSTM neural network, in: M. Ganzha, L.A. Maciaszek, M. Paprzycki (Eds.), Proceedings of the 2019 Federated Conference on Computer Science and Information Systems, FedCSIS 2019, Leipzig, Germany, September 1-4, 2019, in: Annals of Computer Science and Information Systems, vol. 18, 2019, pp. 35–41.

[45] M. Hermans, B. Schrauwen, Training and analysing deep recurrent neural networks, Adv. Neural Inf. Process. Syst. 26 (2013).

[46] L. Men, N. Ilk, X. Tang, Y. Liu, Multi-disease prediction using LSTM recurrent neural networks, Expert Syst. Appl. 177 (2021) 114905.

[47] T. El Idrissi, A. Idri, Deep learning for blood glucose prediction: CNN vs LSTM, in: O. Gervasi, B. Murgante, S. Misra, C. Garau, I. Blecic, D. Taniar, B.O. Apduhan, A.M.A.C. Rocha, E. Tarantino, C.M. Torre, Y. Karaca (Eds.), Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1–4, 2020, Vol. 12250, Springer, 2020, pp. 379–393.

[48] K. Li, J. Daniels, C. Liu, P. Herrero, P. Georgiou, Convolutional recurrent neural networks for glucose prediction, IEEE J. Biomed. Heal. Informat. 24 (2) (2019) 603–613.

[49] K.S. Kalyan, A. Rajasekharan, S. Sangeetha, AMMU: a survey of transformer-based biomedical pretrained language models, J. Biomed. Informat. 126 (2022) 103982.

[50] S. Karita, X. Wang, S. Watanabe, T. Yoshimura, W. Zhang, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N.E.Y. Soplin, R. Yamamoto, A comparative study on transformer vs RNN in speech applications, in: IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019, IEEE, 2019, pp. 449–456.

[51] S.-M. Lee, D.-Y. Kim, J. Woo, Glucose transformer: Forecasting glucose level and events of hyperglycemia and hypoglycemia, IEEE J. Biomed. Heal. Informatics 27 (3) (2023) 1600–1611.

[52] T. Zhu, L. Kuang, C. Piao, J. Zeng, K. Li, P. Georgiou, Population-specific glucose prediction in diabetes care with transformer-based deep learning on the edge, IEEE Trans. Biomed. Circuits Syst. (2024).

[53] T. Zhu, T. Chen, L. Kuangt, J. Zeng, K. Li, 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.

[54] T. Lemishko, A. Landi, A comparative analysis of LSTM versus PatchTST in predictive modeling of asset prices, 2024, Available At SSRN 4793111.

[55] T. Dai, B. Wu, P. Liu, N. Li, J. Bao, Y. Jiang, S. Xia, Periodicity decoupling framework for long-term series forecasting, in: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024, 2024.

[56] W. Terry, J. Lee, A. Kumar, Time series analysis in acid rain modeling: Evaluation of filling missing values by linear interpolation, Atmospheric Environ. (1967) 20 (10) (1986) 1941–1943.

[57] A. Blum, Freestyle libre glucose monitoring system, Clin. Diabetes 36 (2) (2018) 203–204.

[58] M. Intelligence, US continuous glucose monitoring market - growth, trends, COVID-19 impact, and forecasts (2023–2028), 2023, URL https://www.mordorintelligence.com/industry-reports/us-continuous-glucose-monitoring-market. (Accessed: 3 January 2025).

[59] T. Yang, X. Yu, N. Ma, R. Wu, H. Li, An autonomous channel deep learning framework for blood glucose prediction, Appl. Soft Comput. 120 (2022) 108636.

[60] M. Jaloli, W. Lipscomb, M. Cescon, Incorporating the effect of behavioral states in multi-step ahead deep learning based multivariate predictors for blood glucose forecasting in type 1 diabetes, BioMedInformatics 2 (4) (2022) 715–726.

[61] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J.E. Gonzalez, I. Stoica, Tune: A research platform for distributed model selection and training, 2018, arXiv preprint arXiv:1807.05118.

[62] B.E. Flores, The utilization of the wilcoxon test to compare forecasting methods: A note, Int. J. Forecast. 5 (4) (1989) 529–535.

[63] P. Cihan, Comparative performance analysis of deep learning, classical, and hybrid time series models in ecological footprint forecasting, Appl. Sci. 14 (4) (2024).

[64] A. Nath, D. Deb, R. Dey, An augmented subcutaneous type 1 diabetic patient modelling and design of adaptive glucose control, J. Process Control 86 (2020) 94–105.

[65] E.S. Scott, R.T. McGrath, A.S. Januszewski, D. Calandro, A.A. Hardikar, D.N. O'Neal, G. Fulcher, A.J. Jenkins, Hba1c variability in adults with type 1 diabetes on continuous subcutaneous insulin infusion (CSII) therapy compared to multiple daily injection (MDI) treatment, BMJ Open 9 (12) (2019) e033059.

[66] J.A. García-García, A. Reding-Bernal, J.C. López-Alvarenga, Cálculo del tamaño de la muestra en investigación en educación médica, InvestigaciÉn EducaciMédica 2 (8) (2013) 217–224.

[67] A. Deniz-Garcia, H. Fabelo, A.J. Rodriguez-Almeida, G. Zamora-Zamorano, M. Castro-Fernandez, M.d.P. Alberiche Ruano, T. Solvoll, C. Granja, T.R. Schopf, G.M. Callico, et al., Quality, usability, and effectiveness of mhealth apps and the role of artificial intelligence: current scenario and challenges, J. Med. Internet Res. 25 (2023) e44030.

[68] WARIFA Project, WARIFA: Personalized risk prediction for non-communicable diseases, 2024, URL https://www.warifa.eu/es/home-es/. (Accessed: 18 November 2024).

[69] V. Bolón-Canedo, L. Morán-Fernández, B. Cancela, A. Alonso-Betanzos, A review of green artificial intelligence: Towards a more sustainable future, Neurocomputing (2024) 128096.

[70] A.S. Luccioni, A. Hernandez-Garcia, Counting carbon: A survey of factors influencing the emissions of machine learning, 2023, arXiv preprint arXiv:2302.08476.

[71] C. Marling, R. Bunescu, The OhioT1DM dataset for blood glucose level prediction: Update 2020, in: CEUR Workshop Proceedings, Vol. 2675, NIH Public Access, 2020, p. 71.

[72] B. Zhang, B. Haddow, A. Birch, Prompting large language model for machine translation: A case study, in: International Conference on Machine Learning, PMLR, 2023, pp. 41092–41110.

[73] Y. Zhu, Z. Wang, J. Gao, Y. Tong, J. An, W. Liao, E.M. Harrison, L. Ma, C. Pan, Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data, 2024, arXiv preprint arXiv:2402.01713.

[74] K. Zhou, J. Yang, C.C. Loy, Z. Liu, Learning to prompt for vision-language models, Int. J. Comput. Vis. 130 (9) (2022) 2337–2348.

[75] P.B. Nguyen, M.P. Menden, R.W. Holl, A. Hungele, Leveraging pretrained large language model for prognosis of type 2 diabetes using longitudinal medical records, 2025, pp. 2002–2025, MedRxiv.

[76] N. Makarov, M. Bordukova, R. Rodriguez-Esteban, F. Schmich, M.P. Menden, Large language models forecast patient health trajectories enabling digital twins, 2024, pp. 2007–2024, MedRxiv.

[77] S. Abdullahi, K.U. Danyaro, A. Zakari, I.A. Aziz, N.A.W.A. Zawawi, S. Adamu, Time-series large language models: A systematic review of state-of-the-art, IEEE Access (2025).

[78] F.J. Lara-Abelenda, D. Chushig-Muzo, A.M. Wägner, M. Tayefi, C. Soguero-Ruiz, Interpretable and multimodal fusion methodology to predict severe hypoglycemia in adults with type 1 diabetes, Eng. Appl. Artif. Intell. 144 (2025) 110142.