Università degli Studi di Milano-Bicocca

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di Laurea Magistrale in Data Science

Advancements in Diabetes Severity Prediction: A Study of Deep Learning Personalized Approaches

Relatore: Prof. Paolo Napoletano

Co-relatore: Dr. Flavio Piccoli

Tesi di Laurea Magistrale di:

Niccolò Puccinelli

Matricola 881395

Anno Accademico 2022-2023

Contents

1

Introduction

A 2022 modelling study [1] projected that in 2021 approximately 8.4 million individuals globally were affected by type 1 diabetes (T1D). The substantial impact of T1D is set to expand significantly, particularly in regions with limited resources, indicating a rapid escalation. Already being a significant global health concern [2], projections anticipate a surge with a range of 13.5 to 17.4 million individuals living with T1D by 2040.

Type 1 diabetes is often referred to as juvenile diabetes mellitus [3], due to its typical start before the age of 25. This condition results from an insufficient production of insulin by the pancreas, a vital hormone responsible for managing blood sugar levels and aiding sugar utilization within the body. The deficiency in insulin production arises from the mistaking of the insulin-producing beta cells in the pancreas as foreign entities, triggering a destructive immune response against them.

Insulin helps transport glucose from the bloodstream into cells to be used for energy, and people with type 1 diabetes require regular injections or insulin pump therapy to survive, keeping blood glucose levels within a target range. Insulin injections (i.e., bolus insulin) involve the use of a syringe, insulin pen, or insulin vial and needle to manually inject insulin into the subcutaneous tissue. An insulin pump is a small, battery-operated device that continuously delivers basal insulin (a steady low dose) throughout the day, and allows users to administer bolus doses before meals or to correct high blood sugar levels. Insulin is delivered through a tiny, flexible tube inserted under the skin.

Type 1 diabetes is a lifelong condition with no known cure. Therefore, early identification and early treatment are crucial in managing the disease. Diagnosis encompasses evaluating both blood sugar and insulin levels, while the therapeutic approach is aimed at sustaining blood sugar within a near-normal range. If not properly treated, type 1 diabetes can precipitate to several complications, including skin problems, gastroparesis, diabetic nephropathy (chronic kidney disease [4]), diabetic retinopathy and neuropathy [5].

Recent advancements in wearable sensor technology [6], [7], together with Internet of Things (IoT) methodologies [8] and predictive algorithms [9], have facilitated the monitoring of patients' conditions. Specifically, the continuous and accurate prediction of Blood Glucose Concentrations (BGC) has emerged as a crucial element for effective disease management [10]. This real-time monitoring allows for prompt and preventive interventions, effectively mitigating the potential for life-threatening occurrences [11].

The advent of specialized novel technologies for monitoring and collecting patient data has opened doors for harnessing the power of Machine Learning (ML) and Deep Learning (DL) techniques to estimate BGC levels [12]. In particular, the ability of DL techniques to uncover complex patterns has made this type of model a suitable fit for this kind of prediction. Specifically, different architectures such as Convolutional Neural Networks (CNNs) [13] and Recurrent Neural Networks (RNNs) [14], [15] have been investigated, with the latter proving to be the best approach for this type of task. Nonetheless, the challenge of effectively training RNNs [16], due to the widely recognized vanishing gradient issue, has underscored the necessity for an architectural approach capable of mitigating this problem. Consequently, a significant portion of the research centered around RNNs [12] has been dedicated to harnessing the potential of Long Short-Term Memory (LSTM) units [17].

The prediction horizon (PH) is another key factor in forecasting BGC levels. To mitigate health risks and prevent potential hyperglycemic episodes, users must anticipate when to administer insulin boluses by proactively assessing future needs. Naturally, a longer prediction horizon offers greater user utility, but this comes at the cost of reduced model performance over time. Existing research [12] primarily focused on predicting BGC levels within two specific PHs, namely 30 and 60 minutes. However, Shuvo et al. [18] have expanded this scope to encompass prediction horizons of 90 and 120 minutes. For a visual representation of the general pipeline approach, please refer to Figure 1.1 below.

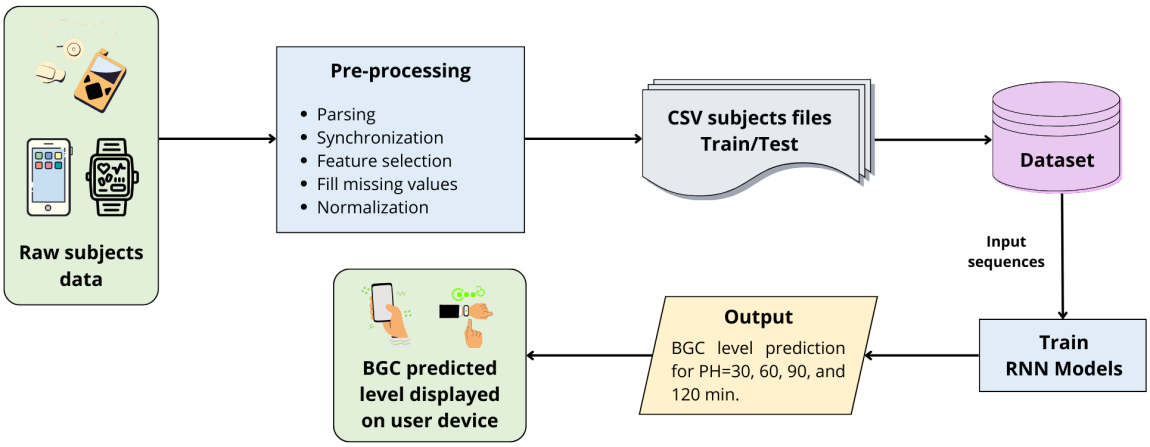


Fig. 1.1: General pipeline for BGC level prediction. Starting with raw subjects data originating from user devices, the selected pre-processed features are input into RNN models for training and subsequent prediction of BGC across different PHs. Finally, users can conveniently access these predictions on their devices.

In the realm of glucose level prediction, the (few) existing datasets exhibit notable shortcomings. They are characterized by incompleteness, heterogeneity, and biases, all of which collectively limit the effectiveness of predictive models and their ability to encompass a wide range of scenarios [19]. The OhioT1DM dataset represents the actual foremost advancement in the field [20]. It contains eight weeks of glucose measurements coupled with a diverse array of supplementary features, all attributed to 12 subjects with heterogeneous characteristics. Given its widespread adoption, the OhioT1DM dataset (refer to Chapter 3 for comprehensive insights) has been selected for this master's thesis work. Nevertheless, the dataset presents a series of challenges, including the substantial number of missing values. Thus, a multitude of data processing techniques were employed. The objective was to systematically evaluate their effectiveness and subsequently select the final dataset by adopting the methodology that yielded the most optimal performance.

The inherent challenge encountered by tasks of this nature revolves around the ability to generalize. Indeed, every user possesses unique characteristics and habits that are difficult to model without a more personalized approach. Therefore, recent advances in the field suggest that tailored approaches (such as

incremental learning [21]) significantly increase accuracy on each patient. Relative to the subjects of the OhioT1DM dataset, Shuvo et al. [18] propose a customization strategy to refine the deep-learning models based on gender and then individual subjects, achieving state-of-the-art performance.

The main contributions of this master's thesis work are therefore aimed at addressing these challenges, focusing on the following key aspects:

- The effects of different pre-processing strategies on prediction performance were evaluated.
- A comparative analysis was conducted between two customization strategies and a general strategy with no customization at all.
- A novel customization strategy, called *Threetask*, was conceived. This strategy refines customization by anchoring it to the median BGC level of the patient. Despite employing a less granular approach than the established reference model, this new architecture outperforms previous approaches.
- The effectiveness of an incremental learning procedure in enhancing model performance was assessed.
- The integration of a supplementary dataset was evaluated to measure its impact on the overall predictive capabilities.
- Each technique has been evaluated on four different prediction horizons: 30-min, 60-min, 90-min and 120-min.
- The diverse approaches and methodologies were analyzed across both regression and classification tasks to comprehend their influence and effectiveness.

This work is structured as follows: Chapter 2 presents an exhaustive review of prior literature pertinent to the subject, to the best extent of our knowledge. Chapter 3 provides a detailed description of the dataset used, while Chapter 4 systematically assesses the diverse pre-processing strategies employed. Chapter 5 and Chapter 6 present the various analyses conducted on regression and classification tasks, respectively, with a focus on the remarkable efficacy of the *Threetask* model. Finally, Chapter 7 summarizes and concludes the research.

2

Related Work

In recent years, much research has been done in the field of blood glucose prediction through data-driven strategies, with an increasing impact of DL models trained on clinical datasets, prominently featuring the OhioT1DM dataset.

Initial research efforts have primarily revolved around the utilization of Autoregressive Moving Average (ARIMA) [22] and Unobserved Components Model (UCM) [23] models for the purpose of modeling and predicting time series data. ARIMA models are constructed by combining three fundamental components. Firstly, the AutoRegression (AR) component scrutinizes the connection between the present data point in the time series and its historical values. Secondly, the Integration (I) component delineates the number of differentiation operations required to render the data stationary. Stationarity, in this context, denotes a state where essential statistical properties like mean and variance remain constant over time. Lastly, the Moving Average (MA) element explores the association between the current data point and past prediction errors or residuals. In contrast, the Unobserved Components Model (UCM) is a more adaptable and intricate framework that decomposes a time series into numerous unobserved components. These components encompass the level, which characterizes the underlying long-term trend within the time series, seasonal, responsible for capturing recurring patterns or seasonality in the data, and cycle, which accommodates longer-term cyclical patterns that may not be strictly seasonal.

Notably, for this type of task, there has been a discernible shift from ARIMA [24] and UCM models [25] to more advanced techniques such as Random Forest and XGBoost [26], and even neural networks, which have demonstrated their effectiveness surpassing traditional ML approaches [18].

For example, Freiburghaus et al. [27] proposed a hybrid Convolutional Recurrent Neural Networks (CRNN) architecture. Within this framework, the input time series undergoes a process of feature selection via pooling layers that keep only the highest value in the window (i.e., max pooling). These features are then fed into an LSTM layer to capture temporal dependencies, and a fully connected layer is used to obtain the final regression outcome. This approach yielded remarkable performance in terms of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), with RMSE = 17.45 mg/dL, MAE = 11.22 mg/dL for 30-min PH and RMSE = 33.67 mg/dL, MAE = 23.25 mg/dL for 60-min PH.

Another notable study conducted by Butt et al. [28] presented a method for converting self-reported features (carbohydrates and insulin bolus) into continuous variables. This innovative approach was paired with a general stacked Bi-LSTM model, achieving RMSE = 16.81 mg/dL for 30-min PH and RMSE = 28.25 mg/dL for 60-min PH.

Interestingly, Beauchamp et al. [29] proposed a different approach to carbohydrate and bolus recommendations within the context of type 1 diabetes management. The neural architecture, employed for both recommendations, combines LSTM and deep residual networks (ResNets) [30], trained using a combination of Mean Squared Error (MSE) and MAE loss functions.

Zhu et al. [31] explored Generative Adversarial Networks (GANs) [32] for predicting BGC levels. Their approach involved training the GAN on the OhioT1DM dataset, selecting blood glucose measurements, meal consumption records, and the corresponding insulin doses. In this setup, the generator utilizes an RNN to compute BGC predictions, while the auxiliary discriminator employs a one-dimensional CNN to distinguish

between the predictive and real BGC values. The proposed algorithm achieved an average RMSE of 18.34 mg/dL, with a MAE of 13.37 mg/dL for the 30-min PH, and an average RMSE of 32.31 mg/dL, with a MAE of 24.20 mg/dL for the 60-min PH.

In [33], the authors executed a two-step process for model training. Initially, they conducted a pre-training phase using data from the *Tidepool* [34] organization and the six subjects from the 2018 OhioT1DM dataset. Subsequently, the model undergoes fine-tuning using data from subjects in 2020. Their neural network architecture is designed with network blocks that have a unique capability: they produce both a forecast and a backcast, which essentially reconstructs the input of the block. The backcast is then subtracted from the block's input, creating a residual signal that is subsequently fed as input to the next block in the architecture. This process continues through the network blocks. Finally, at the last layer, the forecasts generated by each block are combined to produce the ultimate prediction.

Nemat et al. [35] introduced an alternative approach, offering two distinct methods, both founded on stacked regression and data fusion of BGC and activity data, to forecast blood glucose levels in patients with type 1 diabetes. In the first method, the authors leveraged histories of BGC data, appended with the average activity data within the same timelines, to train three base regression models: a Multilayer Perceptron (MLP), an LSTM network, and a Partial Least Squares Regression (PLSR) method, which condenses predictors into a smaller set for regression purposes. In contrast, in the second method the histories of BGC and activity data were trained independently with the same base regression models. In both methods, the predictions produced by the base regression models serve as features for constructing a unified model. This combined model is then employed to make the final predictions, achieving an average RMSE of 19.09 mg/dL, with a MAE of 13.77 mg/dL for PH = 30min, and an average RMSE of 33.55 mg/dL, with a MAE of 25.11 mg/dL for PH = 60min.

Finally, Joedicke et al. [36] harnessed the power of four genetic programming variants to build white-box models for predicting BGC levels. The findings from this research, compared with classical methods including multi-variate linear regression, random forests, and ARIMA models, demonstrated the competitive edge of genetic programming in terms of MAE and RMSE across prediction horizons spanning 30 and 60 minutes.

Despite the good performance, these studies exhibit practical limitations for two main reasons. Firstly, the training and testing phases were confined to subsets of subjects. For instance, in the case of Butt et al. [28], only three subjects (570, 588 and 563) who consistently documented carbohydrate intake were considered. Secondly, the real-world applicability of these approaches is limited by the substantial training data necessary to fine-tune models for the subject under consideration. Even when employing the approach pursued by Freiburghaus et al. [27], involving pre-training on a cohort of six subjects, followed by subsequent transfer learning and fine-tuning on the remaining six individuals, the cumulative training dataset still encompasses measurement data equivalent to over a month's worth of data. Thus, with this methodology, half of the subjects are not involved in model testing.

In order to enhance the applicability of the models to real-world scenarios, the researchers' focus began to turn towards personalization. By accounting for individual variations and adapting to changing circumstances,

personalized models can provide more accurate and tailored recommendations for insulin dosages, dietary choices, and lifestyle modifications, ultimately improving the patient's quality of life and health outcomes. In this context, the papers by Daniels et al. [13] and Shuvo et al. [18] presented a similar approach, based on the multiple branching of the network, involving the gender of the subjects, and then tailoring to each individual characteristics. Daniels et al. evaluated the performance of their CRNN model on a subset of six subjects, while Shuvo et al. employed a stacked LSTM architecture encompassing all of the 12 subjects available in the dataset. The latter work achieved currently state-of-the-art results (MAE = 10.64 mg/dL, RMSE = 16.06 mg/dL for 30-min PH, MAE = 22.07 mg/dL, RMSE = 30.89 mg/dL for 60-min PH, MAE = 30.16 mg/dL, RMSE = 40.51 mg/dL for 90-min PH, MAE = 36.36 mg/dL, RMSE = 47.39 mg/dL for 120-min PH).

Nevertheless, the process of customization still necessitates the accumulation of several weeks' worth of data to fine-tune the model for a specific individual. Furthermore, there is a relevant gap in the literature involving the comparison between general and custom architectures. Indeed, the true impact of personalization can only be comprehensively assessed through a comparative analysis of the two paradigms: personalized and generalized. Lastly, it is important to note that the personalized approach developed by Daniels et al. and Shuvo et al. is linked to user-specific attributes, such as the gender, which might potentially hold limited relevance in the context of glucose prediction.

Therefore, the primary objective of this research centers on conducting a thorough assessment of the diverse forms of personalization alongside a generalized approach. Nonetheless, it is crucial to emphasize that the outcomes will be presented based on evaluations performed using data from entirely new subjects (i.e., without pre-training). This deliberate methodology aims to simulate a pragmatic scenario of immediate applicability.

3

The OhioT1DM dataset

This study leverages the OhioT1DM clinical dataset [20], initially introduced by Ohio University in 2018 for the first Blood Glucose Level Prediction (BGLP) Challenge and updated in 2020 for the second BGLP Challenge. To protect the data contributors and to ensure that the data are used only for research purposes, a Data Use Agreement (DUA) is required. Since its second release in 2020, which added six more subjects to the cohort, this dataset has gained significant popularity as a cornerstone for BGC prediction, as highlighted by its inclusion into numerous research papers [12]. Indeed, it provides a comprehensive collection of attributes associated with type 1 diabetes patients, which, as outlined in the latest diabetes literature [37], significantly influence blood glucose fluctuations.

The OhioT1DM dataset comprises a total of 12 subjects, six of which were recorded in 2018 and six in 2020. Each subject's folder contains eight weeks' worth of Continuous Glucose Monitoring (CGM) of BGC, with readings captured at five-minute intervals with Medtronic Enlite CGM sensors. There are two distinct insulin pump models in use within the study cohort: the 530G model, which is utilized by nine subjects, and the 630G model, adopted by three subjects. Moreover, the dataset contains optional finger stick glucose (FS) measurements, directly reported by the patients themselves. Besides, it includes insulin, physiological sensor and life-event data self-reported via a smartphone application.

To ensure privacy, subjects were de-identified according to the Safe Harbor method, a standard specified by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Therefore, subjects were assigned random patient identification (PID) codes. Moreover, all dates for each individual were shifted by the same random amount of time into the future. This adjustment preserved the original days of the week and times of day, ensuring continuity in the new timeframe. However, the months were shifted, deliberately eliminating the potential influences of seasonality or holidays on the data analysis.

Notably, the subjects from 2018 and 2020 exhibit a primary distinction in the sensor bands employed: Basis Peak for 2018 and Empatica Embrace for 2020. These two types of sensors capture physiological information with variations in granularity and nature of measured observations. Moreover, both bands indicate the times they detected that the wearer was asleep, and this information is included when available.

The subjects are heterogeneous and differ in terms of age, gender, and disease severity, significantly enhancing the model's capacity for generalization across a wider spectrum of individuals. Each data contributor has a pair of XML files: one designated for training and development data, and another dedicated to testing data. This configuration amounts to a grand total of 24 XML files, with two files assigned to each of the 12 contributors.

The dataset is provided already divided into training and testing subsets, maintaining an approximate ratio of 75% for training and 25% for testing data. Table 3.1 provides a clear summary of the dataset.

Table 3.1: Details of the OhioT1DM dataset: year cohort, gender, age range, PID sensor band type, insulin pump model, and number of training and test samples for each data contributor.

Year	Gender	Age	PID	Sensor band	Pump model	Train samples	Test samples
2018	Female	40-60	559	Basis	530G	10796	2514
2018	Male	40-60	563	Basis	530G	12124	2570
2018	Male	40-60	570	Basis	530G	10982	2745
2018	Female	40-60	575	Basis	530G	11866	2590
2018	Female	40-60	588	Basis	530G	12640	2791
2018	Female	40-60	591	Basis	530G	10847	2760
2020	Male	40-60	540	Empatica	630G	11947	2884
2020	Male	40-60	544	Empatica	530G	10623	2704
2020	Male	20-40	552	Empatica	630G	9080	2352
2020	Female	20-40	567	Empatica	630G	10858	2377
2020	Male	40-60	584	Empatica	530G	12150	2653
2020	Male	60-80	596	Empatica	530G	10877	2731

The dataset provides a total of 20 features for each subject in XML format:

1. **patient**: Patient ID number.
2. **glucose_level**: Continuous glucose monitoring measurements, recorded every five minutes.
3. **finger_stick**: Blood glucose concentration values obtained by patient self-monitoring.
4. **basal**: Rate at which basal insulin is continuously infused.
5. **temp_basal**: Temporary basal insulin rate that supersedes the patient's normal basal rate. When the value is 0, this indicates that the basal insulin flow has been suspended. At the end of a temp basal, the basal rate goes back to the normal basal rate.
6. **bolus**: Insulin delivered to the patient, typically before a meal or when the patient is hyperglycemic.
7. **meal**: Self-reported time and type of a meal, plus the patient's carbohydrate estimate for the meal.
8. **sleep**: Times of self-reported sleep, plus the patient's subjective assessment of sleep quality.
9. **work**: Self-reported times of going to and from work. The intensity value is the patient's subjective assessment of physical exertion.
10. **stressors**: Time of self-reported stress event.
11. **hypo_event**: Time of self-reported hypoglycemic episode.
12. **illness**: Time of self-reported illness.
13. **exercise**: Time and duration, in minutes, of self-reported exercise. The intensity value is the patient's subjective assessment of physical exertion.
14. **basis_heart_rate**: Heart rate, aggregated every five minutes. This data is only available for people who wore the Basis Peak sensor band.

15. **basis_gsr**: Galvanic skin response, aggregated every five minutes for those who wore the Basis Peak sensor, and every minute for those who wore the Empatica Embrace device.
16. **basis_skin_temperature**: Skin temperature (°F), aggregated every five minutes for those who wore the Basis Peak sensor, and every minute for those who wore the Empatica Embrace device.
17. **basis_air_temperature**: Air temperature (°F), aggregated every five minutes. This data is only available for people who wore the Basis Peak sensor band.
18. **basis_steps**: Step count, aggregated every five minutes. This data is only available for people who wore the Basis Peak sensor band.
19. **basis_sleep**: Times when the sensor band reported that the subject was asleep. For those who wore the Basis Peak, there is also a numeric estimate of sleep quality.
20. **acceleration**: Magnitude of acceleration, aggregated every minute. This data is only available for people who wore the Empatica Embrace sensor band.

Despite the availability of numerous features, current advancements, as shown in Figure 3.1, and including state-of-the-art research [18], have predominantly focused on the use of the following four primary features:

- **glucose_level** (BGC).
- **finger_stick** (FS).
- **bolus** (BI).
- **meal** (C).

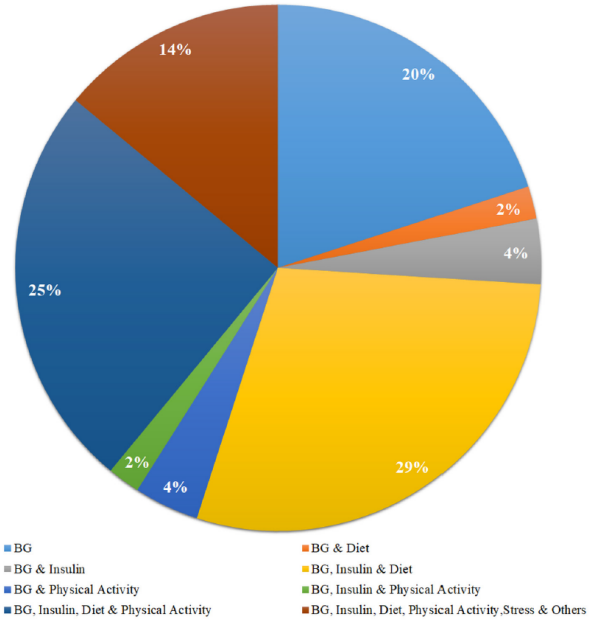


Fig. 3.1: Input feature set contribution in recent studies for diabetes prediction. Image from Butt et al. [28].

To ensure consistency and facilitate the assessment and comparison of methodologies, we have also chosen to employ these identical four features.

The details of the features are summarized in Table 3.2.

Table 3.2: Feature details for the OhioT1DM dataset: name, frequency type, recording source, and year cohort including each feature.

Feature	**Frequency**	**Recording source**	**Subjects cohort**
glucose_level	Periodic	Medtronic	2018 + 2020
finger_stick	Event	Self-reported	2018 + 2020
basal	Event	Self-reported	2018 + 2020
temp_basal	Event	Self-reported	2018 + 2020
bolus	Event	Self-reported	2018 + 2020
meal	Event	Self-reported	2018 + 2020
sleep	Event	Self-reported	2018 + 2020
work	Event	Self-reported	2018 + 2020
stressors	Event	Self-reported	2018 + 2020
hypo_event	Event	Self-reported	2018 + 2020
illness	Event	Self-reported	2018 + 2020
exercise	Event	Self-reported	2018 + 2020
basis_heart_rate	Periodic	Basis	2018
basis_gsr	Periodic	Basis + Empatica	2018 + 2020
basis_skin_temperature	Periodic	Empatica	2020
basis_air_temperature	Periodic	Basis	2018
basis_steps	Periodic	Basis	2018
basis_sleep	Periodic	Basis + Empatica	2018 + 2020
acceleration	Periodic	Empatica	2020

Additionally, the dataset comes with a user-friendly visualization tool known as OhioT1DM Viewer, released to participants in the second BGLP Challenge. This tool facilitates the debugging and the exploration of subject-specific data by opening XML files associated with each subject, providing a comprehensive display of their daily integrated data. For example, if a system produces an inaccurate blood glucose level prediction at a specific moment in time, a closer examination of the data corresponding to that moment could reveal the underlying cause. For instance, the subject might have forgotten to report a meal or might have been feeling ill or stressed. Figure 3.2 provides a screenshot representing an illustrative feature overview from subject 559.

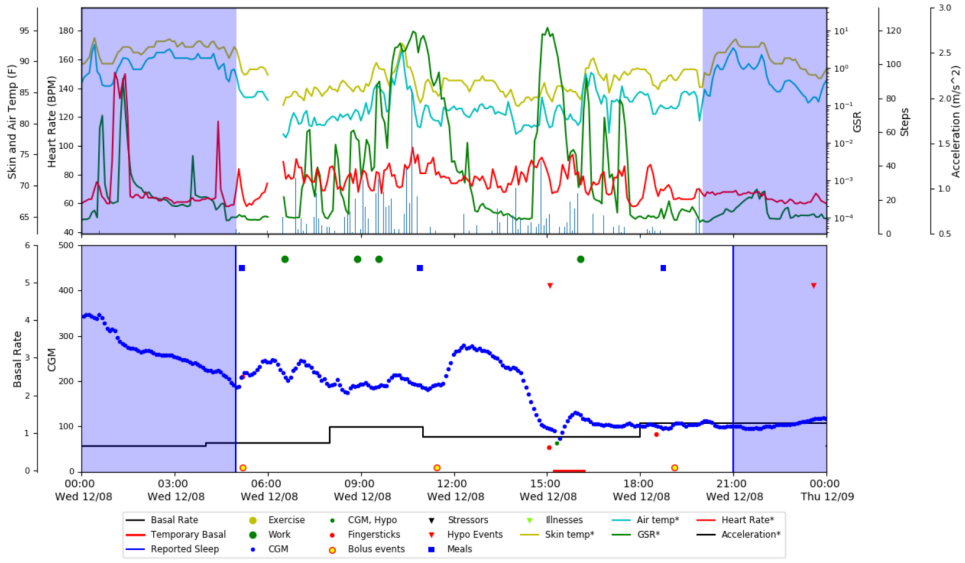


Fig. 3.2: OhioT1DM Viewer - feature visualization for patient 559. The two graphs represent the time series data comprising self-reported features, continuously monitored BGC, and data from wearable devices. The shaded blue regions within the graphs denote the user's sleep periods.

3.1 Parsing & synchronizing

First of all, data parsing and temporal synchronization tasks were performed. The raw data are initially presented in XML format, with two files for each subject, representing the training and the test folds, respectively. Therefore, these 24 files have been converted into the more manageable CSV format. Concurrently, the extracted features have been aligned with the primary column containing the BGC readings. The BGC observations were recorded at five-minute intervals, except in cases of missing data points.

A critical aspect of the synchronization process was to ensure that each recorded BGC timestamp was harmonized with subsequent observations occurring within a maximum time window of four minutes. By adopting to this approach, we realistically accounted for the fact that information pertaining to the remaining features was unavailable before the detection of blood glucose levels.

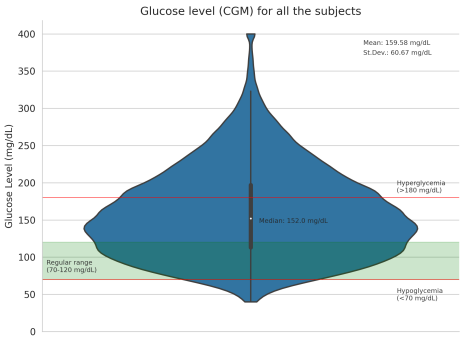
This dataset, parsed and synchronized without further processing techniques, is referred to as *Preprocessed* and constitutes the foundation for the subsequent exploration phase and advanced research.

3.2 Data exploration

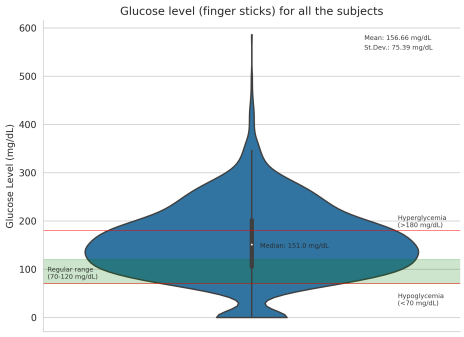
To attain a holistic understanding of the data and the task itself, we conducted an extensive phase of data exploration. The main focus of the exploration included the distributions of diverse observations and the identification of missing values.

First, violin plots were created regarding the general distribution of each column, without taking into account missing values. Below are the plots for the four features that were used (Figure 3.3), namely *glucose_level*, *finger_stick*, *bolus* and *meal*.

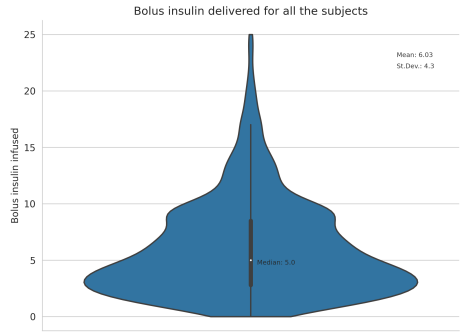
3.2.1 Distributions of the variables



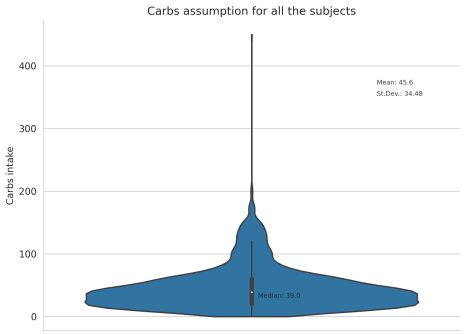
(a) Blood glucose concentration levels - the green range indicates the reasonable range for people without diabetes, the red lines represent the danger thresholds.



(b) Finger stick measurements - the green range indicates the reasonable range for people without diabetes, the red lines represent the danger thresholds.



(c) Bolus insulin delivered.



(d) Carbs assumption.

Fig. 3.3: Variables distribution for all the subjects - violin plot with mean, median and standard deviation.

The depicted plots in Figure 3.3a and Figure 3.3b reveal a similar scenario concerning blood glucose levels, albeit with a notable distinction in the finger stick graph. In this chart, a greater number of observations

cause a noticeable upward extension of the violin plot. This could be due to the presence of outliers or to the inclusion of patients who may have inconsistently recorded their BGC levels via finger stick measurements. The majority of measurements surpass the established normal range for individuals without diabetes, underscoring a significant departure from typical values. Specifically, the median value is around 150 mg/dL, accompanied by a slightly elevated mean. These values accentuates the condition of patients within the OhioT1DM dataset, highlighting a substantial prevalence of observations that exceed the critical threshold of 180 mg/dL. These thresholds and ranges are referenced from [38] and visually represented in Figure 3.4 for further clarity.

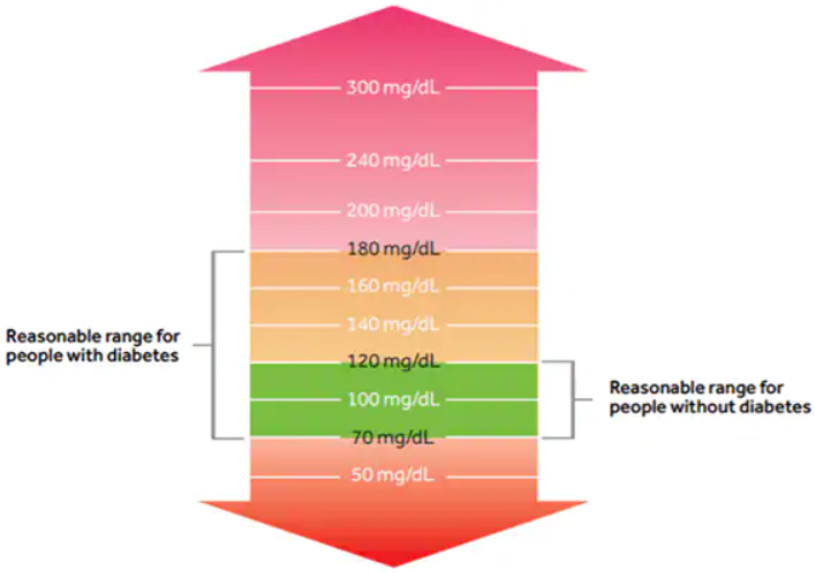


Fig. 3.4: Glucose levels with ranges for people with and without diabetes. Image from Medtronic [38].

Regarding bolus insulin values, the central tendency is represented by a median of 5 units. However, it's noteworthy that there are many relatively high values (probably due to injections in critical situations), with the effect of slightly elevating the mean to approximately 6.02 units.

In the case of carbohydrate values, a notable shift of the mean value, from the baseline median of 37.0 units to around 44.78 units, can be noticed. This change could be attributed to the influence of some irregular meal, giving rise to higher readings. However, it's important to acknowledge that the meticulousness with which these data were recorded varied among subjects and the number itself is actually an estimate of carbohydrate intake, resulting in a certain degree of inconsistency and bias.

3.2.2 Missing values

A significant aspect of the analysis presented in this research, as well as in the broader body of literature [12], centers around the examination and potential resolution of missing data points. Self-reported attributes, such as FS, BI, and C, inevitably introduce this challenge, while missing BGC values can be attributed to various factors, including power interruptions, device changes in CGM systems, and connectivity issues.

Figure 3.5 provides a comprehensive heatmap summarizing the distribution of missing values across all columns and subjects in the dataset, excluding BGC, for which a separate analysis was conducted.

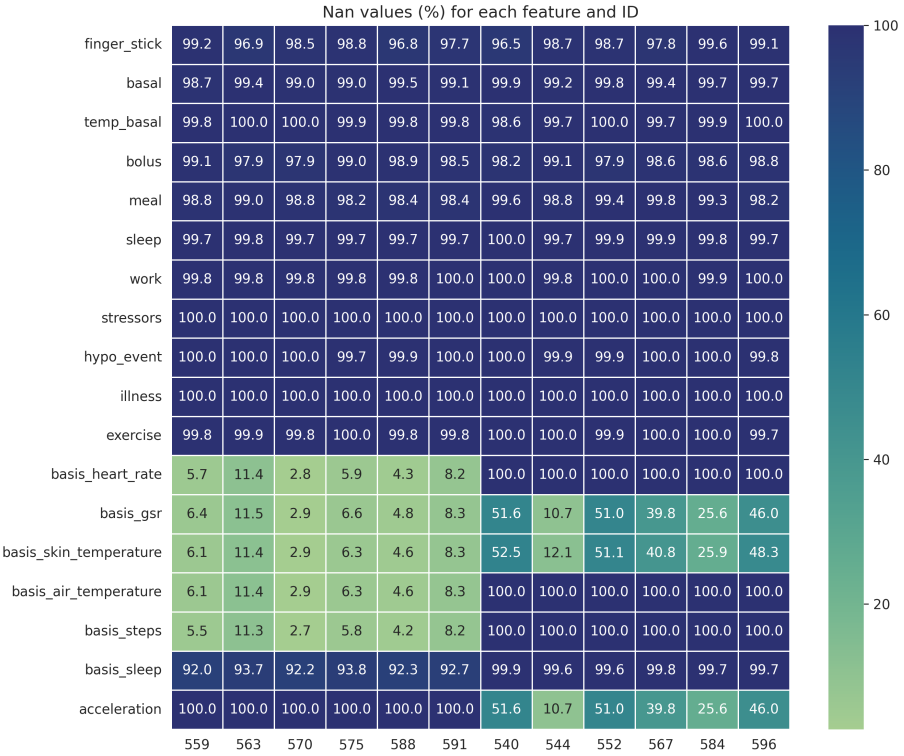


Fig. 3.5: Missing values distribution heatmap (%) across subjects and features, excluding BGC.

The dataset contains a wide array of self-reported features, from those related to finger stick measurements to exercise efforts. It is important to note that due to the nature of self-reporting, it is to be expected that the number of missing values would be so high. However, it is crucial to highlight that among these self-reported features, at least three are notably underrepresented and nearly unusable. Specifically, the variables *stressors*, *hypo_events*, and *illness* are reported so infrequently within the dataset that their inclusion in larger-scale analyses may not yield reliable results (refer to Figure 3.6 for a visual representation of this issue).

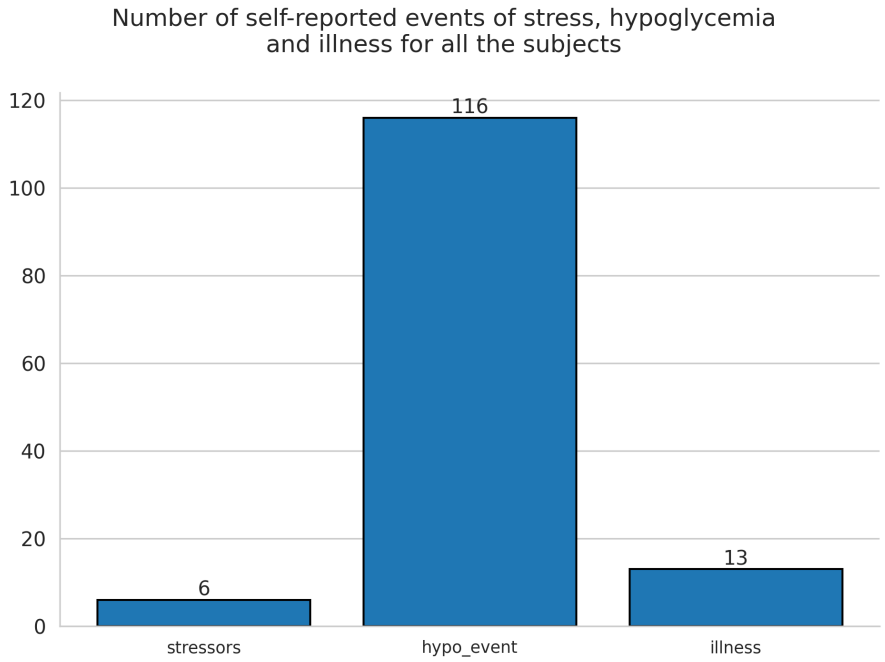


Fig. 3.6: Occurrences of the features *stressors*, *hypo_event* and *illness* across all subjects.

The physiological features, collected with the Basis Peak and Empatica Embrace devices, also have several missing values. The primary reason for these gaps is user-initiated device removal, especially during sleep (as indicated by the *basis_sleep* variable). Furthermore, it's worth highlighting specific limitations related to each device. The Basis Peak device lacks observations regarding *acceleration*, while the Empatica Embrace device was unable to detect *basis_heart_rate*, *basis_air_temperature*, and *basis_steps*. The latter device, used by subjects in the 2020 court study, exhibits, proportionally, a higher number of missing values, nearing 50% for subjects 540, 552, and 596. The combination of a high percentage of missing values and the inherent heterogeneity of the devices used in data collection raises concerns about the practical utility of these features. Furthermore, in a real-world application, continuous device wear by the subjects would be necessary, which is often inconvenient and impractical. These factors collectively suggest that caution should be exercised when attempting to derive meaningful insights or conclusions from these particular features in the dataset.

Values related to blood glucose monitoring require a separate discussion. Reconstructing the data into five-minute time intervals, we uncovered a total of 189,535 observations, with 23,002 of them missing, representing approximately 12.14% of the overall dataset. However, it's essential to note that the distribution of these missing values is not uniform across subjects; it varies significantly, as shown in Figure 3.7. For

instance, subject 588 has only 3.49% missing values, whereas subject 552 exhibits a much higher proportion, with 24.78% of values missing.

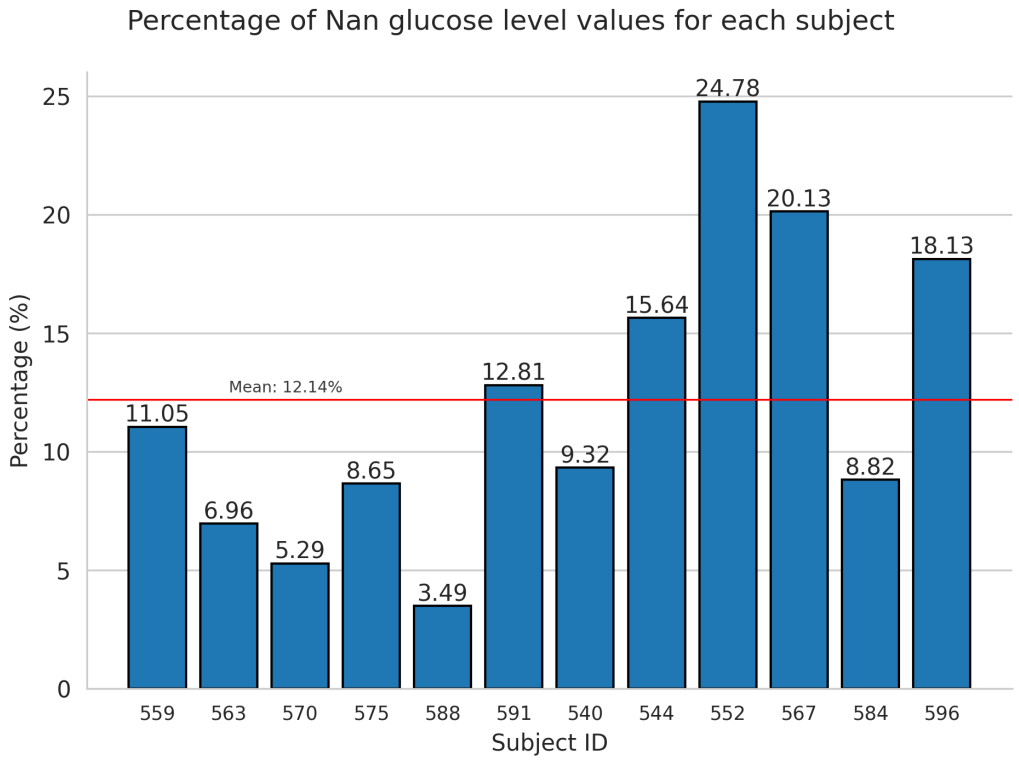


Fig. 3.7: BGC missing values (%) for each subject.

Another significant challenge related to the BGC's missing observations is the presence of substantial gaps within the time series of individual subjects. In practice, common techniques like linear interpolation or padding can be effective for filling in short gaps. However, when compared with longer intervals, these methods tend to yield approximate results. A notable portion of these gaps within the dataset extends well beyond two hours, with some even spanning entire days, making it virtually impossible to impute values without adding bias and distortion. Consequently, it often becomes necessary to consider the entire chunk of the time series as unusable. For instance, subject 591 has 966 consecutive missing values, stretching from 5:43 a.m. on December 26, 2021 to 2:16 p.m. on December 29, 2021. Such long gaps make it practically unfeasible to restore the integrity of the data without compromising its quality.

Finally, as shown in Figure 3.8, the occurrence of missing values appears to be influenced by the time of day. This pattern is probably due to the change of the CGM devices, a process that typically occurs in the morning, which is in fact the time of day with the highest incidence of missing values.

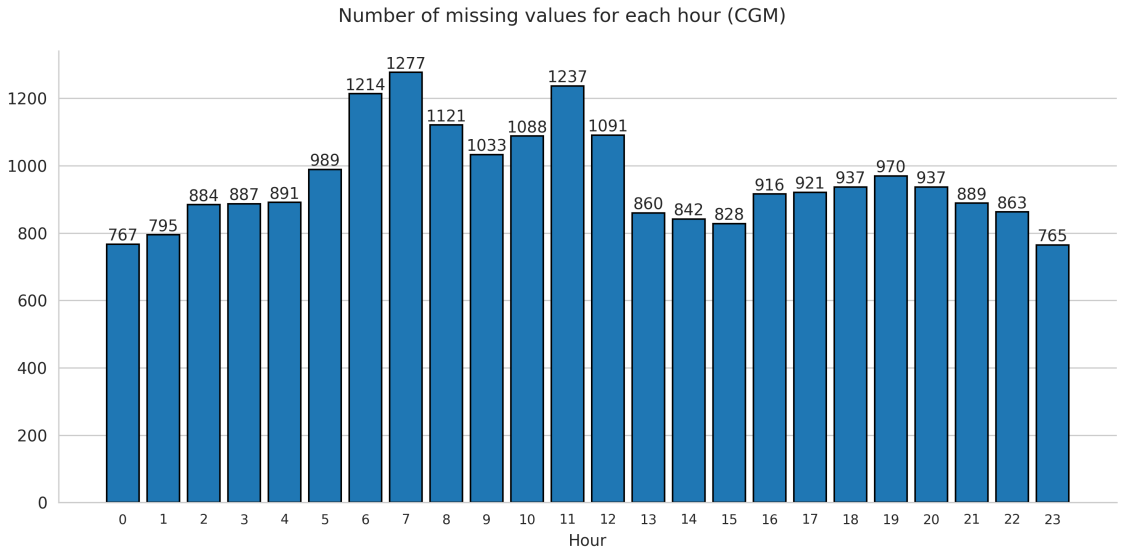


Fig. 3.8: BGC missing values for each hour of the day.

Collectively, these factors severely limit the dataset's usability, with the effectiveness of techniques for imputing missing values primarily confined to addressing minor gaps. For a more detailed analysis of the imputation's strategies and their implications, please refer to Chapter 4.

4

Pre-processing strategies

This chapter deals with the study of different dataset pre-processing methodologies, which were applied to the *Preprocessed* dataset defined in the preceding chapter. The primary objective of this investigation was to assess the impact of different pre-processing strategies on the final performance. Each version differs in terms of data filtering, treatment of missing values, potential transformations, and the introduction of novel features. To evaluate the performance of each dataset, we applied the same methodology employed by the current state-of-the-art work [18]. For an in-depth understanding of this process, please refer to Chapter 5. The subsequent pages provide a comprehensive overview of the pre-processing techniques employed, taking as an initial reference the paper by Shuvo et al. [18].

4.1 Basic

This is the basic version of the dataset. Essentially, it is a re-implementation of the pre-processing methods outlined in the reference paper, but without addressing the missing values associated with the BGC, which were, in practice, discarded. This version served exclusively for future imputation of missing values. In fact, for the subsequent variants of the dataset, gaps in the test set lasting less than 30 minutes were filled using predictions derived from the model trained on the *Basic* dataset, where missing values were indeed omitted.

Therefore, starting from the *Preprocessed* dataset, the initial step involved the removal of missing values associated with BGC. Subsequently, for the self-reported features (namely FS, BI, and C), any missing values were filled with zeros. At this point, to ensure the same scale for all features, normalization was performed, thus guaranteeing the proper learning of the model.

Finally, still following the work by Shuvo et al. [18], to mitigate artifacts, outliers and distorted CGM readings, a two-hour Gaussian filter with standard deviation $\sigma = 1$ has been applied, specifically employing the function `gaussian_filter1d` from the *Scipy* library. Figure 4.1 provides a clear representation of the filter application on BGC values for subject 570.

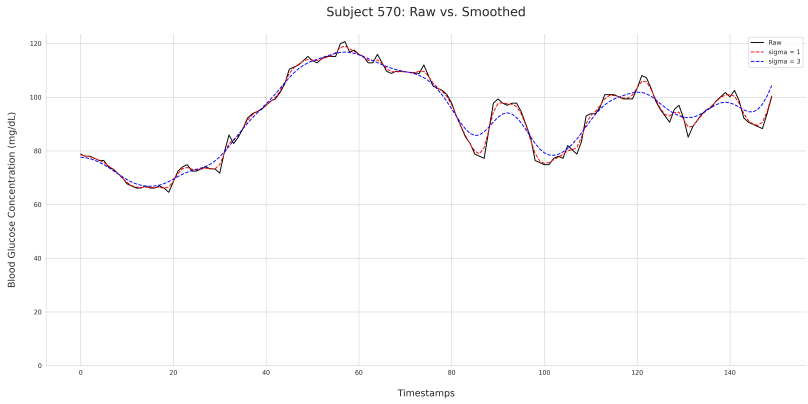


Fig. 4.1: Gaussian filter applied on BGC values from subject 570 with two different σ values. $\sigma > 1$ causes missing hypo- and hyperglycemic events.

4.2 Smoothed

This dataset version has been generated by re-implementing the procedure in [18], following each step as described in the paper. In practice, it differs from *Basic* only in the handling of BGC missing values, which has been done before the normalization and smoothing steps.

In the training set, missing BGC values were imputed through linear interpolation from adjacent glucose values if the interval was less than one our, otherwise samples were discarded to avoid artifacts and incorrect prediction trajectories. For the test set, extrapolation has been implemented in two different conditions for each two-hour sliding window, depending on the gap g between data points:

$$\begin{cases} \text{using the model's predictions} & \text{if } g \leq 30 \text{ min} \\ \text{linear interpolation} & \text{otherwise} \end{cases} \tag{4.1}$$

Regarding the model predictions for $g \leq 30$ minutes, training was logically performed on *Basic*, i.e., the dataset with the missing values discarded.

This dataset represents the basis for the next pre-processing strategies.

4.3 Averaged

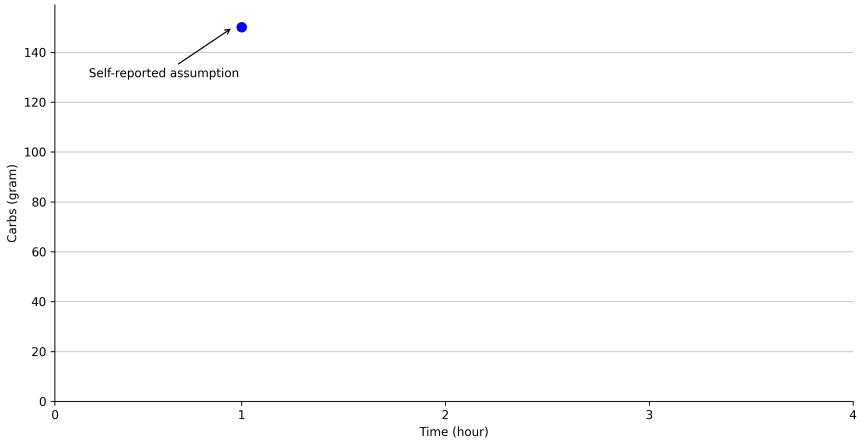
This particular version diverges from *Smoothed* in its approach to filling missing data. In this case, in fact, significant data gaps exceeding a 30-minute threshold were filled by imputing the average values from corresponding time intervals on other days. The imputation of missing values using this methodology was exclusively applied to the training set, while for the test set we employed the same previous extrapolation approach.

4.4 Active_carbs

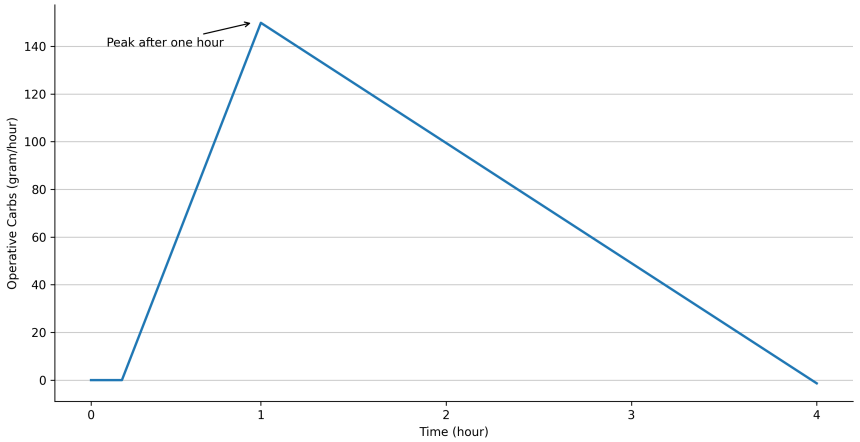
The data processing for this dataset adheres to the same *Smoothed* methodology, with the notable exception of the two features linked to insulin injections (BI) and carbohydrate intake count (C), which were made continuous following the work by Butt et al. [28].

First of all, to convert self-reported features into continuous data, it was necessary to synchronize them with CGM, ensuring they share the same five-minute frequency as observations related to BGC.

Regarding carbohydrate intake, it's important to note that its influence on blood glucose levels is not limited to the precise timestamp of data recording. Instead, it extends over a longer period of time. To illustrate, Kraegen et al. [39] demonstrated that, after a meal, glucose level typically starts to increase around 15 minutes after consumption, peaking in 60 minutes, and then gradually decreases over the subsequent three hours until it reaches a stable state. Consequently, the continuous estimate of carbohydrate is given by the combination of both the rise and decay phases, as depicted in Figure 4.2.



(a) Self-reported carbs assumption.



(b) Operative carbs.

Fig. 4.2: Carbohydrate conversion from self-reported events to operative. The reported value is spread over four hours according to Butt et al. [28].

Therefore, by re-implementing the procedure of the reference paper [28], the operative carbohydrates were estimated as follows:

1. Assuming five-minute reference intervals, for the first three samples (i.e., the first 15 minutes) the value is equal to zero.

2. Next, the operative carbs start rising at a 11.1% rate, until the value reaches its peak at the 12th sample (i.e., 60 minutes after the meal).
3. Then, the value starts decreasing at a 2.8% rate, until it finally reaches zero after 48 samples (i.e., three hours after the peak).

Unlike basal insulin, which has a slow onset, bolus insulin is rapid-acting and is typically administered before meals to counteract rising blood glucose levels. It's essential to note that the action of bolus insulin is also time-dependent. In adults, it reaches its peak effectiveness approximately 75 minutes after administration. To capture the effect of insulin activity, the reference paper employed an exponential decay curve known as "Insulin On Board" (IOB) [40]. However, this curve primarily models the decline of insulin, assuming that it is entirely absorbed at the initial timestamp (Figure 4.3).

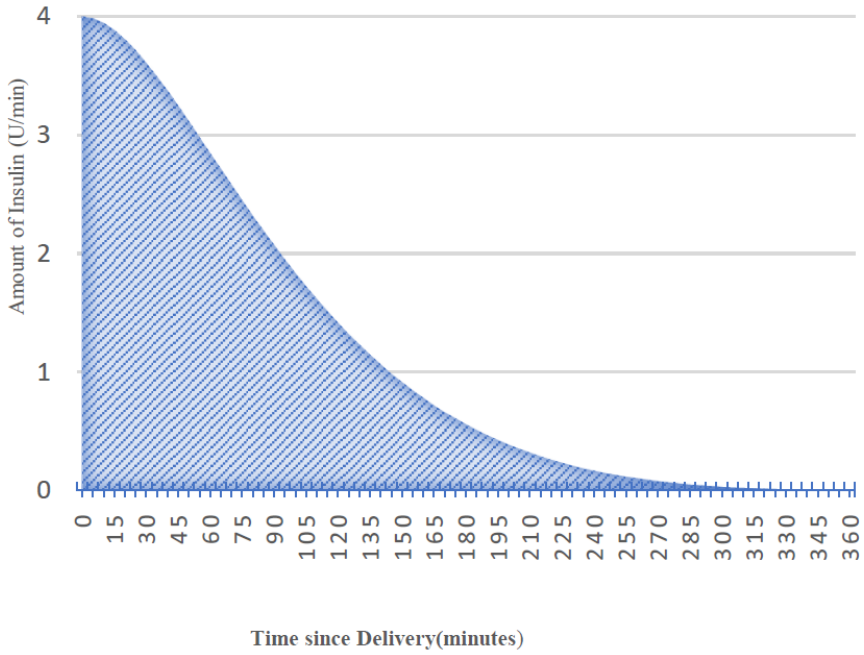


Fig. 4.3: Insulin On Board curve (remaining units of insulin per minute). Image from Butt et al. [28].

Therefore, we opted to incorporate the process of insulin absorption into the transformation, i.e., not exclusively considering the rate of decay. This led us to consider the Insulin Activity (IA) curve [40], which is defined by the following mathematical formula:

$$IA(t) = \frac{S}{\tau^2} * t * \frac{1-t}{t_d} * e^{\frac{-t}{\tau}} \tag{4.2}$$

where:
t is the specific timestamp considered.
t_d is the duration of the effect of insulin.
τ is the time constant of the exponential decay.
a is the Rise time factor.
S is the Auxiliary scale factor.

$$\tau = t_p * (1 - \frac{t_p}{t_d})/(1 - 2 * \frac{t_p}{t_d}) \tag{4.3}$$

t_p is the peak activity time.

$$a = 2 * \frac{\tau}{t_d} \tag{4.4}$$

$$S = 1/(1 - a + (1 + a) * e^{\frac{-t_d}{\tau}}) \tag{4.5}$$

After applying the formula to the self-reported data regarding the subject's bolus insulin injections, the resulting continuous feature curve will closely resemble the one illustrated in Figure 4.4.

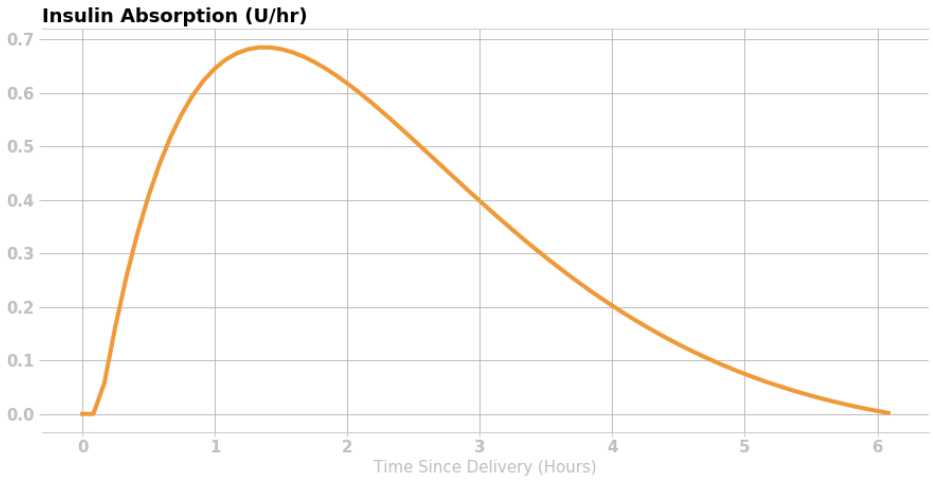


Fig. 4.4: Insulin Activity curve (absorption of units of insulin per hour). Image from LoopDoc [40].

Clearly, as we will see in the next chapters, this type of processing is highly dependent on the consistency and meticulousness with which each individual patient recorded the BI and C values throughout the 8-week data collection period.

4.5 Timed

To address missing values, especially substantial gaps, an attempt was made to introduce, in *Smoothed*, a timestamp-related column *time* for each observation. Theoretically, this additional information could have been leveraged by the model to enhance the accuracy of predicting glucose levels at specific times. Regrettably, this pre-processing approach did not yield any discernible benefits.

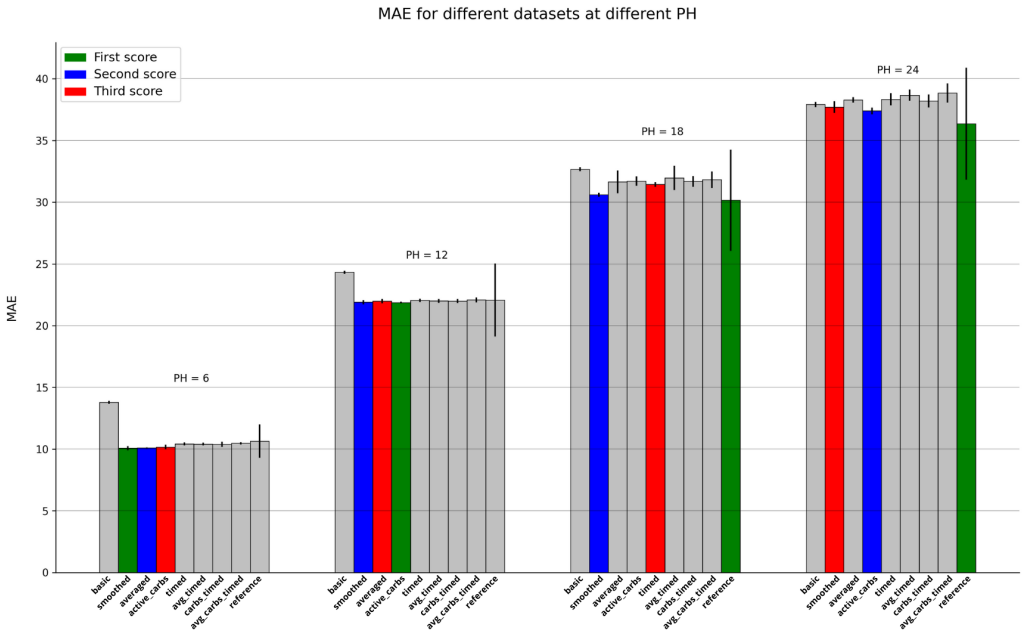
4.6 Combinations

Finally, several combinations of the above datasets were tried, including:

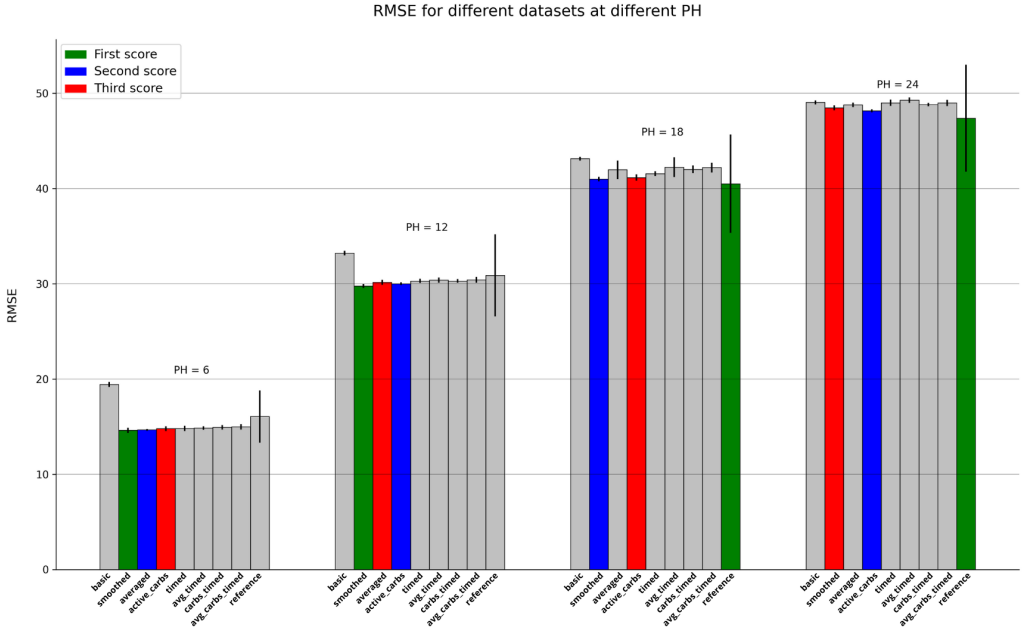
- *Averaged_timed*: dataset *Averaged* with time-related column.
- *Averaged_carbs*: dataset *Averaged* with self-reported features BI and C converted into continuous values following the methodology described above and in [28].
- *Carbs_timed*: dataset *Active_carbs* with time-related column.
- *Averaged_carbs_timed*: combination of all three approaches: *Averaged*, *Active_carbs* and *Timed*.

Combining the various datasets also brought no improvement. Consequently, the forthcoming efforts focused on a comparative analysis of the three most promising datasets: *Smoothed*, *Averaged*, and *Active_carbs*.

Figure 4.5 presents an overview of the performance across all datasets, as measured by MAE and RMSE, including the comparison with the reference [18]. The methodology used for the evaluation is the same as in the reference paper [18], and is explained in the next chapter. It is evident that for shorter time horizons (30 minutes and 60 minutes), the manually generated datasets generally outperform the reference, whereas the latter performs better for 90 minutes and 120 minutes. Notably, the results from the reference exhibit a significantly higher standard deviation across the five different runs of the model. These disparities can be likely attributed to variations in the implementation process, encompassing parsing, synchronization, and pre-processing strategies. In fact, our approach aligned with the guidelines outlined in the reference paper [18], which, however, did not provide the source code for reference purposes.



(a) MAE mean and standard deviation values. In green the best score, in blue the second and in red the third.



(b) RMSE mean and standard deviation values. In green the best score, in blue the second and in red the third.

Fig. 4.5: MAE and RMSE for each dataset manually generated and reference [18]. The values, computed for each PH, represent the mean among five different training iterations.

5

Glucose level estimation

This chapter focuses on the challenge of estimating diabetes levels, which essentially entails a regression task. First, Section 5.1 describes the work done on the different datasets generated, which were tested on the current state-of-the-art *Multitask* model [18]. Subsequently, the three most promising datasets were used to evaluate the impact of customization on the predictions, as detailed in Section 5.2. Following this, we fine-tuned the most promising model and reassessed its performance in Section 5.3. Lastly, we made an endeavor to employ an incremental learning approach, that is discussed in Section 5.4.

5.1 Deep Multitask model

All datasets generated were tested on the *Multitask* model proposed by Shuvo et al. [18]. From these evaluations, we selected the three datasets that exhibited the highest performance, aligning with the current state-of-the-art benchmarks in the literature. Therefore, following the reference paper, the model was manually re-implemented using the popular *PyTorch* library [41].

This approach is based on the premise that Multitask Learning (MTL) enhances overall model performance by focusing on individual tasks through an implicit attention mechanism [42]. It operates by learning features collectively across all subjects. The suggested *Multitask* model is developed through multiple network branching. Initially, it splits into two clusters based on the gender of the subjects. Subsequently, it further divides into a number of final layers equal to the total number of subjects, that is, a subject-specific final layer at the end of the architecture for each patient.

Specifically, the architecture is a recurrent neural network comprising two stacked LSTM layers with 16 and 32 units, respectively. Following each LSTM layer, a dropout layer with a 0.2 dropout rate was introduced to enhance regularization and prevent overfitting. Next, the network diverges into two distinct clusters: one tailored for male subjects and the other for female subjects. Each cluster incorporates a 128-unit Fully-Connected (FC) layer, followed by a dropout layer with a rate of 0.4. The final layer within the architecture is the subject-specific FC layer, equipped with 12 units, responsible for generating predictions based on each subject's unique characteristics. By maintaining the original hyperparameter values, we obtained approximately 19,000 trainable parameters, aligning with the number explained in the reference. The overall *Multitask* architecture is shown in Figure 5.1.

The input of the network is a four-dimensional (BGC, BI, FS, C) sliding window including the most recent two hours (24 samples) of observations. The model's output predicts the difference in glucose level, i.e., the variation in BGC between t and $t+PH$, at four different PH: 30, 60, 90 and 120 minutes, with a different instance of the model for each time horizon. Ultimately, the output can be obtained by summing the variation computed by the model with the glucose level at time t.

The dataset split employed in this experiment, pertaining to the exploration of optimal pre-processing strategies, aligns with the one utilized by both Shuvo et al. [18] and the original authors [20] of the dataset itself, who provided the data already split between training and testing for each subject.

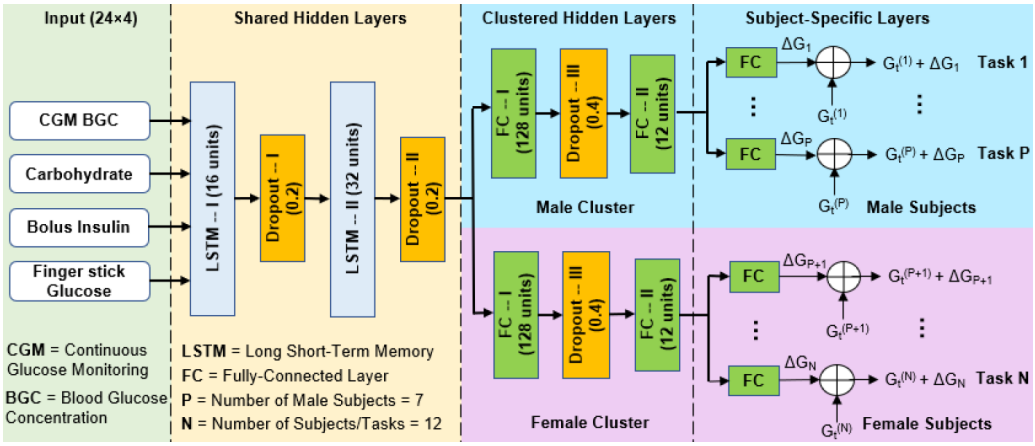


Fig. 5.1: Deep Multitask model architecture. It consists of two LSTM shared hidden layers, two FC clustered hidden layers and subject-specific output layers. Image from Shuvo et al. [18].

The model was trained on each dataset for 250 epochs for 5 runs. In addition, an early stopping procedure with patience equal to 30 and model checkpointing were implemented, using the last week of data for each subject in the training set as validation set. A linear combination of the MAE losses across individual subjects was used as a cost function.

5.1.1 Evaluation metrics

The performance of this model (and further regression models) was evaluated in terms of MAE and RMSE, comparing the two metrics across different datasets and PHs.

MAE is calculated as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5.1}$$

RMSE is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5.2}$$

where:
n is the number of data points.
y_i represents the true values.
$\hat{y}_i$ denotes the predicted values.

5.1.2 Experimental results

Table 5.1 summarizes the results obtained, comparing the three best manually generated datasets, namely *Smoothed*, *Averaged* and *Active_carbs*, with the results achieved by the reference paper [18], denoted as *Reference*.

Table 5.1: Final scores: mean* and standard deviation among 5 different runs. The performance are evaluated both in terms of MAE and RMSE, across all four time horizons.

	PH = 30min		**PH = 60min**		**PH = 90min**		**PH = 120min**	
Dataset	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
Smoothed	10.07±0.17	14.61±0.27	21.90±0.17	29.77±0.21	30.60±0.16	40.99±0.23	37.71±0.47	48.47±0.26
Averaged	10.09±0.05	14.66±0.10	21.98±0.18	30.15±0.26	31.65±0.93	41.97±0.98	38.28±0.22	48.79±0.23
Active_carbs	10.16±0.19	14.79±0.25	21.87±0.07	30.01±0.15	31.71±0.38	41.16±0.32	37.40±0.27	48.16±0.16
Reference**	10.64±1.35	16.06±2.74	22.07±2.96	30.89±4.31	30.16±4.10	40.51±5.16	36.36±4.54	47.39±5.62

*: In green the best score, in blue the second and in red the third.

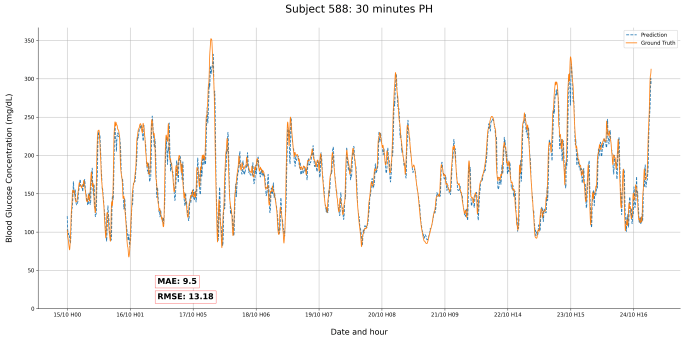
**: MAE and RMSE values from [18].

The outcomes obtained from the various datasets exhibit remarkable similarity and closely align with the results reported in [18]. Specifically, for short and medium-term forecasts (e.g., PH = 90min), the *Smoothed* pre-processing method performs slightly better than other approaches, with *Active_carbs* and *Averaged* alternately occupying the second and third positions in terms of performance. In the case of long-range forecasting (PH = 120min), *Active_carbs* outperforms the other implemented pre-processing strategies.

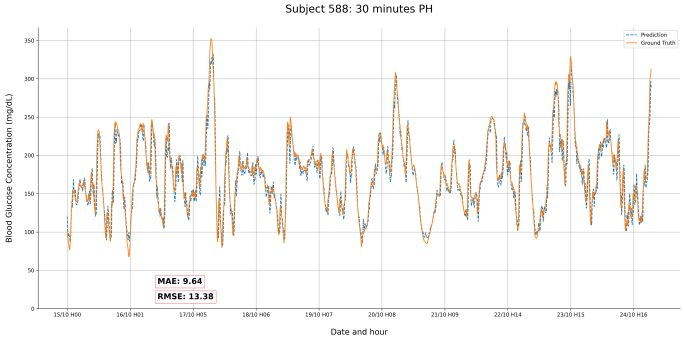
Interestingly, for PH = 30min and PH = 60min, the reference model displays slightly lower performance compared to longer time horizons, while for PH = 90min and PH = 120min our implementation exhibits higher MAE and RMSE for all the three datasets. Nonetheless, our version yields surprisingly more stable results, marked by significantly smaller standard deviations between runs than those reported in the reference paper [18]. This enhanced stability can likely be attributed to several different steps in our approach, including the parsing phase. Moreover, an additional factor that is worth to consider is the validation set. Indeed, there is no clear reference to it in the paper, whereas we took the last week of training data for each subject.

Taking into account the predictions for individual subjects, it becomes intriguing to observe, as previously supposed in Section 4.4, that subjects who maintained a more consistent meal recording habit, particularly 570 and 588, exhibit slightly superior performance on the *Active_carbs* dataset. In Figure 5.2, we present a comparison between the two datasets, specifically for subject 588.

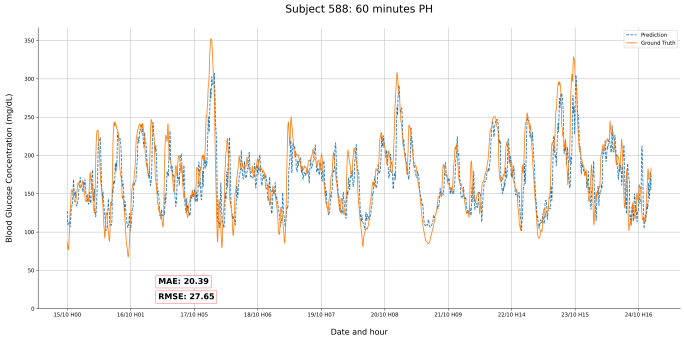
Lastly, it is worth noting a natural decline in overall performance as the time horizon extends. In particular, the model appears to lose its capacity to detect any spikes in the BGC level, even as it continues to track the overall trend. In the next section, the impact of personalization on model performance is evaluated more thoroughly, testing each model on the 3 different datasets selected in this phase.



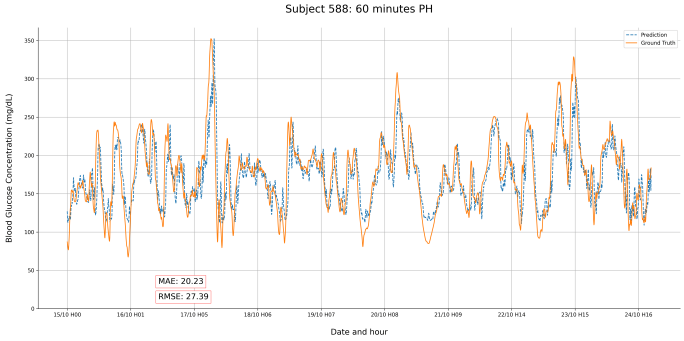
(a) *Smoothed*, MAE = 9.64, RMSE = 13.38.



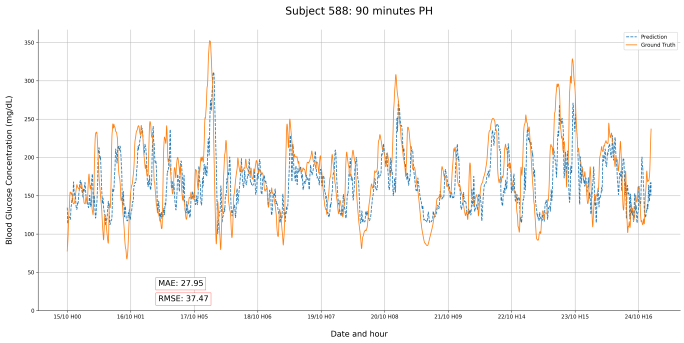
(b) *Active_carbs*, MAE = 9.5, RMSE = 13.28.



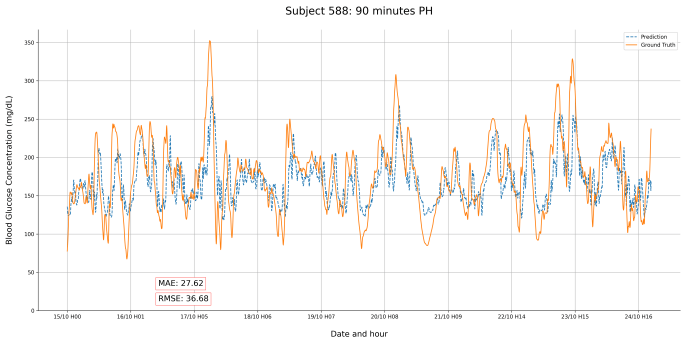
(c) *Smoothed*, MAE = 20.39, RMSE = 27.65.



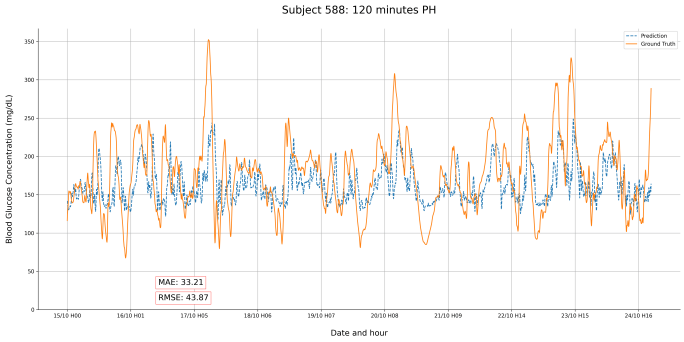
(d) *Active_carbs*, MAE = 20.23, RMSE = 27.39.



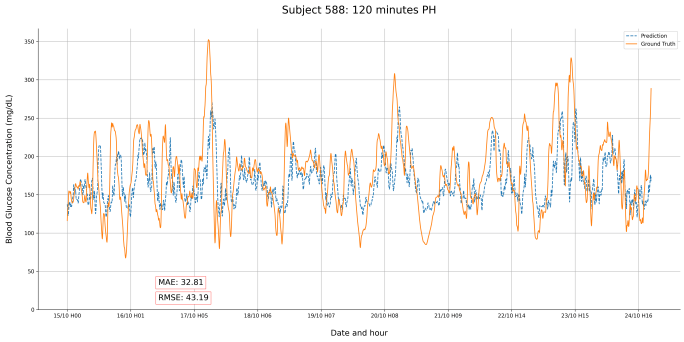
(e) *Smoothed*, MAE = 27.95, RMSE = 37.47.



(f) *Active_carbs*, MAE = 27.62, RMSE = 36.68.



(g) *Smoothed*, MAE = 33.21, RMSE = 43.87.



(h) *Active_carbs*, MAE = 32.81, RMSE = 43.19.

Fig. 5.2: Comparison of predictions for subject 588 on the *Smoothed* and *Active_carbs* datasets, across 30, 60, 90 and 120-min prediction horizons.

5.2 Impact of personalization

The implementation of MTL in a recurrent neural network for personalization purposes yielded significant advancements in predicting BGC levels in terms of MAE and RMSE. Nevertheless, to evaluate the performance of this approach, we conducted a comparative analysis that included the *Multitask* model, the general version of the same, namely *General*, and an additional customized version, namely *Threetask*, based on the median BGC levels of the various subjects within the OhioT1DM dataset.

The three models were tested on all four-time horizons (PH = 30min, 60min, 90min, and 120min) using the best three datasets previously computed, i.e., *Smoothed*, *Averaged* and *Active_carbs*, enhancing the robustness of the evaluation. Additionally, for this task, we adopted another dataset split, namely the Leave-One-Subject-Out Cross-Validation (LOSO-CV) split. This approach entails testing each model on one specific subject while training on the remaining subjects during each iteration. LOSO-CV provides a more realistic representation of real-world scenarios as the models are tested on entirely new subjects during each iteration. The final values of MAE and RMSE were then derived by averaging the results across all iterations, providing a comprehensive evaluation of the models' predictive capabilities across various subjects. Finally, in order to maintain a fair basis for comparison among the models, we kept the architectural parameters and training-related variables consistent with the original *Multitask* model. Collectively, this approach ensured a more robust and unbiased assessment of the models' performance.

5.2.1 General model

The architecture of the *General* model remains identical to the *Multitask* reference, except for the exclusion of the branches responsible for gender-based and subject-specific differentiation in the final FC layer. In practice, the division into branches was made sequential, with no distinction between subjects, as shown in Figure 5.3. All hyperparameters, number of layers, and other options were left unchanged. By maintaining consistency in the core structure and hyperparameter configurations, in addition to ensuring a direct and fair comparison with the customized version, we isolated and distinctly assessed the influence of gender-based and subject-based branches, on both general and individual subjects performance.

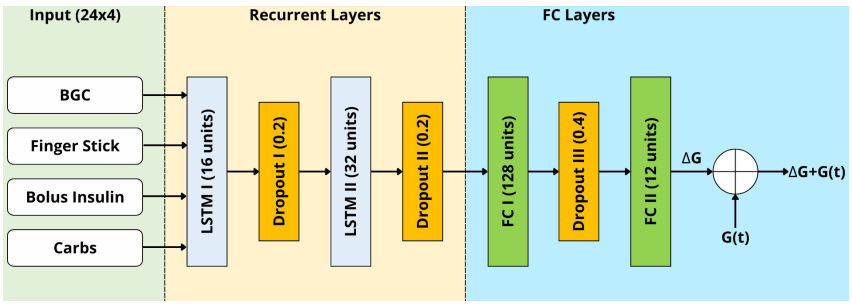


Fig. 5.3: Deep General model architecture. The architecture is the same of the *Multitask* model, with two stacked LSTM hidden layers and two FC hidden layers, but without clusters and subject-specific layers.

5.2.2 Deep Threetask model

During the data exploration phase, we observed a considerable variability in BGC levels among the 12 subjects (refer to Figure 5.4). According to the American Diabetes Association (ADA), a BGC level of less than 140 mg/dL two hours after meals should be considered normal, while for diabetic individuals the acceptable threshold is increased to 180 mg/dL [43]. Therefore, to categorize the subjects into three discrete groups, we have established the following classifications:

- Normal (N). BGC median level is under 140 mg/dL (subjects 563, 575, 540, 552, 596).
- Medium (M). BGC median level is between 140 mg/dL and 180 mg/dL (subjects 559, 588, 591, 544, 567).
- High (H). BGC median level is over 180 mg/dL (subjects 570, 584).

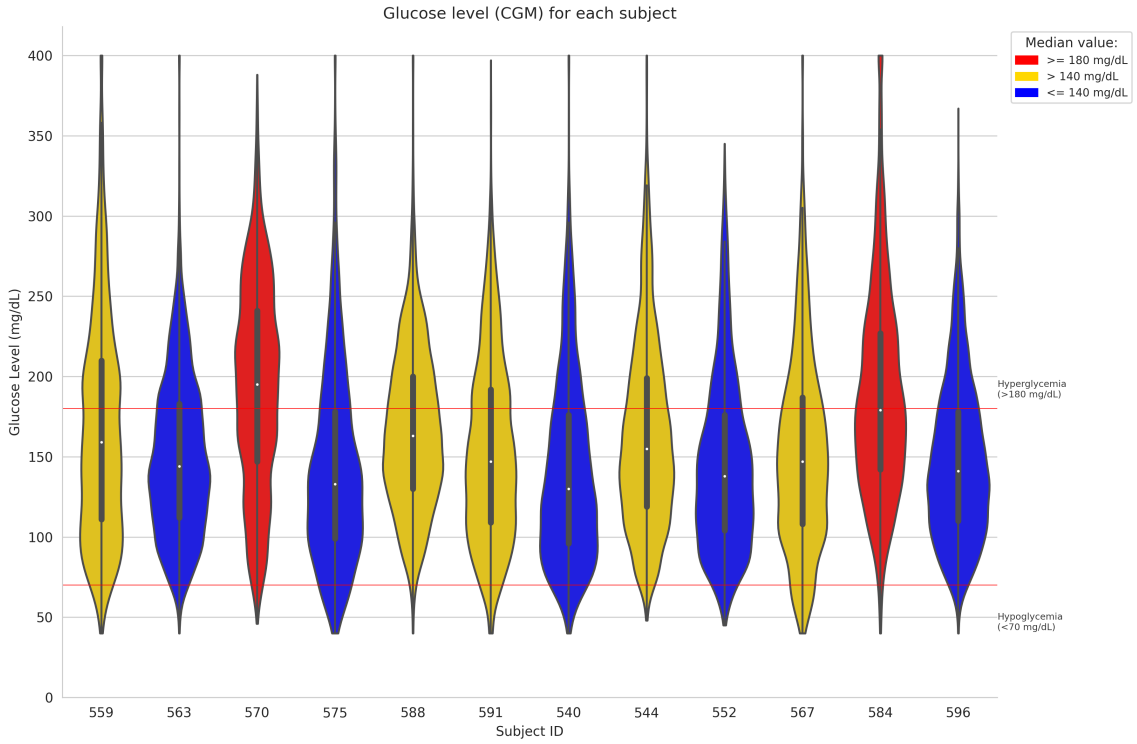


Fig. 5.4: BGC median level for each subject of the OhioT1DM dataset. The color indicates the group (N, M, or H) into which each subject was classified.

The key idea is that the model can implement a branch for each of these ranges. This way, the prediction would be tailored directly to the disease level.

Therefore, the same model as the previous ones was implemented with only one network division layer, which can distinguish median glucose concentration levels by 3 branches placed after the two recurrent layers (i.e., in the same way as the male-female clusters in the *Multitask* model).

As with the *General* model, maintaining the core structure and hyperparameter values unchanged served as a crucial step to ensure a consistent basis for comparing and evaluating the impact of the model with three branches. Figure 5.5 provides a clear representation of the *Threetask* architecture.

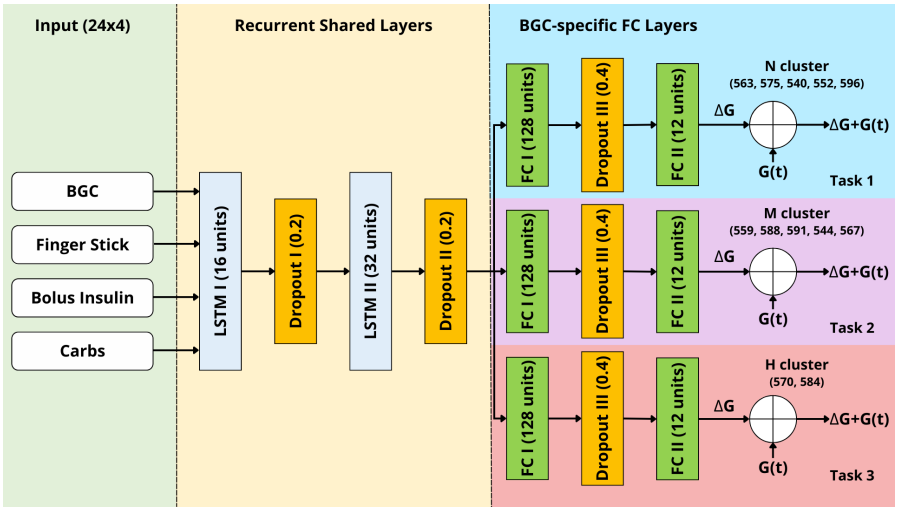


Fig. 5.5: Deep Threetask model architecture. The model has the same general architecture and hyperparameters as the *Multitask* and *General* models, but with different clustering based on the median BGC level of the input subject.

5.2.3 Experimental results

First of all, we conducted an overall performance comparison among the various models across all three datasets, assessing MAE and RMSE on the best dataset (see Table 5.2). It's worth noting that the levels of MAE and RMSE have experienced a slight increase compared to the previous task. This uptick can be attributed to the models being tested on data from subjects that have not been encountered previously. Nevertheless, these values remain comparable to the earlier results, affirming the effectiveness of the approach. Moreover, the outcomes clearly indicate that the performance achieved by the *General* model aligns and slightly overcomes that of the personalized approaches. This observation suggests that the branching mechanism has only a minimal impact on the prediction task. Interestingly, the *Multitask* model exhibits even greater performance degradation than the other models as the prediction horizon extends.

Table 5.2: Final scores: mean* and standard deviation on LOSO-CV across all subjects. The performance of the three models in terms of MAE and RMSE over all four time horizons was evaluated.

	PH = 6		PH = 12		PH = 18		PH = 24	
Model	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
Multitask	11.01±1.58	16.68±2.81	23.53±3.35	32.89±4.98	32.73±4.32	43.87±5.94	40.06±4.87	51.84±6.24
General	11.08±1.64	16.65±2.81	23.62±3.12	32.80±4.63	32.44±4.61	43.16±6.21	38.65±4.96	50.02±6.57
Threetask	11.10±1.53	16.70±2.81	23.44±3.27	32.48±4.84	32.49±4.32	43.19±5.74	38.89±4.68	50.18±5.99

*: In green the best score.

Nonetheless, further insights can be gained by analyzing the performance on each individual subject (see Table 5.3). Upon analysis of the performance of different models, it becomes evident that the *Multitask* model consistently underperforms on individual subjects, especially when considering longer time horizons. In contrast, the overall performance of the *General* model closely resembles that of the *Threetask* model, with the latter only slightly outperforming it on certain subjects. To summarize the results based on the colored boxes in the table, the *Multitask* model has 16 boxes (16.7%), the *General* model has 35 boxes (36.4%), and the *Threetask* model has 45 boxes (46.9%).

Table 5.3: Final scores: best dataset and model* for each subject. The performance are evaluated both in terms of MAE and RMSE, across all four time horizons.

	PH = 6		PH = 12		PH = 18		PH = 24	
PID	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
559	smooth: 12.38	smooth: 19.36	smooth: 27.37	avrg: 38.51	smooth: 38.94	smooth: 52.25	avrg: 46.65	avrg: 60.67
563	avrg: 8.77	avrg: 12.76	smooth: 18.91	avrg: 26.0	avrg: 26.71	avrg: 35.62	smooth: 32.55	smooth: 42.22
570	smooth: 8.96	carbs: 13.45	avrg: 19.21	avrg: 26.57	avrg: 27.84	carbs: 37.17	smooth: 35.72	carbs: 45.86
575	smooth: 10.88	smooth: 17.24	smooth: 23.39	smooth: 33.57	smooth: 33.19	smooth: 44.85	smooth: 40.51	avrg: 53.17
588	avrg: 10.78	avrg: 15.80	carbs: 21.28	carbs: 29.45	carbs: 28.24	avrg: 37.93	smooth: 32.63	avrg: 43.11
591	carbs: 12.26	smooth: 18.91	smooth: 24.79	smooth: 34.61	carbs: 33.32	avrg: 44.10	smooth: 38.40	smooth: 48.91
540	carbs: 11.40	avrg: 16.27	carbs: 25.16	carbs: 34.14	smooth: 33.70	smooth: 44.07	carbs: 39.35	avrg: 50.22
544	avrg: 9.67	avrg: 13.87	carbs: 20.85	smooth: 28.84	smooth: 31.67	smooth: 41.65	carbs: 39.04	carbs: 50.52
552	avrg: 10.04	avrg: 15.03	smooth: 21.87	smooth: 30.28	avrg: 29.9	avrg: 39.77	smooth: 34.95	smooth: 45.47
567	avrg: 12.09	avrg: 19.08	smooth: 26.74	carbs: 37.19	avrg: 36.97	avrg: 48.42	carbs: 43.51	avrg: 55.51
584	avrg: 13.77	carbs: 21.50	carbs: 28.04	carbs: 39.37	carbs: 37.44	carbs: 50.54	avrg: 44.30	avrg: 58.04
596	smooth: 9.24	smooth: 13.89	smooth: 19.22	smooth: 26.20	carbs: 25.89	avrg: 34.74	carbs: 31.64	carbs: 41.35

*: In red if the best score belongs to the *Multitask*, in blue for the *General* and in green for the *Threetask*.

Upon closer examination, the *General* model appears to excel in handling subjects 563, while the *Threetask* model performs exceptionally well on subjects 540, 544, 552, and 584. Although the overall performance of the *Threetask* model appears nearly identical to the *General* model, its true strength lies in its ability to generalize over a larger number of subjects not encountered during training. Additionally, it's noteworthy that subject 563 is situated remarkably close to the next median BGC level threshold, which could potentially mislead the neural network's predictions. However, it's important to note that when the *Threetask* model makes errors, they tend to be more substantial than those made by its general counterpart.

Overall, the *General* and *Threetask* models exhibit striking similarities, with their respective strengths and weaknesses often compensating for each other. Consequently, the classification task (as outlined in Chapter 6) has been conducted by comparing these two models, omitting the *Multitask* model due to its inferior personalization performance.

Finally, to gain a more precise understanding of the distinctions among the models, we conducted a statistical test to assess the equality of expected MAE values across subjects. In this process, we initially assessed the normality assumption for each distribution using the Shapiro-Wilk [44] test. Following that, we applied the t-test [45] and the paired t-test [46] from the *Scipy* [47] library, using a significance level α of 0.05 (corresponding to a 95% confidence level). The primary distinction between these two tests lies in their applicability. The t-test is employed to compare two independent groups, whereas the paired t-test is designed for scenarios where the two groups are interrelated or dependent on one another.

t-test

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{2}{n}}} \tag{5.3}$$

where:
$\bar{X}_1$ and $\bar{X}_2$ are the sample means of the two groups.
n is the sample size of both the groups, assuming that they are equal.

$$s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}} \tag{5.4}$$

$s_{X_1}^2$ and $s_{X_2}^2$ are the sample variances of the two groups.

paired t-test

$$t = \frac{\bar{X}_D}{s_D / \sqrt{n}} \tag{5.5}$$

where:
$\bar{X_D}$ is the mean of the differences between all pairs.
s_D is the standard deviation of the differences between all pairs.
n is the number of paired observations.

The analysis, which encompassed comparisons of the three models across all time horizons, provided insightful results. Firstly, the t-test upheld the null hypothesis for each comparison, meaning that the expected values are statistically equivalent. However, when we conducted the paired t-test between the *Multitask* model and both the *General* and *Threetask* models at PH = 24, the null hypothesis was rejected. This rejection may indicate a greater disparity in performance as the time horizon increases, with the *Multitask* model exhibiting inferior performance compared to the other two. Nevertheless, it's essential to note that this outcome underscores the limited impact of personalization in the context of this specific task.

5.3 Fine-tuning

To further optimize performance and assess the potential for enhancing our approach, we conducted fine-tuning on the *Threetask* model, which has demonstrated slightly superior performance compared to the *General* model, particularly outperforming the *Multitask* model. To achieve this goal, we utilized the *Smoothed* dataset, as there is no notable difference between the datasets in terms of performance, and the *Smoothed* version offers the lowest level of complexity.

Given the complexity in performing a comprehensive grid search for all hyperparameters, we opted for a random search with 30 iterations. Following the same LOSO-CV strategy, we evaluated each trial on the average MAE and RMSE over all the subjects. This fine-tuning procedure was carried out on the model tailored for the 30-minute time horizon, and subsequently, the optimized parameters were extended to the other time horizons. Among the various combinations explored, it's noteworthy that Gated Recurrent Unit (GRU) [48] layers were tested as an alternative to the original LSTMs. In addition, as activation functions for hidden FC layers, we experimented with both Rectified Linear Unit (ReLU) [49] and its Leaky version [50], with α parameters of 0.01 and 0.001. A detailed summary of the hyperparameters is presented in Table 5.4.

Table 5.4: Hyperparameters tested for the fine-tuning on the *Threetask* model.

Hyperparameters	**Values tested**
LSTM/GRU layers	[1, 2]
LSTM/GRU I units	[16, 32]
LSTM/GRU II units	[32, 64]
Dropout I rate	[0, 0.2, 0.4]
Dropout II rate	[0, 0.2, 0.4]
Dropout III rate	[0, 0.2, 0.4]
FC layers	[1, 2]
FC I units	[32, 64, 128]
FC II units	[12, 32]
Learning rate	[0.0001, 0.001]

The best combination of hyperparameters found during the search was the following:

- Layer type: GRU.
- Number of stacked recurrent layers: 2.
- GRU I units: 16.
- GRU II units: 32.
- Number of fully-connected layers: 1.
- Dropout I rate: 0.
- Dropout II rate: 0.4.
- Dropout III rate: 0.4.
- FC units: 64.
- FC activation function: LeakyReLU with $\alpha = 0.01$.
- Learning rate: 0.001.

Table 5.5 shows, in terms of MAE and RMSE values, the comparison between the original *Threetask* model and the fine-tuned version, into which the best hyperparameter combination has been substituted.

Table 5.5: Original vs. Fine-tuned *Threetask* model scores: mean* and standard deviation on LOSO-CV across all four time horizons.

	PH = 6		**PH = 12**		**PH = 18**		**PH = 24**	
Threetask model	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
Original	11.10±1.53	16.70±2.81	23.44±3.27	32.48±4.84	32.49±4.32	43.19±5.74	38.89±4.68	50.18±5.99
Fine-tuned	10.78±1.57	16.38±2.85	23.16±3.32	32.27±4.65	32.34±4.39	43.18±5.75	38.56±5.27	50.23±6.55

*: In green the best score.

Fine-tuning slightly improved model performance, except for RMSE at PH = 120min. In this specific time frame, the standard deviation also increases and, in general, the extent of improvement becomes progressively less conspicuous as the time horizon increases. This phenomenon can likely be attributed to the hyperparameter optimization being carried out primarily on the more immediate time horizon of 30 minutes.

In the subsequent and concluding section, we delve into the incremental learning approach, with the primary objective to enhance performance by progressively training the model on the individual subject.

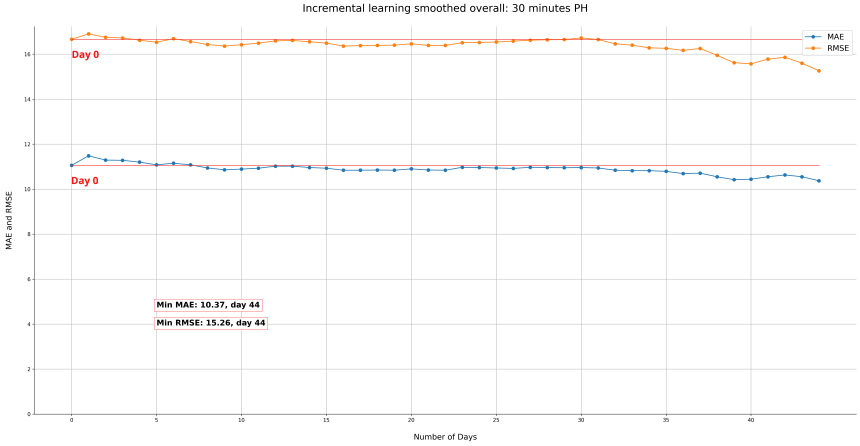
5.4 Incremental learning

Incremental learning [51] is a ML approach where a model is continuously updated and improved over time as new data becomes available. Instead of training the model from scratch each time a new data is acquired, incremental learning builds upon existing knowledge, adapting and refining the model's parameters

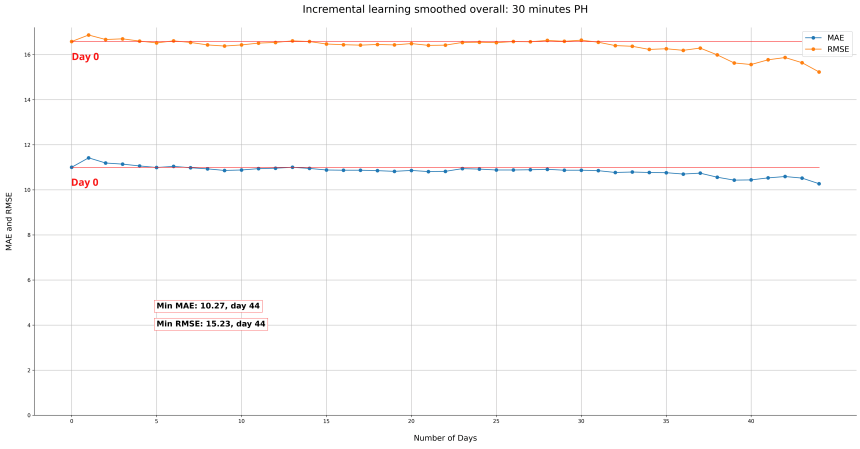
to incorporate the latest information. This process allows the model to stay up-to-date and potentially improve its performance as more data are processed, making it particularly useful for this kind of task, where data referred to each individual subject evolves and accumulates over time.

The procedure was implemented for both the *General* and *Threetask* models, in order to assess how each model performs when subjected to this particular approach. Furthermore, we applied incremental learning to each of the three previously selected datasets, yielding nearly identical results across them. Consequently, we present the overall outcomes for the *Smoothed* dataset.

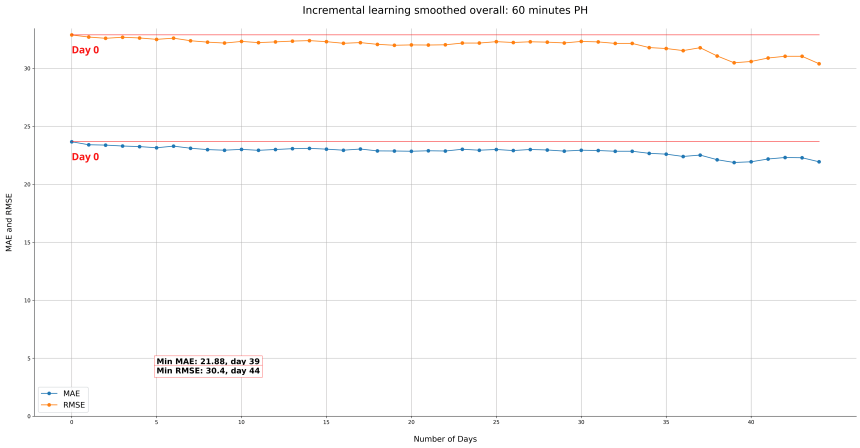
Figure 5.6 illustrates the progression of MAE and RMSE values as the number of days incorporated into the training dataset increases, with a maximum of 45 days considered, for both the *General* and *Threetask* models. The red lines on the graphs, beginning at day zero, indicate the performance before applying the procedure. They serve as a baseline for evaluating the overall effectiveness of the approach.



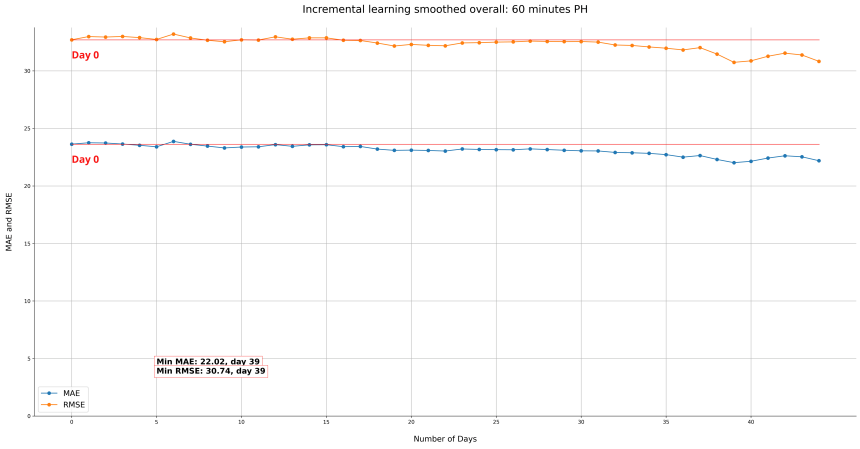
(a) PH = 30min, *General* model.



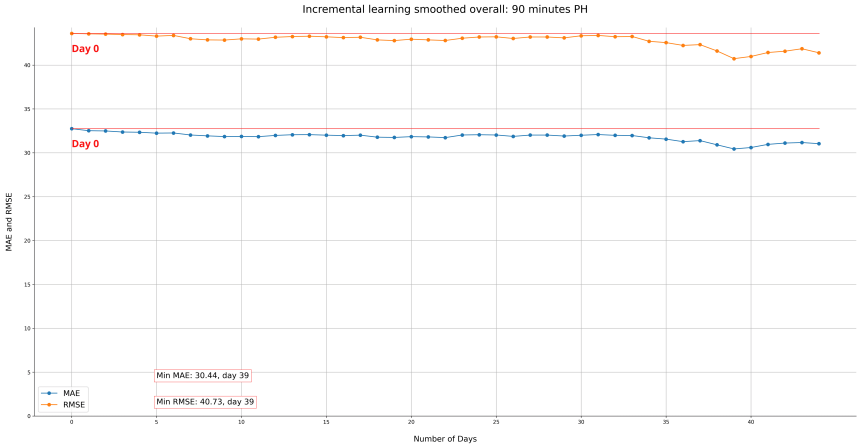
(b) PH = 30min, *Threetask* model.



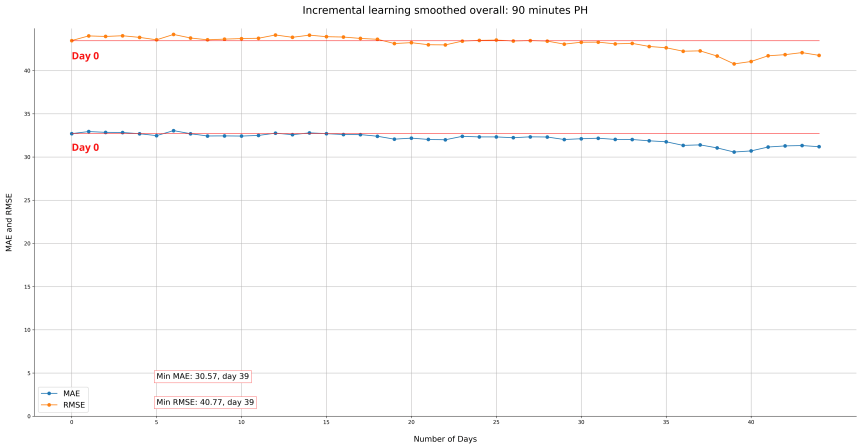
(c) PH = 60min, *General* model.



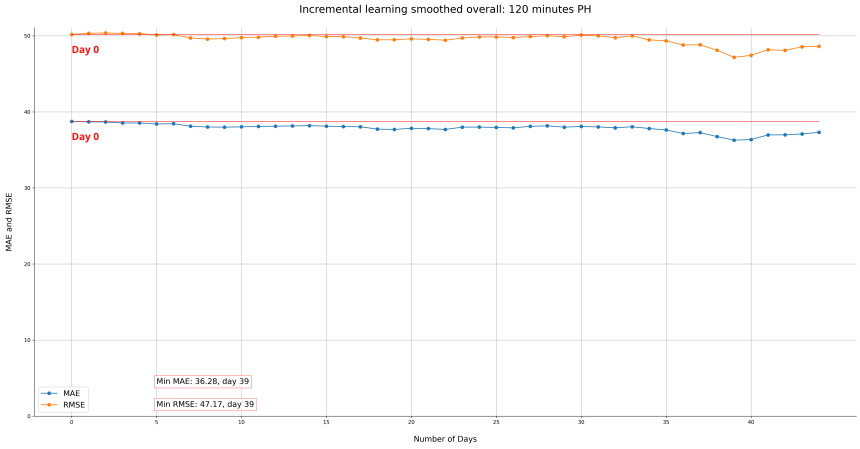
(d) PH = 60min, *Threetask* model.



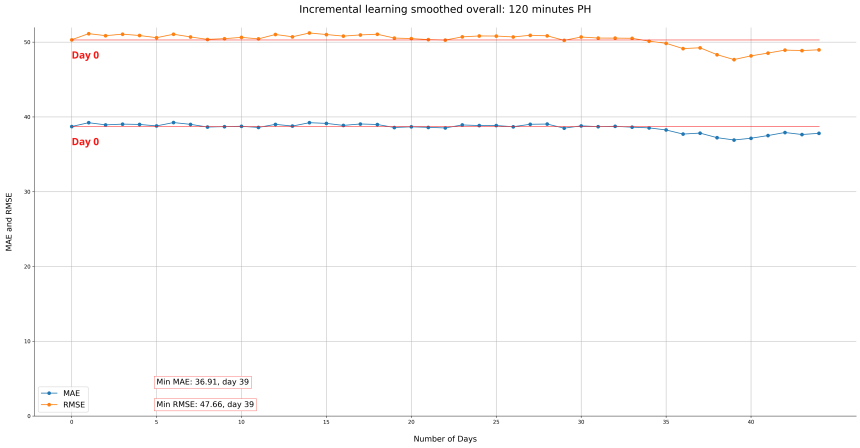
(e) PH = 90min, *General* model.



(f) PH = 90min, *Threetask* model.



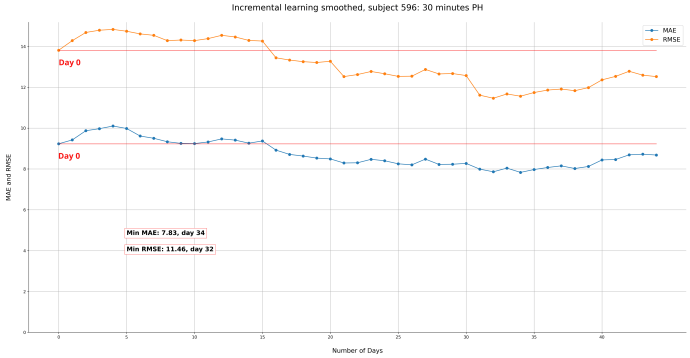
(g) PH = 120min, *General* model.



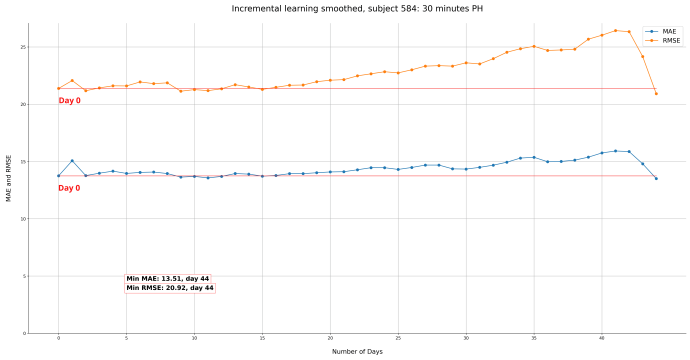
(h) PH = 120min, *Threetask* model.

Fig. 5.6: Incremental learning across all time horizons between *General* and *Threetask* models. MAE and RMSE values are reported as the days included in the training set increase.

Upon closer examination of individual subjects, it becomes apparent that this procedure proved effective for some subjects while potentially degrading performance for others. Furthermore, no significant distinctions emerge between the two models, implying that the impact of customization is minimal in this scenario. Figure 5.7 illustrate this contrast with subjects 584 and 596 at 30min PH, using the *General* model. Subject 584 experienced notable improvement in the first days, whereas subject 596 demonstrated a less favorable outcome. Hence, when we average the MAE and RMSE values across all subjects, we arrive at the nearly constant curves depicted above.



(a) PH = 30min, subject 596.



(b) PH = 30min, subject 584.

Fig. 5.7: Incremental learning across 30-min PH. MAE and RMSE values are reported for subjects 584 and 596 on *General* model.

In summary, the incremental learning procedure yielded marginal enhancements in model performance.

Therefore, we approached the problem from another point of view. The next chapter focuses on BGC level classification, during which multiple strategies were evaluated, including the integration of another dataset featuring 16 additional subjects.

6

Glucose level classification

In this chapter, we present the results of our investigation into the impact of personalization for the classification task. This approach provides an alternative perspective for assessing the issue, and it holds potential advantages for end users. In fact, end users could perceive the predictive model's results not solely as numerical values, but rather as categorized classes corresponding to specific levels of risk.

Initially, we conducted a comparative analysis of the *Smoothed*, *Averaged*, and *Active_carbs* datasets over a 30-minute time horizon for both binary and multi-class classification. However, we observed that the results were nearly identical (Table 6.1), leading us to select the *Smoothed* dataset for use in other time horizons, due to its lower complexity. Regarding the models under consideration, we compared the fine-tuned *General* and *Threetask* models. The *Multitask* model, which exhibited inferior performance, was consequently excluded from further consideration.

To begin, Section 6.1 focuses on the binary classification task, where we established the BGC risk threshold based on the insights from [38]. Moving forward, we divided the range of BGC values into quartiles, leading to a multi-class classification with four distinct classes. We also explored an approach that solely utilized BGC as a feature. These steps are discussed in Section 6.2. Finally, as detailed in Section 6.3, we incorporated data from 10 days of BGC measurements for 16 additional subjects, which were integrated from the BIG IDEAs dataset [52], [53].

Table 6.1: Binary and multi-class classification on 30-min PH: overall macro-averaged accuracy scores (%) and standard deviations, computed on both *General* and *Threetask* models across the three datasets.

	Binary		**Multi-class**	
Dataset	**General**	**Threetask**	**General**	**Threetask**
Smoothed	70.17±3.51%	78.75±2.09%	87.42±2.96%	87.42±3.07%
Averaged	70.42±3.40%	78.75±2.17%	88.00±2.68%	87.75±2.38%
Active_carbs	70.33±3.70%	78.67±2.01%	87.83±2.79%	87.75±2.74%

6.1 Binary classification

The commonly established risk threshold for BGC level stands at 180 mg/dL [38], [43]. Consequently, as part of the initial classification task, measurements were automatically categorized in a binary manner to determine whether this threshold had been surpassed or not. It's important to note that this classification leads to an inherent imbalance, with a higher proportion falling into the "below-threshold" class (65.5%) compared to the "above-threshold" class (35.5%). Nonetheless, it's essential to recognize that this distribution may vary among individual subjects and is indicative of their unique medical conditions.

The classification procedure closely paralleled that of the regression task, with a few key distinctions. The data were labeled binarily, consequently, in the final stage of the classification model, a sigmoid activation function [54] was employed to map the model's output to a range between 0 and 1. Therefore,

binary_crossentropy, which measures the dissimilarity between predicted probabilities and actual class labels, was used as loss function. Finally, macro-averaged accuracy was used as evaluation metric. This metric assesses the model's performance by treating each class independently and subsequently averaging the accuracy scores for all classes. Specifically:

$$\text{Macro-avg accuracy} = \frac{acc_1 + acc_2 + ... + acc_N}{N} \tag{6.1}$$

where:
N is the total number of classes.

$$\text{acc}_i = \frac{TP_i + TN_i}{TP_i + FP_i + TN_i + FN_i} \tag{6.2}$$

TP_i is the number of True Positives for class i.
FP_i is the number of False Positives for class i.
TN_i is the number of True Negatives for class i.
FN_i is the number of False Negatives for class i.

6.1.1 Experimental results

Table 6.2 displays the performance of *General* and *Threetask* models for each subject, as measured by their macro-averaged accuracy scores.

Table 6.2: Binary classification: macro-averaged accuracy scores (%) for each subject across all time horizons on *Smoothed* dataset.

	PH = 6		**PH = 12**		**PH = 18**		**PH = 24**	
PID	**General**	**Threetask**	**General**	**Threetask**	**General**	**Threetask**	**General**	**Threetask**
559	91%	91%	82%	83%	71%	73%	62%	62%
563	86%	87%	77%	78%	66%	67%	58%	55%
570	93%	93%	86%	86%	78%	82%	73%	79%
575	86%	87%	75%	77%	67%	68%	62%	65%
588	87%	87%	78%	79%	71%	73%	63%	60%
591	88%	88%	79%	78%	67%	70%	59%	63%
540	86%	87%	78%	79%	72%	73%	65%	68%
544	90%	90%	80%	80%	70%	70%	62%	63%
552	85%	84%	78%	78%	71%	71%	64%	64%
567	81%	81%	68%	69%	64%	65%	58%	57%
584	88%	85%	77%	76%	68%	62%	62%	47%
596	88%	89%	77%	79%	71%	73%	62%	57%
Overall	87.42%	87.42%	77.92%	78.5%	69.67%	70.58%	62.50%	61.67%

The performance of both architectural models is remarkably similar and comparable, particularly when considering the 30-minute time horizon. Notably, the *Threetask* model exhibits a slight performance advantage for PH = 60min and PH = 90min, following the trends observed in the previous BGC level estimation task. Across the subjects, it is apparent that subjects 567 and 584 exhibit the lowest model performance. Thus, these subjects also present the highest proportion of missing BGC values, accounting for 24.78% and 20.13% of the total, respectively. Conversely, subjects 559, 570, and 544 demonstrate the highest levels of accuracy among the subjects studied. Furthermore, it's worth noting that, consistent with the observations made during the regression task, the *Threetask* model, while generally performing slightly better across most patients, tends to produce significantly larger errors when it does make mistakes, compared to its general counterpart. Specifically, it is worth highlighting a significant decline in accuracy for subject 584 at PH = 120min, where accuracy drops notably from 62% to 47%.

Next, t-tests and paired t-tests were applied in a manner similar to the regression task, comparing the distributions of the accuracy of the patients for each time horizon between the *General* and *Threetask* models, with a significance level of 0.05 (i.e., 95% confidence level). While the results from both tests revealed statistical equivalence for all combinations, it's essential to note that the assumption of normality, as assessed by the Shapiro-Wilk test, was rejected for the distribution of relative accuracy in the case of the *General* model at the 120-minute time horizon. This outcome may slightly diminish the reliability of the related result.

This task served as our initial focal point for classification. As we progressed into the subsequent work, the complexity of the task was increased by moving into a multi-class classification framework encompassing four distinct classes. Furthermore, we conducted a series of different tests, with the aim of evaluating the impact of personalization more thoroughly.

6.2 Multi-class classification

To offer prospective end users a more comprehensive insight into their BGC levels, we approached the classification task with increased complexity, embracing a multi-class perspective. In particular, we divided the BGC levels into four distinct classes based on quartiles derived from the complete distribution of glucose levels across all subjects within the OhioT1DM dataset. The entire distribution spans a broad range, extending from a minimum of 40 mg/dL to a maximum of 400 mg/dL. This categorization resulted in the following class descriptions:

- Class 0 (low risk): BGC $\leq$ 113 mg/dL.
- Class 1 (medium risk): 113 mg/dL < BGC $\leq$ 151 mg/dL.
- Class 2 (high risk): 151 mg/dL < BGC $\leq$ 197 mg/dL.
- Class 3 (very high risk): BGC > 197 mg/dL.

By using quartiles to establish class boundaries, we have achieved an overall balanced distribution among these classes, even though individual subjects may not exhibit the same proportions.

Similar to the previous task for glucose level estimation, our modeling methodology remained consistent. In this instance, we continued to work with the *General* and *Threetask* models, training and evaluating them on the *Smoothed* dataset using Leave-One-Out Cross-Validation across four different time horizons (30, 60, 90, and 120 minutes).

The process paralleled that of the binary classification task, but with the distinction of being a multi-class classification problem. To accommodate this, we employed the *categorical_crossentropy* loss function, which measures the dissimilarity between predicted probability distributions and the actual categorical labels. Lower values of *categorical_crossentropy* signify a more accurate alignment between predictions and real labels. Moreover, we applied the softmax activation function [55] in the final layer of the models. This function transforms a vector of numerical values into a probability distribution, assigning probabilities to different classes. The sum of the probabilities is always equal to 1, and at the end of the process the class with the highest probability has been selected as the output target.

6.2.1 BGC-only predictions

Another critical aspect to assess in predicting BGC levels in diabetic patients is the practical utility of supplementary features, of which many are offered by the OhioT1DM dataset. In most prior research [28], additional self-monitored features, including finger-stick glucose levels (FS), insulin bolus records (BI), and meal-related carbohydrate values (C), have been used alongside BGC levels. However, as highlighted in Section 3.2 on exploration, few subjects carefully reported the values of these features, especially the carbohydrate intake.

To cover and analyze this point, we conducted the same multi-class classification task in a univariate way, focusing solely on BGC levels from the *Smoothed* dataset. To ensure a fair comparison, we retained the *General* and *Threetask* models while keeping the training methodology unchanged. This univariate approach not only simplified both training and the overall task but may also enhance practical applicability. Should the obtained results demonstrate comparability to those achieved with the inclusion of the three additional features, it would indicate that predicting BGC levels can be effectively conducted without patients having to meticulously record this supplementary information. This streamlining of the process could enhance its real-world feasibility.

The following subsection assesses and provides insights into the outcomes of the two distinct approaches —standard and BGC-only— examining their individual performance from the personalization perspective, as well as offering a comparative analysis between them.

6.2.2 Experimental results

Similar to binary classification, Table 6.3 below presents macro-averaged accuracy scores for each subject, distinguishing between the *General* and the *Threetask* models.

Table 6.3: Multi-class classification: macro-averaged accuracy scores (%) for each subject across all time horizons on *Smoothed* dataset.

	PH = 6		**PH = 12**		**PH = 18**		**PH = 24**	
PID	**General**	**Threetask**	**General**	**Threetask**	**General**	**Threetask**	**General**	**Threetask**
559	70%	78%	53%	61%	46%	50%	38%	42%
563	74%	82%	56%	64%	46%	51%	36%	41%
570	76%	82%	61%	58%	56%	37%	45%	27%
575	69%	79%	53%	61%	42%	49%	36%	41%
588	68%	77%	55%	60%	45%	49%	36%	41%
591	67%	77%	55%	58%	44%	48%	37%	40%
540	70%	78%	52%	60%	42%	48%	36%	40%
544	70%	80%	54%	64%	44%	51%	37%	39%
552	72%	79%	57%	61%	47%	49%	41%	40%
567	63%	77%	50%	58%	38%	45%	36%	37%
584	68%	75%	52%	52%	41%	34%	34%	23%
596	75%	81%	61%	66%	51%	54%	39%	44%
Overall	70.17%	78.75%	54.92%	60.25%	45.17%	47.08%	37.58%	37.92%

The results obtained are remarkably intriguing, not primarily due to performance, which tends to deteriorate notably beyond the 60-minute time horizon, but because of the important and clearly visible impact of customization on this task. The *Threetask* model consistently enhances the overall performance of the *General* model, achieving an increase of over 8% in accuracy for the PH = 30min time horizon and approximately 5% for PH = 60min. The *Threetask* model outperforms the *General* model in predicting nearly all patients across all time horizons. Nevertheless, as previously noticed when conducting the estimation task, it's worth highlighting that, although the model generally surpasses its general counterpart, its errors tend to be more conspicuous (e.g., a -19% accuracy drop for subject 570 at PH = 90min). Furthermore, the performance gap between the two architectures narrows as the time horizon extends, underscoring the *Threetask* model's greater efficacy for shorter time horizons.

Statistical tests were once again employed. Initially, the normality of the data was assessed using the Shapiro-Wilk test. Subsequently, paired t-tests and t-tests were conducted to compare the accuracy distributions for each time horizon. The outcomes of these two tests outline a statistically significant difference for the 30 and 60-minute time horizons, whereas for the 90 and 120-minute horizons, the null hypothesis of equality cannot be rejected. However, it is important to note that the test results for PH = 90min for the *Threetask* model and PH = 120min for both models might be influenced by the rejection of the null hypothesis of normality.

Continuing our analysis, Table 6.4 below provides an overview of the classification outcomes for each subject and time horizon, with the sole consideration of BGC level as a training feature.

Table 6.4: Multi-class classification BGC-only: macro-averaged accuracy scores (%) for each subject across all time horizons on *Smoothed* dataset.

	PH = 6		**PH = 12**		**PH = 18**		**PH = 24**	
PID	**General**	**Threetask**	**General**	**Threetask**	**General**	**Threetask**	**General**	**Threetask**
559	69%	78%	52%	61%	46%	49%	38%	42%
563	74%	82%	57%	64%	46%	51%	35%	41%
570	77%	83%	60%	65%	56%	40%	47%	27%
575	69%	79%	52%	61%	43%	49%	36%	43%
588	69%	77%	54%	59%	46%	49%	36%	40%
591	67%	76%	55%	58%	43%	48%	38%	40%
540	70%	78%	52%	60%	43%	48%	38%	39%
544	71%	80%	54%	64%	45%	50%	37%	41%
552	72%	79%	57%	62%	47%	50%	42%	42%
567	63%	78%	50%	57%	39%	45%	36%	36%
584	69%	77%	54%	50%	42%	33%	33%	29%
596	76%	80%	61%	65%	50%	54%	41%	43%
Overall	70.50%	78.92%	54.83%	60.50%	45.50%	47.17%	38.08%	38.58%

Interestingly, the results closely mirror those achieved when incorporating the other three self-reported features. In fact, a marginal improvement is observed in overall accuracy across the table, except for the *General* model at PH = 60min. This striking similarity in performance between the two distinct approaches prompts several inquiries about the actual utility of the additional features, opening avenues for future research aimed at devising a less complex and more user-friendly approach. Thus, users would only need to passively wear the BGC level control device to receive the predictions. Furthermore, the impact of customization on the overall model appears to be nearly identical to the previous scenario, resulting in higher macro-averaged accuracy, particularly for time horizons of 30 and 60 minutes.

Statistical t-tests, conducted on the accuracy distributions across different time horizons for both the standard and the BGC-only approaches, affirm the null hypothesis of identical average expected values for each combination.

The following section focuses on the integration of the BIG IDEAs dataset [52], containing BGC level data from 16 subjects, recorded every five minutes over a span of 10 consecutive days. This approach aims to assess the influence of incorporating new patients into the model training process.

6.3 Data integration

After conducting the assessment of customization's influence on both binary and multi-class classification problems, as well as exploring the impact of incorporating self-reported features to enhance model perfor-

mance, we arrived at a notable finding: the BGC-only approach exhibits effectiveness comparable to the conventional approach (i.e., with the addition of the three self-reported features FS, BI and C).

With this insight in mind, we proceeded to investigate the potential gains associated with the integration of an additional dataset. Our aim was to discern whether the inclusion of data from additional subjects could yield improvements in the final performance of our models. To begin this investigation, we initially carried out data pre-processing and integration for 16 subjects from the BIG IDEAs dataset. Subsequently, we developed a novel model named *Fourtask* based on insights gained from this new dataset. The *Fourtask* model adheres to a similar philosophy as our previous *Threetask* model, with the noteworthy inclusion of an extra branch to further categorize patients.

The results are showcased through a comparison of macro-averaged accuracy among the *General*, *Threetask*, and *Fourtask* models. This assessment aims to gauge not only the influence of incorporating new subjects but also the impact of employing two distinct personalization strategies in contrast to the generalized model.

6.3.1 The BIG IDEAs dataset

The dataset under consideration comprises continuous measurements of BGC levels from 16 distinct patients over a span of 10 consecutive days. More specifically, the patients, seven males and nine females, wore a Dexcom 6 continuous glucose monitor and an Empatica E4 wristband. Moreover, they received a standardized breakfast meal every day. Research data collected include physiological measurements from wearable devices such as heart rate, accelerometry, and electrodermal conductance. All data are time-shifted by date to prevent identification.

The Dexcom G6 device measured interstitial glucose concentration (mg/dL) every five minutes, while the Empatica E4 measured photoplethysmography (PPG), electrodermal activity (EDA), skin temperature, and tri-axial accelerometry, resulting in a total of seven features. However, for the specific task of dataset integration, our focus was solely on the feature associated with blood glucose monitoring.

The data are provided already synchronized in CSV format, maintaining a consistent BGC level recording every five minutes. Consequently, there was no requirement for parsing or additional synchronization steps. Regarding pre-processing, we applied the same methodology as utilized for the *Smoothed* version of the primary OhioT1DM dataset. Finally, it's worth noting that there are no missing values in the BIG IDEAs dataset.

To gain a more comprehensive understanding of disease severity and the distribution of BGC values among different patients, Figure 6.1 offers insightful violin plots illustrating the BGC distributions for each individual.

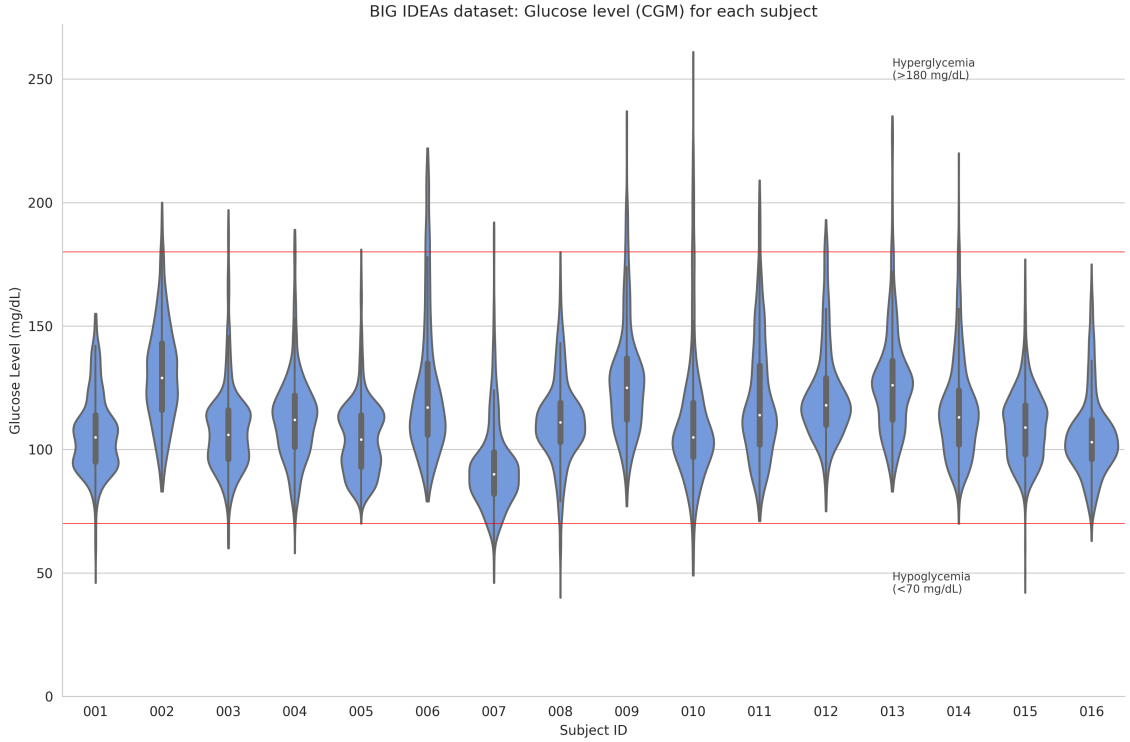


Fig. 6.1: BGC median level for each subject of the BIG IDEAs dataset.

In contrast to the subjects in the OhioT1DM dataset, patients in the BIG IDEAs dataset generally exhibit lower levels of disease severity. Among them, only patients with IDs 002, 009, and 013 display a median BGC level surpassing 120 mg/dL, whereas patient 007 demonstrates a notably lower median value of 90 mg/dL.

6.3.2 Deep Fourtask model

Building upon the *Threetask* model's rationale, we applied a similar approach to categorize subjects from both the OhioT1DM and BIG IDEAs datasets into four distinct groups, each corresponding to a different branch within the neural network. However, due to the contrasting clinical data available in these two datasets, we tried a different categorization approach. Since the patients in the BIG IDEAs dataset have a much lower median BGC level, compared to the OhioT1DM dataset, we opted for a more data-driven strategy. Instead of relying on predefined thresholds provided by the ADA [43], we leveraged the distribution of BGC level values. To achieve this, we segmented the BGC level distribution into quartiles and assigned patients in both datasets to their respective classes accordingly.

This approach led to the establishment of the following categorization:

- Low (L). BGC median level is under 113 mg/dL (subjects 001, 003, 004, 005, 006, 007, 008, 010, 011, 012, 014, 015, 016 from BIG IDEAs).
- Normal (N). BGC median level is between 113 mg/dL and 151 mg/dL (subjects 563, 575, 540, 552, 596 from OhioT1DM and subjects 001, 002, 009, 013 from BIG IDEAs).
- Medium (M). BGC median level is between 151 mg/dL and 197 mg/dL (subjects 559, 588, 591, 544, 567 from OhioT1DM).
- High (H). BGC median level is over 197 mg/dL (subjects 570, 584 from OhioT1DM).

To maintain a fair and consistent basis for comparison, the hyperparameters and core architectural structure remained unchanged, as illustrated in Figure 6.2. The training and evaluation methodology also remained consistent with the *General* and *Threetask* models. The results of the experiment are reported in the following subsection.

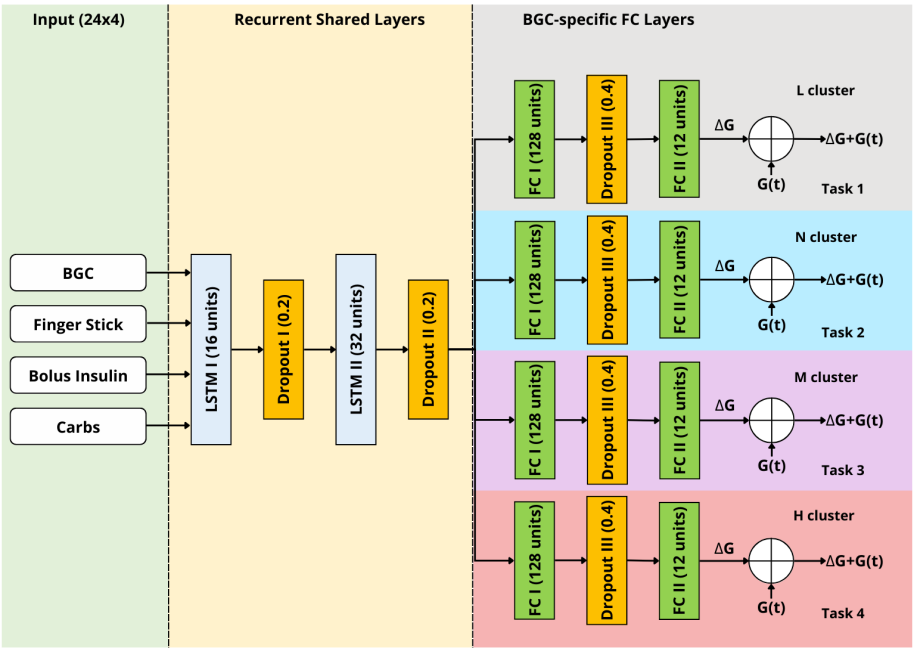


Fig. 6.2: Deep Fourtask model architecture. The model has the same general architecture and hyperparameters as the *Multitask* and *General* models, but with a different clustering strategy based on the quartiles of the BGC distributions among the various subjects.

6.3.3 Experimental results

At this point, we assessed and compared the performance of the *General*, *Threetask*, and *Fourtask* models using the dataset comprising patients from OhioT1DM and BIGIDEAs. It's important to note that subjects

from the BIGIDEAs dataset were exclusively utilized for model training, while the evaluation pertained to the subjects within the primary OhioT1DM dataset. The results for the *Fourtask* model, compared to the other models, expressed in terms of macro-averaged accuracy, allow to assess the impact of employing a distinct personalization strategy on overall performance. Worth noting is that, following the integration of the two datasets, these predictions are based solely on the BGC feature. The LOSO-CV results for the *General*, *Threetask* and *Fourtask* models on the integrated dataset are shown in Table 6.5

Table 6.5: Multi-class classification BGC-only: macro-averaged accuracy scores (%) for each subject across all time horizons. The models are trained on the *Smoothed* integrated dataset, comprising subjects from OhioT1DM and BIG IDEAs.

	PH = 6			**PH = 12**			**PH = 18**			**PH = 24**		
PID	**General**	**3-task**	**4-task**	**General**	**3-task**	**4-task**	**General**	**3-task**	**4-task**	**General**	**3-task**	**4-task**
559	69%	78%	78%	54%	60%	60%	45%	50%	49%	39%	42%	43%
563	74%	82%	82%	56%	64%	63%	44%	50%	51%	36%	39%	41%
570	77%	83%	83%	60%	67%	70%	54%	39%	60%	46%	28%	50%
575	69%	79%	79%	60%	60%	61%	42%	50%	51%	36%	43%	42%
588	70%	77%	76%	52%	60%	58%	43%	49%	48%	36%	41%	40%
591	67%	77%	76%	53%	58%	58%	44%	48%	48%	38%	41%	42%
540	70%	78%	78%	54%	60%	60%	43%	48%	48%	36%	39%	39%
544	71%	80%	80%	52%	64%	64%	44%	51%	50%	36%	42%	41%
552	72%	79%	79%	54%	62%	61%	47%	50%	50%	44%	41%	42%
567	62%	77%	77%	58%	56%	57%	39%	34%	45%	37%	37%	37%
584	68%	74%	75%	49%	51%	56%	43%	33%	44%	36%	22%	37%
596	75%	81%	81%	53%	65%	65%	50%	53%	54%	41%	44%	43%
Overall	70.33%	78.75%	78.67%	54.67%	60.58%	61.08%	44.83%	46.25%	49.83%	38.42%	38.25%	41.42%

In general, the personalized approach still exhibits superior performance when compared to the *General* model across all time horizons, except for the *Threetask* model at PH = 120min. The introduction of new patients led only to a slight increase in performance for both the *General* and *Threetask* models, and the results remain largely comparable for each prediction horizon. Furthermore, the *Threetask* model continues to exhibit the same issue identified earlier: despite its superior overall performance, when it makes errors, those errors tend to be quite substantial. For instance, it registers a significant -18% accuracy deviation for subject 570 at PH = 120min, compared to the *General* model. Overall, it's worth noting that the *General* and *Threetask* models did not appear to derive significant performance benefits from the inclusion of the BIG IDEAs dataset to augment the OhioT1DM dataset.

Statistical tests (t-test and paired t-test) were conducted for each time horizon. These tests were applied to both the *General* and *Threetask* models, with and without the inclusion of patients from the BIG IDEAs

dataset. Notably, all of these tests concluded by accepting the null hypothesis, indicating that the average expected values of the accuracy vectors with and without integration were statistically equal in all cases.

Regarding the *Fourtask* approach, it surpasses the overall performance of both the *General* and *Threetask* models for all time horizons exceeding 30 minutes, for which its accuracy aligns closely with that of the *Threetask* model. Importantly, the incorporation of an additional branch and the distinct subject partitioning based on quartiles within the OhioT1DM and BIG IDEAs datasets effectively address the issue of high magnitude errors observed in the *Threetask* model. This approach mitigates severe drops in accuracy, and even in the worst-case scenarios, the performance remains in line with that of the *General* model.

In summary, the personalized approach (and notably the *Fourtask* model), proves to be more effective overall, while the incorporation of subjects from the BIG IDEAs dataset yielded very limited improvements. Table 6.6 provides a comprehensive overview of the results obtained through the different approaches in terms of macro-averaged accuracy.

Table 6.6: Overall multi-class classification results: macro-averaged accuracy scores (%) and standard deviation across subjects on *Smoothed* dataset, for each prediction horizon.

	Standard		**BGC-only**		**BGC-only - integrate BIG IDEAs**		
PH	**General**	**Threetask**	**General**	**Threetask**	**General**	**Threetask**	**Fourtask**
30-min	70.17±3.51%	78.75±2.09%	70.50±3.71%	78.92±1.98%	70.33±3.79%	78.75±2.38%	78.67±2.39%
60-min	54.92±3.28%	60.25±3.49%	54.83±3.21%	60.50±4.07%	54.67±3.35%	60.58±4.15%	61.08±3.77%
90-min	45.17±4.54%	47.08±5.58%	45.50±%4.15	47.17±%5.37	44.83±3.76%	47.17±5.46%	49.83±4.00%
120-min	37.58±2.81%	37.92±6.05%	38.08±3.57%	38.58±5.09%	38.42±3.33%	38.25±6.31%	41.42±3.25%

The final results yield the following findings:

- **Degradation in performance**: Overall, performance for this classification task begins to deteriorate significantly as early as the 60-minute relative time horizon.
- **Customization's impact**: Customization plays a key role, with the *Threetask* model generally outperforming the *General* model across most subjects. However, it occasionally exhibits more noticeable errors for specific patients as the time horizon increases.
- **Limited effect of additional features**: The inclusion of additional features (FS, BI, and C) doesn't seem to have a substantial impact on model performance. The BGC-only approach stands out for its ability to reduce task complexity, and enhance user-friendliness without noticeable declines in performance.
- **Marginal gain from BIG IDEAs dataset**: The introduction of subjects from the BIG IDEAs dataset results in only a slight improvement in performance. However, the *Fourtask* model's tailored approach on the integrated dataset outperforms the other models, effectively mitigating the *Threetask* model's error magnitudes for certain individuals as the time horizon increases.

7

Conclusions

Efficient blood glucose forecasting plays a crucial role in the continuous monitoring of diabetic patients. Understanding the dynamics of blood glucose concentration is essential for effectively managing diabetes, ensuring proper treatment, and gaining insights into the progression of the disease.

This thesis contributes to the pursuit of accurate glucose monitoring. By leveraging data from a diverse group of 12 subjects sourced from the OhioT1DM dataset, we first selected the four widely used features from previous works. These factors encompassed continuous glucose monitoring (BGC), finger stick measurements (FS), as well as self-reported data on bolus injected insulin (BI) and carbohydrate intake (C) values. While the patients themselves provided information on the latter three features, the BGC data was collected continuously via insulin pumps.

Subsequently, we implemented various pre-processing techniques, drawing inspiration from contemporary advancements in the field. Notably, we transformed self-reported events concerning bolus insulin injections and carbohydrate intake into continuous features. Additionally, we addressed the challenge of handling the numerous missing values within the OhioT1DM dataset, which in several circumstances span entire days. To assess the efficacy of our pre-processing strategies, we employed the state-of-the-art Deep Multitask model. This neural network comprises two stacked LSTM layers and utilizes a dual-branch approach. The training process is personalized, initially considering patient gender, and subsequently adapting to each subject's unique characteristics. Ultimately, we chose the *Smoothed*, *Averaged*, and *Active_carbs* datasets for further analysis and experimentation.

The utilization of personalized models for predicting blood glucose levels undoubtedly represents a promising approach in this regard. Nonetheless, current research lacks effective comparison between the customization strategies and the generic version. To address this gap, we conducted a glucose level estimation task using Leave-One-Subject-Out Cross-Validation. The evaluations for the regression task were conducted on three distinct datasets, namely the previously mentioned *Smoothed*, *Averaged*, and *Active_carbs*, and we assessed the MAE and RMSE values across four different time horizons (30, 60, 90, and 120 minutes), and employing three different models: a *General* model, the *Multitask* model, and a novel customized model referred to as *Threetask*. The *Threetask* model's approach leverages the median BGC level specific to each patient for personalized predictions.

While the *General* model not only matched but, in several instances, even outperformed the *Multitask* model that served as our reference, the *Threetask* model exhibited an overall performance comparable to the *General* model, whereas displaying a superior adaptability to the unique characteristics of individual subjects. It's worth noting, however, that when the *Threetask* model does make errors, the magnitude of those errors is notably larger compared to the *General* model. This suggests that while the *Threetask* model excels in delivering personalized predictions, it may occasionally face challenges that result in more substantial prediction errors. Furthermore, it is important to underscore that, throughout the study, we maintained consistency and ensured the comparability of results, by keeping the hyperparameters, number, and type of layers unchanged. This approach was chosen to guarantee an equitable and meaningful comparison with the state-of-the-art *Multitask* model.

To conclude the regression task, we proceeded with fine-tuning the most promising *Threetask* model using a random search approach, aiming to further improve its performance. During the testing, we noticed that the utilization of GRUs instead of LSTMs proved advantageous.

Additionally, we experimented with an incremental learning procedure in an attempt to enhance the model's customization to individual patient data. Regrettably, the technique did not achieve optimal results.

Subsequently, the research extended into a classification task. Following an evaluation of the *General* and *Threetask* models in a binary classification task using the 180 mg/dL blood glucose level threshold, the models were subjected to a more intricate challenge: a multi-class classification problem achieved by dividing the BGC distribution into quartiles. Specifically, we assessed the performance in terms of macro-averaged accuracy across four different time horizons (30, 60, 90, and 120 minutes). In this scenario, the customized *Threetask* model consistently outperformed the *General* model across all time horizons, particularly on the 30-minute relative horizon.

Moving forward, we investigated the influence of the supplementary self-reported FS, BI, and C features on model performance by comparing them with a BGC-only approach. Interestingly, we discovered that the BGC-only approach yielded consistent results, comparable to those obtained through the standard approach. This finding carries a multitude of advantages, including reduced complexity, the potential to seamlessly integrate other datasets solely based on BGC values, and enhanced user-friendliness as end users are relieved from the task of recording additional variables' values.

Lastly, we assessed the impact of incorporating 10 days' worth of BGC data from 16 subjects within the BIG IDEAs dataset into the training process. While the improvement observed was marginal, the novel personalized *Fourtask* model, based on the same key idea of the median BGC values as the *Threetask* approach, but with a different split, surpassed the performance of both the *General* and *Threetask* models. Importantly, it effectively mitigated the issue of the *Threetask* related to the magnitude of prediction errors for certain subjects as the time horizon extended.

In conclusion, the study conducted on the OhioT1DM dataset underscores the substantial impact of personalization in forecasting the BGC levels of individuals with type 1 diabetes. The ability for users to possess prior knowledge of glucose levels across various time horizons holds the potential to proactively avoid hyperglycemic events through timely and informed preventive measures.

Future work

The promising findings of this thesis open the door to several avenues for future research:

- **Efficient imputation of missing values**: The OhioT1DM dataset contains numerous missing values, particularly concerning BGC levels. These gaps in data often span entire days, making them unsuitable for straightforward methods like linear interpolation. Implementing an efficient imputation technique has the potential to significantly improve the predictions.
- **New personalization approaches**: Investigate alternative methods for personalizing the deep learning models to further improve their performance in glucose level estimation and classification tasks.
- **Integration of subjects**: Explore the integration of different subjects to enhance the generalizability of the models and their applicability to a broader population of individuals with type 1 diabetes.
- **Other machine learning techniques**: Experiment with different machine learning techniques, such as reinforcement learning or unsupervised learning, to discover new insights and potentially improve the performance of the models in predicting and classifying glucose levels.
- **Long-term glucose level prediction**: Extend the current models to predict glucose levels over longer time horizons, which could be beneficial for individuals with type 1 diabetes in planning their insulin dosages and managing their blood sugar levels more effectively.
- **Integration with wearable devices**: Investigate the feasibility of integrating the developed models with wearable devices, such as continuous glucose monitors or insulin pumps, to provide real-time glucose level predictions and personalized recommendations for insulin dosages.

References

[1] T. Robinson, S. Linklater, F. Wang, S. Colagiuri, C. Beaufort, K. Donaghue, D. Magliano, J. Maniam, T. Orchard, P. Rai, and G. Ogle, "Global incidence, prevalence, and mortality of type 1 diabetes in 2021 with projection to 2040: a modelling study," *The Lancet Diabetes & Endocrinology*, vol. 10, 09 2022.

[2] G. N. Sierra, "The global pandemic of diabetes," *African Journal of Diabetes Medicine*, vol. 17, no. 11, pp. 4–8, 2009.

[3] A. Mandal, "What is type 1 diabetes?" News Medical Life Sciences, 07 2023, last accessed: 10th Aug 2023. [Online]. Available: https://www.news-medical.net/health/What-is-Type-1-Diabetes.aspx

[4] T. K. Chen, D. H. Knicely, and M. E. Grams, "Chronic kidney disease diagnosis and management: a review," *Jama*, vol. 322, no. 13, pp. 1294–1304, 2019.

[5] W. T. Cade, "Diabetes-related microvascular and macrovascular diseases in the physical therapy setting," *Physical therapy*, vol. 88, no. 11, pp. 1322–1335, 2008.

[6] W. Sun, Z. Guo, Z. Yang, Y. Wu, W. Lan, Y. Liao, X. Wu, and Y. Liu, "A review of recent advances in vital signals monitoring of sports and health via flexible wearable sensors," *Sensors*, vol. 22, no. 20, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/20/7784

[7] A. J. Perez and S. Zeadally, "Recent advances in wearable sensing technologies," *Sensors*, vol. 21, no. 20, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/20/6828

[8] R. Abderahman, R. Karim, T. Horst, A. Andrea, A. Salem, A. Yaser, and I. Mohammad, "The internet of things (iot) in healthcare: Taking stock and moving forward," *Internet of Things*, vol. 22, p. 100721, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2542660523000446

[9] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "Ai in health and medicine," *Nature Medicine*, pp. 1–8, 2022.

[10] Y. Mei, "Modeling and control to improve blood glucose concentration for people with diabetes," Ph.D. dissertation, Iowa State University, 2017.

[11] M. Baig, H. Gholamhosseini, A. Moqeem, F. Mirza, and M. Lindén, "A systematic review of wearable patient monitoring systems – current challenges and opportunities for clinical adoption," *Journal of Medical Systems*, vol. 41, 06 2017.

[12] B. Kerstin, B. Razvan C., M. Cindy, and W. Nirmalie, Eds., *Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data co-located with 24th European Conference on Artificial Intelligence, KDH@ECAI 2020, Santiago de Compostela, Spain & Virtually, August 29-30, 2020*, ser. CEUR Workshop Proceedings, vol. 2675. CEUR-WS.org, 2020. [Online]. Available: https://ceur-ws.org/Vol-2675

[13] J. Daniels, P. Herrero, and P. Georgiou, "Personalised glucose prediction via deep multitask networks," in *KDH@ECAI*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221793894

[14] R. Bevan and F. Coenen, "Experiments in non-personalized future blood glucose level prediction," in *KDH@ECAI*, 2020.

[15] G. Cappon, L. Meneghetti, F. Prendin, J. Pavan, G. Sparacino, S. Del Favero, and A. Facchinetti, "A personalized and interpretable deep learning based approach to predict blood glucose concentration in type 1 diabetes," in *KDH@ECAI*, 2020.

[16] R. Pascanu, T. Mikolov, and Y. Bengio, "Understanding the exploding gradient problem," *CoRR*, vol. abs/1211.5063, 2012. [Online]. Available: http://arxiv.org/abs/1211.5063

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

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

[19] Z. Angehrn, L. Haldna, A. S. Zandvliet, E. Gil Berglund, J. Zeeuw, B. Amzal, S. A. Cheung, T. M. Polasek, M. Pfister, T. Kerbusch *et al.*, "Artificial intelligence and machine learning applied at the point of care," *Frontiers in Pharmacology*, vol. 11, p. 759, 2020.

[20] C. Marling and R. Bunescu, "The ohiot1dm dataset for blood glucose level prediction: Update 2020," *CEUR workshop proceedings*, vol. 2675, pp. 71–74, 09 2020.

[21] H. Amrani, D. Micucci, and P. Napoletano, "Personalized models in human activity recognition using deep learning," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 01 2021, pp. 9682–9688.

[22] S. Ho and M. Xie, "The use of arima models for reliability forecasting and analysis," *Computers & Industrial Engineering*, vol. 35, no. 1, pp. 213–216, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360835298000667

[23] A. C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1990.

[24] R. H. Shumway and D. S. Stoffer, "Arima models," *Time series analysis and its applications: with R examples*, pp. 75–163, 2017.

[25] R. McShinsky and B. Marshall, "Comparison of forecasting algorithms for type 1 diabetic glucose prediction on 30 and 60-minute prediction horizons," in *KDH@ECAI*, 2020.

[26] A. Bhimireddy, P. Sinha, B. Oluwalade, J. W. Gichoya, and S. Purkayastha, "Blood glucose level prediction as time-series modeling using sequence-to-sequence neural networks," *CEUR workshop proceedings*, 2020. [Online]. Available: https://par.nsf.gov/biblio/10188463

[27] J. Freiburghaus, A. Rizzotti, and F. Albertetti, "A deep learning approach for blood glucose prediction of type 1 diabetes," in *Proceedings of the Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data co-located with 24th European Conference on Artificial Intelligence (ECAI 2020), 29-30 August 2020, Santiago de Compostela, Spain*, 2020.

[28] H. Butt, I. Khosa, and M. A. Iftikhar, "Feature transformation for efficient blood glucose prediction in type 1 diabetes mellitus patients," *Diagnostics*, vol. 13, no. 3, 2023. [Online]. Available: https://www.mdpi.com/2075-4418/13/3/340

[29] J. Beauchamp, R. C. Bunescu, and C. Marling, "A general neural architecture for carbohydrate and bolus recommendations in type 1 diabetes management," in *KDH@ ECAI*, 2020, pp. 43–47.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[31] T. Zhu, X. Yao, K. Li, P. Herrero, and P. Georgiou, "Blood glucose prediction for type 1 diabetes using generative adversarial networks," in *CEUR Workshop Proceedings*, vol. 2675, 2020, pp. 90–94.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Advances in neural information processing systems*, vol. 27, 2014.

[33] H. Rubin-Falcone, I. Fox, and J. Wiens, "Deep residual time-series forecasting: Application to blood glucose prediction," in *KDH@ ECAI*, 2020, pp. 105–109.

[34] "Tidepool," last accessed: 26th Sep 2023. [Online]. Available: https://www.tidepool.org

[35] H. Nemat, H. Khadem, J. Elliott, and M. Benaissa, "Data fusion of activity and cgm for predicting blood glucose levels," in *Knowledge Discovery in Healthcare Data 2020*, vol. 2675. CEUR Workshop Proceedings, 2020, pp. 120–124.

[36] D. Joedicke, G. Kronberger, J. M. Colmenar, S. M. Winkler, J. M. Velasco, S. Contador, and J. I. Hidalgo, "Analysis of the performance of genetic programming on the blood glucose level prediction challenge 2020," in *KDH@ ECAI*, 2020, pp. 141–145.

[37] A. Z. Woldaregay, E. Årsand, S. Walderhaug, D. Albers, L. Mamykina, T. Botsis, and G. Hartvigsen, "Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes," *Artificial Intelligence in Medicine*, vol. 98, pp. 109–134, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365717306218

[38] Medtronic, "Symptoms of diabetes," 11 2022, last accessed: 31th Aug 2023. [Online]. Available: https://www.medtronicdiabetes.com/about-diabetes/symptoms-and-complications

[39] E. Kraegen, D. Chisholm, and M. E. McNamara, "Timing of insulin delivery with meals," *Hormone and Metabolic Research*, vol. 13, no. 07, pp. 365–367, 1981.

[40] LoopDoc, "Glucose prediction," last accessed: 3rd Sep 2023. [Online]. Available: https://loopkit.github.io/loopdocs/operation/algorithm/prediction

[41] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[42] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, pp. 41–75, 1997.

[43] A. Campbell, "Diabetes self-management," June 2019, last accessed: 27th Sep 2023. [Online]. Available: https://www.diabetesselfmanagement.com/managing-diabetes/blood-glucose-management/blood-sugar-chart/#:~:text=The%20American%20Diabetes%20Association%20recommends%20that%20the%20blood%20sugar%201,level%20in%20someone%20with%20diabetes

[44] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[45] Scipy, "ttest_ind function," last accessed: 8th Sep 2023. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html

[46] ——, "ttest_rel function," last accessed: 8th Sep 2023. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html

[47] ——, "Scipy library," last accessed: 8th Sep 2023. [Online]. Available: https://scipy.org/

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

[49] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[50] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[51] J. He, R. Mao, Z. Shao, and F. Zhu, "Incremental learning in online scenario," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 926–13 935.

[52] P. Cho, J. Kim, B. Bent, and J. Dunn, "Big ideas lab glycemic variability and wearable device data (version 1.1.2)," *PhysioNet*, 2023.

[53] B. Bent, P. J. Cho, M. Henriquez, A. Wittmann, C. Thacker, M. Feinglos, M. J. Crowley, and J. P. Dunn, "Engineering digital biomarkers of interstitial glucose from noninvasive smartwatches," *NPJ Digital Medicine*, vol. 4, no. 1, p. 89, 2021.

[54] N. Sridhar, "The generalized sigmoid activation function: Competitive supervised learning," *Information Sciences*, vol. 99, no. 1, pp. 69–82, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025596002009

[55] J. Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," *Advances in neural information processing systems*, vol. 2, 1989.

Acknowledgements

First and foremost, I wish to express my heartfelt gratitude to my thesis advisor, Professor Paolo Napoletano. His dedication to guiding me through this thesis, providing consistent feedback, and, above all, his humane approach have been invaluable.

I would also like to extend my appreciation to Dr. Flavio Piccoli for his unwavering availability and generous support, which has been fundamental in my thesis pathway.

I am deeply indebted to my classmates, who have made this academic voyage both engaging and rewarding. Their assistance, constructive feedback, and unwavering moral support have been of immense significance.

I extend my gratitude to Matilde, the person who has been closest to me over the past two years. Her understanding and her immovable belief in me have been pillars of strength.

I also want to express my heartfelt thanks to the friends who took time out of their busy schedules to visit me during these past two years. Their presence has been incredibly meaningful.

Lastly, I am profoundly thankful to my family for their support, understanding, and motivation throughout these past years. Their encouragement has been my driving force.