On the Use of Personalized Models for Blood Glucose Concentration Prediction

Niccolò Puccinelli
University of Milano-Bicocca
Milan, Italy
n.puccinelli@campus.unimib.it

Flavio Piccoli
University of Milano-Bicocca
Milan, Italy
flavio.piccoli@unimib.it

Paolo Napoletano
University of Milano-Bicocca
Milan, Italy
paolo.napoletano@unimib.it

2023 IEEE 13th International Conference on Consumer Electronics - Berlin (ICCE-Berlin) | 979-8-3503-2415-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCE-BERLIN58801.2023.10375621

***Abstract*—Patients with type-1 diabetes need to constantly monitor blood glucose concentration (BGC) level to stay in a healthy range. Consumer devices for BGC monitoring can be integrated with machine and deep learning techniques so that glucose level can be forecast and promptly provided to the patient.**

Recent advancements in the field suggest the use of a customization step based on each subject for blood concentration prediction. However, there is no comparison with other customization strategies and more importantly, there is no quantitative analysis on the benefits of such a customization.

In this paper: (1) we evaluate the impact of several preprocessing strategies on the performance; (2) we conduct a comparative analysis between 2 different customization methods and a general purpose strategy with no customization at all, and finally, (3) we propose a new personalization technique, called *Threetask*, that performs slightly better than other strategies on the majority of the patients, especially in the 60- and 90-minutes horizon.

Experiments have been conducted on the OhioT1DM dataset which contains eight weeks of continuous monitoring of Blood Glucose Concentration from 12 subjects.

***Index Terms*—Blood Glucose Concentration estimation, Deep Learning, Diabetes of Type 1**

I. Introduction

As the prevalence of type-1 diabetes mellitus continues to pose a significant global health concern [1], the continuous and accurate prediction of Blood Glucose Concentrations (BGC) has become a crucial factor for effective disease management [2]. To address this critical need, recent advancements in the field of wearable sensors [3], [4] leveraging Internet of Things (IoT) techniques [5] in conjunction with predictive algorithms [6] have enabled continuous monitoring of the patient's conditions. This real-time monitoring allows for prompt and preventive interventions, thereby minimizing the risk of life-threatening events [7]. Over the years, machine- and deep-learning techniques have demonstrated high accuracy and reliability in estimating BGC levels [8]. Various regression approaches have been explored in the context of type-1 diabetes prediction, ranging from traditional linear Autoregressive Integrated Moving Average (ARIMA) models [9] and Unobserved Components Model (UCM) models [10] to more advanced Machine Learning (ML) techniques such as Random Forest and XGBoost [11] as well as Deep Learning (DL) [12].

In recent years, DL has proven to be exceptionally effective, surpassing the capabilities of conventional ML approaches. The ability of DL models to automatically extract and combine relevant patterns from complex data made them particularly well-suited for this challenge. Several architectures, such as Convolutional Neural Networks (CNNs) [13] and Recurrent Neural Networks (RNNs) [14], [15] have been investigated. Among them, RNNs, in particular those employing Long Short-Term Memory (LSTM) units, have emerged as the most effective approach due to their ability to address the vanishing gradient problem [16]. However, these data-driven techniques face several challenges.

Firstly, available datasets often suffer from biases, heterogeneity, and incompleteness, making it difficult to cover all possible scenarios and resulting in limited availability of labeled data [17]. In this context, Butt et al. [18] propose a method for addressing sampling consistency, filling of missing samples, and filtering. In particular, they propose a transformation of self-reported features (carbohydrate and insulin bolus) into continuous variables.

Secondly, generalization remains a concern due to the unique characteristics and habits of individual users. Recent advancements in the field, in fact, suggest that a post-training personalization helps to significantly improve the detection accuracy on each patient. Rather than employing a one-size-fits-all approach, personalized models are tailored to account for the unique physiological and behavioral characteristics of individual patients. A recent paper by Shuvo et al. [12] propose a customization strategy to refine the deep-learning models based on gender and then individual subjects. However, this strategy has not been compared with other personalization strategies and also with a general purpose strategy with no personalization at all. In this context, personalization and customization ad interchangeable words.

To address the aforementioned limitations of current state of the art, in this work:

- we firstly investigate the impact of several pre-processing strategies on the performance;
- we conduct a comparative analysis between 2 different personalization techniques and a general strategy with no customization at all, and finally
- we propose a new personalization technique, called *Threetask*, that outperforms previous methods in the majority of the patients, especially in the 60- and 90-minutes horizon.

TABLE I
DETAILS OF THE OHIOT1DM DATASET. IT IS COMPOSED OF 12 PATIENTS AFFECTED BY TYPE-1 DIABIETES, COLLECTED IN TWO DIFFERENT TIMES, SIX IN 2018 AND SIX IN 2020.

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

All the experiments have been conducted on the OhioT1DM dataset [19] which contains eight weeks of continuous monitoring of Blood Glucose Concentration from 12 subjects.

II. DATASET

The OhioT1DM clinical dataset is widely employed by several works [8] for BGC prediction. It contains eight weeks' worth of Continuous Glucose Monitoring (CGM) of BGC, recorded every five minutes, and optional finger stick glucose (FG) measured directly by the patients. Besides, it includes insulin, physiological sensor and self-reported life-event data. The dataset, whose first version in 2018 consisted of 6 subjects, was then updated in 2020 with data from 6 more subjects. In order to mantain privacy, subjects were assigned a random patient identification (PID). The 2018 and 2020 subjects differ mainly in the type of sensor band used: Basis Peak for 2018 and Empatica Embrace for 2020. These two types of sensors are responsible for physiological data and differ in granularity and type of observations measured. Data are provided already divided between training and testing, with a ratio of about 75%-25%, respectively. Table I provides a clear summary of the dataset.

Although numerous features are available for use, current advancements [18], including state-of-the-art research [12], have predominantly focused on utilizing the following the four primary features:

- *glucose_level* (BGC): Continuous glucose monitoring data, recorded every 5 minutes.
- *finger_stick* (FS): Blood glucose values obtained through self-monitoring by the patient.
- *bolus* (BI): Insulin delivered to the patient, typically before a meal or when the patient is hyperglycemic.
- *meal* (C): The self-reported carbohydrate estimate for the meal.

To ensure consistency and facilitate the evaluation and comparison of various customization approaches, we have also chosen to adopt these same four features.

III. IMPACT OF DATA PRE-PROCESSING STRATEGIES

Data pre-processing strategies have an impact on the performance. In this section, we perform an in-depth analysis of existing techniques to address missing content at different temporal gaps, normalization and data filtering.

Firstly, data parsing and temporal synchronization were performed. The data, initially in XML format, were converted into CSV format, and the features were synchronized with the main column corresponding to BGC readings, which were recorded every 5 minutes (except for missing values). Specifically, we ensured that each BGC timestamp was associated with posterior observations that occurred within a maximum delay of 4 minutes. By adopting this approach, we realistically accounted for the fact that we do not possess information about the other features prior to the blood glucose detection.

A. Pre-processing strategies

With the aim of studying the effects of pre-processing on the final performance, we propose four versions of the dataset, each one obtained with a different strategy to handle missing data, normalization and data filtering. Following, we describe these versions in details.

Smoothed: this is the version of the training and test datasets achieved by reimplementing the pre-processing procedure as described by Shuo et al. [12]. The original dataset presents several missing values associated with BGC measurements. For training, missing values are imputed through linear interpolation from adjacent values if the interval is less than one hour, otherwise they are discarded. For testing, extrapolation is implemented in two different conditions for each two-hour sliding window, depending on the gap g between data points:

$$\begin{cases} \text{using model's predictions} & \text{if } g \leq 30 \text{ min} \\ \text{linear interpolation} & \text{otherwise} \end{cases} \tag{1}$$

For the three self-reported features (FS, BI, C), missing values are replaced with 0. After filling the missing values, data is normalized. To mitigate artifacts, outliers and distorted CGM readings, following [12], we applied a Gaussian filter with standard deviation $\sigma = 1$.

Averaged: this version differs from *Smoothed* by the technique used to fill missing data. In this case, in fact, large holes of data, more than 30 min, are filled with the average of the values referring to the same time interval of the other days.

Active carbohydrates (Active_carbs): Dataset preprocessed following the methodology explained in Butt et al. [18] with the addition that the variables BI and C are converted to continuous values.

B. Deep Multitask model

Following [12], we use a Recurrent Neural Network (RNN) comprising two stacked LSTM layers with 16 and 32 units, respectively. Following each LSTM layer, a dropout layer with a dropout rate of 0.2 is incorporated to enhance regularization

TABLE II
FINAL SCORES: MEAN* AND STANDARD DEVIATION AMONG 5 DIFFERENT RUNS

	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

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

and prevent overfitting. Beyond this point, the network diverges into two distinct clusters: one tailored for male subjects and another for female subjects. Each cluster includes a 128-unit Fully-Connected (FC) layer, followed by a dropout layer with a rate of 0.4. The final layer in the architecture is the subject-specific FC layer with 12 units, responsible for making predictions based on the unique characteristics of each subject. The total number of trainable parameters of the model is approximately 19,000 parameters. The original values of the hyperparameters have been adopted.

The input of the network is a four-dimensional (BGC, BI, FG, C) sliding window including the last two hours (24 samples) of observations. The output of the model is the prediction of the difference in glucose level at different Prediction Horizons (PH), i.e. the variation in BGC between *t* and *t+PH*, where the time horizon tested is 30, 60, 90 and 120 minutes. In the end, the output is obtained by summing the variation computed by the model to the glucose level at time *t*.

The dataset split used for these experiments is the same provided in [19] that is a 4-folds cross validation split.

The model was trained on each dataset for 250 epochs (implementing early stopping with patience equal to 30 and model checkpointing) for 5 runs. A linear combination of the MAE losses across individual subjects is used as a cost function.

C. Evaluation metrics

The performance of the model was evaluated by MAE and RMSE, comparing the metrics on different datasets and all 4 time horizons.

MAE is calculated as:

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

RMSE is calculated as:

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

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

D. Experimental results

Table II summarizes the results obtained. The outcomes obtained from the various datasets exhibit remarkable similarity and closely align with the results reported in [12]. For short and mid-range forecast (such as for PH $\leq$ 90min) the smoothed pre-processing performs slightly better than other approach with active_carbs and average being alternatively at second and third place. For long-range forecast (PH = 120 min) active_carbs performs better than the other pre-processing strategies.

In the next section, we will evaluate more thoroughly the impact of personalization on model performance, testing each model on the 3 different datasets obtained in this phase.

IV. PERSONALIZATION

The use of a personalized multitask recurrent neural network yielded significant advancements in predicting blood glucose concentration levels in terms of MAE and RMSE values. However, to ensure a comprehensive evaluation of its performance, we compare the Deep Multitask model, which serves as the reference, against two other models: a generalized version of the same model with no personalization at all, and a customized model based on the median BGC level of the various subjects in the OhioT1DM dataset. To robustly validate the performance of the models, we conducted tests on all four time horizons (PH = 30 min, 60 min, 90 min and 120 min) using the three datasets previously computed. Additionally, for this task, we adopted another dataset split, namely the Leave-One-Subject-Out Cross-Validation (LOSO-CV) split, that considers the test on one given subject and the training on the remaining ones. This approach provides a more realistic representation of real-world scenarios as the models are tested on entirely new subjects during each iteration. This is crucial in the Consumer Electronics field because it measures the degree of adaptability of the method to a new customer. The final values of MAE and RMSE are then derived by averaging the results across all iterations, providing a comprehensive evaluation of the model predictive capabilities across various subjects. This ensures a more robust and unbiased assessment of their performance.

A. General model

The architecture of the general model remains identical to the reference, except for the exclusion of the branches responsible for distinguishing between males and females and

TABLE III
FINAL SCORES: MEAN* AND STANDARD DEVIATION ON LOSO-CV

	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.

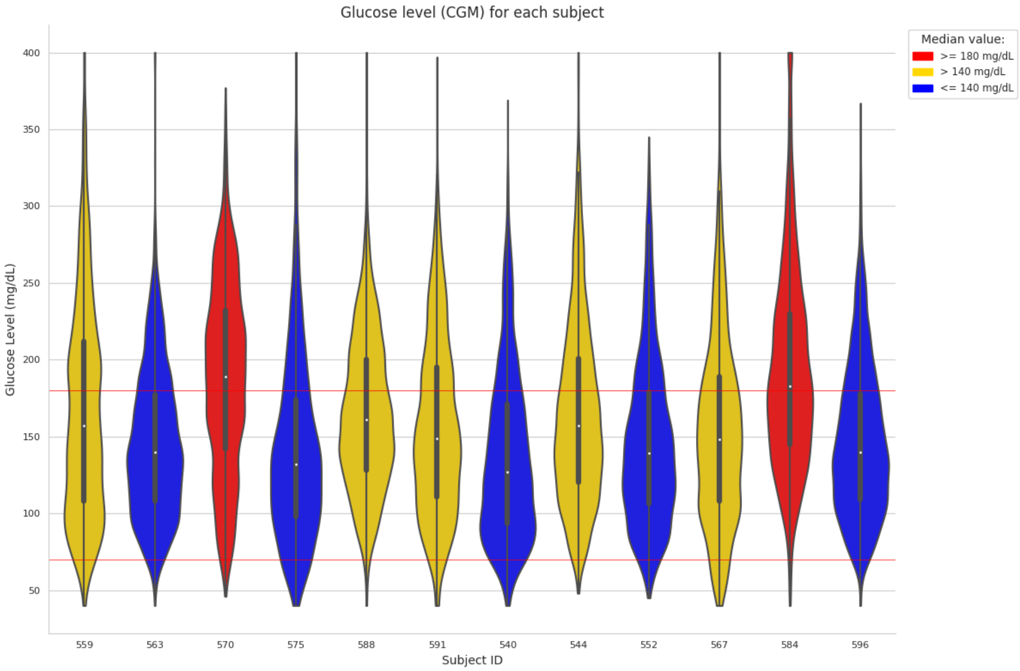

Fig. 1. BGC median level for each subject.

between each subject in the final FC layer. All hyperparameters, number of layers, and other options remain unchanged from the multitask model. By preserving the core structure and hyperparameter settings, we can make a direct comparison between the general model and the reference multitask model. This allows us to isolate the impact of the additional branches used for gender-based and subject-based differentiation.

B. Deep Threetask model

During the data exploration phase, we observed a considerable variability in BGC levels among the 12 subjects (Fig. 1). To categorize the subjects more precisely:

- 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).

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).

C. Experimental results

First of all, we compared the overall performance between the various models over all the three datasets, considering MAE and RMSE on the best dataset (table III). The levels of MAE and RMSE have slightly increased compared to the previous task, which can be attributed to the models being tested on data from subjects that have never been seen before. Nevertheless, these values remain well within an acceptable range and are still comparable to the previous results, affirming the effectiveness of the approach. Moreover, the outcomes clearly reveal that the performance achieved by the general model aligns and slighlty overcomes that of the personalized approaches, indicating that the branching mechanism has a minimal impact on the prediction task. The Multitask model

TABLE IV
FINAL SCORES: BEST DATASET AND MODEL* FOR EACH SUBJECT

	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.

shows even greater performance degradation than the other models as the prediction horizon increases.

Nonetheless, further insights can be gained by analyzing the performance on each individual subject (table IV). Upon analysis of the performance of different models, it is 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 is quite similar to 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%).

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 seems 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, subject 563 is situated remarkably close to the next median BGC level threshold and this proximity may potentially mislead the neural network's predictions. However, it is worth noting that when the Threetask model makes errors, they tend to be more significant than those made by its General counterpart.

V. Conclusions

The application of personalized models for blood glucose level prediction undoubtedly represents a promising approach. However, in this study, we have demonstrated that the general counterpart of the model not only matches but, in several instances, even surpasses the performance of the Multitask model considered as the reference. Our evaluations were conducted on three distinct datasets, carefully selected based on the Multitask model, and we assessed the MAE and RMSE values across four different time horizons (30, 60, 90, and 120 minutes).

Additionally, we developed a novel customized model known as Threetask, which leverages the median BGC level of individual patients. This Threetask model displayed an overall performance comparable to the general model, while exhibiting a superior ability to adapt to the unique characteristics of specific subjects. In practical terms, the Threetask model outperformed the general model for individual subjects in most cases. However, it is crucial to note that when the Threetask model makes errors, the difference in error magnitude compared to the general model is significant. This suggests that while the Threetask model excels at personalized predictions, it may occasionally encounter challenges that result in more substantial errors.

Lastly, it is crucial to highlight 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 adopted to facilitate a fair comparison with the state-of-the-art Multitask model. However, during our testing, we noticed that the utilization of Gated Recurrent Units (GRUs) instead of LSTMs proved advantageous, leading to further error reduction.

Considering these promising findings, the future research could involve starting from either the general model or the threetask model and conducting an extensive search for the optimal combination of hyperparameters. This pursuit aims to achieve even better performance enhancements, benefiting individuals managing their health and advancing the field of personalized medical modeling.

Acknowledgement

This work was partially funded by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

This work was also partially funded by the project ON Foods. Funder: Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 - Call for tender No. 341 of 15 March 2022 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU. Award Number: Project code PE00000003, Concession Decree No. 1550 of 11 October 2022 adopted by the Italian Ministry of University and Research, CUP D93C22000890001, Project title "ON Foods - Research and innovation network on food and nutrition Sustainability, Safety and Security – Working ON Foods".

References

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

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

[3] 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

[4] 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

[5] A. Rejeb, K. Rejeb, H. Treiblmaier, A. Appolloni, S. Alghamdi, Y. Alhasawi, and M. Iranmanesh, "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

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

[7] 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.

[8] K. Bach, R. C. Bunescu, C. Marling, and N. Wiratunga, 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

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

[10] 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.

[11] 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

[12] 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.

[13] J. Freiburghaus, A. Rizzotti-Kaddouri, and F. Albertetti, "A deep learning approach for blood glucose prediction and monitoring of type 1 diabetes patients," in *KDH@ECAI*, 2020.

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

[17] 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.

[18] 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

[19] 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.