University of Milano-Bicocca
Master's degree session in Data Science
19 October 2023

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

Niccolò Puccinelli

University of Milano-Bicocca
n.puccinelli@campus.unimib.it

Supervisor:
Prof. Paolo Napoletano
University of Milano-Bicocca
paolo.napoletano@unimib.it

Co-supervisor:
Dr. Flavio Piccoli
University of Milano-Bicocca
flavio.piccoli@unimib.it

Type-1 diabetes

- **Global** health concern [1], set to **expand** significantly [2].

- Deficiency in insulin production arising from the insulin-producing beta cells recognition as **foreign entities**, triggering a destructive immune response against them.

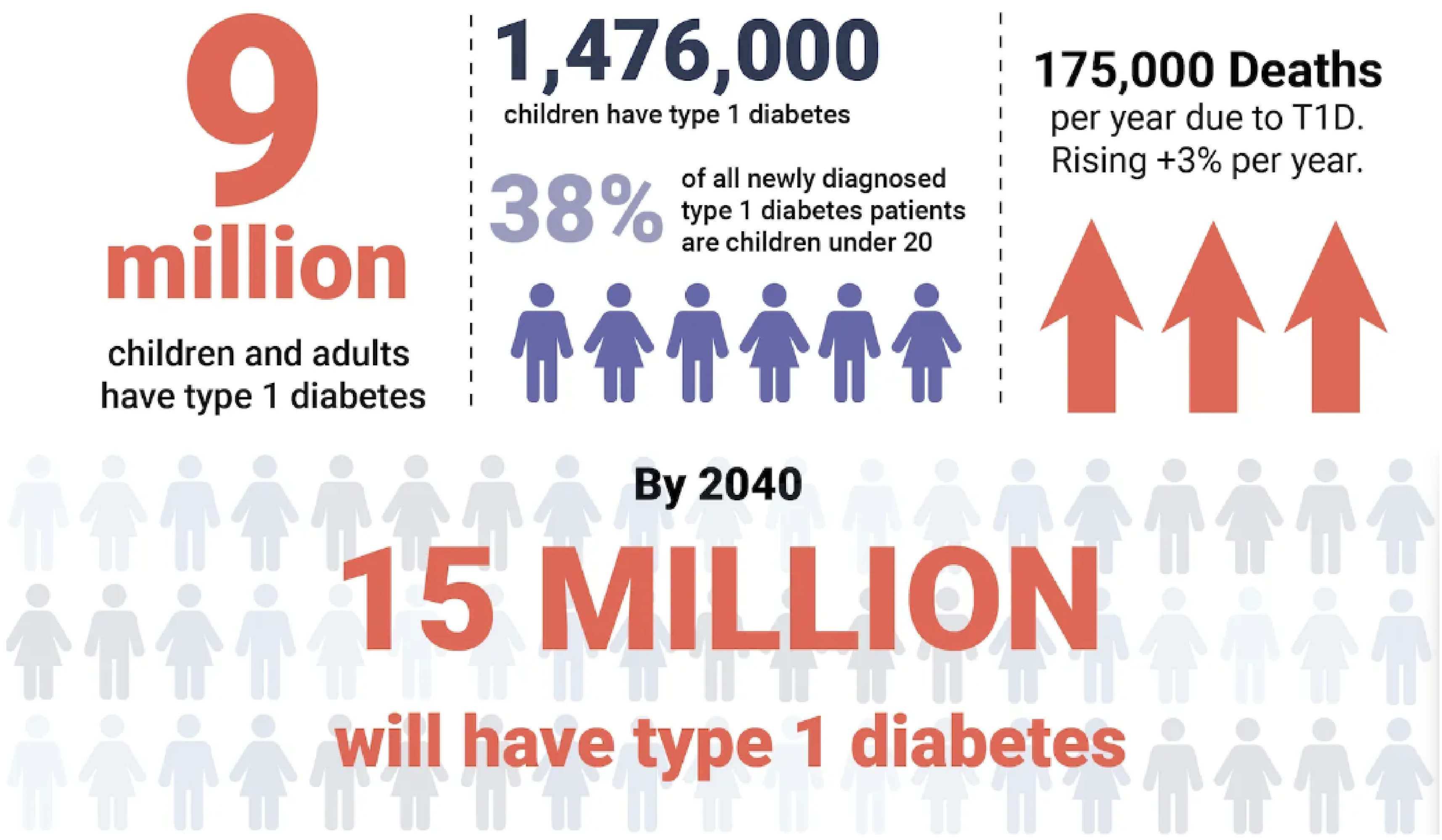


Image credits: [3]

Main symptoms

Excessive thirst and hunger, fatigue, partial vision loss, sudden weight loss, frequent urination.

[4]

Complications

Kidney failure, neuropathy, amputations, heart attacks, socio-psychological problems. **Need for insulin injections**.

Introduction

- Crucial need for continuous and accurate **Blood Glucose Concentration** (BGC) **prediction** [5].

- Advancements in **wearable sensors** and **IoT** techniques [6], [7], [8].

- **Real-time monitoring**: prompt interventions, minimizing life-threatening events.

- In recent years, **machine- and especially deep-learning** techniques demonstrated increasing accuracy and reliability in estimating BGC levels [9].

Multiple **challenges** in data-driven BGC prediction.

- **Biases**, **heterogeneity**, **incompleteness** of available datasets.
- **Personalization**: unique characteristics and habits.

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Main contributions

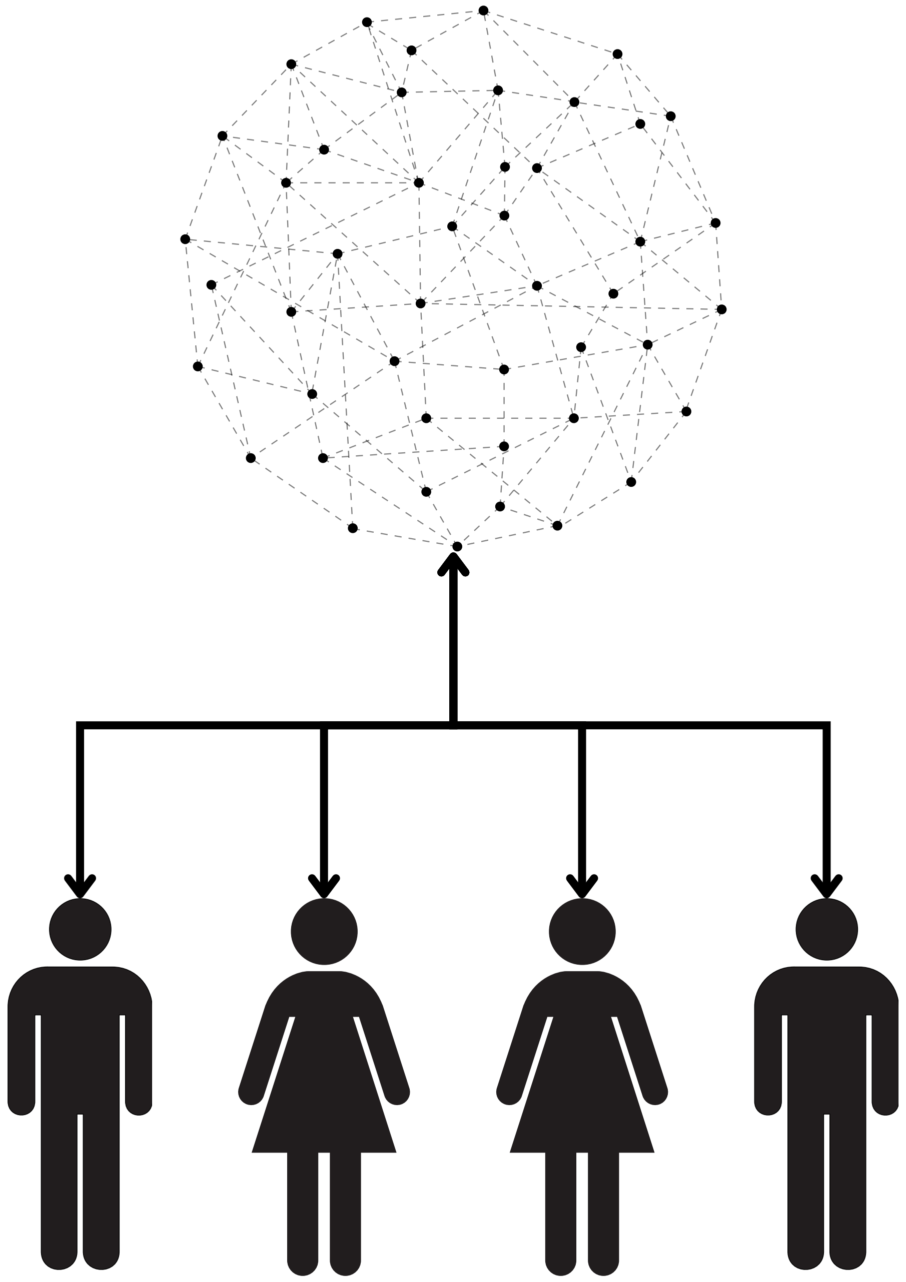
- Investigating the impact of several **pre-processing strategies** on the performance.

- Conducting a comparative analysis between 2 different **personalization techniques** and a **general** strategy with no customization at all.

- Proposing a new personalization technique, called ***Threetask***, that outperforms previous methods in the majority of the patients.

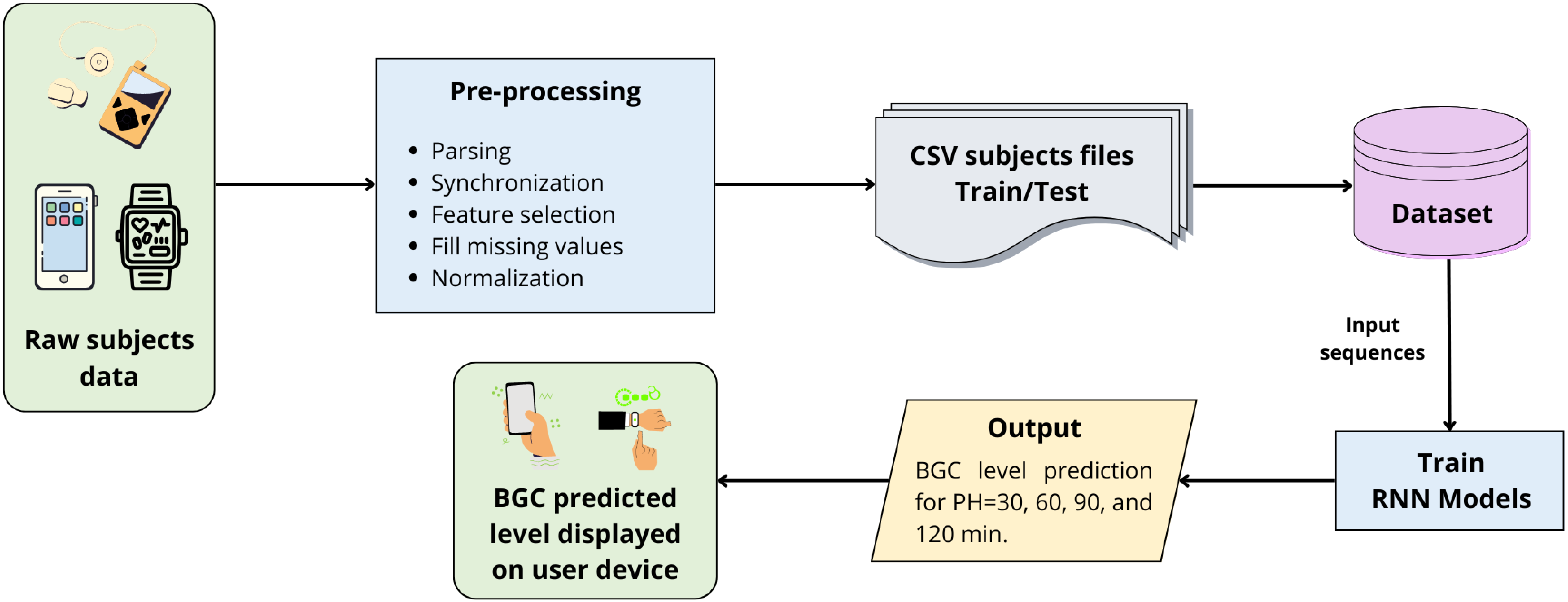
- Studying the influence of **additional features and subjects** on overall and subject-specific performance.

- Analyzing diverse approaches and methodologies across both **regression** and **classification** tasks, on multiple **Prediction Horizons** (PHs): 30 min, 60 min, 90 min and 120 min.



Model personalization

General pipeline



The OhioT1DM clinical dataset

- Experiments conducted on **OhioT1DM** clinical dataset [10], **widely employed** by several works.
- **Eight weeks'** worth of **Continuous Glucose Monitoring** (CGM) **of BGC levels** from 12 **heterogeneous** patients, recorded every **five minutes**.
- Optional **finger stick** glucose (FG) measured directly by the patients.
- **Insulin**, **physiological sensor** and **self-reported** life-event data.
- **Training-test split** already provided (~75%-25%).

Dataset detials

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

Selecting features

- Current advancements [9], including state-of-the-art research [12], predominantly focused on utilizing the following the **four primary features**:

- **glucose_level** (BGC): Continuous glucose monitoring of BGC data, recorded every five 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.

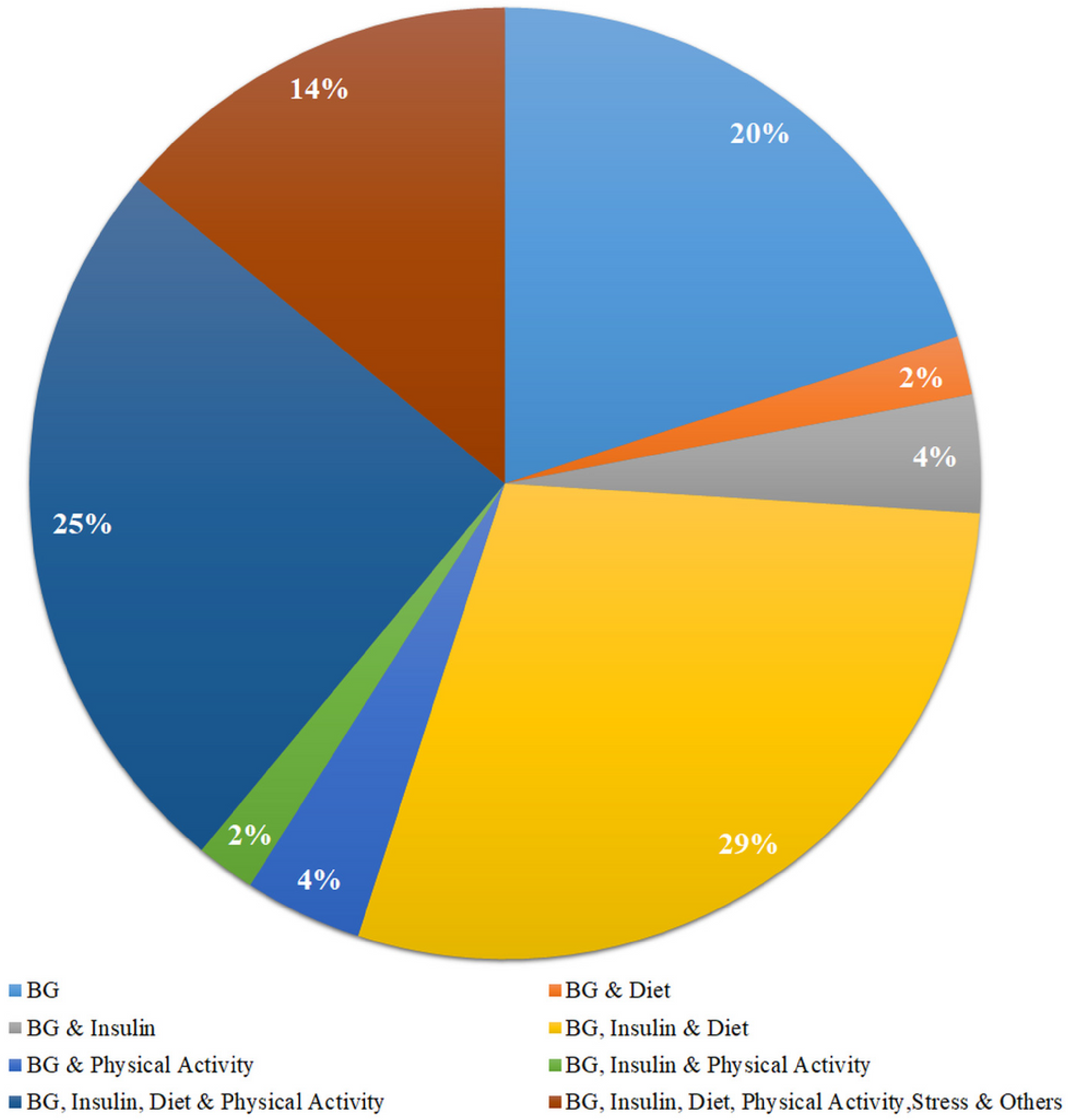


Image credits: [9]

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Impact of data pre-processing strategies

Pre-processing strategies

- Several pre-processing strategies and combinations tested. **Best three** datasets:

- **Smoothed**: Pre-processing reimplementation of state-of-the-art research by Shuvo et al. [12].

Missing values

$$\begin{cases} \text{linear interpolation} & \text{if } g \leq 120 \text{ min} \\ \text{discarded} & \text{otherwise} \end{cases}$$

Training

$$\begin{cases} \text{using the model's predictions} & \text{if } g \leq 30 \text{ min} \\ \text{linear interpolation} & \text{if } 30 \text{ min} > g \leq 120 \text{ min} \\ \text{discarded} & \text{otherwise} \end{cases}$$

Test

- **Averaged**: Same as standard, but large holes of data in the **training set** (> 30 min) are filled with the average of the values referring to the same time interval of the other days.

- **Active_carbs**: Dataset preprocessed following the methodology explained in Butt et al. [9], i.e., the variables BI and C are converted to **continuous** values.

Deep Multitask model

- Reimplementation of the neural network architecture developed by Shuvo et al. [12].

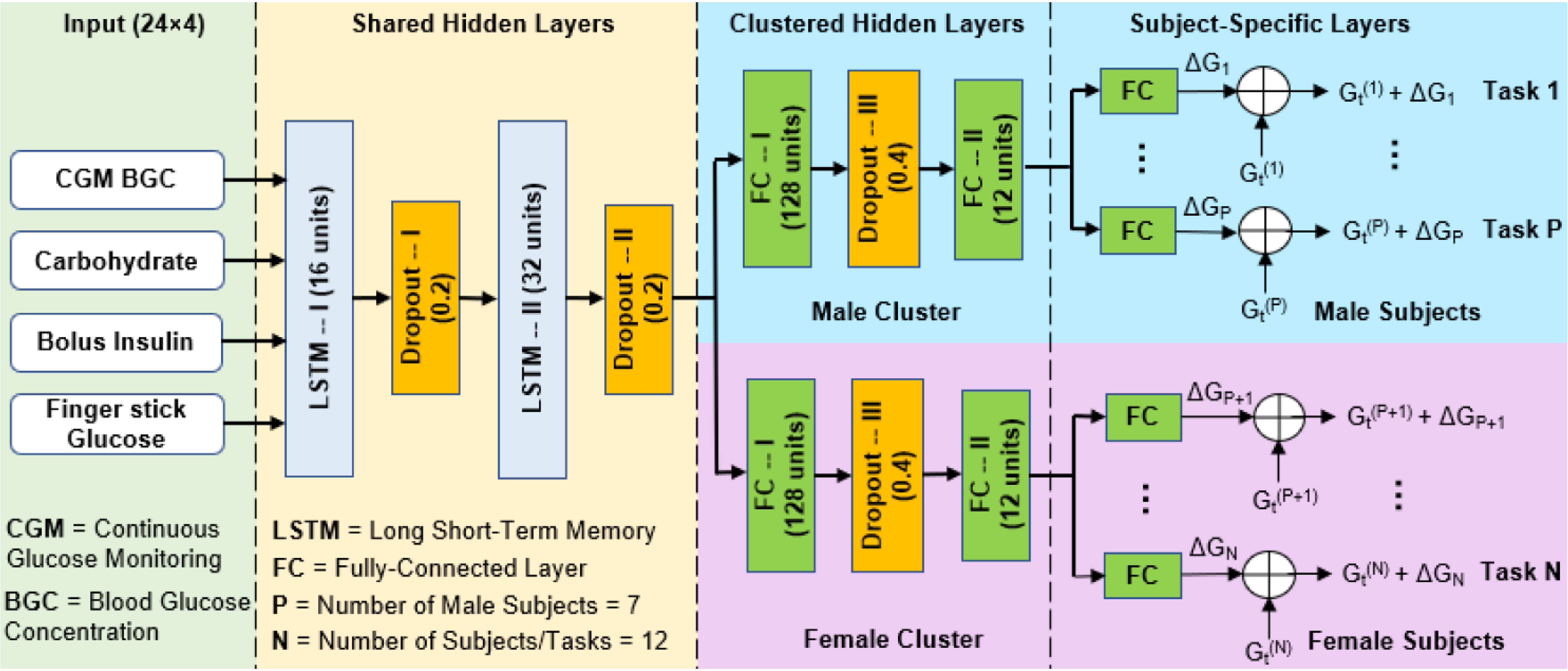


Image credits: [12]

Experimental results

- Testing pre-processing strategies on the **Deep Multitask** model.
  - Same methodology employed in [12]: training-test split as already provided (**~75%-25% for each subject**).
  - **Regression** task on **4 different Prediction Horizons** (PH).
  - Input: **two-hour** sliding window.
  - Evaluating Mean Absolute Error (**MAE**) and Root Mean Squared Error (**RMSE**) 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
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.

- Remarkable similarity with the results reported in [12].

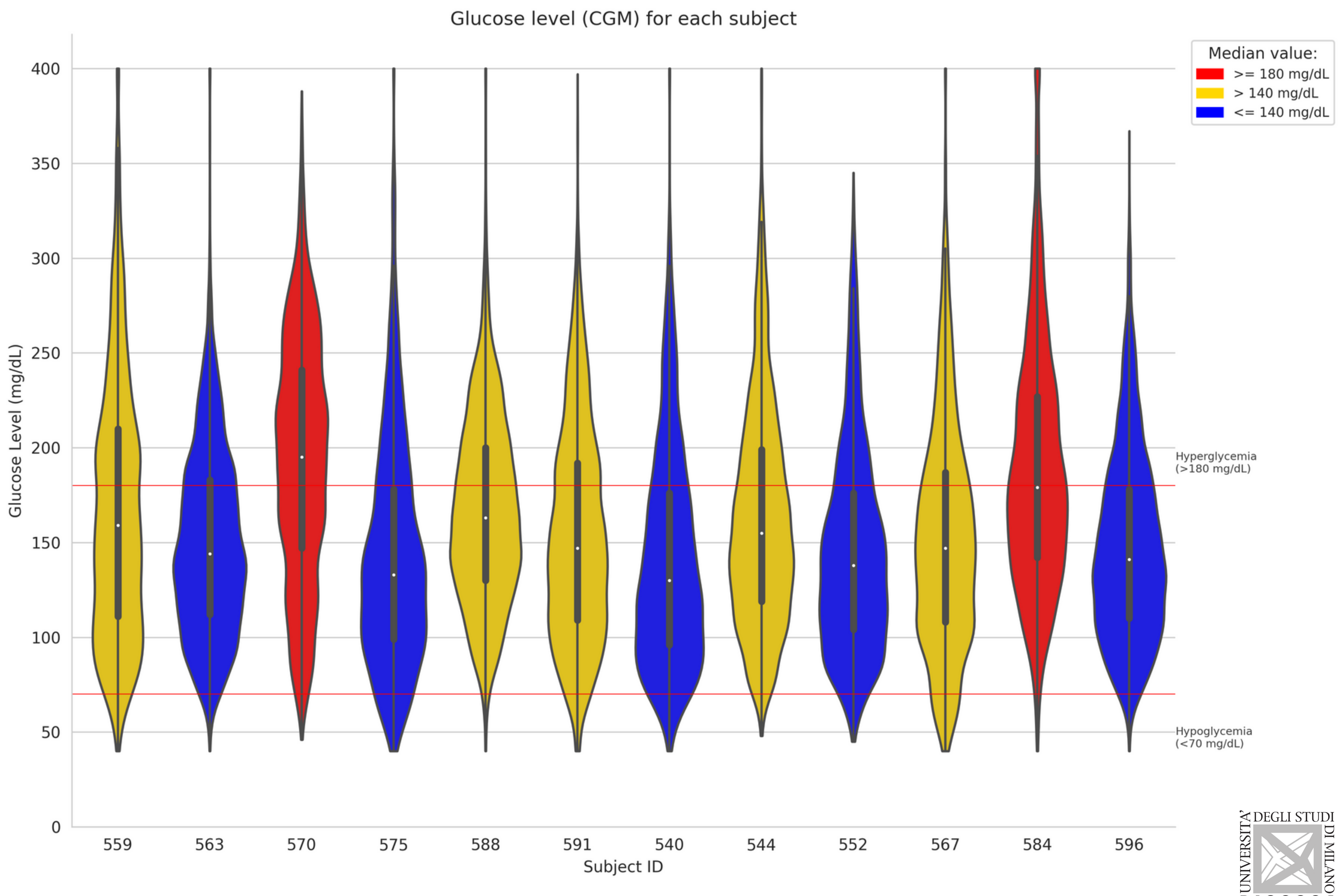
Impact of personalization

Deep Threetask model

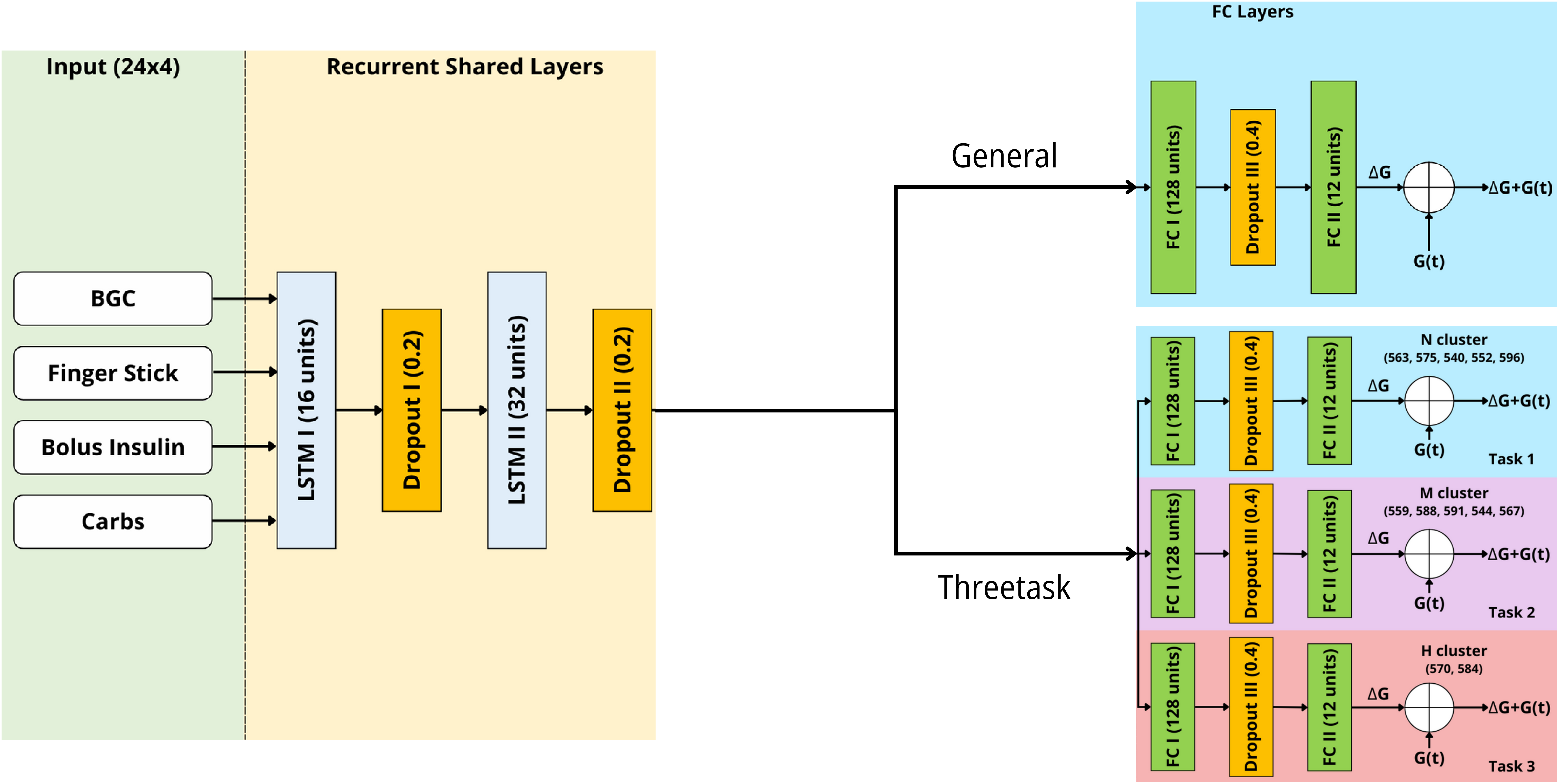
- Considerable **variability in BGC levels** among the 12 subjects observed during data exploration.

- Implementing a **branch** for each range (according to [13]):
  - **Normal** (N). BGC median level under 140 mg/dL.
  - **Medium** (M). BGC median level between 140 mg/dL and 180 mg/dL.
  - **High** (H). BGC median level over 180 mg/dL.

- **Prediction tailored directly to the disease level**.



Personalized vs. general model



Experiments

- Comparing performance between ***Multitask***, ***General***, and ***Threetask***.
- Conducting tests on **all four time horizons** (PH = 30 min, 60 min, 90 min and 120 min).
- Leveraging the **three datasets** previously computed (smoothed, averaged, active_carbs).
- Adopting **Leave-One-Subject-Out Cross-Validation** (LOSO-CV) split.
    - Providing a more realistic representation of real-world scenarios as the models are tested on **entirely new subjects** during each iteration.
- Final values of MAE and RMSE derived by averaging the results across all subjects.
- Maintaining **same core structure and hyperparameter configuration** for each model.
- **Robust** and **unbiased** assessment of the performance of the models.

Experimental results

- Overall, MAE and RMSE **slightly increased** compared to the previous task (models are now tested on data from subjects that have never been seen before), but the range is **still comparable**.
- *Multitask* consistently **underperforms** (16 cells), *General* (35 cells) is similar to *Threetask* (45 cells).
- *Threetask* generalizes over a **larger number of subjects**, but when it makes errors they tend to be **more significant** than those made by *General*.

	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*.

Classification task

Experiments

- Evaluating **personalization** on several different **classification tasks**, mainly conducted on *General* and *Threetask* models.

- Similar performance of the three datasets, thus we employed the standard **Smoothed**.

- Same core structure and configuration for both models to ensure **fair comparison**.

- Similar methodology as regression task: **LOSO-CV** split and evaluation on all four time horizons (PH = 30 min, 60 min, 90 min and 120 min).

- Evaluation metric: **macro-averaged accuracy**.
    - Treating each class **independently** and subsequently averaging the accuracy scores for all classes.

- Experimenting on **binary** and **multi-class** classification.

Experiments

- **Binary classification**, based on the commonly established risk threshold for BGC level **180mg/dL** [13], [14].

- **Multi-class classification**: obtaining overall balanced distribution by leveraging **quartiles**.

  Class 0 (low risk): BGC $\leq$ 113 mg/dL.

  Class 1 (medium risk): 113 mg/dL $<$ BGC $\leq$ 151 mg/dL.

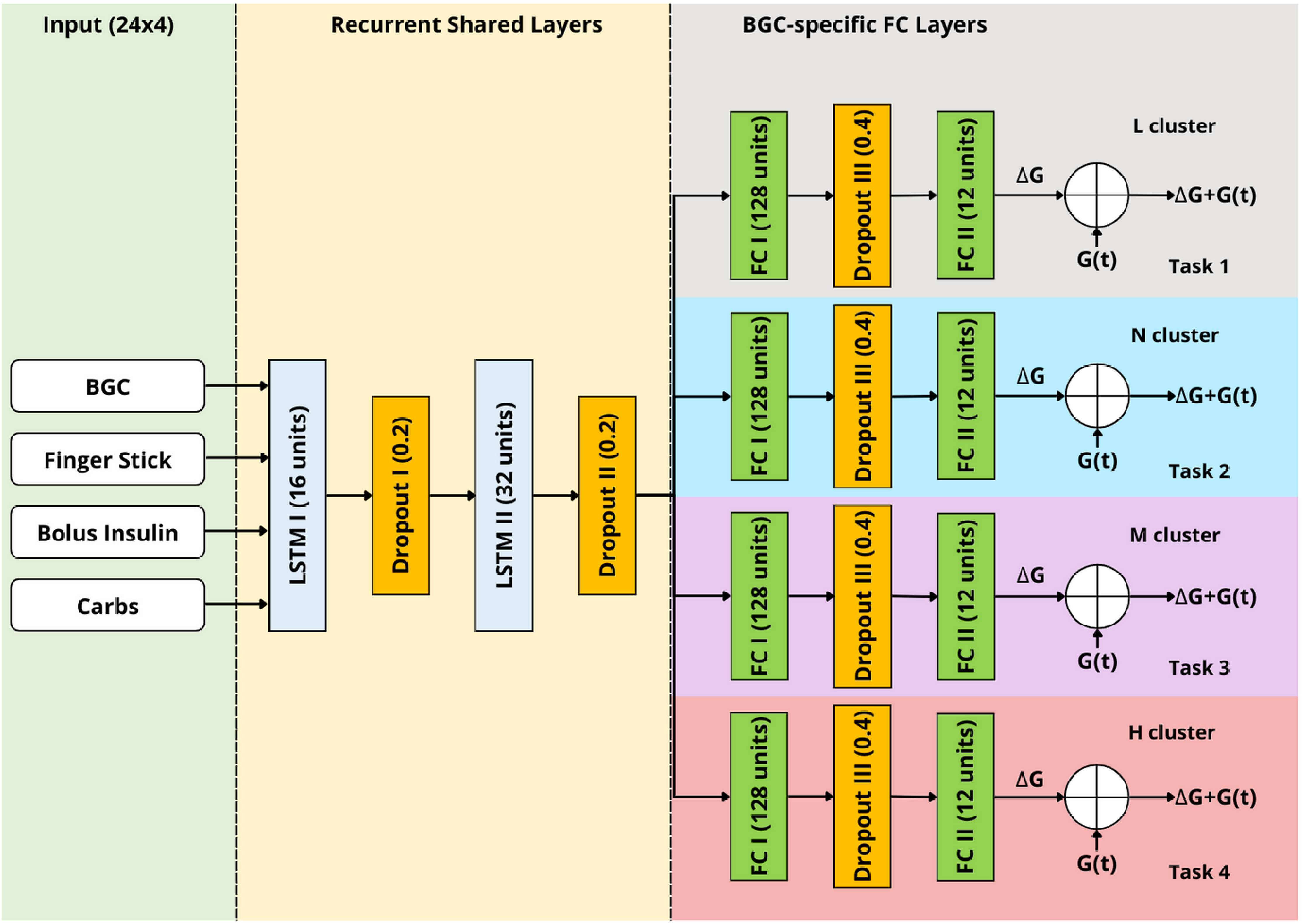
  Class 2 (high risk): 151 mg/dL $<$ BGC $\leq$ 197 mg/dL.

  Class 3 (very high risk): BGC $>$ 197 mg/dL.

  - **Standard**: evaluating personalization impact.
  - **BGC-only**: evaluating effectiveness of additional features.
  - **Integrate**: evaluating influence of adding additional subjects (**BIG IDEAs** dataset [14]) to the training process.

Deep Fourtask model

- BIG IDEAs subjects show generally **lower BGC values**, thus we added an **additional branch** (L).



Lower BGC values cluster

Experimental results

- Binary classification shows **almost equal performance** for *General* and *Threetask* models.

- **Multi-class classification**:
  - *Threetask* model **outperforms** *General*, especially on shorter time horizons.
  - Standard and BGC-only approaches performance are closely aligned: additional features demonstrate **limited usefulness**.
  - The integration of the new subjects has **limited effectiveness**. Nevertheless, the *Fourtask* model **outperforms** *Threetask* model, mitigating the problem related to error magnitude.

	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%**

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Conclusions

- **Personalized models** application for BGC level prediction represents a promising approach.

- However, 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.

- Developed a **novel customized model** (*Threetask*), leveraging the **median BGC** level of patients.
    - Slightly outperformed the *General* model for individual subjects in most cases in regression task.
    - Consistently outperformed the general model in multi-class classification task.

- Demonstrated **limited effectiveness** of additional features.

- Further increased performance through the ***Fourtask*** approach.

- The integration of more and more subjects, a more efficient imputation of missing values, and **new personalization approaches** based on the distribution of patients' BGC levels could **further improve** performance.

References

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

[2] 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.

[3] https://www.thejdca.org/publications/report-library/archived-reports/2022-reports/view-this-email-in-your-browser-click-here-to-unsubscribe-the-growing-global-burden-of-t1d.html.

[4] Standl, E., Khunti, K., Hansen, T. B., & Schnell, O. (2019). The global epidemics of diabetes in the 21st century: Current situation and perspectives. European journal of preventive cardiology, 26(2_suppl), 7-14.

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

[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.

[7] A. J. Perez and S. Zeadally, "Recent advances in wearable sensing technologies," Sensors, vol. 21, no. 20, 2021. [Online]. Available.

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

References

[9] 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.

[10] 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.

[11] 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.

[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] A. Campbell, "Diabetes self-management," June 2019, last accessed: 27th Sep 2023. [Online]. Available.

[14] 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.

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

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Thank you for your attention

n.puccinelli@campus.unimib.it