Smart Health 38 (2025) 100616

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

Glucose data interpretation in pediatric diabetes using an artificial intelligence approach

Giovanni Paragliola [a], Sara Campanella [b], Valentino Cherubini [c], Valentina Tiberi [c], Paola Pierleoni [b], Alberto Belli [b], Antonio Iannilli [c], Lorenzo Palma [b,*]

[a] ICAR-CNR, Via Pietro Castellino 111, Naples, 80131, Italy
[b] Università Politecnica delle Marche, Information Engineering Department, Via Brecce Bianche 12, Ancona, 60131, Italy
[c] Department of Women's and Children's Health, "G. Salesi" Hospital, Via Filippo Corridoni 11, Ancona, 60123, Italy

ARTICLE INFO

Keywords:
Autoencoder
Classification
Clustering
Glucose data
Diabetes
Patient profiling

ABSTRACT

Semi-automatic solutions to monitor and treat diabetes have been recently developed, including insulin pumps and continuous glucose monitoring devices. Integrating computational techniques with electrical, communication, and information systems offers significant opportunities. However, no decision support systems are capable of adequately managing and analyzing the data provided by these devices. As a result, the high specificity and complexity of the information generated cannot be effectively utilized in everyday clinical practice. Therefore, this paper proposes an artificial-intelligent-based approach to identify distinct patterns within the glucose readings of pediatric diabetic patients. The objectives are twofold: first, to cluster the data employing a dimensionality reduction technique based on autoencoders, and second, to classify the data using the labels derived from the clustering phase to profile the glycemic trends. Furthermore, the blind evaluation conducted by medical professionals on the clustering results has offered crucial clinical validation to the work carried out. The results highlight the effectiveness and reliability of the proposed approach, achieving a classification performance with accuracy values up to 98%. The data reduction step was fundamental to speed up the subsequent processes while improving the metrics. The medical evaluation allowed us to improve the work by finding a correspondence between experimental results and clinical value.

1. Introduction

Diabetes is a chronic, metabolic disorder identified by impaired insulin secretion or action (Craig, Hattersley, & Donaghue, 2009). Whenever glucose homeostasis is not maintained, resulting in hyperglycemia, diabetes is diagnosed. Although type 1 diabetes (T1D) can be diagnosed at any age, it is the most common autoimmune disease in children, with most cases occurring during childhood (5–7 years) or near puberty (10–14 years) (Grasso & Chiarelli, 2024). In addition, the incidence of childhood diabetes is increasing in many countries, and considering geographical differences in trends, the annual growth rate is estimated to be around 3% (Patterson et al., 2019).

As there is no known cure for diabetes, people living with it must manage their blood sugar levels to live healthier lives. Intensive insulin therapy has limitations, including an increased incidence of hypoglycemia and the need for frequent non-automated glucose monitoring (Boughton & Hovorka, 2024). As a result, new technologies have been developed to improve adherence to the therapy

* Corresponding author.
E-mail address: l.palma@staff.univpm.it (L. Palma).

https://doi.org/10.1016/j.smhl.2025.100616
Received 17 May 2025; Received in revised form 9 August 2025; Accepted 5 October 2025
Available online 11 October 2025

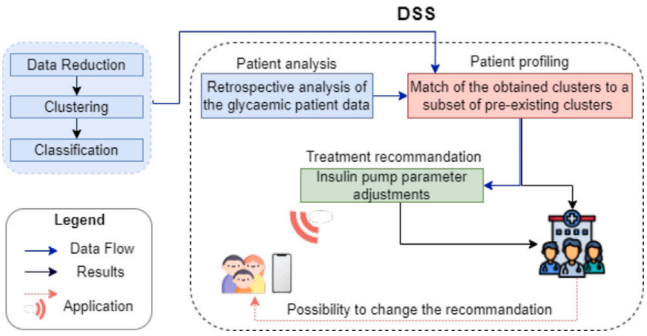


Fig. 1. Proposed framework for a Decision Support System (DSS). The blue box represents what we have implemented in this work, and which part of the proposed framework was done. The blue arrows represent the data flow, starting from data collection and continuing through its use in training and testing the DSS (data analysis). The black arrows indicate the pathway of the results generated by the DSS (results visualization). In contrast, the red arrows show who receives these results and how they are used (final deployment).

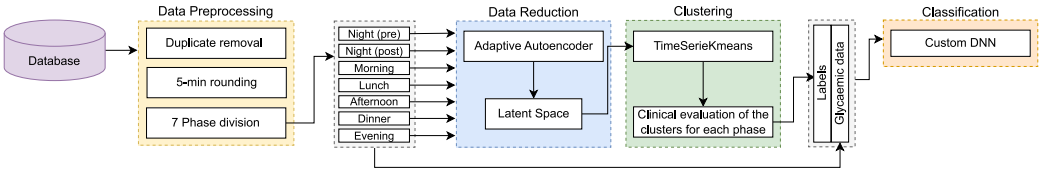


Fig. 2. Flowchart of the proposed study.

such as continuous subcutaneous insulin infusion (CSII), also known as insulin delivery with pumps. It uses only short- or rapid-acting insulin, reducing administration variability and glucose fluctuations. Advanced pump technology now mimics physiological demands accurately. Continuous glucose monitoring (CGM) has been integrated with controlled insulin delivery in basal and bolus modes, providing real-time glycemic control and early hypoglycemia detection. This combination, known as an artificial pancreas (AP) or hybrid closed-loop system, has proven effective, especially for children and adolescents, as an alternative to traditional treatment (Campanella, Paragliola, Cherubini, Pierleoni, & Palma, 2024). Despite all these improvements, a hybrid system still requires users to input meal times and request on-demand boluses (Campanella et al., 2022) while, in a fully automated closed-loop system, there should be no need for notifications about meals or exercise (Lal, Ekhlaspour, Hood, & Buckingham, 2019).

From CGM data to insulin pump parameters and bolus information, an artificial pancreas generates and manages a wide variety of data (Zhu, Li, Herrero, & Georgiou, 2021). For example, the CGM collects 288 glucose readings daily, or nearly 3.5 million data points yearly. Due to the lack of quantitative techniques for analyzing the data produced by these devices, which prevents the use of the strong specificity and richness of these data, this information is not fully utilized in routine clinical practice or daily diabetes management. Using the mentioned data, techniques can be developed to categorize CGM patterns and their correlation with insulin pump parameters. This will allow access to details of daily variations and optimize treatment outcomes for personalized therapy.

Artificial intelligence (AI) medical applications are widespread, including in diabetology (Arnia, Saddami, Roslidar, Muharar, & Munadi, 2024; Cheng, Zhu, Li, & Xu, 2023). The lives of diabetic patients, the work of healthcare professionals, and the state of the healthcare system can all be impacted and improved by these technologies (Ahmed, Ali, Masud, & Naznin, 2024; Ellahham, 2020; Gautier, Ziegler, Gerber, Campos-Náñez, & Patek, 2021; Makroum, Adda, Bouzouane, & Ibrahim, 2022; Xie & Wang, 2019).

Based on these considerations, this work presents an AI-based approach to support the clinical decision-making process for profiling pediatric diabetic patients using CGM data. As reported in Fig. 1, our work (the blue block) is part of a broader framework aimed at analyzing many glycemic trends and clustering them to identify an ideal profile in which glycemia remains within optimal ranges. These clusters are then compared to patients' glycemic data to understand how much they deviate from the ideal profile. Based on these comparisons, insulin pump setting adjustments are suggested and sent to the healthcare professional for validation and the patient. The proposed approach uses time-series-based clustering to identify distinct patterns and trends within the glucose readings of diabetic patients. It incorporates additional steps beyond clustering to enhance pattern recognition for patient profiling. Specifically, a dimensionality reduction technique based on autoencoders is applied to improve clustering results by compressing the data into a lower-dimensional space, preserving essential features while eliminating noise, thus facilitating more accurate clustering. In addition, a deep learning-based classifier is used to analyze the data based on the clustering result. This classifier uses the patterns identified during clustering to perform more detailed and accurate data classifications. The results demonstrate the effectiveness and reliability of the proposed approach, achieving highly satisfactory classification performance with accuracy values reaching up to 98%. Moreover, the clustering results are robust at all stages, thanks to the application of the autoencoder, which helps improve the consistency and stability of the identified clusters.

2. Related works

Although extensive research has been conducted, our literature review emphasizes a significant gap: clustering-based classification methods are absent in diabetology (Campanella et al., 2024). This observation stands out, as no comparable approaches have been identified in the current scientific literature, making this an underexplored and potentially innovative study area (Campanella et al., 2024). However, similar methods have shown significant efficacy in other biomedical fields, offering tailored therapies and a deeper understanding of disease diversity. Xu and colleagues (Xu, Nwe, & Guan, 2014) propose a cluster-based analysis technique that considers intersubject variation to quantify stress using physiological markers. Similarly, Khalaf et al. (2020) focus their work on individual variations in physiological responses rather than group-level comparisons to predict Challenge versus Threat.

In the work (Amit, Gavriely, & Intrator, 2009), they propose a framework for computational analysis that can be used to distinguish between different heart sound morphologies and categorize them according to physiological conditions. Hierarchical clustering and two classification algorithms (K-nearest-neighbor and discriminant analysis) form the basis of the analysis framework. Various new techniques have been used to apply different classifiers to classify the sleep state in EEG signals. In the paper (Al-Salman, Li, Oudah, & Almaged, 2023), they propose a framework based on the k-means algorithm to cluster the wavelet coefficients of each level of sleep. To determine the different stages, the extracted features are then sent to the least squares support vector machine (LS-SVM) classifier. Clustering-based classification has been successfully used in motion tracking, especially in the context of emotion recognition as stated by the works (Chen et al., 2022; Yang, Cai, & Hu, 2022). In the field of diabetology, the pursuit of personalized therapy and individualized monitoring has led to the exploration of data clustering techniques applied to glycemic data. By clustering these data, distinct patterns and trends specific to each individual's glucose levels can be identified (Biagi et al., 2019; Contreras, Quirós, Giménez, Conget, & Vehi, 2016; Mao et al., 2022; Tao et al., 2021). In the work proposed by Hall et al. (2018), the focus was on the characterization of glucose patterns, aiming to identify the actual types of genotypes depending on the level of glucose variability using spectral clustering using a 2.5 hour glucose record. Another work was done by Kahkoska et al. (2019) groups eight features derived from CGM data using self-organizing maps. They identified three clusters of young people with T1D and elevated glycated hemoglobin over their target population. Finally, the work by Lobo, Farhy, Shafiei, and Kovatchev (2021) presents a data-driven method for identifying Ω, a limited number of representative daily profiles (motifs), to match almost any daily CGM profile generated by a patient to one of the motifs in Ω. However, while clustering has been widely used to analyze glycemic patterns, current approaches remain largely exploratory. To date, no study has effectively combined data reduction techniques—such as dimensionality reduction or feature selection—with clustering-based classification. This integrated approach could enhance the interpretability and efficiency of diabetes data analysis, paving the way for more sophisticated and scalable personalized treatment strategies.

In terms of comparison with other CGM data interpretation, it is worth noting that traditional rule-based and statistical metrics, such as time in range, mean absolute error (MAGE) and coefficient of variation, while clinically established, oversimplify temporal dynamics and fail to uncover latent glycemic behavior patterns derived from full CGM trajectories. These limitations have been highlighted by Cui, Goldfine, Quinlan, James, and Sverdlov (2023) and David et al. (2025). More advanced functional data analysis (FDA) approaches have modeled glucose curves as smooth functions and identified phenotypes via functional principal component analysis (FPCA). These approaches offer improved temporal resolution and heterogeneity analysis but often rely on strong smoothness assumptions and lack the flexibility to scale non-linearly across pediatric data (Gecili et al., 2020). Supervised time-series classification models, such as logistic regression, ARIMA, recurrent neural networks, and CNNs, have primarily been applied to CGM data for event prediction (e.g. hypoglycaemia forecasting), but they depend on predefined labels and are not suited to identifying patient-specific glycemic profiles in an exploratory manner.

In contrast, our pipeline combines unsupervised clustering with autoencoder-based nonlinear dimensionality reduction to discover latent, interpretable glycemic profiles. This is followed by classification for real-time patient stratification, enabling prospective application in clinical settings.

There is currently only one publication that uses methods similar to ours for data reduction and clustering but it is not applied to the same population. This rarity underscores how unique and innovative our approach is. Lim, Cho, and Kim (2022) proposed a multi-task disentangled variational autoencoder (VAE) to study the properties of latent representations (LR) and to use LR for various tasks such as temporal clustering, glucose prediction, and event detection.

2.1. Our contribution

The literature analysis indicates that no existing approaches are comparable to the one we propose. The novelty of our work lies in the application of a dedicated pipeline to analyze glycemic data from pediatric diabetic patients using hybrid closed-loop devices. Our approach allows us to identify characteristic glycemic profiles, providing valuable insights into glucose regulation patterns. Furthermore, we have structured our pipeline into distinct phases to gain an initial understanding of how therapy can be adapted at different times of the day. This phased approach enables a more detailed and personalized assessment of glycemic trends, potentially improving treatment optimization. Additionally, our findings have undergone medical validation, ensuring that the results are clinically relevant and applicable to real-world therapeutic adjustments. Our work can serve as a patient monitoring tool. Specifically, once typical glycemic profiles are identified for each phase of the day—based on historical data from all patients using the autoencoder and clustering methods—optimal and non-optimal patterns can be defined. When analyzing an individual patient, their glycemic trends in each phase can then be assigned to one of these predefined clusters through a classification process. This allows for a deeper understanding of the patient's daily habits and glycemic control, and enables the use of this information to fine-tune insulin pump settings, thereby improving overall treatment management.

Table 1
Schematic representation of the data matrix.

ID	Phase	Day	Glycemia
458974	Night (pre)	1	[156, 160, 161, …, 187]
458974	Night (post)	1	[100, 102, 101, …, 120]
458974	Morning	1	[130, 135, 141, …, 180]
458974	Lunch	1	[181, 181, 184, …, 150]
458974	Afternoon	1	[156, 159, 162, …, 187]
458974	Dinner	1	[190, 194, 197, …, 255]
458974	Evening	1	[260, 261, 265, …, 300]
458974	Night (pre)	2	[298, 295, 289, …, 200]
[…]	[…]	[…]	[…]
874563	Evening	365	[200, 208, 210, …, 239]

3. Materials and methods

This section outlines the entire pipeline process, represented in the flowchart in Fig. 2.

First, we used preprocessing techniques to clean, organize, and normalize the data. This ensures that the data format is suitable for further analysis. We used an autoencoder to reduce the computational burden and highlight the most relevant features for more efficient downstream processing. After the autoencoder provided a reduced-dimensional representation, it was fed into a clustering algorithm. This step made exploring patterns easier, allowing us to identify different patterns or trends in the dataset. A classifier uses the insights gained from the clustering process to categorize data points into different classes. The classifier uses the patterns identified during clustering to make informed predictions. Combining clustering and classification enhances the ability to effectively interpret and act on the grouped patterns, resulting in improved classification performance. In the Supplementary Materials, there is additional information about the methodology and the results.

3.1. Dataset

Data from 99 patients wearing the Dexcom G6® CGM and the Tandem™ t:slim insulin pumps were provided by the "G. Salesi" Children's Hospital of Ancona. We selected only those who activated the Control-IQ algorithm, reducing the cohort to 91 young individuals out of the initial 99. The Control-IQ algorithm analyzes CGM data and automatically adjusts insulin delivery based on preset insulin parameters to maintain glucose levels within a target range of 70–180 mg/dL. It can increase or decrease basal insulin delivery and administer corrective insulin boluses as needed. The data were collected over one year (01/01/2022–31/12/2022), and the patients have a mean age of 14 ± 4.73 years. For privacy reasons, the hospital provided us with raw data—already stripped of all personal information except for the year of birth and gender—after obtaining approval from the Ethics Committee (protocol 2020/439) and signed informed consent.

3.2. Pre-processing

Any lines with duplicate glycemia readings at the same timestamp have been removed to avoid any potential errors, and the minutes have been adjusted to multiples of five (for instance, from 12:38 to 12:40). We then included a column for each patient that took the day into account to provide information about the number of days we had available for each patient. To effectively capture all the underlying characteristics of glucose fluctuations, we separated the day into seven distinct phases. Based on clinical experience, our collaborating physicians have found that this division into distinct phases not only facilitates improved glycemic control for both patients and clinicians but also provides a structured foundation for the subsequent configuration of insulin pump profiles necessary for insulin delivery (Cherubini et al., 2021; Eissa, Good, Elliott, & Benaissa, 2020). Specifically:

1. $Morning$, from 06:35 to 12:00 (65 samples);
2. $Lunch$, from 12:05 to 16: (47 samples);
3. $Afternoon$, from 16:05 to 19:00 (35 samples);
4. $Dinner$, from 19:05 to 22:00 (35 samples);
5. $Evening$, from 22:05 to 23:55 (22 samples);
6. $Night\ (pre)$, from 00: 00 to 2:00 (24 samples);
7. $Night\ (post)$, from 02:05 to 06:30 (53 samples).

Because of the dawn phenomenon, which involves variations in glycemic fluctuations based on the patient's age and whether they are in the pubertal phase, which happens about 2:00 a.m., we chose to split the night into *pre* and *post* phases. Due to a few missing days, the data was not always consecutive; therefore, we had to build a brief phase in the evening. We subsequently collected the glucose readings, based on the phase, in a single array and added an ID for each patient. This process was carried out for all patients, and Table 1 shows an example of the final structure. Each row represents a specific measurement phase for a given patient on a particular day. The table includes four key pieces of information: the patient's unique ID, the phase of the day when the

Table 2
Number of signals across phases.

Phase	N° segment
Night (pre)	22755
Night (post)	22831
Morning	22835
Lunch	22795
Afternoon	22726
Dinner	22722
Evening	22025

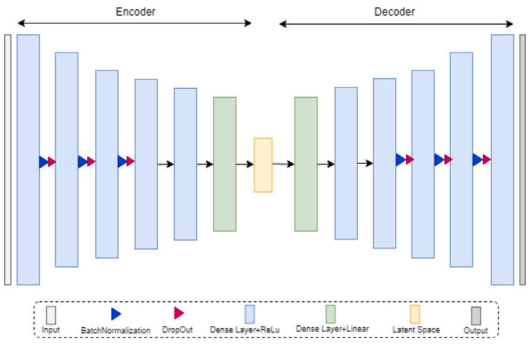


Fig. 3. Schematic representation of the adopted autoencoder.

measurement was taken, the corresponding day, and the array of the recorded glycemic values. The values in the "Glycemia" column represent the extreme measurements recorded within each phase. Specifically, the first value corresponds to the earliest timestamp in that phase, while the last value represents the latest. For example, patient 458974 had glycemic measurements recorded on Day 1 across multiple phases. For example, for patient 458974 on Day 1, glycemic measurements recorded during the *Night (pre)* phase range between 156 and 187, where 156 is the glycemic value at 00:00, and 187 is the value at 01:55. Similarly, in the *Dinner* phase, glycemia ranges from 190 to 255, corresponding to the first and last recorded values for that period. Table 2 shows the total number of segments for each phase.

3.3. Data reduction: Autoencoder

The primary function of an autoencoder is to encode information into a compressed and meaningful representation and then decode it back so that the reconstructed input is as close as possible to the original (Bank, Koenigstein, & Giryes, 2023) and is used for image classification, object recognition, natural language processing, network security, and so on (Li, Pei, & Li, 2023). A powerful application is data reduction: it searches for a projection technique that converts high-feature space data into low-feature space data. There are two types of dimensionality reduction techniques: linear and nonlinear (Wang, Yao, & Zhao, 2016). Principal component analysis (PCA) is a commonly used technique for data reduction. Although it is a well-established technique for linear dimensionality reduction, it is not applicable in contexts such as CGM analysis, where the underlying data distribution is non-linear and influenced by dynamic biological and behavioral interactions (Zhang, Holt, & Khovanova, 2016). Similarly, although autoencoders are a well-established technique for non-linear dimensionality reduction, their application in CGM data analysis for pediatric T1D remains limited (Afsaneh, Sharifdini, Ghazzaghi, & Ghobadi, 2022). In our work, we employ autoencoders not as a methodological novelty, but as a crucial step to enhance clustering robustness and enable high-fidelity pattern recognition for subsequent patient profiling. This integrated approach represents a novel contribution to AI-based decision support in diabetology.

To model our custom autoencoder, we performed several iterative experiments. Different architectural configurations, activation and loss functions, and optimization algorithms were explored and evaluated. These experiments aimed to identify the combination of parameters and techniques that yielded the most satisfactory data reconstruction and feature extraction results. Fig. 3 shows a scheme of the final version of the autoencoder neural network. We have modeled five key parameters:

- *Number of layers*: The encoder and decoder each have the same number of layers, excluding the input and output layers. The model includes a total of 13 dense layers and 6 layers each for batch normalization and dropout.
- *Bottleneck Size*: This parameter determines the degree of data compression and also serves as a regularization term. This model sets the bottleneck size to 25% of the input size. This value is the result of several experiments aimed at tuning this parameter to achieve the best tradeoff between performance and error minimization.
- *Activation Function*: The ReLU activation function was applied to the first seven dense layers to introduce nonlinearity, which helps to learn complex patterns by effectively ignoring negative values. A linear activation function was used for the output layer.

- *Loss Function*: Since this is a regression problem, we applied the Mean Squared Error (MSE) loss function, which calculates the squared difference between the actual and predicted data.

3.4. Clustering

The goal of clustering is to divide data points into different groups (clusters) according to the characteristics of the data points. K-means is one of the most applied algorithms for clustering and aims to maximize the distances between clusters while minimizing the distances within a cluster (Hartigan & Wong, 1979). Since we are working with time series, we decided to use a variant called Time Series K-means: it can efficiently take advantage of the intrinsic subspace information in a time series data collection. More precisely, to effectively explore the temporal sequence information associated with time series data, the algorithm tries to smooth the weights of adjacent time stamps, and hence the uncovered subspaces become more meaningful for clustering time series data (Huang, Ye, Xiong, Lau, Jiang, & Wang, 2016). It has been implemented thanks to the *TimeSeriesKMeans* from the Python package *tslearn*, with the following input parameters: (i) *n_clusters*: it indicates the number of clusters to create, changing from 2 to 8 each time; (ii) *n_init*: indicates the number of times the K-means algorithm is automatically run with different centroid seeds, with the final result being the best, in terms of inertia, among all *n_init* successive runs. In this work, this parameter was set to 100; (iii) *metric*: it indicates the metric to be used for distance computation and clustering assignment. For our purposes, the Euclidean metric was chosen. Three indexes have been chosen as evaluation criteria to determine the optimal number of clusters. The Silhouette Coefficient (SS) is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each sample (Shahapure & Nicholas, 2020). It was computed using the *tslearn.clustering* library. The Calinski–Harabasz (CH) index, known as the Variance ratio criterion, measures how similar an element is to its cluster to the other clusters. The larger the CH index, the better the clustering, since it means that there is a large dispersion between clusters and a small dispersion between elements of the same cluster (Caliński & Harabasz, 1974). The Davies–Bouldin Index (DB) measures the similarity between clusters and is defined as the ratio of the intra-cluster distances to the inter-cluster distances. The more clusters are farther apart and less dispersed, the lower the score and the better the result (Davies & Bouldin, 1979). Both these metrics have been computed using the *sklearn.metrics* library. Finally, the labels obtained from the clustering process are then used as reference labels for the subsequent classification task.

3.5. Classifier

While the clustering step identifies representative glycemic clusters, which are interpreted as characteristic profiles of patient subgroups, the classification component plays a distinct and essential role in the overall framework. Specifically, the classifier is trained to recognize these profiles by learning the latent structure of the time-series clusters. Once trained, it can assign new, unseen glycemic trajectories to the most appropriate existing cluster, enabling real-time patient stratification without the need to recompute clustering for each new case.

For the classification problem, we adopted the network proposed in Paragliola and Coronato (2018, 2021). A review of the state of the art reveals that no existing approaches employ a pipeline similar to ours. Therefore, we opted to use neural networks that have been successfully tested in other biomedical contexts, adapting them to the specific requirements of our problem. Fig. 4 provides an overview of the proposed hybrid network. At the input layer, the model receives a time series (TS). The architecture consists of three main components:

1. Recurrent Layer: This layer captures temporal dependencies within the TS using a Long Short-Term Memory (LSTM) network, chosen for its ability to model sequential correlations. The LSTM processes a sequence of data with x_t, analyzing each value x_t at time t. Both input and output TS are represented as feature vectors, with dimensions determined by the filter size of the unit cell. The LSTM state at time t is given by:

$$h_t = \sigma(W_h x_t + U_h h_{t-1} + b_h)$$

where σ is the sigmoid function, and W_h, U_h, and b_h are trainable parameters, with h_{t-1} as the previous state. The LSTM consists of a single layer with an output space dimensionality of either 1 or 10 units per cell. Its output is fed into the convolutional layer.

2. Convolutional Layer: This layer extracts the most informative features by applying filters to local input regions. It consists of two stacked blocks, each containing a convolutional network (C) and a max pooling layer (P). The first block has a kernel size of 10 for both C and P, while the second uses a kernel size of 8. The convolutional operation is defined as:

$$y_t = \text{Conv}(x_t, \theta)$$

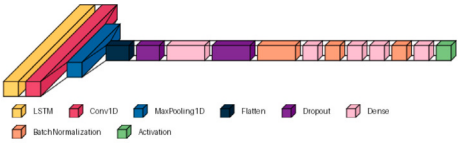
followed by a max pooling step:

$$y'_t = \text{MaxPooling}(y_t, k)$$

where Conv represents the convolutional operator, and k reduces dimensionality. The extracted features, formatted as a vector, are then flattened for processing by the deep layer.

3. Dense Layer: Comprising four fully connected Deep Neural Network (DNN) layers, each has half the neurons of its predecessor. Batch Normalization (BN) and dropout layers follow each fully connected layer, with ReLU as the activation function to prevent gradient saturation. BN enhances convergence and generalization, while dropout mitigates overfitting. A basic fully connected layer is expressed as:

$$y = W x + b$$



(a) CNN Architecture

Layer (type)	Output Shape	Param #
lstm_3 (LSTM)	(None, 310, 1)	24
conv1d_3 (Conv1D)	(None, 310, 2)	10
max_pooling1d_1 (MaxPooling1D)	(None, 155, 2)	0
flatten_1 (Flatten)	(None, 310)	0
dropout_2 (Dropout)	(None, 310)	0
dense_5 (Dense)	(None, 512)	159,232
dropout_3 (Dropout)	(None, 512)	0
batch_normalization_3 (BatchNormalization)	(None, 512)	2,048
dense_6 (Dense)	(None, 128)	65,664
batch_normalization_4 (BatchNormalization)	(None, 128)	512
dense_7 (Dense)	(None, 64)	8,256
dense_8 (Dense)	(None, 32)	2,080
batch_normalization_5 (BatchNormalization)	(None, 32)	128
dense_9 (Dense)	(None, 4)	132
activation_1 (Activation)	(None, 4)	0

(b) CNN Architecture summary

Fig. 4. Model visual representation and summary.

The final layer employs a softmax activation function for multi-class classification. To optimize the model's performance, we conducted a grid search over a defined subset of hyperparameters, systematically exploring the following configurations:

- LSTM output size: {1, 10}
- Neurons per layer: {512, 1024, 2048}
- Learning rate: {0.001, 0.0001, 0.00001}
- Epochs: {1000, 1500, 2000}
- Convolution kernel size: {4, 8, 10}
- Pooling filter size: {2, 4, 8}

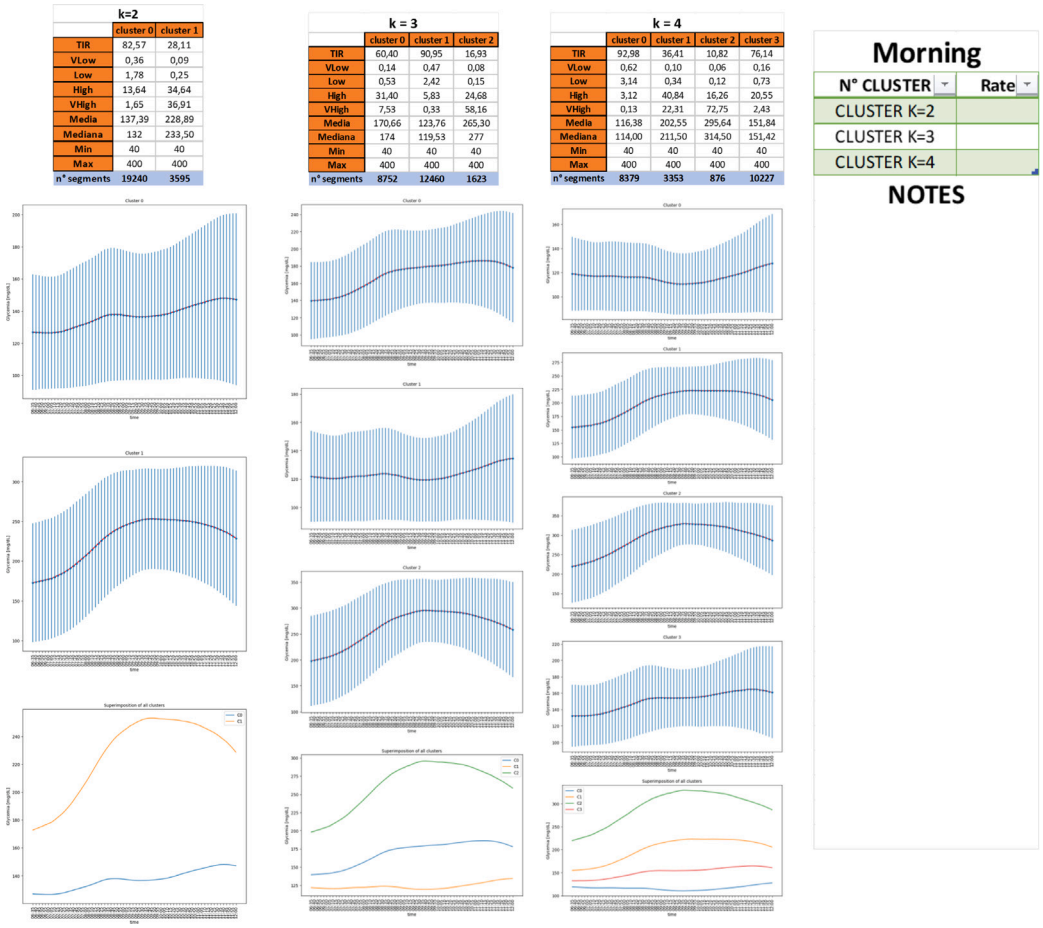
This exhaustive search enabled us to identify the optimal set of hyperparameters that yielded the best model performance. The combination of these hyperparameters was key to fine-tuning the classifier and ensuring its accuracy. The most common metrics derived from the confusion matrix have been chosen to evaluate the classifier: Accuracy, Precision, F1-Score and Sensitivity (Recall) (Powers, 2020). Precision (PREC) is the ratio of correctly classified samples to all samples assigned to that class. Precision is bounded on [0, 1], where 1 represents all correctly predicted samples in the class and 0 represents no correctly predicted samples in the class. The Recall (REC), also known as sensitivity, indicates the rate of positive samples correctly classified and is computed as the ratio between correctly classified positive samples and all samples assigned to the positive class. The F1 score (F1) is the harmonic mean of precision and recall, penalizing extreme values of either. The F1-score is bounded to [0, 1], where 1 represents maximum precision and recall values and 0 represents zero precision and/or recall: The accuracy (ACC) is the ratio between the number of samples that are correctly classified and the overall number of samples in the dataset being assessed.

This step is crucial for future deployment scenarios where data from previously unobserved patients must be rapidly interpreted and profiled. In such cases, the classifier acts as an operational layer for early patient profiling, supporting adaptive therapy decisions based on the glycemic pattern to which the patient most closely resembles. Thus, the classification task transforms the clustering results from a retrospective analysis tool into a prospective decision support system, bridging the gap between data exploration and real-world clinical application.

3.6. Blind clinical evaluation

The obtained clustering results were analyzed by 3 physicians from the "G. Salesi" Hospital, using the blinded interpretation methodology (Boutron et al., 2007) to eliminate any bias. Fig. 5 shows an example of the information provided to the medical team for each phase. The physicians were provided with Excel files containing the glycemic metrics, together with plots of the mean signals and standard deviation of each cluster. In particular, we calculated the metrics described by Battelino et al. (2019):

1. Time in range (TIR): This is defined as the percentage of time that glucose levels are within the target range of 70–180 mg/dL, corresponding to 3.9–10.0 mmol/L. More than 70% of daily readings should fall within this range.
2. Time Above Range (TAR): This is the time spent above the target range (above 180 mg/dL or 10.0 mmol/L). The total number of readings above this threshold should not exceed 25%. Two levels can be distinguished: Level 1 (High), when glucose levels are above 180 mg/dL but below 250 mg/dL, and Level 2 (Very High), when glucose levels are above 250 mg/dL.
3. Time below range (TBR): This is the time spent below the target range (lower than 70 mg/dL or 3.9 mmol/L). These values should not exceed 4% of the total number of readings. Similar to TAR, TBR can be divided into two levels: Level 1 (low) when glucose levels are between 54 and 69 mg/dL, and Level 2 (very low) when glucose levels are below 54 mg/dL.



k=2

	cluster 0	cluster 1
TIR	82,57	28,11
VLow	0,36	0,09
Low	1,78	0,25
High	13,64	34,64
VHigh	1,65	36,91
Media	137,39	228,89
Mediana	132	233,50
Min	40	40
Max	400	400
n° segments	19240	3595

k = 3

	cluster 0	cluster 1	cluster 2
TIR	60,40	90,95	16,93
VLow	0,14	0,47	0,08
Low	0,53	2,42	0,15
High	31,40	5,83	24,68
VHigh	7,53	0,33	58,16
Media	170,66	123,76	265,30
Mediana	174	119,53	277
Min	40	40	40
Max	400	400	400
n° segments	8752	12460	1623

k = 4

	cluster 0	cluster 1	cluster 2	cluster 3
TIR	92,98	36,41	10,82	76,14
VLow	0,62	0,10	0,06	0,16
Low	3,14	0,34	0,12	0,73
High	3,12	40,84	16,26	20,55
VHigh	0,13	22,31	72,75	2,43
Media	116,38	202,55	295,64	151,84
Mediana	114,00	211,50	314,50	151,42
Min	40	40	40	40
Max	400	400	400	400
n° segments	8379	3353	876	10227

Morning

N° CLUSTER	Rate
CLUSTER K=2	
CLUSTER K=3	
CLUSTER K=4	

NOTES

Fig. 5. Example of the information provided to the medical team for the blind evaluation.

We also calculated the Time in Tight Range (TITR), which represents the time spent in a tight target range (between 70 and 140 mg/dL) (Shah et al., 2019). Various statistical parameters, such as mean and median, were calculated for each cluster and phase.

Physicians were tasked with evaluating the best configurations obtained from the clustering results. Based on the available data, they had to provide ratings. They considered factors such as the patient cohort, the use of hybrid closed-loop systems, and their familiarity with the most common and representative configurations. Each physician was assigned a score from 1 to 5, with 5 indicating a highly favorable opinion (representative cluster) and 1 indicating a negative opinion (non-representative cluster). They were also encouraged to include supporting notes to justify their ratings. The six ICC indices were computed to assess the reliability of the three physicians (Koch, 2004). These indices help determine how consistent measurements are, depending on whether raters are fixed or random and whether single or multiple raters are used. ICC1 measures absolute agreement for a single, randomly chosen rater, while ICC2 evaluates consistency among random raters. ICC3 assumes fixed raters and assesses consistency without generalizing beyond them. When multiple raters are involved, the corresponding indices improve reliability by averaging ratings. ICC1k extends ICC1 by considering the mean of multiple raters, leading to higher reliability. ICC2k follows ICC2 but averages multiple random raters, further enhancing consistency. ICC3k builds on ICC3 by averaging fixed raters, reducing variability and increasing reliability.

These indices have been computed using the *pingouin* library and the *intraclass_corr* function.

4. Experiments and results

This section outlines the definition of the training dataset for each component of the proposed pipeline. The original dataset consisted of glucose readings from 91 patients using CGM sensors. The signal was denoised during pre-processing and split into seven phases to capture various glucose change characteristics. The training data is filtered by phase and used to train the autoencoders and clustering algorithms, both unsupervised. It is important to note that the $TimeSeriesKMeans$ algorithm was trained using the autoencoder's compressed output, while the autoencoder itself used the original segments. A test set was created for classifier training

Table 3
Overview of the size reduction for each phase.

Phase	Original size	Reduced size	Average loss
Morning	66	16	0.122
Lunch	48	12	0.146
Afternoon	36	9	0.139
Evening	23	5	0.140
Dinner	36	9	0.154
Night (pre)	24	6	0.139
Night (post)	55	13	0.105

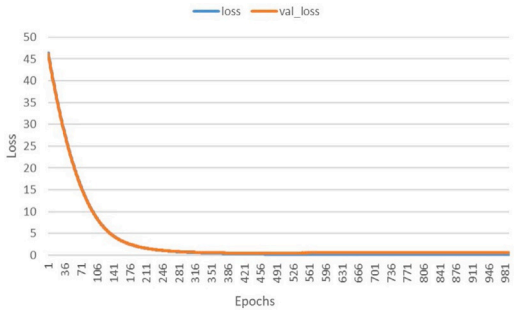


Fig. 6. Generic autoencoder loss representation trend.

by randomly selecting segments from a subset of patients, representing 10% of the original training set. A validation set, 5% of the total, was randomly chosen from the training set. There was no overlap between the patients in the training and test sets. Experiments were conducted using TensorFlow v2.15.0 and Keras v2.15.0, on local clusters consisting of 73 biprocessor systems (16 cores, 512/1024 GB RAM, 1.2 TB storage), and 292 NVIDIA V100 GPUs with 16 GB or 32 GB memory.

4.1. Data-reduction results

This analysis aims to evaluate the performance of autoencoders for data reduction of input features to low-feature-space data in terms of reconstruction accuracy and data compression efficiency. Time series data, consisting of sequential data points typically measured over time intervals, present unique challenges for data compression due to their temporal dependencies. Autoencoder performance in this context is critical for applications that require efficient data storage and transmission without significant information loss.

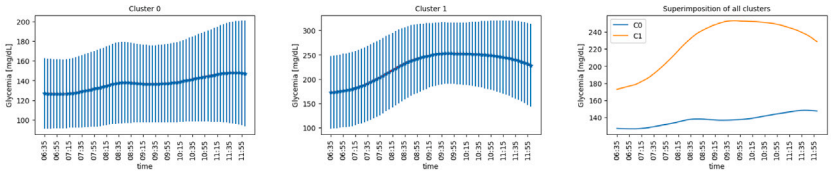
The MSE was used to quantify the quality of the reconstruction. Lower MSE values indicate better reconstruction performance, implying that the autoencoder has effectively captured and retained the essential information from the input data during the compression and decompression process. Fig. 6 shows a qualitative example of the loss progression of the autoencoder during training. The loss curve shows how the error decreases as the model learns over successive iterations. Specifically, this Figure refers to the autoencoder training in the $Night\ (pre)$ phase. Monitoring the loss history helps to understand the learning dynamics of the autoencoder and to diagnose problems such as overfitting or underfitting.

As can be seen, the training error curve and the validation error curve overlap almost perfectly, indicating the absence of overfitting and the effectiveness of the training process.

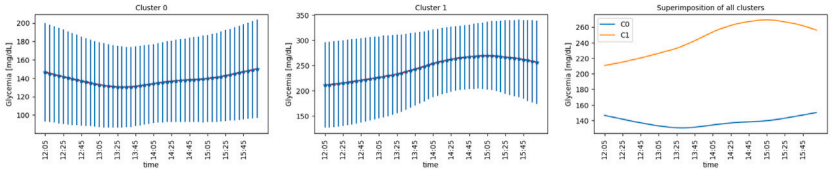
The autoencoder is designed to define a latent space for encoded signals equal to 25% of the original dimension for each phase. This means that the autoencoder compresses the input data to a representation of one-fourth of the original data's size. This reduction is significant for practical purposes because it allows for more efficient storage and faster processing while preserving the most important features of the original data. Table 3 shows the original and reduced dimensions of the data, along with the average error at the end of the training process. The Table shows how effectively the autoencoder compresses the data and achieves reconstruction accuracy. It shows that the average error obtained at the end of the autoencoder training process is between 0.105 and 0.154. This low error, expressed as MSE, confirms the quality of the training. An MSE in this range indicates that the autoencoder is highly capable of reconstructing the input data from its compressed form with minimal loss of information. Such performance underscores the potential of autoencoders for data reduction tasks and highlights their applicability in scenarios where maintaining high reconstruction fidelity is critical.

4.2. Clustering results and clinical evaluation

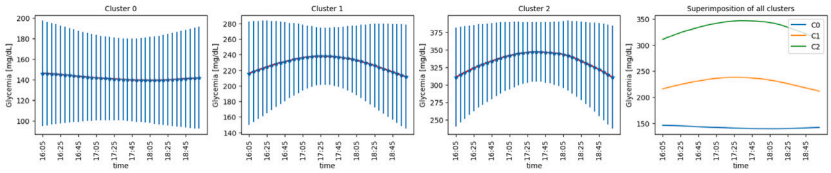
Once the data is reduced, the clustering algorithm comes into play. We applied the elbow method, varying k from 2 to 8, to determine a suitable number of clusters. We applied the *kneed* library and the function named $KneeLocator$ (Satopaa, Albrecht,



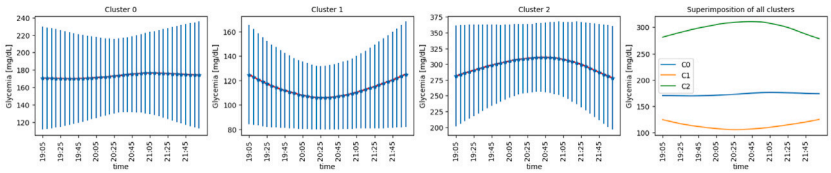
(a) Morning phase for k=2.



(b) Lunch phase for k=2.



(c) Afternoon phase for k=3.



(d) Dinner phase for k=3.

Fig. 7. Glycemic patterns identified as the most typical ones by the blind clinical examination: Morning, Lunch, Afternoon and Dinner phases.

Irwin, & Raghavan, 2011): given x and y arrays, the function attempts to identify the knee/elbow point of a line fit to the data. The knee/elbow is defined as the point of the line with maximum curvature. We used the elbow method to reduce the number of configurations of k to test, given that it does not always identify the optimal k: in fact, it can be ineffective when there is no clear inflexion point, when clusters have varying shapes and densities, or when it tends to overestimate k. Furthermore, due to the nature of our data, we cannot solely rely on statistical metrics at this stage. Therefore, we saw that decay was constant for all phases after $k = 5$ and so we evaluated only three configurations ($k = 2, 3, 4$).

The results of the index calculations for the three tested setups are presented in Table 4. In contrast, Table 5 shows the numerosity of each phase and configuration. Looking at Table 4, we can see that the clustering quality indices change as K increases. The SS tends to decrease with higher K values, suggesting that fewer clusters may provide a clearer separation between groups. Similarly, the CH is higher for lower values of K, indicating a better-defined clustering structure. The DB remains relatively stable, with slightly higher values for K = 2, which suggests that the clusters are fairly compact. Table 5 presents the number of data points in each cluster and it can be observed that with K = 2, the clusters are quite imbalanced, especially in the afternoon and late-night periods. As K increases to 3 or 4, the distribution of points across clusters becomes more even, allowing for a more detailed segmentation of the data. Notably, the fluctuations in cluster size during nighttime periods suggest that behavior in these time slots might be more diverse. Moreover, we performed a statistical analysis by computing the Kruskal–Wallis H statistic and corresponding p-values (Table 6). For each phase, we averaged the signals within each cluster and then used these aggregated values to assess differences across groups. The resulting p-values indicate that, for most phases, there are statistically significant differences between clusters, suggesting that the clustering captures meaningful distinctions in the signal characteristics across different times of day.

From a clinical point of view, we computed the metrics discussed in Section 3.6 and the mean signals of each cluster (Table 7 and Figs. 7, 8). We calculated the mean signal of each phase to have a visual representation of trends in that cluster. Clinicians observed that in daily phase graphs (morning, afternoon, evening), there is a clear distinction between normoglycemic (cluster 0) and dysglycemic patterns (other clusters). Clustering highlights different patient behaviors: those who manage therapy well (e.g., timely boluses) maintain a high TIR, while poor carbohydrate counting or missed boluses lead to hyperglycemia. In complex phases like afternoon, evening, and night, increasing the number of clusters helps identify specific trends. Afternoon difficulties often persist into dinner and night, making these phases critical. Lunch also shows a clear separation between good and poor management, emphasizing the need for personalized glycemic strategies. To assess the clinicians' scores, we calculated the ICC indices (Table 8.

Table 4
Indexes of each cluster.

	Morning	Lunch	Afternoon	Dinner	Evening	Night (pre)	Night (post)
K = 2							
SS	0.66	0.60	0.75	0.62	0.53	0.55	0.64
CH	35652	32102	53938	31671	35185	34935	32977
DB	0.57	0.62	0.47	0.58	0.64	0.60	0.58
K = 3							
SS	0.55	0.56	0.62	0.57	0.57	0.59	0.58
CH	45149	48761	61662	49887	54742	52759	48506
DB	0.57	0.53	0.54	0.52	0.52	0.50	0.52
K = 4							
SS	0.55	0.55	0.55	0.55	0.55	0.55	0.58
CH	59478	60202	78477	59758	65762	57670	57575
DB	0.52	0.53	0.54	0.54	0.52	0.52	0.50

Table 5
Number of signals per each cluster.

N° clusters	Morning	Lunch	Afternoon	Dinner	Evening	Night (pre)	Night (post)
K = 2							
Cluster 0	19246	18363	2734	19011	13669	9629	19250
Cluster 1	3595	4432	19992	3711	8356	13126	3581
K = 3							
Cluster 0	8752	11488	16768	12181	8001	7857	9890
Cluster 1	12460	2186	4526	8123	3402	3109	2122
Cluster 3	1623	9121	1432	2418	10622	11789	10819
K = 4							
Cluster 0	8379	10006	7311	10318	1947	2798	905
Cluster 1	3353	7647	972	1411	8387	5864	10357
Cluster 2	876	1222	12298	6816	6493	3931	8762

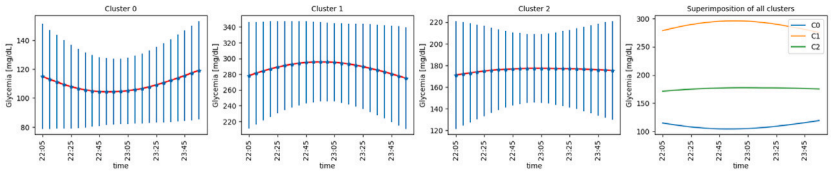
Table 6
Kruskal–Wallis test results for each phase of the day.

Fase	Kruskal–Wallis H	p-value
Morning	3.87	0.0493
Lunch	7.84	0.0051
Afternoon	24.37	5.11×10^{-6}
Evening	60.46	7.45×10^{-14}
Night (Pre)	63.12	1.96×10^{-14}
Night (Post)	17.22	0.00018
Dinner	57.59	3.12×10^{-13}

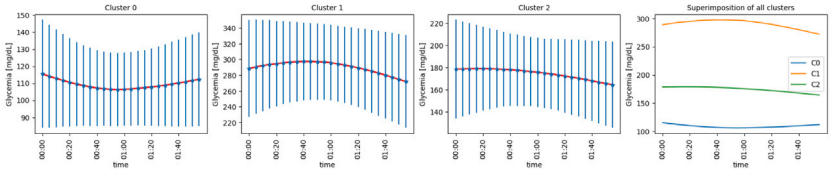
The ICC values reveal significant variations in reliability across different periods. *Lunch* and *Night Pre* exhibit perfect agreement (ICC = 1), indicating highly consistent ratings with no variability. *Afternoon* also shows excellent reliability (ICC1 = 0.90, ICC1k = 0.96), suggesting strong agreement among raters. *Morning* and *Evening* have moderate reliability (ICC1 = 0.56, ICC1k = 0.79), meaning there is some variation in ratings, though averaging raters improves consistency. *Dinner* has the lowest reliability (ICC1 = 0.44, ICC1k = 0.70), suggesting notable inconsistencies among raters. *Night Post* retains strong agreement (ICC1 = 0.76, ICC1k = 0.90), though slightly lower than *Night Pre*. Across all periods, averaging raters (ICC1k, ICC2k, ICC3k) significantly enhances reliability, reinforcing the importance of using multiple ratings for more stable measurements. The wide range of ICC values suggests that some periods are more prone to variability, potentially due to external factors affecting rater consistency.

4.3. Classifier results

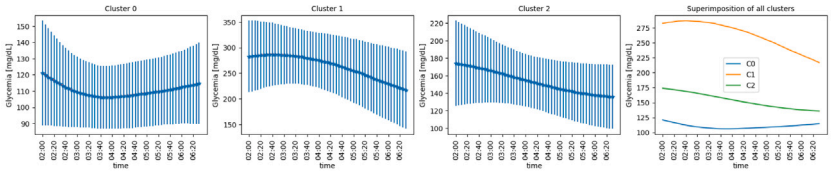
This subsection presents the classifier's results for k values of 2, 3, and 4. Table 8 provides a detailed overview of the classification process, emphasizing the clinical perspective in defining the class settings for each phase. Greater weight was given to the clinical criteria to decide which configuration was more suitable for this study. It results in a two-class setting for the noon and morning phases and a three-class setting for the remaining phases. This approach ensures that the classification aligns more closely with clinical relevance and practical applicability. For each phase, a global evaluation that includes all classes and a by-class assessment that reports metrics specific to each class is provided. An overview of the results confirms the quality of the model, with highly satisfactory accuracy values ranging from 87% in the *Evening* phase to 98% in the *Lunch* phase. This accuracy range highlights



(a) Evening phase for k=3.



(b) Night (pre) phase for k=3.



(c) Night (post) phase for k=3.

Fig. 8. Glycemic patterns identified as the most typical ones by the blind clinical examination: Evening, Night pre and post phases.

Table 7
Glycemic indexes for each phase (C0 = Cluster 0; C1 = Cluster 1; C2 = Cluster 2).

Metric	K = 2				K = 3														
	Morning		Lunch		Afternoon			Dinner			Evening			Night (pre)			Night (post)		
	C0	C1	C0	C1	C0	C1	C2	C0	C1	C2	C0	C1	C2	C0	C1	C2	C0	C1	C2
TIR (%)	82,56	28,10	77,44	18,74	77,83	14,27	2,73	56,43	90,10	3,57	92,39	1,37	55,66	94,16	0,80	59,30	96,15	8,70	79,27
VLow (%)	0,35	0,08	0,73	0,06	0,46	0,06	0,00	0,17	1,24	0,00	1,21	0,00	0,03	1,42	0,00	0,03	0,75	0,04	0,15
Low (%)	1,77	0,25	3,08	0,25	2,11	0,12	0,00	0,67	5,04	0,02	5,18	0,00	0,11	3,90	0,00	0,08	2,47	0,07	0,44
High (%)	13,26	34,64	16,62	34,12	18,28	51,41	5,88	35,14	3,45	18,00	1,19	23,11	40,94	0,49	20,27	38,18	0,57	35,29	18,55
Vhigh (%)	1,65	36,90	2,12	46,81	1,29	34,11	93,14	7,57	0,14	78,38	0,015	75,50	3,23	0,00	78,91	2,39	0,03	55,87	1,56
TITR (%)	57,21	12,15	52,99	7,81	48,80	4,01	0,25	25,75	75,23	1,12	81,71	0,19	15,94	86,86	0,14	15,71	89,90	2,20	39,77
Mean	137,38	228,88	137,89	246,97	141,50	229,46	334,25	173,32	112,71	299,15	108,84	289,11	176,15	108,8	290,10	174,00	110,05	262,64	152,34
Median	132	233	132	245	139	229	341	169	109	298	107	283	174	109	284	171	110	257	147

Table 8
ICC indexes results.

Phase	ICC1	ICC2	ICC3	ICC1k	ICC2k	ICC3k
Dinner	0.44	0.41	0.36	0.7	0.68	0.63
Afternoon	0.90	0.90	0.86	0.96	0.96	0.95
Evening	0.56	0.53	0.43	0.79	0.77	0.69
Morning	0.56	0.53	0.43	0.80	0.77	0.70
Lunch	1.00	1.00	1.00	1.00	1.00	1.00
Night pre	1.00	1.00	1.00	1.00	1.00	1.00
Night post	0.76	0.75	0.66	0.90	0.90	0.85

the model's ability to perform consistently well across different day phases. Furthermore, the model's consistency is underscored by the remarkably high values observed for recall, precision, and F1 scores. Of particular note is the F1 score, which ranges from 0.84 at $Night(post)$ to 0.99 for $Lunch$, indicating robust performance across different metrics. A more detailed analysis shows that the phases with the highest consistency are $Lunch$, $Afternoon$, $Night$ $(Post)$, and $Morning$, where the overall evaluation metrics exceed 0.90. During these phases, the model achieves high accuracy and shows strong recall and precision, resulting in F1 scores that consistently average 0.94. These high scores across multiple metrics underscore the model's ability to classify data across different phases, ensuring reliable performance. The robustness of the model is further demonstrated by its ability to maintain high performance under varying conditions. For example, the consistently high recall values indicate that the model effectively identifies relevant instances in each phase, minimizing false negatives. Similarly, the precision values indicate that the model accurately distinguishes between relevant and irrelevant instances, reducing the rate of false positives.

Table 9 reports the classification metrics (Precision, Recall, F1-score, Accuracy) for each phase and class, based on the clinically selected clustering configurations. Overall, the classifier demonstrates high performance across most phases, with global F1-scores

Table 9
Classifier results for the clinically chosen phases.

Phase		Precision	Recall	F1	Accuracy
Lunch	Global	0.98	0.98	0.98	98%
	0	0.98	1.00	0.99	
	1	1.00	0.96	0.98	
Afternoon	Global	0.95	0.94	0.94	94%
	0	0.90	1.00	0.95	
	1	0.97	0.82	0.89	
	2	1.00	0.98	0.99	
Night (post)	Global	0.95	0.94	0.94	95%
	0	0.96	1.00	0.98	
	1	1	0.72	0.84	
	2	0.93	0.96	0.95	
Dinner	Global	0.95	0.94	0,94	94%
	0	0.90	1	0.95	
	1	1	0.94	0.97	
	2	1	0.44	0.62	
Morning	Global	0.98	0.97	0.97	97%
	0	0.97	1.00	0.99	
	1	1.00	0.86	0.92	
Evening	Global	0.89	0.87	0.85	87%
	0	0.98	0.99	0.99	
	1	1	0.36	0.53	
	2	0.81	0.99	0.89	
Night (pre)	Global	0.91	0.90	0.88	90%
	0	0.99	0.99	0.99	
	1	1.00	0.19	0.32	
	2	0.85	0.99	0.99	

between 0.85 and 0.98 and accuracies ranging from 87% to 98%. Phases modeled with k = 2 (*morning*, *lunch*) show high consistency, with balanced Precision and Recall close to 1.00 across both classes, resulting in optimal values of F1-scores and accuracy (97%–98%). In phases with k = 3, the global metrics remain strong, but class-level variability increases. Class 1 in *Night (pre)* and *Evening* shows a drop in Recall (0.19 and 0.36 respectively), despite high Precision (1.00), suggesting that the classifier struggles to detect the minority cases, due to class imbalance or intra-cluster overlap while class 0 exhibits high Recall and F1, indicating robust identification of the dominant cluster across phases. The *Afternoon* and *Night (post)* phases, despite having three classes, maintain high F1-scores for all clusters ($\geq$0.84), showing effective discrimination even in more complex temporal segments. The divergence between Precision and Recall in some classes highlights potential trade-offs in sensitivity vs. specificity, which may require further tuning or data balancing in future iterations.

5. Discussion

Our work aims to combine the power of dimensionality reduction by autoencoders with the effectiveness of clustering and classification to build a robust system for analyzing and recognizing patterns in glycemic data.

First, we preprocessed the data, then we divided the day into several sub-phases to better identify the impact of certain elements, such as food consumption, over the day. This is the first time such a detailed temporal segmentation has been used in this particular setting.

Another aspect of the proposed methodology, which is innovative as no state-of-the-art solutions propose it, is using an autoencoder as a data reduction technique. By reducing the dimensionality of the original data in the first step, the subsequent clustering process becomes simpler, more manageable, and less susceptible to noise.

The autoencoder's results demonstrate the model's effectiveness in compressing time series data into a significantly smaller latent space while maintaining a high degree of reconstruction accuracy. The low MSE values confirm the ability to learn a compact representation of the data that preserves the essential features needed for reconstruction. Although encoder training is demanding and time-consuming, requiring intervals ranging from several to tens of hours, the resulting performance improvement is substantial. The time consumption is phase-dependent: as the size of the original signal increases, the number of parameters in the network also increases, resulting in longer training times. This significant improvement in performance metrics justifies the significant time investment required during the training phase.

Both segmentation and data reduction had a massive and positive impact on the time and quality of the clustering algorithm. We used the $TimeSeriesKmeans$ method, which allows us to group similar time series by considering the shape and trend of the series rather than individual data points. It uses specific time series distance metrics, such as Dynamic Time Warping (DTW), to measure the similarity between time series. We tested both DTW and Euclidean distance, but the results showed no significant differences. Since each phase's segments were all the same length, we decided to use Euclidean distance instead of DTW to save computing time.

From a statistical point of view, the SS, CH, and DB show consistent and robust clustering performance over different time intervals. The clustering analysis shows optimal results in the afternoon. For k = 2, Silhouette Scores (0.53–0.75) and high Calinski–Harabasz Index values indicate well-separated, compact clusters, while the Davies–Bouldin Index (0.47–0.64) suggests good clustering. Similar trends are observed for k = 3 and k = 4, indicating consistent clustering performance across different times and cluster numbers. Since the metrics did not highlight a predominance of one of the configurations and considering the final application of this work, we decided to also take into consideration the clinical information together with the more purely methodological ones. We believe that for the ultimate purpose and application of this work, it is essential to integrate both clinical information and methodological aspects. This comprehensive approach ensures that the findings are not only methodologically sound but also clinically relevant, thereby enhancing the overall impact and applicability of the research in real-world settings. Dimensionality reduction proves crucial for enhancing both the efficiency and effectiveness of the analysis. For example, the clustering performance for the afternoon phase, evaluated without the use of an autoencoder, shows notably lower results across all metrics. The SS drops from 0.47 for $k = 2$ to 0.31 for $k = 4$, indicating reduced clustering quality. SS shows an average improvement of 75% using the autoencoder-based approach. Moreover, the application of the autoencoder significantly reduces computation time for all the phases, with average simulation durations decreasing from approximately 4 hours to just 30 minutes, highlighting substantial gains in both performance and efficiency. Reading the comments made by the clinicians, they agree that in the *morning* phase graph (Fig. 7(a)), the division between a normoglycemic trend (cluster 0) and a dysglycemic one (cluster 1) can be seen. Two different patient behaviors can be observed: in cluster 0, some patients correctly inject boluses before meals; despite the high standard deviation, the glycemia remains within the optimal range (TIR = 82.56%). On the other hand, cluster 1 shows those who do not correctly count carbohydrates or miss boluses. This can be seen by the fact that their curve always rises and never returns to the optimal range (TIR = 28.10% vs. TAR = 71%). Compared to the results obtained for the same phase but with k = 3, we can understand how the clustering algorithm works: it tries to divide the dysglycemic patterns into more groups to isolate the euglycemic ones. This is an important aspect, especially for the *evening* and *night* (*post*) phases, where it is necessary to increase the number of clusters to better understand the factors that can lead to a particular behavior. For example, cluster 0 of the *evening* phase is the better one, since the TITR and TIR are very high (81.71% and 92.39%). This may be due to an early dinner, proper management of therapy, or the young age of the patients who fall into this cluster. On the contrary, the other two are associated with bad behavior, especially cluster 1. Cluster 2 is the most numerous, and its trend can be seen as a group of people who have a late dinner with incorrect bolus management. Similar considerations can be made for *night* (*post*), where we can appreciate the same trends for all the clusters. For the *afternoon* phase, k = 3 was chosen by the doctors. The same recurrent pattern of the above phases is visible, where there is a more specific division for the dysglycemic trends. Here, cluster 0 represents the desirable one, even if the TIR is not the best. For the other two, there is a great predominance of hyperglycemic episodes, with cluster 2 being the worst with a TAR above 90%. This may be due to the different activities that characterize the afternoon, such as school, physical activity, and snacks, and the lack of ability to manage them. These difficulties can also be appreciated in the *dinner* and *night* (*pre*) phases, since it is the continuation of the afternoon and therefore the effects of the therapy management are still visible. The persistence of these difficulties in both phases suggests that the challenges of glycemic management during the afternoon have prolonged effects, making these phases particularly critical for monitoring and adjusting therapy. Dietary habits and evening routines further add to the complexity of glycemic management, requiring ongoing attention and adaptive strategies from patients and their physicians. For the *lunch* phase, the doctors agreed on the results for k = 2. As for the morning, we can see two distinct patterns of euglycemia (cluster 0) and dysglycaemia (cluster 1). The first is characterized by a small percentage of TBR, mainly in the first hours (12:00–13:00), probably due to an incorrect basal rate setting or a very early lunch bolus. These findings underscore the importance of tailored glycemic management strategies, especially during key phases such as lunch, afternoon, dinner, and night.

There is a certain convergence between the choices made by the medical team and the statistical results: in fact, the configuration for k = 4 is the one that had the lowest scores from the physicians and is the one with the clustering metrics slightly worse than k = 2 and k = 3. For the two remaining configurations, obtaining the doctors' evaluation was crucial. This feedback underscored the importance of validating results from a clinical perspective in the biomedical field, ensuring that the outcomes are meaningful and applicable in practice. While the metrics for k = 2 and k = 3 were comparable, it was only through the medical interpretation that we could derive significance from the results. This clinical insight ultimately guided our decision on which configuration to adopt, reinforcing that statistical measures alone are insufficient without contextual medical understanding to make the final choice. Based on this combined evaluation, we ultimately tested the classifier using k = 2 for the morning and lunch phases, as this configuration aligned best with the medical interpretation and provided meaningful clusters for those specific periods. For the other phases, however, the configuration with k = 3 was selected, as it captured more nuanced distinctions in the data confirmed to be clinically relevant by the physicians. This hybrid approach highlights the importance of adapting model parameters not only based on algorithmic performance, but also on domain expertise and practical utility in real-world biomedical applications.

Finally, we use the labels obtained from clustering to classify the data. A closer analysis shows that the phases with the highest consistency are *Lunch*, *Afternoon*, *Night* (*Post*), *Dinner*, and *Morning*, where the overall evaluation metrics are above 0.90. In these phases, the metrics for recall, precision, and F1 score each reach an average value of 0.94, indicating a high level of consistency and reliability in the model's performance during these periods. Table 9 shows the classification results from the clinician's point of view. These results reflect the traditional, expert-driven approach to evaluating patient data. However, to provide a comprehensive evaluation, we also performed a data-driven analysis for each phase using different cluster settings, specifically k = 2, k = 3, and k = 4 clusters. The results of these data-driven analyses are presented in Tables 4, 5 and 6 in the Supplementary Materials.In particular, the data-driven perspective with k = 2 clusters is the best setting, achieving remarkable accuracy ranging from 97% to 99%. Looking at the averages, the k = 2 setting has an average accuracy of 98%, a significant improvement of 5,1% percentage

points over the clinical perspective, which has an average accuracy of 93%. This significant increase highlights the effectiveness of the data-driven approach in classifying patient data. The average accuracy for the k = 3 and k = 4 settings are slightly lower, at 91% and 90%, respectively. Although these accuracies are somewhat lower than those achieved by the clinical perspective, they still reflect strong performance in the classification tasks. This indicates that the model maintains high classification standards even with more complex clustering. Again, the recall, precision, and F1 metrics confirm the quality of the results. All of these metrics have high average values above 0.90, confirming the robustness and reliability of the data-driven approach. A high recall indicates that the model successfully identifies a large proportion of relevant instances, while a high precision indicates that the majority of identified instances are relevant. The F1 score, which is the harmonic mean of recall and precision, further validates the balanced performance of the model.

The classification performances highlight an important aspect regarding the proposed model's support of the clinician's decision-making process. Indeed, the data-driven approach demonstrates superior classification capabilities regarding patient profiling compared to the clinician's perspective. This suggests that integrating data-driven insights could improve clinical decision-making by providing more accurate and nuanced patient profiles. Although slightly worse, the k = 3 and k = 4 settings could also be evaluated for possible study. Despite their slightly lower accuracy, these settings still perform well on all evaluation metrics, making them viable options for further exploration. Evaluating these settings could provide additional insight into the model's performance and the potential trade-offs between different levels of data clustering complexity. The analysis demonstrates the effectiveness of the data-driven approach, particularly with the k = 2 setting, in improving classification accuracy and supporting clinical decision-making.

For honest discussion, it is important to acknowledge that the model's performance is lower than desired in a few cases. Specifically, during the dinner phase, the classifier struggles with the second class, demonstrating a lack of recognition and, consequently, lower performance metrics. While the overall performance in most phases remains robust and highly satisfactory, the disparity observed in the dinner phase highlights an area where the model's classification capabilities are less effective. This issue indicates a potential need for further refinement and adaptation of the model to better address the unique characteristics and challenges of different phases. These results highlight the model's effectiveness in accurately classifying data across different phases, demonstrating its reliability and potential utility in clinical applications. Table 9 shows the performance metrics of the classifier for each clinically validated phase, based on the selected number of clusters. Overall, the classifier demonstrated strong performance across most phases, with global F1-scores ranging from 0.85 to 0.98 and accuracies between 87% and 98%. Both the morning and lunch phases, which were modeled using k = 2, performed quite well, with global F1-scores of 0.97 and 0.98, respectively. This demonstrated that the k = 2 clustering configuration was adequate and in line with clinical expectations. For the phases modeled with k = 3—including afternoon, dinner, evening, and the night segments—results showed slightly more variability in class-level performance. Recall scores for various classes in the night (pre) and evening phases were notably lower (class 1 had 0.19 and 0.36 recall, respectively), indicating potential difficulties in differentiating between specific behavioral states during these periods. The classifier nonetheless demonstrated strong global performance, demonstrating its ability to adapt even in more complex clustering situations. Remarkably, the classifier maintained good accuracy even while its recall was lower for several minority classes. This implies that when it assigns an instance to these classes, it is generally correct, but somewhat cautious. In clinical settings, such as alert systems, where false positives are more harmful than false negatives, this behavior could be preferred.

These findings reinforce the benefit of combining data-driven metrics with clinical judgment when determining the optimal number of clusters per phase. The mixed use of k = 2 and k = 3, guided by medical evaluation, enabled the classifier to maintain both statistical soundness and clinical relevance.

5.1. Limitations

Despite the promising results, some limitations must be highlighted. To generate a subset of optimal profiles that can match a greater number of patients, it is essential to extend the cohort of participants evaluated. Increasing the sample size will improve the reliability and applicability of the results to broader populations. It would also allow for considering other variables such as physical activity and diet. Moreover, the relatively small number of physicians currently involved in the evaluation process limits the overall significance of the findings. However, this study represents an important first step, as the metrics obtained are promising and suggest a solid foundation for further investigation.

Other limitations concern the AI aspects of the proposed approach. The main issue revolves around adopting a uniform autoencoder and classifier for all phases. This approach introduces a lack of specificity, as each phase has unique structural characteristics that should ideally be considered when designing the neural network architecture. Although the model can adapt to the varying lengths of the phases, specific models for each phase are essential to capture the data's intrinsic characteristics better. Addressing this limitation requires developing phase-specific models that can adapt to each phase's idiosyncrasies while maintaining overall coherence and consistency across the system. Such an approach would involve adapting the architecture and parameters of the autoencoder and classifier better to match each phase's specific characteristics and requirements.

6. Conclusion

Using clustering algorithms, this study characterizes glucose profiles to find recurring patterns throughout the day, generalizes the most common glycemic behaviors, and trains a classifier to detect them automatically. To our knowledge, this is the first effort to develop such a pipeline using non-simulated glucose profiles. This innovative method provides a more realistic representation of

actual glucose fluctuations, which holds great promise for improving individualized diabetes care and treatment plans. Combining autoencoders and clustering significantly accelerated the clustering process, resulting in more stable classifier results. The metrics also confirm this, where the average autoencoder loss is less than 0.15 and the classifier accuracies are above 89%. Future research should use larger data sets and explore associations between glucose trends and patient factors such as socioeconomic status, medical conditions, age, and boluses. Educating caregivers about data sharing is critical for accurate, personalized diabetes management. From an AI perspective, addressing the lack of specificity in current approaches is a step toward maximizing the utility and effectiveness of AI-based solutions in personalized therapy and healthcare.

Insulin pump settings can be improved and tailored to each patient's specific glycemic patterns by detecting distinctive glucose trends throughout the day. This optimization reduces the chance of hyperglycemia and hypoglycemia while enhancing overall glucose management. These developments may eventually result in more stable blood sugar levels, which may lessen diabetes-associated difficulties and enhance the lives of young patients and their carers.

CRediT authorship contribution statement

Giovanni Paragliola: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Conceptualization. **Sara Campanella:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Data curation. **Valentino Cherubini:** Writing – review & editing, Visualization, Validation, Investigation, Data curation, Conceptualization. **Valentina Tiberi:** Writing – review & editing, Visualization, Validation, Resources, Formal analysis, Data curation. **Paola Pierleoni:** Writing – review & editing, Visualization, Project administration, Investigation, Funding acquisition, Conceptualization. **Alberto Belli:** Writing – review & editing, Visualization, Validation, Resources, Project administration, Investigation, Funding acquisition, Formal analysis. **Antonio Iannilli:** Writing – review & editing, Visualization, Validation, Data curation. **Lorenzo Palma:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Lorenzo Palma reports financial support was provided by European Union. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was partially supported by the European Union - Next Generation EU under the Italian National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.3, CUP E63C22002070006, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART").

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.smhl.2025.100616.

Data availability

The authors do not have permission to share data.

References

Afsaneh, E., Sharifdini, A., Ghazzaghi, H., & Ghobadi, M. Z. (2022). Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: A comprehensive review. *Diabetology and Metabolic Syndrome*, *14*(1), http://dx.doi.org/10.1186/s13098-022-00969-9.

Ahmed, B. M., Ali, M. E., Masud, M. M., & Naznin, M. (2024). Recent trends and techniques of blood glucose level prediction for diabetes control. *Smart Health*, *32*, Article 100457. http://dx.doi.org/10.1016/j.smhl.2024.100457.

Al-Salman, W., Li, Y., Oudah, A. Y., & Almaged, S. (2023). Sleep stage classification in EEG signals using the clustering approach based probability distribution features coupled with classification algorithms. *Neuroscience Research*, *188*, 51–67.

Amit, G., Gavriely, N., & Intrator, N. (2009). Cluster analysis and classification of heart sounds. *Biomedical Signal Processing and Control*, *4*(1), 26–36. http://dx.doi.org/10.1016/j.bspc.2008.07.003.

Arnia, F., Saddami, K., Roslidar, R., Muharar, R., & Munadi, K. (2024). Towards accurate diabetic foot ulcer image classification: Leveraging CNN pre-trained features and extreme learning machine. *Smart Health*, *33*, Article 100502. http://dx.doi.org/10.1016/j.smhl.2024.100502.

Bank, D., Koenigstein, N., & Giryes, R. (2023). Autoencoders. In *Machine learning for data science handbook: Data mining and knowledge discovery handbook* (pp. 353–374). Springer.

Battelino, T., Danne, T., Bergenstal, R. M., Amiel, S. A., Beck, R., Biester, T., et al. (2019). Clinical targets for continuous glucose monitoring data interpretation: Recommendations from the international consensus on time in range. *Diabetes Care*, *42*(8), 1593–1603.

Biagi, L., Bertachi, A., Giménez, M., Conget, I., Bondia, J., Martín-Fernández, J. A., et al. (2019). Individual categorisation of glucose profiles using compositional data analysis. *Statistical Methods in Medical Research*, *28*(12), 3550–3567.

Boughton, C. K., & Hovorka, R. (2024). The role of automated insulin delivery technology in diabetes. *Diabetologia*, *67*(10), 2034–2044.

Boutron, I., Guittet, L., Estellat, C., Moher, D., Hróbjartsson, A., & Ravaud, P. (2007). Reporting methods of blinding in randomized trials assessing nonpharmacological treatments. *PLoS Medicine*, *4*(2), Article e61.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics. Theory and Methods*, *3*(1), 1–27.

Campanella, S., Paragliola, G., Cherubini, V., Pierleoni, P., & Palma, L. (2024). Towards personalized AI-based diabetes therapy: A review. *IEEE Journal of Biomedical and Health Informatics*, 1–16. http://dx.doi.org/10.1109/JBHI.2024.3443137.

Campanella, S., Sabbatini, L., Cherubini, V., Tiberi, V., Marino, M., Pierleoni, P., et al. (2022). Machine learning approach for care improvement of children and youth with type 1 diabetes treated with hybrid closed-loop system. *Electronics*, *11*(14), 2227.

Chen, L., Wang, K., Li, M., Wu, M., Pedrycz, W., & Hirota, K. (2022). K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human–robot interaction. *IEEE Transactions on Industrial Electronics*, *70*(1), 1016–1024.

Cheng, H., Zhu, J., Li, P., & Xu, H. (2023). Combining knowledge extension with convolution neural network for diabetes prediction. *Engineering Applications of Artificial Intelligence*, *125*, Article 106658. http://dx.doi.org/10.1016/j.engappai.2023.106658.

Cherubini, V., Rabbone, I., Berioli, M. G., Giorda, S., Lo Presti, D., Maltoni, G., et al. (2021). Effectiveness of a closed-loop control system and a virtual educational camp for children and adolescents with type 1 diabetes: A prospective, multicentre, real-life study. *Diabetes, Obesity and Metabolism*, *23*(11), 2484–2491.

Contreras, I., Quirós, C., Giménez, M., Conget, I., & Vehi, J. (2016). Profiling intra-patient type I diabetes behaviors. *Computer Methods and Programs in Biomedicine*, *136*, 131–141.

Craig, M. E., Hattersley, A., & Donaghue, K. C. (2009). Definition, epidemiology, and classification of diabetes in children and adolescents. *Pediatric Diabetes*, *10*, 3–12.

Cui, E. H., Goldfine, A. B., Quinlan, M., James, D. A., & Sverdlov, O. (2023). Investigating the value of glucodensity analysis of continuous glucose monitoring data in type 1 diabetes: An exploratory analysis. *Frontiers in Clinical Diabetes and Healthcare*, *4*, http://dx.doi.org/10.3389/fcdhc.2023.1244613.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2), 224–227.

Eissa, M. R., Good, T., Elliott, J., & Benaissa, M. (2020). Intelligent data-driven model for diabetes diurnal patterns analysis. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), 2984–2992.

Ellahham, S. (2020). Artificial intelligence: The future for diabetes care. *The American Journal of Medicine*, *133*(8), 895–900. http://dx.doi.org/10.1016/j.amjmed.2020.03.033.

Gautier, T., Ziegler, L. B., Gerber, M. S., Campos-Náñez, E., & Patek, S. D. (2021). Artificial intelligence and diabetes technology: A review. *Metabolism*, *124*, Article 154872.

Gecili, E., Huang, R., Khoury, J. C., King, E., Altaye, M., Bowers, K., et al. (2020). Functional data analysis and prediction tools for continuous glucose-monitoring studies. *Journal of Clinical and Translational Science*, *5*(1), http://dx.doi.org/10.1017/cts.2020.545.

Grasso, E. A., & Chiarelli, F. (2024). Type 1 diabetes and other autoimmune disorders in children. *Pediatric Diabetes*, *2024*(1), Article 5082064.

Hall, H., Perelman, D., Breschi, A., Limcaoco, P., Kellogg, R., McLaughlin, T., et al. (2018). Glucotypes reveal new patterns of glucose dysregulation. *PLoS Biology*, *16*(7), Article e2005143.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108.

Huang, X., Ye, Y., Xiong, L., Lau, R. Y., Jiang, N., & Wang, S. (2016). Time series k-means: A new k-means type smooth subspace clustering for time series data. *Information Sciences*, *367–368*, 1–13. http://dx.doi.org/10.1016/j.ins.2016.05.040.

Kahkoska, A. R., Adair, L. A., Aiello, A. E., Burger, K. S., Buse, J. B., Crandell, J., et al. (2019). Identification of clinically relevant dysglycemia phenotypes based on continuous glucose monitoring data from youth with type 1 diabetes and elevated hemoglobin A1c. *Pediatric Diabetes*, *20*(5), 556–566.

Khalaf, A., Nabian, M., Fan, M., Yin, Y., Wormwood, J., Siegel, E., et al. (2020). Analysis of multimodal physiological signals within and between individuals to predict psychological challenge vs. threat. *Expert Systems with Applications*, *140*, Article 112890.

Klonoff, D. C., Bergenstal, R. M., Cengiz, E., Clements, M. A., Espes, D., Espinoza, J., et al. (2025). CGM data analysis 2.0: Functional data pattern recognition and artificial intelligence applications. arXiv:2505.07885.

Koch, G. G. (2004). Intraclass correlation coefficient. *Encyclopedia of Statistical Sciences*.

Lal, R. A., Ekhlaspour, L., Hood, K., & Buckingham, B. (2019). Realizing a closed-loop (artificial pancreas) system for the treatment of Type 1 diabetes. *Endocrine Reviews*, *40*(6), 1521–1546. http://dx.doi.org/10.1210/er.2018-00174.

Li, P., Pei, Y., & Li, J. (2023). A comprehensive survey on design and application of autoencoder in deep learning. *Applied Soft Computing*, *138*, Article 110176. http://dx.doi.org/10.1016/j.asoc.2023.110176.

Lim, M. H., Cho, Y. M., & Kim, S. (2022). Multi-task disentangled autoencoder for time-series data in glucose dynamics. *IEEE Journal of Biomedical and Health Informatics*, *26*(9), 4702–4713.

Lobo, B., Farhy, L., Shafiei, M., & Kovatchev, B. (2021). A data-driven approach to classifying daily continuous glucose monitoring (CGM) time series. *IEEE Transactions on Biomedical Engineering*, *69*(2), 654–665.

Makroum, M. A., Adda, M., Bouzouane, A., & Ibrahim, H. (2022). Machine learning and smart devices for diabetes management: Systematic review. *Sensors*, *22*(5), 1843.

Mao, Y., Tan, K. X. Q., Seng, A., Wong, P., Toh, S.-A., & Cook, A. R. (2022). Stratification of patients with diabetes using continuous glucose monitoring profiles and machine learning. *Health Data Science*.

Paragliola, G., & Coronato, A. (2018). Gait anomaly detection of subjects with Parkinson's disease using a deep time series-based approach. *IEEE Access*, *6*, 73280–73292.

Paragliola, G., & Coronato, A. (2021). An hybrid ECG-based deep network for the early identification of high-risk to major cardiovascular events for hypertension patients. *Journal of Biomedical Informatics*, *113*, Article 103648. http://dx.doi.org/10.1016/j.jbi.2020.103648.

Patterson, C. C., Karuranga, S., Salpea, P., Saeedi, P., Dahlquist, G., Soltesz, G., et al. (2019). Worldwide estimates of incidence, prevalence and mortality of type 1 diabetes in children and adolescents: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes Research and Clinical Practice*, *157*, Article 107842.

Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. CoRR, arXiv:2010.16061.

Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops* (pp. 166–171). IEEE.

Shah, V. N., DuBose, S. N., Li, Z., Beck, R. W., Peters, A. L., Weinstock, R. S., et al. (2019). Continuous glucose monitoring profiles in healthy nondiabetic participants: A multicenter prospective study. *The Journal of Clinical Endocrinology & Metabolism*, *104*(10), 4356–4364.

Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics* (pp. 747–748). IEEE.

Tao, R., Yu, X., Lu, J., Shen, Y., Lu, W., Zhu, W., et al. (2021). Multilevel clustering approach driven by continuous glucose monitoring data for further classification of type 2 diabetes. *BMJ Open Diabetes Research and Care*, *9*(1), Article e001869.

Wang, Y., Yao, H., & Zhao, S. (2016). Auto-encoder based dimensionality reduction. *Neurocomputing*, *184*, 232–242, RoLoD: Robust Local Descriptors for Computer Vision 2014.

Xie, J., & Wang, Q. (2019). A personalized diet and exercise recommender system for type 1 diabetes self-management: An in silico study. *Smart Health*, *13*, Article 100069. http://dx.doi.org/10.1016/j.smhl.2019.100069.

Xu, Q., Nwe, T. L., & Guan, C. (2014). Cluster-based analysis for personalized stress evaluation using physiological signals. *IEEE Journal of Biomedical and Health Informatics*, *19*(1), 275–281.

Yang, M., Cai, C., & Hu, B. (2022). Clustering based on eye tracking data for depression recognition. *IEEE Transactions on Cognitive and Developmental Systems*.

Zhang, Y., Holt, T. A., & Khovanova, N. (2016). A data driven nonlinear stochastic model for blood glucose dynamics. *Computer Methods and Programs in Biomedicine*, *125*, 18–25.

Zhu, T., Li, K., Herrero, P., & Georgiou, P. (2021). Deep learning for diabetes: A systematic review. *IEEE Journal of Biomedical and Health Informatics*, *25*(7), 2744–2757. http://dx.doi.org/10.1109/JBHI.2020.3040225.