Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Evaluating Time Series Classification Models for Nocturnal Hypoglycemia: From Predictive Performance to Environmental Impact

FRANCISCO J. LARA-ABELENDA[1], DAVID CHUSHIG-MUZO[1], CARMELO BETANCORT ACOSTA[2,3], ANA M. WÄGNER[2,3], CONCEIÇÃO GRANJA[4], and CRISTINA SOGUERO-RUIZ,[1],

[1]Department of Signal Theory and Communications, Telematics and Computing Systems, Rey Juan Carlos University, Madrid, Spain (e-mail: francisco.lara@urjc.es; david.chushig@urjc.es; cristina.soguero@urjc.es)
[2]Endocrinology and Nutrition Department, Complejo Hospitalario Universitario Insular Materno-Infantil, CHUIMI, Las Palmas de Gran Canaria, Spain
[3]Instituto Universitario de Investigaciones Biomédicas y Sanitarias, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain (e-mail: carmelo.betancort@fpct.ulpgc.es; ana.wagner@ulpgc.es)
[4]Norwegian Centre for E-health Research, University Hospital of North Norway, Tromsø, Norway (email: conceicao.granja@ehealthresearch.no)

Corresponding author: Lara-Abelenda, Francisco J. (e-mail: francisco.lara@urjc.es).

This work was supported by the European Commission through the H2020-EU.3.1.4.2, European Project WARIFA (Watching the risk factors: Artificial intelligence and the prevention of chronic conditions) under Grant Agreement; and by the Spanish federal grants PID2019-107768RA-I00 & PID2023-149457OB-I00 (all funded by the agency AEI/10.13039/501100011033). The study sponsors have not been involved in any stage of the study.

ABSTRACT Type 1 Diabetes (T1D) is an autoimmune condition that results in an insulin deficiency. People with T1D require the administration of exogenous insulin to maintain target glucose levels. However, insulin therapy can cause hypoglycemic episodes, which occur when blood glucose levels fall below 70 mg/dL. Nocturnal Hypoglycemia (NH) occurs while the individual is asleep and can lead to different clinical complications. Developing predictive approaches to predict NH before sleep could reduce these episodes and mitigate acute complications. While numerous models exist for Time Series Classification (TSC), their use for NH prediction remains limited. This study evaluates 14 different TSC models for NH prediction, assessing their performance by evaluating classification metrics, computational time, and environmental impact (measured by energy consumption and CO_2 emissions). The approaches include distance-based, convolutional-based, deep learning, dictionary-based, feature-based, shapelet-based, and interval-based methods. We employed glucose data from 52 individuals with T1D. Experimental results showed that interval-based and feature-based approaches achieved the best predictive performance, obtaining the highest Area Under the Curve Operator (AUCROC) of 0.703. Additionally, both demonstrated low environmental impact due to their short computational time. However, substantial differences in environmental impact were observed depending on the approach. Distance-based methods and deep learning approaches exhibited the highest environmental impact. This paper provides key insights into the effectiveness of TSC models for NH prediction, highlighting the trade-off between model performance and environmental impact.

INDEX TERMS Nocturnal hypoglycemia, type 1 diabetes, continuous glucose monitoring, time series classification, carbon footprint, green machine learning, environmental impact.

I. INTRODUCTION

Over the past several years, the incidence and prevalence of Type 1 Diabetes (T1D) have increased worldwide [1]. In 2021, T1D affected an estimated 8.4 million people globally, and approximately 510,000 new cases were reported [2]. T1D is an autoimmune disease caused by the destruction of pancreatic islet beta cells that leads to either absolute or partial insulin deficiency [3]. To maintain target glycemic levels, commonly between 70 and 180 mg/dL, exogenous insulin is administered through either multiple daily injections or continuous subcutaneous infusion pumps [4].

Hypoglycemia is a side effect of insulin therapy that occurs when blood glucose falls below 70 mg/dL for at least 20 minutes [5]–[7]. Nocturnal Hypoglycemia (NH) refers to a hypoglycemic episode that occurs while a person is asleep at night and it is particularly common among individuals with T1D [8]. According to an international consensus [5]–[7], NH is defined as hypoglycemia that occurs while an individual is asleep. NH is associated with clear disturbances in sleep and negatively affects subjective sleep quality. Sleep disturbances can also lead to insulin resistance and poor glycemic control [9]. Therefore, predicting NH may help individuals with T1D to better manage glucose control, reduce hypoglycemic events and improve patient's quality of life.

Technological advances in Continuous Glucose Monitoring (CGM) devices, which measure interstitial glucose, have significantly improved the tracking, control and management of glucose levels over time, improving patient's quality of life and preventing clinical complications [10]. The emergence of Machine Learning (ML) models has created valuable opportunities to enhance healthcare applications, supporting clinical decision-making. More specifically, in diabetes research, several ML models have been proposed to forecast glucose levels over different prediction horizons (PH) [11]–[13], including the prediction of potential episodes of severe hypoglycemia [14], [15]. Several studies have employed ML-based models to predict hypoglycemia [16], but a few have been proposed for predicting NH [17]–[21]. Most approaches first perform a statistical feature extraction from CGM data, and then, different ML models are employed to predict NH. In the literature, a variety of Time Series Classification (TSC) methods, including distance-based, dictionary-based, shapelet-based, and interval-based approaches have shown excellent results in other applications [22], [23], but these remain unexplored for NH prediction.

In the past decade, the resurgence of Artificial Neural Networks (ANNs) and the advent of Deep Learning (DL) models have brought great milestones, with high predictive results in multiple domains [24], [25]. However, the current trend in DL focuses on enhancing model performance by increasing their size (through additional layers and neurons), resulting in a higher number of parameters and floating-point operations. This increases the computational cost for training and inference, requiring significant memory and energy resources, and leading to substantial energy consumption, water usage for cooling data centers, and increased greenhouse gas emissions [26]–[28]. To address this, tools that assess energy consumption and carbon emissions of ML/DL models [29], [30] are crucial. These tools promote the creation of more sustainable models, aligning with green artificial intelligence, which prioritizes reducing computational costs while maintaining a balance between efficiency and predictive performance [27].

Therefore, this paper aims to provide a comprehensive evaluation of different models to predict the occurrence of NH episodes. The evaluation is conducted across two key categories: *(ii)* predictive performance; and *(ii)* environmental impact. To achieve this, we train 14 different models across seven categories (two models per category). The next methods were evaluated: *(1) distance-based methods* (Proximity Forest (PF) [31] and Dynamic Time Warping with K-Near Neighbours (DTW-KNN)) [32]); *(2) convolutional-based models* (Random Convolutional Kernel Transform (ROCKET) [33] and T-Rep [34]); *(3) DL-based models* Inception [35] and LITE [36]; *(4) dictionary-based techniques* (WEASEL_V2 [37] and Multiple Representations Sequence Miner (MrSQM) [38]); *(5) feature-based techniques* (with features extracted using GlucoStats, tsfresh [39]); *(6) shapelet approaches* (Random Scalable and Accurate Subsequence Transform (RSAST)) [40] and Random Dilated Shapelet Transform (RDST) [41]; *(7) interval-based methods* (QUANT [42], the Randomized-Supervised Time Series Forest (r-STSF) [43]). We employed CGM data from 52 individuals with T1D and collected at the Complejo Hospitalario Insular-Materno Infantil de Las Palmas de Gran Canaria. To the best of our knowledge, this is the first study that evaluates multiple TSC models for predicting NH in individuals with T1D, analyzing both predictive performance and environmental impact.

The primary contributions of this paper are as follows:

- Implementation of novel TSC methods from seven different categories to predict the occurrence of NH using CGM data belonging to T1D people.
- Evaluation of the performance of 14 different models by analyzing their performance and computational efficiency.
- Analysis of the environmental impact of each of the 14 models by computing their energy consumption and CO_2 emissions.

This paper is organized as follows. A review of related work is presented in Section II. The dataset description and preprocessing are described in Section III, whereas the proposed methodology and TSC models are further detailed in Section IV. Experimental results are shown in Section V, and finally, discussion and conclusions are presented in Section VI and Section VII, respectively.

II. RELATED WORK

In the literature, numerous studies have employed ML and DL models to predict hypoglycemia [16], [46], but few have investigated NH prediction. This section provides an overview of state-of-the-art methods used for NH prediction, with a particular emphasis on datasets from individuals with T1D. A summary of these methods is shown in Table 1.

Vu *et al.* [17] employed a large dataset consisting of extracted features from CGM data over one million nights. The Random Forest (RF) model was used to predict NH within a 6-hour PH, achieving an Area Under the Curve Operator (AUCROC) of 0.90 for early night (from midnight to 03:00 AM) and 0.75 for late night (from 03:00 AM to 06:00 AM). In a similar way, Mosquera-Lopez *et al.* [18] extracted 59 features from CGM data and information about insulin, and meals from 124 individuals, with a total of 22,804 nights. The predictions were made during sleep between 00:00 and 05:59 (approximately 6 hours of PH). Support vector regression and a decision-theoretic criterion were used to predict overnight minimum glucose levels and NH, obtaining an AUCROC of 0.86. Jensen *et al.* [19] combined CGM data with information about meals, insulin usage, and demographics from a dataset

TABLE 1: A comparative analysis of methodologies presented in the literature for predicting nocturnal hypoglycemia in individuals with diabetes. These works were sorted by publication date, ordered from the earliest to the latest. Note that samples refer to the number of nights of CGM data considered for the study.

Study	Year	# people	# samples	Prediction horizon	Models used	Data modality	Dataset
Vu *et al.* [17]	2019	10,000	1,000,000	6 hours	RF	CGM statistics, insulin	Private
Mosquera-Lopez *et al.* [18]	2020	134	22,804	6 hours	SVR	CGM statistics insulin, meals	Tidepool platform
Jensen *et al.* [19]	2020	463	4,721	6 hours	LDA	CGM statistics insulin, meals	Clinical trial, daily living
Berikov *et al.* [20]	2022	406	6,451	15 minutes, 30 minutes	RF, LASSO MLP	CGM statistics	Private hospital
Mosquera-Lopez *et al.* [21]	2024	366	44.154	8 hours	SVR, ENN	CGM statistics	Glooko, T1DEXI
Kozinetz *et al.* [44]	2024	380	380	30 minutes	CNNs	CGM	Private hospital
Leutheuser *et al.* [45]	2024	13	66	8 hours	RF, LR, MLP, RNN	CGM statistics	Private hospital

Description of acronyms: Continuous Glucose Monitoring (CGM), Convolutional Neural Networks (CNN), Evidential Neural Network (ENN), Linear Discriminant Analysis (LDA), Least Absolute Shrinkage and Selection Operator (LASSO), Linear Regression (LR), Multilayer Perceptron (MLP), Recurrent Neural Network (RNN), Random Forest (RF), Support Vector Regressor (SVR), The Type 1 Diabetes and Exercise Initiative (T1DEXI).

of 463 people with T1D, with a total of 4,721 nights. By combining the linear discriminant analysis with feature forward selection, the authors achieved an AUCROC of 0.79 for NH prediction within a 6-hour PH.

Berikov *et al.* [20] analyzed CGM data from 406 adults, employing RF, Least Absolute Shrinkage and Selection Operator (LASSO), and the Multilayer Perceptron (MLP) for NH prediction in PHs of 15 and 30 minutes. The models obtained AU-ROC values of 0.97 and 0.94, respectively. The extracted features included various glucose dynamics statistics, such as the coefficient of variation and low blood glucose index. Mosquera *et al.* [21] extracted descriptive features from CGM data across different time frames before sleep (*e.g.,* previous 7 nights, daytime or 30 minutes). The Evidential Neural Network was developed to predict NH, achieving an AUCROC of 0.80 and 0.71 for 0–4 hours versus. 4–8 hours, both after bedtime. Kozinetz *et al.* [44] employed CGM data from 380 T1D individuals to train multiple DL and ML models. The models predicted NH for 30-minute PH, achieving an F1-score of 0.86. Finally, Leutheuser*et al.* [45] extracted statistics from CGM data and combined them with physiological data from 13 children over 66 days, specifically on days when they performed physical exercise. Logistic Regression (LR), RF, and MLP models were trained. The best-performing model was RF which achieved an AUCROC of 0.752. However, the high standard deviation values suggested that the small sample size significantly impacted the model's stability.

Following the review of the literature, we identified two research gaps in the existing methodologies. First, most approaches used statistical features extracted from CGM data to predict NH [17]–[21]. After feature extraction, several ML-based models such as RF, LASSO, LR and MLP were considered to predict NH. Only two studies explored the use of DL-based models on CGM data [44], [45]. In [44], the authors used Convolutional Neural Networks (CNNs), whereas in [45], recurrent neural networks were considered. The literature review for time-series classification revealed that various methodologies beyond statistical feature extraction and traditional ML approaches have been explored in different areas such as biomedical applications, human activity recognition, cybersecurity and others [23]. However, most TSC models remain underexplored for NH prediction. Second, none of the previous methods provided an analysis of time and energy consumption as well as CO_2 emissions. The increasing computational demands of ML and DL models require evaluating not only the predictive performance but also the environmental impact and carbon footprint. Despite the growing awareness of the environmental impact of AI, there is a lack of studies quantifying these factors in the context of TSC, and more particularly for NH prediction.

III. DATASET DESCRIPTION AND PREPROCESSING

This section presents the dataset used and describes the preprocessing stage. We employed CGM data from 52 participants with T1D, collected at Complejo Hospitalario Insular-Materno Infantil de Las Palmas de Gran Canaria. The participants had a mean age of 46.4 years with a standard deviation (STD) of 10.3. The sample was gender-balanced, consisting of 50% women and 50% men. CGM data were captured using the FreeStyle Libre 2 and FreeStyle Libre 3. Participants signed written, informed consent for the use of their data. The Provincial Ethics Committee of Las Palmas deemed the study exempt from assessment because it fell outside the Biomedical Research regulation.

Patients were invited to participate in this study by the physicians belonging to the WARIFA project in scheduled visits at the diabetes clinic. WARIFA is an international project that aims to build AI-based and personalized models for discovering risk factors and predicting Non-Communicable Diseases (NCDs). Several authors of this study are members of the WARIFA project. CGM data from these participants were securely stored on a server accessible only through a virtual private network, with credentials restricted to authorized personnel involved in the WARIFA project. The raw data were retrieved in comma-separated values (CSV) format from the LibreView platform by one of the authors of this study.

The preprocessing stage started with the training/test division of the dataset. We considered 42 and 10 individuals for training and testing, respectively. Then, multiple samples were created for each user. Each sample consisted of daytime and nighttime periods, where nighttime was defined as the interval between midnight and 6 AM, and daytime covered the period from 12 PM to midnight. Ideally, each patient has associated 365 samples, with paired daytime and nighttime recordings for each day of the year.

Due to the presence of missing values in the CGM data, we applied the following criteria: samples with more than 10% missing values during daytime or missing values during the nighttime were excluded. If a given day had more than 10% missing data, the following night was also discarded. Similarly, if any missing data were detected during the night, the previous day of CGM data was removed. Therefore, only a total of 8,875 samples were extracted from the 52 users, with an average of 170 samples per user. Then, linear interpolation [47] was considered to impute missing values (with a maximum of 2 hours of CGM data) during the daytime period owing to its efficacy in short-term periods [48].

Finally, the daytime period is used as input for our TSC models, whereas the nighttime is employed to define the target class. This class is determined based on the occurrence of NH, defined as at least 30 consecutive minutes of interstitial glucose levels below 70 mg/dL ($\leq$ 3.9 mmol/L) during the nighttime interval. As a result, in 6,990 samples the daytime period was not followed by an NH event, while in 1,885 samples an occurrence of an NH event was detected.

IV. METHODS

In this section, we introduce the notation, proposed methodology as well as the predictive models considered to predict NH in T1D people.

A. NOTATION

Let an input dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, consisting of N patients, where $\mathbf{x}_i$ represents glucose values recorded by a CGM device during daytime, and y_i indicates the presence or absence of a NH event during the following night. Each sample $\mathbf{x}_i$ is represented as a vector $\mathbf{x}_i \in \mathbb{R}^{1\times T}$, where T denotes the number of time steps. These time steps correspond to a sequence of temporally ordered observations, forming a TS. In this study, we set $T = 48$, so we only take into account 12 hours of CGM to make the prediction. Regarding the label class, each sample in y_i is determined by:

$$y_i = \begin{cases} 1, & \text{if patient } i \text{ developed NH,} \\ 0, & \text{otherwise.} \end{cases}$$

To thoroughly quantitatively assess the predictive performance of the models, we employed a variety of classification metrics, including AUCROC, recall, and specificity [49]. These were computed based on the prediction of both positive and negative classes, using the following categories: true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), true negative rates (TNR), and false positive rates (FPR).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{AUCROC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR})$$

$$\text{FPR} = \frac{FP}{FP + TN}$$

$$\text{TPR} = \frac{TP}{TP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

B. PROPOSED WORKFLOW

In this study, we develop a data-driven approach for predicting NH in individuals with T1D using various TSC methods. A schematic of the workflow is shown in Figure 1. Our approach consists of the following five stages.

1) *Train/test division:* To ensure adequate training and evaluation, we randomly split the dataset (consisting of 52 participants) into a training subset and a test subset, with 42 and 10 participants, respectively.
2) *Data preprocessing:* Due to the CGM devices used (FreeStyle Libre 2 or 3), data presented different acquisition frequencies and formats. To standardize data, we resampled them to a uniform 15-minute interval. Next, we split CGM data into two periods: daytime (12 PM to 12 AM) and nighttime (12 AM to 6 AM). Samples with any missing values during the night or more than 10% missing values during the day were discarded. Finally, a linear interpolation was applied to fill gaps in the CGM time series.
3) *NH detection:* An NH event was determined as any sample where CGM recordings dropped below 70 mg/dL for 30 minutes during the nighttime period. Samples meeting this criterion were labeled as ‘1’ (NH event), while all other samples were labeled as ‘0’ (no NH event).
4) *Model training:* Prior to training, we applied a random under-sampling technique to ensure a balanced dataset. We employed various ML-based methods for NH prediction, spanning multiple model architectures across seven categories (described in the previous section).
5) *Performance evaluation:* A comprehensive evaluation of different TSC models was performed using various

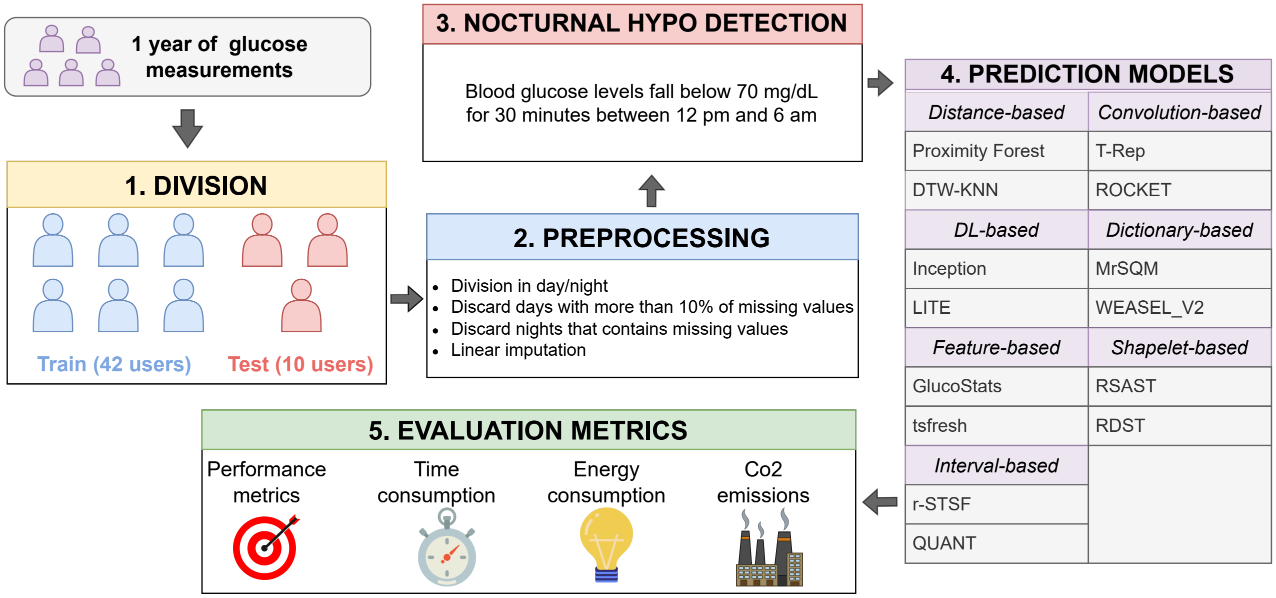


FIGURE 1: Workflow for predicting NH with data-driven models. The next acronyms are used in the figure: Dynamic Time Warping with K-Near Neighbours (DTW-KNN), Random Scalable and Accurate Subsequence Transform (RSAST), the Randomized-Supervised Time Series Forest (r-STSF), Random Convolutional Kernel Transform (ROCKET), Random Dilated Shapelet Transform (RDST), Multiple Representations Sequence Miner (MrSQM).

classification metrics, including AUCROC, specificity, and recall. We also analyzed the computational cost (time consumption in minutes) and the environmental impact (by estimating the energy consumption (kWh) and CO_2 emissions) during the training phase through CodeCarbon tool [50].

The proposed workflow is run five times (each with a different seed). As a result, the final outcomes are reported as the mean and STD across the five iterations. This approach ensures better generalization of the models and provides a more stable evaluation. The code associated with this methodology is available in the following link: https://github.com/ai4healthurjc/NH_evaluation.git

C. TIME SERIES CLASSIFICATION MODELS

In this subsection, we further detail the TSC models employed for NH prediction, categorized into seven model types.

1) Feature-based methods

Feature-based methods extract descriptive statistics as features from time series to be used in tabular classifiers. These features can summarize the whole TS or small periods of the whole time series. These methods are characterized as series-to-vector transformations. To extract glucose statistics, we considered the Python libraries: Tsfresh [39] and GlucoStats [51].

Tsfresh is used for extracting statistics from time series and provides a total of 794 descriptive features [39]. After feature extraction, a matrix is generated, with rows identifying samples and columns representing the extracted features. Some examples of the computed features include the absolute power of the time series, the maximum value, the sum of changes, and various correlation measures, among others. For classification models, the dimensionality of extracted features is crucial to avoid redundancy, as too many irrelevant features can impair the model's ability to generalize beyond the training set, leading to overfitting [52]. To address this, tsfresh includes a feature selection algorithm based on statistical hypothesis testing. The test is configured depending on the type of supervised machine learning problem (classification/regression) and the feature type (categorical/continuous). As a result, the number of extracted features is reduced, and only those considered important are used as input for classification. In our case, the classification model employed is an RF classifier, an ensemble method based on decision trees that achieves strong performance with tabular data [53].

GlucoStats is a library designed to extract and visualize meaningful statistical features from CGM data. It is implemented in Python with a focus on modularity, parallelization, and extensibility. GlucoStats enables the extraction of 59 statistical features, categorized into six main categories: time-related statistics, statistics related to the number of observations within different glucose ranges, descriptive statistics, risk assessment of hypoglycemia and hyperglycemia, and glucose variability metrics. Additionally, the library allows for the segmentation of TS data into different windows, enabling a more detailed analysis. Instead of computing features from the complete time series, they can be extracted per window. In our case, the TS was divided into four windows: 12 PM to 6 PM, 6 PM to 9 PM, 9 PM to 11 PM, and 11 PM to 12 AM. As a result, the final feature matrix consists of 59 features per window, resulting in a total of 236 features. After feature extraction, we applied the Relief algorithm [54] to reduce the number of features and prior training an LR model for NH prediction.

2) Shapelet-based models

Shapelet-based classification models measure the similarity between a shapelet and time series by using this similarity as a discriminatory feature for classification [55]. A shapelet is defined as a time-series subsequence that is in some sense maximally representative of a class and enables TSC based on local, phase-independent similarity in shape. The main advantage of these approaches is that shapelets are interpretable, preserving model comprehensibility. In this study, we applied two shapelet-based models: RDST [41] and RSAST [40].

RDST introduces a novel TS shapelet approach by incorporating the concept of dilation. Dilation is applied during shapelet formulation and influences the features extracted from the distance vector between a shapelet and a time series. Regarding the distance function, each value of the shapelet is compared to a dilated subsequence of the input time series. This dilation allows shapelets to be non-continuous, enabling them to match either non-contiguous patterns or contiguous ones. By focusing on key points of a pattern without covering it entirely, RDST enhances flexibility in pattern recognition. Additionally, to reduce computational time, the search for shapelets is performed randomly.

RSAST is a method proposed to address the weaknesses of SAST [56] and reduce its computational time. SAST is a shapelet-based method that improves computational efficiency and reduces complexity by selecting only a few instances per class from the dataset. Although SAST achieves faster computation compared to STC, its complexity remains cubic concerning the length of the instances, which can be problematic for long time series. To enhance the scalability of SAST, RSAST eliminates the need to explore every possible set of subseries in a training dataset by employing a stratified sampling technique combined with statistical tools, significantly reducing the search space for shapelets. These statistical tools include analysis of variance (ANOVA) [57] and the autocorrelation function [58]. As a result, RSAST substantially reduces computation time while preserving both accuracy and interpretability.

3) Interval-based methods

Interval-based methods randomly split time series into multiple intervals or sub-series (commonly of fixed offsets), compute statistics for these intervals, and then combine these statistics to train an ensemble of predictive models [59]. Most interval-based approaches include a random selection for choosing intervals, where the same random interval locations are used across every time series. These methods are generally fast and memory-efficient [59]. In this paper, two interval-based methods were considered, including QUANT [42] and r-STSF [43].

QUANT uses quantile-based features to capture the underlying distribution of TS values, integrating this information with implicit temporal localization achieved through the segmentation of intervals. Essentially, the method assumes that discriminative class information is embedded in the distributional characteristics of values at different temporal locations. By adjusting the number of quantiles computed within each interval, the approach offers a tunable trade-off between representational detail and computational efficiency, as quantile extraction is both straightforward and resource-efficient [42]. In other words, QUANT encodes a prior that class can be distinguished based on the distribution of values in different locations of a time series. Quantiles allow for representing the distribution of values in an interval in more or less detail (i.e., by computing more or fewer quantiles), and are simple and efficient to compute [42]. Additionally, the feature extraction is performed on the first order differences, second order differences, and a Fourier transform of the input series along with the original series

r-STSF is a tree-based ensemble model that builds trees using features derived from statistics over randomly selected intervals. Instead of relying on the raw TS representation, this model extracts features from the periodogram and autoregressive representation. Additionally, r-STSF employs a stochastic optimization approach and an ensemble of binary trees to select a set of interval features with high discriminating power from the high-dimensional interval feature space. The trees in the ensemble are constructed in a randomized manner following the extra-trees algorithm [60], which reduces the variance of the ensemble and improves classification performance. It is important to note that the process of extracting candidate interval features is repeated a predefined number of times. For an ensemble of 100 trees, the algorithm extracts 10 sets of candidate discriminatory intervals. Finally, each randomized tree is built using a number of randomly selected interval features from the set of candidates, and TSC is performed based on these features.

4) Convolution-based models

Over the last years, CNNs have demonstrated outstanding performance in predictive tasks using image and time series [61]. In these algorithms, convolution and pooling operations are used to extract features from time series, and then these are passed through an MLP or linear classifiers for classification tasks. In this study, we implemented two convolutional-based models: ROCKET [33] and T-Rep [34].

ROCKET transforms time series using a large number of convolutional kernels. Kernels with random length, weights, bias, dilation, and padding are used to create different activation maps. In particular, the kernel dilation has a critical impact on the high performance achieved by ROCKET. These maps are summarized by two pooling operators: *(i)* Proportion of Positive Values (PPV), which computes the percentage of positive values; *(i)* and Global Maximum Pooling (GMP), which extracts the maximum value from the activation map. For each time series, $2k$ features are extracted, where k is the number of kernels. Finally, the transformed features are used to train a linear classifier (ridge regression classifier) to perform the TSC task.

T-Rep is a self-supervised method for learning fine-grained representations of both univariate and multivariate time series [34]. It transforms time series into latent representations aiming to enhance performance in subsequent tasks such as forecasting, anomaly detection and classification. We define a *time-embedding* as a vector representation of time, obtained as the output of a learned function that encodes time series characteristics such as trend, periodicity, and distribution shifts. These *time-embeddings* improve the model's resilience to missing data and enhance its performance when dealing with finite-state systems and non-stationary data. The temporal structure of the embeddings is learned through pretext tasks in self-supervised learning. Consequently, T-Rep integrates time embeddings within the feature-extracting encoder, enabling the model to capture detailed time-related dynamics. The main component of the encoder consists of a temporal CNN-based encoder, which is composed of two layers of one dimensional dilated convolutions. After the feature extraction stage, a fully connected two-layer MLP with ReLU activations is used for TSC.

5) DL-based models

The remarkable results obtained by DL models in different domains have motivated the development of TSC models based on ANN-based architectures. These models automatically learn discriminative feature representations from raw time series data, allowing them to obtain excellent results in subsequent tasks. In this study, we have evaluated the performance of two DL-based models for NH prediction: InceptionTime [35] and LITE [36].

InceptionTime, developed by [62], obtains the final classification decision following an ensemble of the predictions of five Inception networks, with each network contributing equally to the final output. The architecture of an Inception network classifier includes two residual blocks, where each block is composed of three Inception modules instead of traditional fully convolutional layers. Each residual block's input is transferred via a shortcut linear connection to be added to the next block's input. After the residual blocks, for multivariate time series, a Global Average Pooling (GAP) layer is applied. Finally, a fully connected softmax layer, with a number of neurons equal to the number of classes in the dataset, is used for the TSC task.

LITE is a variation of InceptionTime that consists of only 2.34% of the total parameters of the original InceptionTime model while achieving comparable performance. This efficiency is inspired by convolutional approaches such as ROCKET. In this case, a bottleneck operation is performed to reduce the number of parameters. The bottleneck operation consists of 1D convolutions with a unit kernel size. Furthermore, several boosting techniques, including multiplexing, dilation, and custom filters, are applied to enhance performance.

6) Distance-based models

Distance-based methods classify time series through pairwise similarities between different time series and using a specific distance metric. The methods considered were PF [31] and DTW-KNN [32].

PF uses a range of distance metrics to categorize TS according to their similarity to 'exemplar' time series [31]. PF algorithm, which builds an ensemble of classification trees with 'splits' using the proximity of a given TS T to a set of reference time series: if T is closer to the first reference time series, then it goes to the first branch if it is closer to the second reference time series, then it goes to the second branch, and so on. Proximity Forest integrates 11 TS measures for evaluating similarity. At each node, a set of reference series is selected, one per class, together with a similarity measure and its parameterization. These selections are made stochastically.

DTW is an effective method for estimating the optimal alignment between TS and sequence elements, enabling the measurement of a global distance between patterns. Its main goal is to determine the minimal warping path on an element-wise cost matrix given a predefined cost function. To achieve this, the sequences are "warped" non-linearly to assess their similarity while being invariant to non-linear variations in the time dimension. KNN classifier combined with DTW has demonstrated strong effectiveness for TSC [63], particularly due to its ability to handle non-linear mappings. Therefore, we have implemented this approach in our study. Lastly, we perform KNN with the distance matrix achieved with the DTW method to perform the TSC task.

7) Dictionary-based models

Dictionary-based methods utilize phase-independent subsequences by sliding a window over TS. Instead of measuring the distance to a subsequence, as in shapelet-based approaches, each window is transformed into a word, and the frequency of occurrence of repeating patterns is recorded. Then, a classifier is trained using the occurrence of words as input to make the TSC. In this study, we have selected WEASEL_V2 [37] and MrSQM [38] as representative examples of this category of method.

WEASEL_V2 is an improved version of WEASEL [64]. WEASEL extracts and normalizes subsequences of varying lengths from a time series. Then, the Fourier transform is applied to approximate these subsequences, and an ANOVA F-test is used to select the real and imaginary Fourier coefficients that best separate TS from different classes. These selected Fourier values are then discretized into words using a symbolic feature aggregation method. Next, a large sparse dictionary is built from the words across all chosen window lengths. To reduce the size of the dictionary and remove irrelevant words, a Chi-squared test is applied. Finally, using the occurrence of the remaining words (TF-IDF [65]), a RIDGE regression classifier is trained to perform the TSC. WEASEL_V2 introduces two major improvements: (1) it incorporates a fixed dilation, similar to convolutional models, by applying a dilation mapping in the window creation process; and (2) it creates a dense dictionary limited to 256 words. To increase diversity and reduce bias in the dense dictionary, WEASEL_V2 employs an ensemble approach over multiple parameter configurations, using randomization. As a result, this model achieves better performance compared to the original WEASEL.

MrSQM [38] is a dictionary-based method that relies on multiple symbolic representations. It consists of three main steps: (1) symbolic transformation; (2) feature selection; and (3) classification model. In the first step, the TS is transformed into words using Symbolic Aggregate approXimation [66] or symbolic feature aggregation. After reducing the number of words, a novel supervised symbolic feature selection approach is applied in the all-subsequence space by adapting a Chi-square bound developed for discriminative pattern mining. Finally, with the selected features, a LR model is trained to perform the TSC task. LR is chosen due to its accuracy, scalability, model transparency, and ability to provide well-calibrated prediction probabilities.

D. ENVIRONMENTAL IMPACT AND CARBON FOOTPRINT WITH CODECARBON

Recently, several energy estimation tools such as CodeCarbon, CarbonTracker and PowerTop have been proposed for measuring the environmental impact of ML/DL models [27]. These tools primarily rely on Intel Running Average Power Limit and NVIDIA Management Library to measure the energy consumption of applications, which are considered reliable for measuring CPU and GPU power consumption. According to several experiments of energy estimation tools [67], CodeCarbon showed the most precise accuracy compared to physical wattmeters. In this study, we measured the environmental impact using version 2.8.3 [50]. CodeCarbon is an open-source Python package designed to track power consumption by monitoring the total system power usage during computational tasks. It uses hardware-level metrics (if available) or software-level estimations to fetch the energy drawn by the CPU, GPU, and RAM. By aggregating the energy consumption of these components, CodeCarbon calculates the corresponding carbon footprint based on the machine's geographic location, using region-specific carbon intensity data for electricity generation. These carbon intensity values are sourced from the global energy mix file in the CodeCarbon repository, which contains per-country data from Our World In Data. If this data is unavailable, it defaults to a static value of 475 gCO2eq/KWh. The goal of CodeCarbon is to measure and track the carbon footprint of ML model training and inference, supporting efforts to make ML processes more environmentally sustainable by providing insights into energy consumption and emissions.

V. RESULTS

In this section, we present the results of different TSC methods for NH prediction in T1D patients. We first present the experimental setup, then an extended comparison of the predictive performance of ML-based models for NH prediction, and finally, an analysis of time consumption, energy consumption and CO_2 emissions is presented.

TABLE 2: Hyperparameter values explored for each one of the 14 evaluated approaches.

Approach	Hyperparameter	Values/options
PF	number_trees	$\{50, 100, 150\}$
	number_splitters	$\{3, 5, 8\}$
	max_depth	$\{2, 4, 6\}$
	min_samples_plit	$\{2, 4, 6\}$
DTW-KNN	number of neighbors	$[20, 150]$
T-Rep	epochs	$\{5, 10, 20, 50\}$
	C	$\{1e^{-6}, 1e^{-5}, 1e^{-4}, 1^{-2}$ $1^{-1}, 1, 5, 10, 20\}$
	solver	liblinear, saga
ROCKET	number_kernels	$\{5000, 10000, 20000\}$
	max_dilations_per_kernel	$\{16, 32, 64\}$
	n_features_per_kernel	$\{2, 4, 6\}$
Inception	depth	$\{4, 6, 8\}$
	number_filters	$\{16, 32, 64\}$
	kernel_size	$\{20, 40, 60\}$
	n_conv_per_layer	$\{3, 5, 8\}$
	number_epochs	$\{50, 75, 100\}$
LITE	number_classifiers	$\{3, 5, 8\}$
	number_filters	$\{16, 32, 64\}$
	kernel_size	$\{20, 40, 60\}$
	number_epochs	$\{50, 75, 100\}$
MrSQM	strat	RS, SR
	features_per_rep	$\{300, 500, 700\}$
	selection_per_rep	$\{1000, 1500, 2000\}$
WEASEL_V2	min_window	$\{4, 8, 10\}$
	max_feature_count	$\{10000, 20000, 30000\}$
GlucoStats	C	$\{1e^{-6}, 1e^{-5}, 1e^{-4}, 1^{-2}$ $1^{-1}, 1, 5, 10, 20\}$
	solver	*liblinear*, *saga*
	windowing_param	$[[0, 1, 0, 0], [0, 2, 0, 0],$ $[0, 3, 0, 0], [0, 6, 0, 0]]$
tsfresh	default_fc_parameters	minimal, efficient, comprehensive
	relevant_feat_extractor	False, True
RSAST	n_random_points	$\{5, 10, 20\}$
	nb_inst_per_class	$\{5, 10, 20\}$
RDST	max_shapelets	$\{500, 1000, 3000, 5000\}$
	distance	DTW, Euclidean, Manhattan, Minkowski
r-STSF	number_intervals	$\{30, 50, 75\}$
	min_interval_length	$\{3, 5, 7\}$
	number_estimators	$\{100, 200, 300\}$
QUANT	interval_depth	$\{4, 6, 8\}$
	quantile_divisor	$\{4, 6, 8\}$

A. EXPERIMENTAL SETUP

For the implementation of the models, Python 3.10 was employed. For DTW-KNN, we used dtwParallel [68], whereas for T-Rep and GlucoStats, we used the GitHub repositories in github.com/imprs/TS-Rep and github.com/ai4healthurjc/glucostats. For the remaining algorithms, we relied on the aeon library (version 1.0.0) [59]. The hyperparameter values selected for the models considered in this study are detailed in Table 2. To determine the optimal hyperparameter values for each classification method, we employed GridSearchCV with a three-fold cross-validation over the training subset. Model training and hyperparameter tuning were performed using all available CPUs from two AMD EPYC 7713P 64-core processors.

TABLE 3: Performance comparison of different times series classification models used for predicting NH. The best-performing values for each metric are highlighted in bold.

Type	Model	Accuracy	Specificity	Recall	AUCROC
Distance-based	PF	0.719±0.041	0.736±0.059	0.661±0.060	0.699±0.020
	DTW-KNN	0.702±0.029	0.719±0.043	0.637±0.048	0.678±0.014
Convolution-based	T-Rep	0.648±0.049	0.636±0.069	0.703±0.077	0.670±0.024
	ROCKET	0.694±0041	0.709±0.060	0.639±0.056	0.674±0.012
DL-based	Inception	0.670±0.011	0.681±0.024	0.628±0.049	0.655±0.016
	LITE	0.732±0.032	0.761±0.061	0.623±0.086	0.692±0.017
Dictionary-based	MrSQM	0.593±0.017	0.596±0.027	0.577±0.044	0.587±0.017
	WEASEL_V2	0.727±0.025	0.763±0.030	0.580±0.041	0.670±0.024
Feature-based	GlucoStats	0.740±0.026	0.763±0.042	0.643±0.068	**0.703±0.021**
	tsfresh	0.712±0.029	0.723±0.044	0.671±0.048	0.697±0.013
Shapelet-based	RSAST	0.581±0.018	0.583±0.035	0.567±0.045	0.576±0.008
	RDST	0.610±0.045	0.582±0.068	**0.720±0.056**	0.656±0.025
Interval-based	r-STSF	**0.741±0.026**	**0.767±0.041**	0.640±0.056	**0.703±0.018**
	QUANT	0.723±0.026	0.747±0.043	0.653±0.060	0.697±0.012

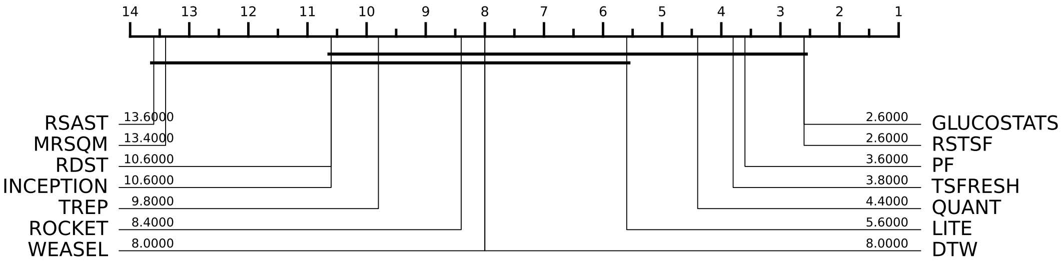


FIGURE 2: Critical difference plot on test accuracy for different time series classification models used for nocturnal hypoglycemia prediction. Best models are placed to the right.

B. CLASSIFICATION RESULTS

Table 3 showed that interval-based models achieved the highest overall performance across multiple evaluation metrics. In particular, r-STSF exhibited the highest value of accuracy (0.741±0.026), AUCROC (0.703±0.018) and specificity (0.767±0.041). Among feature-based models (extracted using GlucoStats) demonstrated reasonable performance, achieving an accuracy of 0.740±0.026 and an AUCROC of 0.703±0.021, which are comparable to r-STSF. This indicated that handcrafted feature extraction techniques can be highly effective in TSC, and more particularly, for NH prediction.

Dictionary-based and shapelet-based methods showed mixed results. WEASEL_V2 achieved 0.727±0.025 in accuracy, but its recall was 0.580±0.041 (lower than other methods). Shapelet-based approaches (RSAST and RDST) obtained the worst accuracy and recall values. Convolution-based and DL-based methods (ROCKET, T-Rep, Inception, and LITE) performed moderately well, with LITE achieving an accuracy of 0.732±0.032. However, the AUCROC values were generally lower than the best-performing interval- and feature-based models, suggesting that they may struggle with distinguishing between classes in this specific dataset. The results indicated that interval-based and feature-based approaches are the most effective for NH prediction, with r-STSF and glucose statistics (extracted using GlucoStats) achieving the best balance across accuracy, specificity, recall, and AUCROC.

Figure 2 shows a critical difference (CD) diagram, which ranks the performance of various time-series classification models based on their performance over the 5 iterations of experiments. The values for the CD diagram are computed by ranking the models based on their AUCROC scores for each iteration and then computing the mean position over the five iterations. Consequently, a lower value indicates better performance, as it means the model consistently ranked among the top performers throughout the five iterations. The models at the right of the diagram exhibit superior rankings, whereas those on the left show relatively lower performance. The results indicate that GlucoStats and r-STSF achieve the highest rankings, both with an average rank of around 2.6, confirming their effectiveness as top-performing approaches. Notably, both models ranked among the top three performers in each of the five iterations. Other high-ranking models include PF, tsfresh, and QUANT, reinforcing the strong performance of feature-based and interval-based methods. Conversely, methods such as RSAST, MrSQM, and RDST are positioned towards the left, with RSAST exhibiting the lowest ranking (13.6). This suggests that shapelet-based and certain dictionary-based approaches underperform compared to other methodologies. Notably, DL-based models (e.g., Inception and LITE) achieve mid-range rankings, indicating competitive but not superior performance. The distance-based DTW method ranks relatively low, further supporting the observation that alternative approaches capture more discriminative time-series representations. Overall, the CD diagram highlights that feature-based and interval-based methods consistently outperform other approaches, aligning with the results observed in the tabular performance metrics.

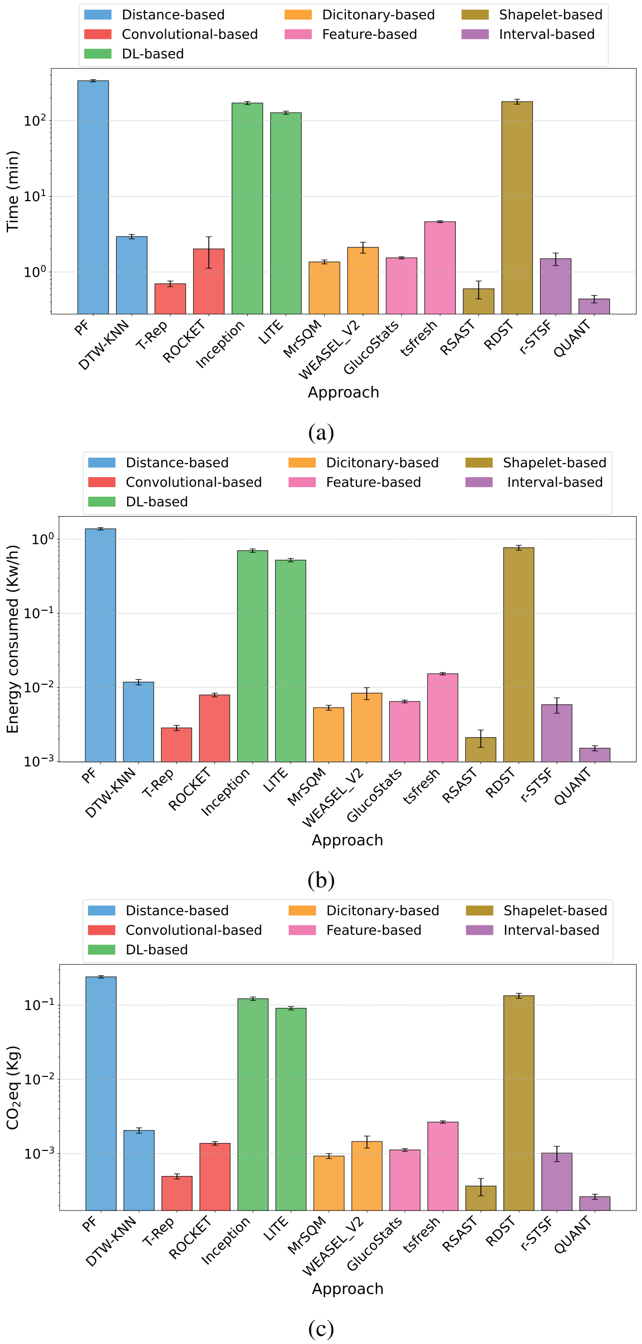
C. TIME CONSUMPTION, ENERGY CONSUMPTION AND CO_2 EMISSIONS OF TSC MODELS

Figure 3 presents a comparative analysis of the computational time consumption, energy consumption and CO_2 emissions for different TSC models used for NH prediction and categorized into 7 different types of approaches. The x-axis represents different classification approaches categorized into methodological families, including distance-based, convolutional-based, DL-based, dictionary-based, feature-based, shapelet-based, and interval-based methods.

In Figure 3 (a), the y-axis displays the computational time in minutes on a logarithmic scale, allowing for a clearer distinction between models with significantly different execution times. The results reveal that the PF approach, RDST, and both DL-based approaches (Inception and LITE) exhibit the highest computational times, significantly surpassing all other methods. This aligns with their known inefficiency in large-scale TSC tasks, requiring more than 100 minutes of training. The remaining approaches demonstrate a considerably lower computational time. DTW-KNN, ROCKET, tsfresh, and WEASEL_V2 show moderate time consumption, slightly exceeding three minutes. Finally, T-Rep, MrSQM, GlucoStats, RSAST, r-STSF, and QUANT complete their computations in less than 90 seconds. Notably, the QUANT approach achieves the lowest computational time, making it the most efficient method in this comparison. Overall, these results highlight the trade-offs between computational efficiency and model complexity in TSC for NH prediction.

The pattern observed in energy consumption was similarly followed in the environmental impact, which is measured by energy consumption (kW/h) (see Figure 3 (b)) and CO_2 emissions (see Figure 3 (c)). Note that the y-axis in both figures is presented on a logarithmic scale to facilitate comparison across approaches with widely varying resource demands. As shown, the distance-based method named PF and DL-based approaches Inception and LITE exhibited high energy consumption and CO_2 emissions, highlighting the computational cost associated with these architectures. The shapelet-based model RDST ranked among the most energy-intensive and contaminating, aligning with the computational complexity inherent in shapelet extraction. Feature-based methods performed with CGM-derived features extracted using GlucoStats and tsfresh, convolution-based methods (T-Rep and ROCKET), and dictionary-based approaches (MrSQM and WEASEL_V2) consumed an intermediate amount of energy and presented moderate amount of CO_2 emissions.

Interval-based approaches (QUANT and r-STSF) exhibited the lowest energy consumption and CO_2 emissions, and QUANT presented the least energy among all TSC methods. The pattern observed in energy consumption was similarly reflected in carbon emissions (see Figure 3 (c)), reinforcing the direct relationship between energy usage and environmental impact. Overall, these results show that interval-based methods have the least environmental impact, followed by convolutional models, dictionary-based models, and feature-based models. In addition, the results underscore the significant variation in resource consumption across machine learning methodologies.

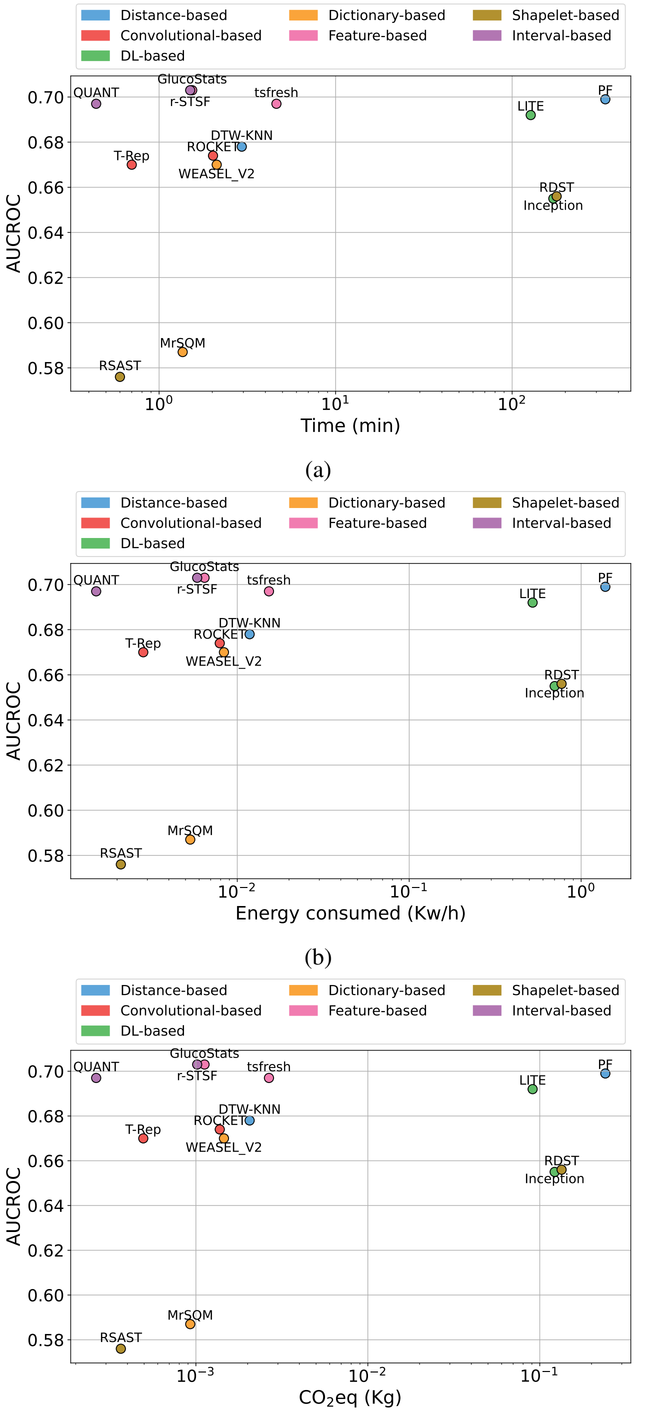


(a)

(b)

(c)

FIGURE 3: Comparison of: (a) time consumption, (b) energy consumption and (c) CO_2 emissions for different TSC models used for nocturnal hypoglycemia prediction.



(a)

(b)

(c)

FIGURE 4: Comparison of the performance of all TSC models based on their mean AUCROC values relative to: (a) time consumption, (b) energy consumption, and (c) CO2 emissions. The best-performing models are positioned in the top left corner.

Figure 4 presents a comparison of model performance based on AUCROC against computational time consumption, energy consumption, and CO_2 emissions for different TSC models. The best models are located in the top left corner, as they achieve the highest performance while maintaining low computational cost and environmental impact. The models in this optimal region include interval-based methods (QUANT and r-STSF) and feature-based models (GlucoStats and tsfresh). These models consistently remain in the same favorable zone for time consumption (Figure 4 (a)), energy consumption (Figure 4 (b)), and CO_2 emissions (Figure 4 (c)). As a result, we can conclude that these approaches balance computational efficiency and high predictive performance. Note that the models located in the bottom left region (indicating poor accuracy but high efficiency) include RSAST and MrSQM, while those in the top right region (high accuracy but low efficiency) include LITE, PF, Inception, and RDST. Therefore, despite offering state-of-the-art performance, these latter models raise environmental concerns due to their high energy consumption and carbon footprint. The rest of the approaches present a moderate performance and environmental impact, offering more energy-efficient alternatives.

VI. DISCUSSION

In this paper, a comprehensive evaluation of 14 different approaches (classified into seven categories) is performed for predicting NH in individuals with T1D. NH is a serious health problem in individuals with T1D and early prevention is crucial to avoid complications. However, most existing studies that aim to predict NH using AI models have primarily relied on feature-based approaches. As a consequence, further research is needed to explore and identify the most effective predictive approaches. Therefore, we presented a comprehensive evaluation of 14 different TSC methods, which were categorized into seven categories, for predicting NH in individuals with T1D. To provide a comprehensive and global evaluation of each one of the approaches, we measured three key parameters: predictive performance, time consumption and environmental impact (energy consumption and CO_2 emissions). This extensive evaluation not only provides insights into model performance but also highlights the potential real-world impact of implementing these models on a larger scale.

Regarding the performance metric, we used the AUCROC to determine the approaches with the highest performance. The best approaches were r-STSF (interval-based) and models that used glucose features extracted using GlucoStats (feature-based), achieving an AUCROC of 0.703 ± 0.018 and 0.703 ± 0.021, respectively. This showed that interval-based approaches effectively capture relevant temporal patterns for the given classification task. Both approaches were closely followed by PF (distance-based) with 0.699 ± 0.020, QUANT (interval-based) with 0.697 ± 0.012, tsfresh (Feature-based) with 0.697 ± 0.013, and LITE (DL-based) with 0.692 ± 0.017. These six models achieved an AUCROC above 69%, making them the most effective in predicting a possible NH event. Overall, the performance metrics indicate that Interval-based and Feature-based approaches are the most effective for this classification task, along with the PF and LITE approaches.

As shown in previous sections, there might be situations where we need extra evaluation metrics in addition to predictive metrics (such as accuracy, specificity and others), to be able to decide the type of ML or DL model. The time consumption in training is associated with the algorithms' computational complexity. While accuracy reflects an algorithm's ability to correctly classify data, computational complexity offers a different dimension of evaluation, which is key for assessing the feasibility of implementing and deploying trained models at a large scale. For instance, in our experiments, although distance-based methods (PF) and DL-based approaches (RDST) reached high accuracy values, they had a considerable computational cost. The training time of these models is by far superior (more than 100 minutes) compared with the other approaches. On the other hand, T-Rep, MrSQM, GlucoStats, RSAST, r-STSF, and QUANT complete their training in less than 90 seconds, highlighting a reduced computational time.

Lastly, training ML and DL models have a significant global environmental impact. This impact needs to be measured for a proper and comprehensive evaluation of an AI model. Therefore, we evaluated the energy consumption and carbon emissions of each of the 14 approaches. The results indicate that interval-based methods have the least environmental impact, followed by convolutional models, dictionary-based models, and feature-based models. The QUANT model achieves the lowest environmental impact, consuming only a mean of 0.27×10^{-3} kg and 0.15×10^{-2} kW/h. Note that distance-based and DL-based models present the highest environmental impact.

To sum up, we evaluated the efficiency of the approaches by comparing their performance metrics, time consumption, and environmental impact to determine the best possible options. As a result, the interval-based models (QUANT and r-STSF) and feature-based models (GlucoStats and tsfresh) are the most efficient, as they achieved the best balance between AUCROC and environmental impact. Conversely, DL-based, distance-based, and shapelet-based approaches, despite achieving a good AUCROC, are not feasible for large-scale implementation due to their high complexity and environmental impact. These findings emphasize the importance of selecting an appropriate model based on both predictive performance and computational feasibility. They provide valuable insights into the trade-offs between classification accuracy and energy efficiency in time-series classification. Lastly, the strong correlation between energy usage and CO_2 emissions highlights the importance of considering computational sustainability when selecting ML models, particularly for large-scale applications.

In the present study, we employed TSC models and achieved a maximum accuracy and AUCROC of 0.741 and 0.703, respectively. While this result is competitive, it is lower than the predictive results presented in several prior studies (see Section II). This difference is mainly due to variations in the number of samples of dataset considered and PHs. For instance, Vu *et al.* and Mosquera-Lopez *et al.* employed large datasets, with one million and 44,000 samples, respectively. The data size can be related to the generalizability and performance of their models. Furthermore, the definition and temporal framing of NH prediction differs across studies. For instance, Berikov *et al.* and Kozinetz *et al.* adopted short-term prediction horizons of 15 to 30 minutes, making predictions for short-term PHs, which can not compared with studies with PHs of several hours. For example, our study establishes the NH prediction over a broader temporal window of 6 hours. Additionally, variations in data modality and the inclusion of other variables such as insulin administration and meal intake can contribute to differences in model performance compared to TSC models. Note that this work is a comparative analysis of TSC models for NH prediction, which aims not only to assess the predictive performance of models but also to determine those models are more feasible for this clinical task, providing a foundation for the trade-off between performance and environmental impact.

It is worth noting that the models developed in this study that reached the highest predictive results will be integrated into a real-world mobile-based application as part of the WARIFA project [69]. The WARIFA application is accessible through smartphones and collects data from various sources, including user-generated and public data for assessing NCD risks and providing personalized lifestyle recommendations. Regarding glucose and NH, users in the mobile-based application will provide CGM data and the data-driven models built into this work will be used for predicting NH, preventing acute clinical events and improving patient's quality of life.

Limitations. Despite the promising results, several limitations should be considered. In this study, the predictive models were trained using a medium-sized dataset, consisting of only a few thousand samples, which can limit the predictive performance. Additionally, datasets were imbalanced, with fewer samples associated with NH events. We considered an under-sampling approach to balance the training subset. This class imbalance may affect the robustness of the models, particularly when applied to more diverse populations or real-world clinical scenarios. Future research can consider other resampling approaches such as oversampling and hybrid methods, including generative adversarial networks to create synthetic data of minority classes. Additionally, we will aim to incorporate additional datasets covering a broader range of demographic cohorts and clinical settings.

VII. CONCLUSIONS

In this paper, a comprehensive evaluation of 14 different approaches (classified into seven categories) is performed for predicting NH in individuals with T1D. This evaluation was carried out by assessing the performance in two categories: predictive and environmental impact (measured by energy consumption and CO_2 emissions) over a dataset of 52 T1D people belonging to the Complejo Hospitalario Insular-Materno Infantil de Las Palmas de Gran Canaria. The experimental results indicate that the most efficient approaches, in terms of balancing AUCROC and environmental impact, are the interval-based and feature-based models. Specifically, the models that achieved the best AUCROC with reduced training time and environmental impact are r-STSF and GlucoStats, achieving an AUCROC of 0.703±0.018 and 0.703±0.021, respectively. Conversely, DL-based, distance-based, and shapelet-based approaches, despite achieving good AUCROC, are not feasible for large-scale implementation due to their high complexity and environmental impact. This research explores the need to determine the best approach for predicting NH events and preventing related health complications. Additionally, it emphasizes the importance of selecting an appropriate model based on predictive performance, computational complexity, and environmental impact, which are relevant factors for large-scale implementation of ML and DL models.

DECLARATION OF CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Vanderniet JA, Jenkins AJ, Donaghue KC. Epidemiology of type 1 diabetes. Current Cardiology Reports. 2022;24(10):1455-65.

[2] Gregory GA, Robinson TI, Linklater SE, Wang F, Colagiuri S, de Beaufort C, et al. Global incidence, prevalence, and mortality of type 1 diabetes in 2021 with projection to 2040: a modelling study. The Lancet Diabetes & Endocrinology. 2022;10(10):741-60.

[3] Herold KC, Delong T, Perdigoto AL, Biru N, Brusko TM, Walker LS. The immunology of type 1 diabetes. Nature Reviews Immunology. 2024;24(6):435-51.

[4] Olafsdottir AF, Polonsky W, Bolinder J, Hirsch IB, Dahlqvist S, Wedel H, et al. A randomized clinical trial of the effect of continuous glucose monitoring on nocturnal hypoglycemia, daytime hypoglycemia, glycemic variability, and hypoglycemia confidence in persons with type 1 diabetes treated with multiple daily insulin injections (GOLD-3). Diabetes Technology & Therapeutics. 2018;20(4):274-84.

[5] Søholm U, Broadley M, Zaremba N, Divilly P, Baumann PM, Mahmoudi Z, et al. The impact of hypoglycaemia on daily functioning among adults with diabetes: a prospective observational study using the Hypo-METRICS app. Diabetologia. 2024;67(10):2160-74.

[6] Divilly P, Zaremba N, Mahmoudi Z, Søholm U, Pollard DJ, Broadley M, et al. Hypo-METRICS: Hypoglycaemia—MEasurement, ThResholds and ImpaCtS—A multi-country clinical study to define the optimal threshold and duration of sensor-detected hypoglycaemia that impact the experience of hypoglycaemia, quality of life and health economic outcomes: The study protocol. Diabetic Medicine. 2022;39(9):e14892.

[7] Divilly P, Martine-Edith G, Zaremba N, Søholm U, Mahmoudi Z, Cigler M, et al. Relationship between sensor-detected hypoglycemia and patient-reported hypoglycemia in people with type 1 and insulin-treated type 2 diabetes: the Hypo-METRICS study. Diabetes Care. 2024;47(10):1769-77.

[8] Yale JF. Nocturnal hypoglycemia in patients with insulin-treated diabetes. Diabetes Research and Clinical Practice. 2004;65:S41-6.

[9] Zhu B, Abu Irsheed GM, Martyn-Nemeth P, Reutrakul S. Type 1 diabetes, sleep, and hypoglycemia. Current Diabetes Reports. 2021;21:1-19.

[10] Poolsup N, Suksomboon N, Kyaw AM. Systematic review and meta-analysis of the effectiveness of continuous glucose monitoring (CGM) on glucose control in diabetes. Diabetology & Metabolic Syndrome. 2013;5:1-14.

[11] Lara-Abelenda FJ, Chushig-Muzo D, Peiro-Corbacho P, Wägner AM, Granja C, Soguero-Ruiz C. Personalized glucose forecasting for people with type 1 diabetes using large language models. Computer Methods and Programs in Biomedicine. 2025;265:108737.

[12] Lee SM, Kim DY, Woo J. Glucose transformer: Forecasting glucose level and events of hyperglycemia and hypoglycemia. IEEE Journal of Biomedical and Health Informatics. 2023;27(3):1600-11.

[13] Xue Y, Guan S, Jia W. BGformer: an improved informer model to enhance blood glucose prediction. Journal of Biomedical Informatics. 2024;157:104715.

[14] Shi M, Yang A, Lau ES, Luk AO, Ma RC, Kong AP, et al. A novel electronic health record-based, machine-learning model to predict severe hypoglycemia leading to hospitalizations in older adults with diabetes: A territory-wide cohort and modeling study. PLoS Medicine. 2024;21(4):e1004369.

[15] Lara-Abelenda FJ, Chushig-Muzo D, Wägner AM, Tayefi M, Soguero-Ruiz C. Interpretable and multimodal fusion methodology to predict severe hypoglycemia in adults with type 1 diabetes. Engineering Applications of Artificial Intelligence. 2025;144:110142.

[16] Tsichlaki S, Koumakis L, Tsiknakis M, et al. Type 1 diabetes hypoglycemia prediction algorithms: systematic review. JMIR Diabetes. 2022;7(3):e34699.

[17] Vu L, Kefayati S, Idé T, Pavuluri V, Jackson G, Latts L, et al. Predicting nocturnal hypoglycemia from continuous glucose monitoring data with extended prediction horizon. In: AMIA Annual Symposium Proceedings. vol. 2019. American Medical Informatics Association; 2019. p. 874.

[18] Mosquera-Lopez C, Dodier R, Tyler NS, Wilson LM, El Youssef J, Castle JR, et al. Predicting and preventing nocturnal hypoglycemia in type 1 diabetes using big data analytics and decision theoretic analysis. Diabetes Technology & Therapeutics. 2020;22(11):801-11.

[19] Jensen MH, Dethlefsen C, Vestergaard P, Hejlesen O. Prediction of nocturnal hypoglycemia from continuous glucose monitoring data in people with type 1 diabetes: a proof-of-concept study. Journal of Diabetes Science and Technology. 2020;14(2):250-6.

[20] Berikov VB, Kutnenko OA, Semenova JF, Klimontov VV. Machine learning models for nocturnal hypoglycemia prediction in hospitalized patients with type 1 diabetes. Journal of Personalized Medicine. 2022;12(8):1262.

[21] Mosquera-Lopez C, Roquemen-Echeverri V, Tyler NS, Patton SR, Clements MA, Martin CK, et al. Combining uncertainty-aware predictive modeling and a bedtime Smart Snack intervention to prevent nocturnal hypoglycemia in people with type 1 diabetes on multiple daily injections. Journal of the American Medical Informatics Association. 2024;31(1):109-18.

[22] Bagnall A, Lines J, Bostrom A, Large J, Keogh E. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery. 2017;31:606-60.

[23] Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. Deep learning for time series classification: a review. Data Mining and Knowledge Discovery. 2019;33(4):917-63.

[24] Shamshirband S, Fathi M, Dehzangi A, Chronopoulos AT, Alinejad-Rokny H. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. Journal of Biomedical Informatics. 2021;113:103627.

[25] Ahmed SF, Alam MSB, Hassan M, Rozbu MR, Ishtiak T, Rafa N, et al. Deep learning modelling techniques: current progress, applications, advantages, and challenges. Artificial Intelligence Review. 2023;56(11):13521-617.

[26] Dhar P. The carbon impact of artificial intelligence. Nat Mach Intell. 2020;2(8):423-5.

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

[28] Cowls J, Tsamados A, Taddeo M, Floridi L. The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. Ai & Society. 2023:1-25.

[29] Ohalete NC, Aderibigbe AO, Ani EC, Ohenhen PE, Akinoso AE. Data science in energy consumption analysis: a review of AI techniques in identifying patterns and efficiency opportunities. Engineering Science & Technology Journal. 2023;4(6):357-80.

[30] Wright D, Igel C, Samuel G, Selvan R. Efficiency is not enough: A critical perspective of environmentally sustainable AI. arXiv preprint arXiv:230902065. 2023.

[31] Lucas B, Shifaz A, Pelletier C, O'Neill L, Zaidi N, Goethals B, et al. Proximity forest: an effective and scalable distance-based classifier for time series. Data Mining and Knowledge Discovery. 2019;33(3):607-35.

[32] Müller M. Dynamic time warping. Information retrieval for music and motion. 2007:69-84.

[33] Dempster A, Petitjean F, Webb GI. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. Data Mining and Knowledge Discovery. 2020;34(5):1454-95.

[34] Fraikin A, Bennetot A, Allassonnière S. T-rep: Representation learning for time series using time-embeddings. arXiv preprint arXiv:231004486. 2023.

[35] Ismail-Fawaz A, Devanne M, Weber J, Forestier G. Deep learning for time series classification using new hand-crafted convolution filters. In: 2022 IEEE International Conference on Big Data (Big Data). IEEE; 2022. p. 972-81.

[36] Ismail-Fawaz A, Devanne M, Berretti S, Weber J, Forestier G. Look into the lite in deep learning for time series classification. International Journal of Data Science and Analytics. 2025:1-21.

[37] Schäfer P, Leser U. WEASEL 2.0: a random dilated dictionary transform for fast, accurate and memory constrained time series classification. Machine Learning. 2023;112(12):4763-88.

[38] Nguyen TL, Ifrim G. Fast time series classification with random symbolic subsequences. In: International Workshop on Advanced Analytics and Learning on Temporal Data. Springer; 2022. p. 50-65.

[39] Christ M, Braun N, Neuffer J, Kempa-Liehr AW. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). Neurocomputing. 2018;307:72-7.

[40] Varela NR, Mbouopda MF, Nguifo EM. RSAST: Sampling Shapelets for Interpretable Time Series Classification; 2023.

[41] Guillaume A, Vrain C, Elloumi W. Random dilated shapelet transform: A new approach for time series shapelets. In: International Conference on Pattern Recognition and Artificial Intelligence. Springer; 2022. p. 653-64.

[42] Dempster A, Schmidt DF, Webb GI. Quant: A minimalist interval method for time series classification. Data Mining and Knowledge Discovery. 2024:1-26.

[43] Cabello N, Naghizade E, Qi J, Kulik L. Fast, accurate and explainable time series classification through randomization. Data Mining and Knowledge Discovery. 2024;38(2):748-811.

[44] Kozinetz RM, Berikov VB, Semenova JF, Klimontov VV. Machine Learning and Deep Learning Models for Nocturnal High-and Low-Glucose Prediction in Adults with Type 1 Diabetes. Diagnostics. 2024;14(7):740.

[45] Leutheuser H, Bartholet M, Marx A, Pfister M, Burckhardt MA, Bachmann S, et al. Predicting risk for nocturnal hypoglycemia after physical activity in children with type 1 diabetes. Frontiers in Medicine. 2024;11:1439218.

[46] Duckworth C, Guy MJ, Kumaran A, O'Kane AA, Ayobi A, Chapman A, et al. Explainable machine learning for real-time hypoglycemia and hyperglycemia prediction and personalized control recommendations. Journal of Diabetes Science and Technology. 2024;18(1):113-23.

[47] Terry W, Lee J, Kumar A. Time series analysis in acid rain modeling: Evaluation of filling missing values by linear interpolation. Atmospheric Environment (1967). 1986;20(10):1941-3.

[48] Rancati S, Bosoni P, Schiaffini R, Deodati A, Mongini PA, Sacchi L, et al. Exploration of Foundational Models for Blood Glucose Forecasting in Type-1 Diabetes Pediatric Patients. Diabetology. 2024;5(6):584-99.

[49] Shalev-Shwartz S, Ben-David S. Understanding machine learning: From theory to algorithms. Cambridge university press; 2014.

[50] Courty B, Schmidt V, Luccioni S, Goyal-Kamal, MarionCoutarel, Feld B, et al.. mlco2/codecarbon: v2.4.1. Zenodo; 2024. Available from: https://doi.org/10.5281/zenodo.11171501.

[51] Peiro-Corbacho P, Lara-Abelenda FJ, Chushig-Muzo D, Wägner AM, Granja C, Soguero-Ruiz C. Glucostats: An Efficient Python Library for Glucose Time Series Feature Extraction and Visual Analysis; 2025. Available at SSRN. Available from: https://ssrn.com/abstract=5203999.

[52] Crespo Márquez A. The curse of dimensionality. In: Digital Maintenance Management: Guiding Digital Transformation in Maintenance. Springer; 2022. p. 67-86.

[53] Parmar A, Katariya R, Patel V. A review on random forest: An ensemble classifier. In: International conference on intelligent data communication technologies and internet of things (ICICI) 2018. Springer; 2019. p. 758-63.

[54] Urbanowicz RJ, Meeker M, La Cava W, Olson RS, Moore JH. Relief-based feature selection: Introduction and review. Journal of biomedical informatics. 2018;85:189-203.

[55] Hills J, Lines J, Baranauskas E, Mapp J, Bagnall A. Classification of time series by shapelet transformation. Data mining and knowledge discovery. 2014;28:851-81.

[56] Mbouopda MF, Nguifo EM. Scalable and accurate subsequence transform for time series classification. Pattern Recognition. 2024;147:110121.

[57] St L, Wold S, et al. Analysis of variance (ANOVA). Chemometrics and intelligent laboratory systems. 1989;6(4):259-72.

[58] Ramsey FL. Characterization of the partial autocorrelation function. The Annals of Statistics. 1974:1296-301.

[59] Middlehurst M, Schäfer P, Bagnall A. Bake off redux: a review and experimental evaluation of recent time series classification algorithms. Data Mining and Knowledge Discovery. 2024:1-74.

[60] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine learning. 2006;63:3-42.

[61] Li G, Jung JJ. Deep learning for anomaly detection in multivariate time series: Approaches, applications, and challenges. Information Fusion. 2023;91:93-102.

[62] Ismail Fawaz H, Lucas B, Forestier G, Pelletier C, Schmidt DF, Weber J, et al. Inceptiontime: Finding alexnet for time series classification. Data Mining and Knowledge Discovery. 2020;34(6):1936-62.

[63] Canhoto JLO, de Mattos Neto PSG, Barbosa TR, da Silva Santos JEM, de Campos IM, Junior GLM, et al. Application of time series analysis to classify therapeutic breathing patterns. Smart Health. 2024;32:100460.

[64] Schäfer P, Leser U. Fast and accurate time series classification with weasel. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management; 2017. p. 637-46.

[65] Zhao G, Liu Y, Zhang W, Wang Y. TFIDF based feature words extraction and topic modeling for short text. In: Proceedings of the 2018 2nd international conference on management engineering, software engineering and service sciences; 2018. p. 188-91.

[66] Lin J, Keogh E, Wei L, Lonardi S. Experiencing SAX: a novel symbolic representation of time series. Data Mining and Knowledge Discovery. 2007;15:107-44.

[67] Bouza L, Bugeau A, Lannelongue L. How to estimate carbon footprint when training deep learning models? A guide and review. Environmental Research Communications. 2023;5(11):115014.

[68] Escudero-Arnanz Ó, Marques AG, Soguero-Ruiz C, Mora-Jiménez I, Robles G. dtwParallel: A Python package to efficiently compute dynamic time warping between time series. SoftwareX. 2023;22:101364.

[69] WARIFA Project. WARIFA: Personalized Risk Prediction for Non-Communicable Diseases; 2024. Accessed: 2024-11-18. Available from: https://www.warifa.eu/es/home-es/.

FRANCISCO J. LARA-ABELENDA received the B.Sc. in Biomedical Engineering from Rey Juan Carlos University and the M.Sc. in Machine Learning for Health from Carlos III University. He is currently working towards a Ph.D. in Explainable and Multimodal Artificial Intelligence with applications in healthcare at Rey Juan Carlos University. His research interests encompass signal processing, natural language processing, explainable artificial intelligence, and multimodal fusion methods.

DAVID CHUSHIG-MUZO received the Ph.D. degree in machine learning applied to healthcare applications from the Rey Juan Carlos University (URJC) in 2022. He worked as a post-doctoral researcher in the WARIFA project by building interpretable risk prediction models. Since 2024, he has been Assistant Professor with the Department of Signal Theory and Communications, Telematics and Computing Systems at URJC. He has coauthored several research papers in international journals and conferences. He has participated as a researcher in public funding projects, mainly related to machine learning models in clinical setting. His main research interests include statistical learning theory, feature selection, data augmentation, unsupervised learning, interpretability methods and multimodal learning.

CARMELO BETANCORT ACOSTA graduated in Medicine at the University of Las Palmas de Gran Canaria (ULPGC) in 2017. In 2018, he started a specialisation in Endocrinology and Nutrition at the Complejo Hospitalario Universitario Insular Materno Infantil (CHU-IMI) de Las Palmas de Gran Canaria, which finished in 2022. After completing the specialist training, he moved to Ciudad Real to work at the Hospital General Universitario de Ciudad Real until 2023, where he was responsible for the Neuroendocrinology Monographic Consultation. From 2023, he has participated in the European project WARIFA: Artificial Intelligence and the Prevention of Chronic Diseases as a member of the research team. He has worked as a specialist in endocrinology and nutrition in both the public (CHU-IMI) and private sectors.

ANA M. WÄGNER is a specialist in Endocrinology and Nutrition, consultant at the Complejo Hospitalario Universitario Insular Materno-Infantil de Gran Canaria, Associate Professor of Medicine. Director of the Biomedical and Health Research Institute (iUIBS) of the University of Las Palmas de Gran Canaria (ULPGC). Her main interest is to improve the lives of people living with diabetes. She is principal investigator of several national and European research projects, mainly with focus on diabetes and on technology and she has authored more than one hundred international publications in indexed journals.

CONCEIÇAO GRANJA is a Senior Researcher at the Norwegian Centre for E-health Research, and an Associate Professor at Nord University. She completed her PhD at the Faculty of Engineering, University of Porto, where she focused on developing a novel simulation-based optimization approach to improve patient admission scheduling. Conceição conducted postdoctoral studies at the Norwegian Centre for Telemedicine, where she studied digital interaction between hospitals and patients in a surgical context. Her research integrates complex scheduling techniques and meta-heuristic optimization algorithms to improve patient outcomes and resource management in healthcare settings.

CRISTINA SOGUERO-RUIZ is a Full Professor at Rey Juan Carlos University. She got the Ph.D. degree in machine learning with applications in healthcare, in 2015, with the Joint Doctoral Program in Multimedia and Communications in conjunction with University Rey Juan Carlos and University Carlos III. She won the Orange Foundation Best Ph.D. Thesis Award by the Spanish Official College of Telecommunication Engineering. She has published several papers in JCR journals and international conferences. She has participated in several research projects (with public and private funding) related to healthcare data-driven machine learning systems. Her current research interests include machine learning, data science, and statistical learning theory.

• • •