IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY SECTION

Received 5 December 2022, accepted 15 January 2023, date of publication 18 January 2023, date of current version 31 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3237992

RESEARCH ARTICLE

Layered Meta-Learning Algorithm for Predicting Adverse Events in Type 1 Diabetes

FEDERICO D'ANTONI[1], LORENZO PETROSINO[1], (Student Member, IEEE), ALESSANDRO MARCHETTI[1], LUCA BACCO[1,4,5], SILVIA PIERALICE[2], LUCA VOLLERO[1], (Member, IEEE), PAOLO POZZILLI[2], VINCENZO PIEMONTE[3], AND MARIO MERONE[1], (Member, IEEE)

[1]Department of Engineering, Research Unit of Computer Systems and Bioinformatics, Universitá Campus Bio-Medico di Roma, 00128 Rome, Italy
[2]Department of Medicine, Unit of Diabetology and Endocrinology, Universitá Campus Bio-Medico di Roma, 00128 Rome, Italy
[3]Department of Engineering, Unit of Chemical-Physics Fundamentals in Chemical Engineering, Universitá Campus Bio-Medico di Roma, 00128 Rome, Italy
[4]ItaliaNLP Laboratory, Istituto di Linguistica Computazionale "Antonio Zampolli," National Research Council, 56124 Pisa, Italy
[5]Research and Development Laboratory, Webmonks S.r.l., 00178 Rome, Italy

Corresponding author: Mario Merone (m.merone@unicampus.it)

This work was supported by Programma Operativo Regionale Fondo Europeo di Sviluppo Regionale (POR FESR) Lazio 2014-2020, Progetto T0002E0001, "Progetti di Gruppi di Ricerca 2020," under Grant A0375E0196.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Campus Bio-Medico University of Rome (11/05/2021), and performed in line with the Declaration of Helsinki.

ABSTRACT Type 1 diabetes mellitus (T1D) is a chronic disease that, if not treated properly, can lead to serious complications. We propose a layered meta-learning approach based on multi-expert systems to predict adverse events in T1D. The base learner is composed of three deep neural networks and exploits only continuous glucose monitoring data as an input feature. Each network specializes in predicting whether the patient is about to experience hypoglycemia, hyperglycemia, or euglycemia. The output of the experts is passed to a meta-learner to provide the final model classification. In addition, we formally introduce a novel parameter, α, to evaluate the advance by which a prediction is performed. We evaluate the proposed approach on both a public and a private dataset and implement it on an edge device to test its feasibility in real life. On average, on the Ohio T1DM dataset, our system was able to predict hypoglycemia events with a time gain of 22.8 minutes, hyperglycemia ones with an advance of 24.0 minutes. Our model not only outperforms presented models in the literature in terms of events predicted with sufficient advance, but also with regard to the number of false positives, achieving on average 0.45 and 0.46 hypo- and hyperglycemic false alarms per day, respectively. Furthermore, the meta-learning approach effectively improves performance in a new cohort of patients by training only the meta-learner with a limited amount of data. We believe our approach would be an essential ally for the patients to control the glycemic fluctuations and adjust their insulin therapy and dietary intakes, enabling them to speed up decision-making and improve personal self-management, resulting in a reduced risk of acute and chronic complications. As our last contribution, we assessed the validity of the approach by exploiting only blood glucose variations as well as in combination with the information of the insulin boluses, the skin temperature, and the galvanic skin response. In general, we have observed that providing other information but CGM leads to slightly lower performances with respect to considering CGM alone.

INDEX TERMS Meta learning, event detection, time series analysis, diabetes.

I. INTRODUCTION

Type 1 diabetes mellitus (T1D) is a chronic autoimmune disease, occurring as a consequence of the organ-specific

The associate editor coordinating the review of this manuscript and approving it for publication was Sung-Min Park.

immune destruction of the insulin-producing β-cells of the pancreas. Such cells are glucose thermostats, sensing glucose and releasing insulin to maintain physiologic glucose levels within a normal range [1]. Once these cells are destroyed, patients with T1D lose blood glucose control, requiring insulin therapy replacement lifetime. The pathophysiology and clinical management of T1D and insulin-treated type 2 diabetes (T2D) differs significantly. T2D is characterized by insulin resistance and a relative lack of insulin. While T2D subjects maintain low glucose variability thanks to their residual pancreatic function, patients with T1D often cannot control their blood sugar levels, even with optimal medical management. Despite structured self-monitoring of blood glucose, these subjects experience frequent hyperglycemia and hypoglycemia events, even asymptomatic, and are at increased risk of diabetic ketoacidosis. These glycemic fluctuations lower the quality of life, potentially leading to recurrent hospitalization and chronic complications that may reduce their life expectancy [2]. Diabetes is a widespread condition and represents a global leading cause of death. Although it has no cure, it can be managed through daily insulin administrations to keep the glycemic level in the euglycemic range, i.e., between 70 and 180 mg/dl [3]. In recent years, the use of Continuous Glucose Monitoring (CGM) devices increased considerably, allowing patients to keep track of their glycemic trend 24 hours a day.

A CGM system typically consists of an adhesive sensor and a display device to collect glucose data. In contrast to traditional finger stick testing, where capillary blood glucose is detected, the electrochemical sensor of CGM measures glucose concentration from the interstitial fluid in the subcutaneous layer. The most modern devices can continuously and wirelessly transmit real-time data to a receiver or smartphone application. However, as glucose must diffuse from the capillaries into the interstitial fluid for reading, there is an approximate lag time of 8–10 minutes between plasma and interstitial concentrations under steady-state conditions, which may increase when glucose levels are rapidly rising or falling [4]. CGM systems have improved glycemic control and represent a useful tool to lower glycated hemoglobin (HbA1c) in diabetic patients [5]. Moreover, evidence suggests that real-time CGM supports insulin-requiring patients in advancing their knowledge of the disease, providing insights into glycemic fluctuation and ameliorating personal self-management capabilities as well as their behavior and quality of life [6]. In some cases, such devices are coupled with an insulin pump, which simulates physiologic pancreatic functioning by injecting small amounts of insulin on demand. The development of new CGM systems is currently a field of research. The solutions adopted in the literature may concern improvements in the sensor system and the data transfer method. Usually, the goodness of a CGM reading is measured with the MARD, Mean Absolute Relative Difference, i.e., the average of the absolute error between all CGM values and matched reference values. Although analytical performance cannot be fully evaluated by a single parameter, systems with an overall MARD < 10% are generally regarded as having good performance [7]. However, such a measure is affected by several factors. Some studies have shown substantial variations of this value throughout the day [8], [9]. Still, it remains one of the most widely used indicators to assess the accuracy of a CGM device. To better understand the advances and evolution of CGM devices, we refer the reader to the recent study in [10].

The quality of life of people with diabetes improves considerably by preventing the blood glucose levels from exceeding the euglycemic range [11], [12]. For this reason, in the last decade researchers have focused their efforts on developing data-driven algorithms capable of predicting future blood glucose levels or incoming adverse events to allow patients to prevent or mitigate them. The main objective of T1D control is to correct hyperglycemia while avoiding hypoglycemia [13].

Although CGM sensors are widely adopted by people with diabetes and despite the considerable recent progress in the frame of blood glucose levels forecasting, hypo- and hyperglycemic events are still frequently reported [13], [14], [15]. Works in the literature attempting to reduce the number of these events suffer from some open issues. First, most models only focus on predicting future blood glucose levels with a regression task [16], [17]. As such, regression predicts future glucose levels regardless of whether they are in the hypoglycemic or hyperglycemic range. It has been proven by recent works that predicting adverse glycemic events using classification rather than regression leads to improved performance [18], [19].

Second, the vast majority of studies focus only on the prediction of hypoglycemia [20], [21], [22], [23], [24], [25]. It is a sensible choice because this condition can arrive unannounced also in severe cases, leading to serious short-term complications. In this regard, in a recent review on machine learning techniques for hypoglycemia prediction, Mujahid et al. [23] stated that *is important to understand that hypoglycemia prediction is blood glucose level prediction in essence*. Nonetheless, most of such works mainly aim at maximizing the true positive rate at the expense of a considerably low precision score, which is often not reported [20], [21] or impossible to compute [19], [22], [26], [27], [28], [29]. Indeed, it is acknowledged that any prediction algorithm has to ''decide'' between raising a lot of alerts to detect all events (good recall, bad precision, a lot of false positives) or trying to minimize the nuisance of the patient (good precision, limited false positives, at the expense of a lower recall). Works focusing on hypoglycemia prediction usually choose the former approach [30], with few exceptions [28]. It reduces patient engagement with the technology.

Third, predicting glycemic excursions, and in particular incoming hypoglycemic events, is a very challenging task. Although a wide literature exists about the prediction of glycemic events, spanning from regressive models [29] to

ensemble models [20] and cutting-edge technologies such as deep neural networks [19], none of such models can fully represent the complex rules lying behind the different glucose dynamics of T1D patients. It also happens because the datasets utilized to build such models are usually limited in size. Recently, meta-learning has proven to be effective in solving and improving the generalization of few-shot tasks that would be unsolvable by training from scratch [31]. A new study from Zhu et al. [17] successfully used model-agnostic meta-learning to enable fast adaptation of a neural network for forecasting future glycemic levels of T1D patients. However, this approach requires a second, patient-personalized fine-tuning phase, which could require weeks of data gathering and manual labeling from the physicians.

Finally, some works focus only on the sample-based approach [21], [22], [27], [28]. This is a limitation, because such an approach may lead to overestimating the performance, generating high recall scores because correctly predicted continuous hypo/hyperglycemic samples count as several true positives, whereas the event may have not been predicted in advance.

For the reasons above, we propose a meta-learning system based on a multi-expert predictive model relying on an event-based approach. The experts consist of either Recurrent Long Short-Term Memory (LSTM) or Convolutional Neural Networks (CNN). We aim to develop a model capable to achieve a good trade-off between the amount of correctly predicted events (i.e., high recall per class) and the number of false alarms (i.e., high precision per class) while evaluating performance on a public dataset. We consider a 30-minute (6-timestamp) prediction horizon (PH) since it would be a sufficient time to warn patients about incoming adverse events [32]. We evaluate the effective advance by which predictions are performed by introducing a parameter α, evaluating performance as α varies. Due to the strong imbalance between the classes, we use a Leave-1-Patient-Out Cross-Validation approach to maximize the number of samples from the minority classes in the discovery set. Such an approach would also provide users with a ready-to-use model which does not require a fine-tuning period on patient-specific data. In addition, we aim to develop a univariate approach to make the predictive models more suitable for real-life applications. By not requiring the user to utilize different devices for data recording, it could be usable by patients that exploit only CGM for therapy while reducing the computational burden required to combine several heterogeneous data. Moreover, previous works have shown that using several input features besides CGM does not improve performance sensitively without a computationally expansive preprocessing [33], [34], which is likely to be avoided when performing tasks on edge devices [35]. Finally, we implement the proposed system on an edge-computing device to evaluate the real-life feasibility and applicability of the proposed approach.

The main contributions of this paper are summarized as follows:

- We propose a layered meta-learning approach, which uses a decision tree as a meta-learner to exploit the predictions from a multi-expert model based on either LSTM or CNN architectures. Each network is an expert in detecting one of the three investigated classes and pursues a univariate approach.
- We introduce a taxonomy of the works in the literature approaching the task of predicting blood glucose levels, which differ from each other by the number of features they exploit, the purpose of their prediction, and the experimental setup used for validation.
- The method is tested in 3 different experimental configurations, evaluating the performance in terms of metrics derived from the confusion matrix with varying values of α, considering an event-based approach.
- We tackle the high imbalance of the dataset by resorting to cost-sensitive learning and using a sample-based approach for the training and validation phases to increase the number of observations from the minority classes in the discovery set.
- The system observed on both subjects using Continuous Subcutaneous Insulin Infusion (CSII) and subjects with multiple daily injections (MDI) shows robust performance, whatever the therapy;
- We present a comparison between our results and those achieved using other well-established methods in the literature; we also test variants of the proposed models which exploit a multivariate approach.
- We implement the proposed approach on an edge device to evaluate the performance in terms of training and inference time.

The remainder of this work is organized as follows: the next section presents the state of the art of glucose levels prediction, with a specific focus on the classification of adverse glycemic events; section II describes the public and the private dataset utilized for validation, as well as the proposed models and the edge device on which they are implemented; section III describes the experimental design for the tests that have been performed, and presents a description of the competitors that we tested in different settings (using CGM alone or with additional input features); section IV describes the results we achieved for the tests, for the edge implementation, and for the comparison with the other models; finally, section V provides concluding remarks and some possible future developments.

A. BACKGROUND

Since the introduction of CGM devices for monitoring blood glucose levels, research efforts moved towards the development of predictive models capable of improving the life quality of people with T1D [36]. In most cases, such models resort to data-driven techniques to predict glucose values in advance [37]. In the frame of blood glucose levels prediction, the works in the literature that exploit data-driven models differ from each other in three main aspects: **Univariate vs**

Multivariate approach. The CGM sensor measurements can be regarded as time series, as they are temporal sequences of evenly spaced data points; straightforwardly, different time-series approaches have been used in the literature to perform predictions given past sequences of glucose levels. In time-series problems, the sequence of input data can be represented by single or multiple attributes, i.e., one variable or a set of variables that vary over time. The former approach is usually referred to as univariate time series (UTS) and the latter as multivariate time series (MTS). UTS approaches exploit only past glucose levels to perform predictions on future glucose levels, whereas MTS approaches exploit the previous knowledge of other features such as injected insulin and carbohydrate intake; **Regression vs Classification.** The former is the most widely adopted approach [16], consisting in forecasting the exact future glucose level given a prediction horizon (PH), i.e., how far forward in time the prediction is performed. The latter instead aims at predicting if an event will occur or not, i.e., if the patient is going to experience hypoglycemia or hyperglycemia in the time window defined by the PH or after a fixed amount of minutes; **Precision Medicine vs *k*-fold Cross Validation.** Precision medicine aims to develop a predictive model suited to patient-specific data, and, as a consequence, the available data from each patient are split into training and test sets. The training set is used to fit the model on data from one subject, and the test set is used to evaluate performance on other data from the same subject. Conversely, k-fold Cross Validation is a statistical technique consisting in splitting the whole dataset, composed of data from several patients, into k subsets to evaluate model performance on the entire dataset by using, in turns, one fold as the test set and the remaining $k-1$ folds as discovery (training and validation) set. In many cases, the Leave-1-Patient-Out Cross Validation is used, a special kind of k-fold Cross-validation in which each fold consists of all the data of a single subject.

With regard to the regression task, future glucose levels forecasting resorts to different types of approaches, including kernel machines [38], forests of decision trees [39], generative models [29], artificial neural networks [19], [40], state-space models [41], time-series and latent variable models [16]. Such models are by far the most investigated in the literature but are limited by the fact that their results, usually reported in terms of Root Mean Squared Error (RMSE), provide no information on whether prediction errors occur in the proximity of hypo/hyperglycemic levels or the euglycemic range. However, the predicted glycemic values can be compared to the original CGM track to compute the amount of adverse glycemic events identified by resorting to a regression-to-classification task [18], [29]. A model has recently been proposed for joint blood glucose levels forecasting and hypoglycemia prediction [19].

A different approach to predict future glycemic events resorts to the classification task: in this case, rather than predicting the exact future glycemic level, the models aim to predict if the patient is going to experience an adverse glycemic event in the next few minutes (most papers consider a PH of 30 minutes). Usually, these studies consider two main classes to be predicted, according to the read CGM value, namely hypoglycemia (CGM $\leq 70\ mg/dl$) and hyperglycemia (CGM $\geq 180\ mg/dl$), although some studies take into account further thresholds, or only focus on predicting a specific event (e.g., postprandial hypoglycemia). Two main approaches fall into this area: **sample-based prediction** [20], [21], [22], [27], [28] in which, at each timestamp, a prediction is performed according to the PH; in this way, each sample is classified, and the model performance is evaluated based on the predictions performed for all the timestamps; and **event-based prediction** [18], [27], [29], in which consecutive timestamps of hypo/hyperglycemia are considered as a single event; a prediction of an event is considered a true positive if an actual event occurs in the next minutes. A summary of the state of the art of glycemic events prediction is reported in Table 1, where we also report the results of our previous work [42] in which we performed a regression-to-classification task.

II. MATERIALS AND METHODS

A. DATASET

A wide-ranging analysis of the state of the art shows that tests are usually performed on private datasets, therefore it is difficult to make a fair comparison between the various algorithms. In other cases, tests are performed on data from in silico patients [43], [44], generated using T1DM simulators [45], [46]; nonetheless, the results achieved using virtual patients usually overestimate the model predictive capability because real-life complications such as physical activity, stress and illness are not taken into consideration [47]. However, a public dataset composed of data from six real subjects is available since 2018 [14] and that has been afterward enlarged with data from six more subjects [15]. To promote a fair comparison with works from other researchers, we perform tests on this publicly available dataset. In addition, we validate the proposed model using data from a private dataset.

1) PUBLIC VALIDATION DATASET (Ohio)

The Ohio T1DM dataset was initially available to participants in the first and second Blood Glucose Level Prediction (BGLP) Challenge in 2018 and 2020 and then became publicly available to other researchers. In this work, we consider the original format [14] and its expansion [15] as a single dataset. It contains eight weeks of data concerning continuous glucose monitoring, insulin, physiological sensor, and self-reported life-events of twelve adults suffering from T1D (five females and seven males, aged between 20 and 60, each using the Medtronic *Enlite*^{TM} CGM sensor and a fitness band), all following a Continuous Subcutaneous Insulin Infusion

TABLE 1. State of the art of the glycemic events prediction task. For each work, we report the main author together with the number of patients in the dataset and the validation strategy, the adopted model, the specific sample-based or event-based approach and, where available, the average classification Recall (*R*) and Precision (*P*) of predictions up to 30 minutes ahead of time for the classes hypoglycemia (Hypo), normoglycemia (Norm) and hyperglycemia (Hyper). We mark as not available (*n/a*) the performance values that were not reported and are not possible to compute.

First author	# Patients	Validation	Model	Approach	Results [%]			
						Hypo	Norm	Hyper
Gadaleta [18]	89	Leave-1-Patient-Out	SVM	Event-Based	R	86	n/a	95
					P	36	n/a	56
Daskalaki [26]	23	Precision Medicine	cARX + RNN	Event-Based	R	100	n/a	100
					P	n/a	n/a	n/a
Cappon [27]	100 in silico	Precision Medicine	XGBoosted Tree	Sample-Based	R	92	76	86
					P	n/a	n/a	n/a
Seo [20]	104	5-fold Cross Validation	Random Forest	Sample-Based	R	89.6	91.3	n/a
					P	38.9	n/a	n/a
Dave [21]	112	10-fold Cross Validation	Random Forest day + Random Forest night	Sample-Based	R	93.7	94.4	n/a
					P	15.1	99.8	n/a
Marcus [28]	11	Precision Medicine	Kernel Ridge Regression	Sample-Based	R	64	96	61
					P	n/a	n/a	n/a
Cichosz [22]	10	Precision Medicine	Linear Logistic Regression	Sample-Based	R	79	99	n/a
					P	n/a	n/a	n/a
Yang [19]	124	Precision Medicine	Long Short-Term Memory (LSTM) classifier	Event-Based	R	92.6	92.5	n/a
					P	n/a	n/a	n/a
Prendin [29]	112	Precision Medicine	Autoregressive Integrated Moving Average (ARIMA)	Event-Based	R	82	n/a	n/a
					P	64	n/a	n/a
Jensen [24]	463	5-fold Cross Validation	Linear Discriminant Analysis (LDA) classifier	Sample-Based	R	73	75	n/a
					P	22	97	n/a
Zhu [17]	49	Holdout Validation	Bidirectional RNN with meta-learning	Event-Based	R	84.1	n/a	n/a
					P	65.6	n/a	n/a
D'Antoni [42]	33	Precision Medicine	ARTiDe Jump NN	Event-Based	R	59.8	n/a	47.2
					P	86.4	n/a	58.0

therapy(CSII). More detailed information about the dataset can be found in [14] and [15].

The dataset is already split into a training and a test set for each patient; however, since we aimed to perform a Leave-1-Patient-Out Cross Validation, we joined the training and the test sets of each patient to make a single fold. The recorded data report many interruptions; plus, two different fitness bands were used in the first and second releases to record physical data.

We decided to pursue a univariate approach, so CGM sensor data is used alone as an input feature of the proposed model. In order to test the multivariate variant of the models, and provide a fair comparison between different approaches, we utilized only the features that are in common between the datasets; furthermore, in order to develop a system as autonomous as possible and to reduce the burden on the patient, we only considered the features collected by sensors and without the direct involvement of the user. After this selection, the four considered features are CGM sensor read values, injected insulin, skin temperature, and galvanic skin response.

2) PRIVATE VALIDATION DATASET (UCBM)

The Unit of Endocrinology and Diabetology of Campus Bio-Medico University (UCBM) Hospital provided anonymized CGM data of five T1D patients (all males), all using Dexcom G5 CGM sensor, aged between 32 and 43 (average 38.6 ± 5), HbA1c between 5.7 and 8.4, weight between 67 and 95 kg, daily insulin requirement per kg between 0.07 and 0.85 UI/Kg/die (average 0.49 ± 0.29). Three patients use CSII, whereas two follow an MDI. Every patient was monitored for a period ranging from 3 to 14 days (average 8 ± 3.8), for a total of 40 days, during which they regularly performed physical activity. Predicting glucose levels of T1D patients during physical activity is particularly tough due to quick variations occurring [48]. It is worth noting that the patients from the UCBM dataset utilize a different CGM sensor than patients from the public dataset.

Informed Consent Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of Campus Bio-Medico University of Rome (11 May 2021). No patient identifiable information was used in this study and only data from which identifying factors have been removed were used for statistical analysis. Informed consent was obtained from all subjects involved in the study.

B. DATA PREPROCESSING

As aforementioned, many disconnections occurred during the data recording period concerning both the CGM sensor and the fitness band. In general, this leads to complications when training a time-series model. To minimize complications and allow a comparison between the performance of the UTS and the MTS approach, we included in the dataset only the timestamps in which all the considered features were available at the same time for at least 12 consecutive timestamps

(60 minutes). Indeed, in this work we found that a size of the input sequence of 6 timestamps (i.e. the latest 30 minutes) provides optimal results. Since a PH of 30 minutes is being considered, consecutively recorded sequences shorter than 60 minutes would not provide a ground truth value to evaluate the effectiveness of the prediction. Also, we excluded from the analysis the 6 timestamps preceding and following a sensor calibration or disconnection, since huge variations of glycemia were present during such events, resulting in noisy data for the model training. Next, we composed a different feature matrix for each patient by joining all the portions of data obtained in this way. No further preprocessing was performed on raw data; the only exception concerns the amount of injected insulin: we added the bolus values to the basal insulin rate at the corresponding timestamps. In this way, we joined the basal insulin and the injected boluses into a single insulin feature.

C. DATA LABELING

To perform a classification task and properly evaluate the model, data labeling is essential. Different approaches have been pursued in the literature for the prediction of glycemic events, spanning from binary classification problems [20], [21], [22] to 4-class problems [18]. In this paper, we approached a three-class classification task, considering classes hypoglycemia, hyperglycemia, and normoglycemia (euglycemia). We chose well-established thresholds to define classes based on CGM values, considering the following formal definition:

$$\begin{cases} \text{Hypoglycemia} & \text{if } CGM \leq 70\ mg/dl \\ \text{Normoglycemia} & \text{if } 70\ mg/dl < CGM < 180\ mg/dl \\ \text{Hyperglycemia} & \text{if } CGM \geq 180\ mg/dl \end{cases}$$

For each sample in the dataset, we observe the subsequent 6 timestamps (30 minutes) and act differently according to the values in that time window:

- if a hypo/hyperglycemic value is in the considered time window, then the sample under observation is labeled either as hypoglycemia or hyperglycemia.
- if the sample under observation falls within the hypo- or hyperglycemic ranges, the sample is labeled either as hypoglycemia or hyperglycemia regardless of the values in the subsequent time window.
- if the sample under observation and all the samples in the considered time window are in the euglycemic range, then the sample is labeled as normoglycemia.

Note that this labeling strategy tends to generate ''alarms'' every time an adverse event is forthcoming or is already happening, whereas it considers as ''normal'' all the other timestamps. It is also why, differently from other works [18], we decided not to consider severe hypo- or hyperglycemia as classes: the proposed model generates an alarm every time an event is predicted or present, regardless of its severity.

In the sample-based approach, after the labeling step, the public dataset includes 5866 hypoglycemia, 67972 euglycemia, and 38175 hyperglycemia samples, corresponding to about 389 days of data. The Imbalance Ratio, defined as the ratio between the number of samples of the most and the least represented class, is $IR = 11.6$. Thus, the dataset presents a high imbalance ($IR \geq 9$) according to the definition given in [49]. The event-based approach presents 413 events of hypoglycemia, 66786 samples of euglycemia, and 1417 events of hyperglycemia, with a consequent $IR = 161.7$. It indicates a strongly imbalanced dataset [49]. The euglycemia cannot be considered an event. According to the physiological meaning and the labeling strategy we chose, we consider all the normoglycemia samples (every single time-step) as independent observations (events) in the event-based approach. Following this strategy, the number of observations is slightly smaller due to data rearrangement during the event-based performance evaluation.

Regarding the private dataset, it includes 819 hypoglycemia, 7113 normoglycemia, and 3221 hyperglycemia samples ($IR = 8.7$), corresponding to 55 events of hypoglycemia, 7044 samples of normoglycemia, and 72 events of hyperglycemia ($IR = 128$).

D. EDGE DEVICES

The increasing development of new, more powerful, dedicated hardware enables the emergence of a branch of artificial intelligence known as inference at the edge [50], [51]. It involves the machine learning models being run directly from a proximity device using data collected from associated sensors. With the growing interest in the telemedicine approach [52], [53], the inference at the edge can enable predictive models that work in real-time with patient data to improve both medical quality and efficiency. For this reason, to date, several works exploit the potential of edge computing not only from a more methodological and general point of view (e.g., [54]) but also in the field of glycemic level prediction. Zhu et al. [55], for example, proposed an Embedded Edge Evidential Neural Network to predict future glycemic levels of adult T1D patients in real-time by exploiting CGM sensor readings and an edge-computing device.

To test the feasibility of the predictive model implementation and utilization on an edge system, we needed to identify the target hardware. Because of its low cost and high computational capabilities, our choice fell on the Raspberry Pi4. The Raspberry Pi4 presents a Broadcom BCM2711 quad-core Arm Cortex A72 of 1.5 GHz processor, with 4 GB of random access memory. Furthermore, to carry out the tests, we used Raspbian OS (a Debian-derived operating system) as the operating system. To limit the experimental time, we chose to carry out these tests using three identical devices. We standardized the data collected during testing and installed the dependencies required to carry out the tests only on one device. Then, the operating system image was copied

over two different memory cards and inserted into the other devices to make them clones of the previous one.

E. METHODS

We now introduce the main contribution of this paper, consisting of a novel approach for predicting adverse glycemic events while generating few false alarms. We propose a meta-learning approach based on a multi-expert system. In particular, we resort to layered meta-learning, in which a base-learner models task-specific characteristics while a meta-learner models the features shared by the tasks [31]. As the base-learner, we utilized a multi-expert system based on a deep neural network architecture. We evaluated two different architectural approaches, one based on recurrent neural networks (LSTM) and the other based on convolutions (CNN). We selected these models because they achieve state-of-the-art performance on tasks related to time-series, including T1D management [23], [30]. The softmax layer output of each expert is passed to a decision tree (the meta-learner). Figure 1 reports the architectural schemes of the two implemented base-learners, while Figure 2 reports the scheme of the entire system.

1) BASE-LEARNER

The base-learner is a multi-expert system consisting of three deep neural networks, either Recurrent with LSTM units or with three convolutional layers. We will refer to these multi-expert models as **ME-LSTM** and **ME-CNN**, respectively. The rationale lies in observing that the overall performance on a skewed dataset may be improved by combining the decisions of three different models [56], each specialized in detecting one of the three classes under examination. In other words, in this phase, the original three-class problem is decomposed into three binary classification problems, and, straightforwardly, a binary relabeling was performed before training each expert. During the training of the single expert, a weighted classification layer provides the final decision. We optimized the LSTM and CNN models through a grid search on the number of hidden layers and the number of nodes for each layer. We report further details in paragraph II-F.

LSTM

In general, recurrent layers of RNNs consist of recurrent cells which are affected by both past states and current inputs. Almost all the exciting results achieved in the latest years with RNNs have been achieved by the LSTM. Thanks to its ability to learn long- and short-term sequence patterns, it is nowadays considered the state-of-the-art model for time-series forecasting and sequence classification [57]. Each LSTM cell consists of three gates. The first two have a role when updating the cell state: the *input* gate decides what part of the new information will be stored, while the *forget* one what information will be thrown away. The third gate, the *output* one, decides what information can be output based on the cell state.

In this work, a single expert consists of the succession of the following layers: a sequence-input layer, which takes as input an $m \times n$ matrix of features, where m is the number of features and n is the number of recent timestamps to be input; a first LSTM layer of n_h hidden units; a second LSTM layer of $\frac{1}{2}n_h$ hidden units; a fully-connected layer of two units (i.e., one for each class investigated by the expert); a two-neurons softmax layer, which takes the network output values between 0 and 1. We report the schematic representation of the expert structure in Figure 1. The proposed model exploits only CGM as an input feature, thus $m = 1$ (univariate approach). In this work, we found that a value of $n = 6$ (i.e., the latest 30 minutes) provided optimal results. The value of n_h for each expert was empirically determined as described in section II-F.

CNN

Apart from a sequence-input layer (the same as in the LSTM case), each CNN expert involves three convolutional layers with different numbers of filters, also called kernels. In the univariate approach, we fix the filter size equal to 1×2. For each layer, each filter slides (with a stride equal to 1) along one direction (the temporal dimension). At each step, a convolution of the samples (time instants) covered by the filter window is applied. In the multivariate approach we fix the filter size equal to 2×2, and each filter slides along the two dimensions.

Given the small size of the kernels, we have chosen not to include pooling layers. We applied, instead, a batch normalization layer [58] after each convolutional layer to standardize their inputs among the samples in each batch.

After the last convolutional layer, a dense layer of 64 nodes with the ReLU activation function and a 2-node dense layer with a softmax activation function provide the expert output.

2) META-LEARNER

Given the outputs of the three experts for an input sample, a straightforward decision strategy could be to compare them and select the class for which its expert model shows the greatest value. We adopt this strategy to evaluate the performance of the base-learners (**ME-LSTM** and **ME-CNN**) models. However, given that each expert is trained separately, it is not ensured that just picking the greatest value between the experts' outputs would provide the best choice for assigning the final label. Looking at the proposed architecture in terms of layered meta-learning, each expert in the base-learner is utilized to model the characteristics that are specific to its binary classification task. This knowledge is exploited by the meta-learner to model the features shared between the binary classification tasks and the 3-class classification task.

The meta-learner utilized in this study is a CART decision tree, a powerful graph-based method used in machine learning. It is a successive model that unites cohesively a series

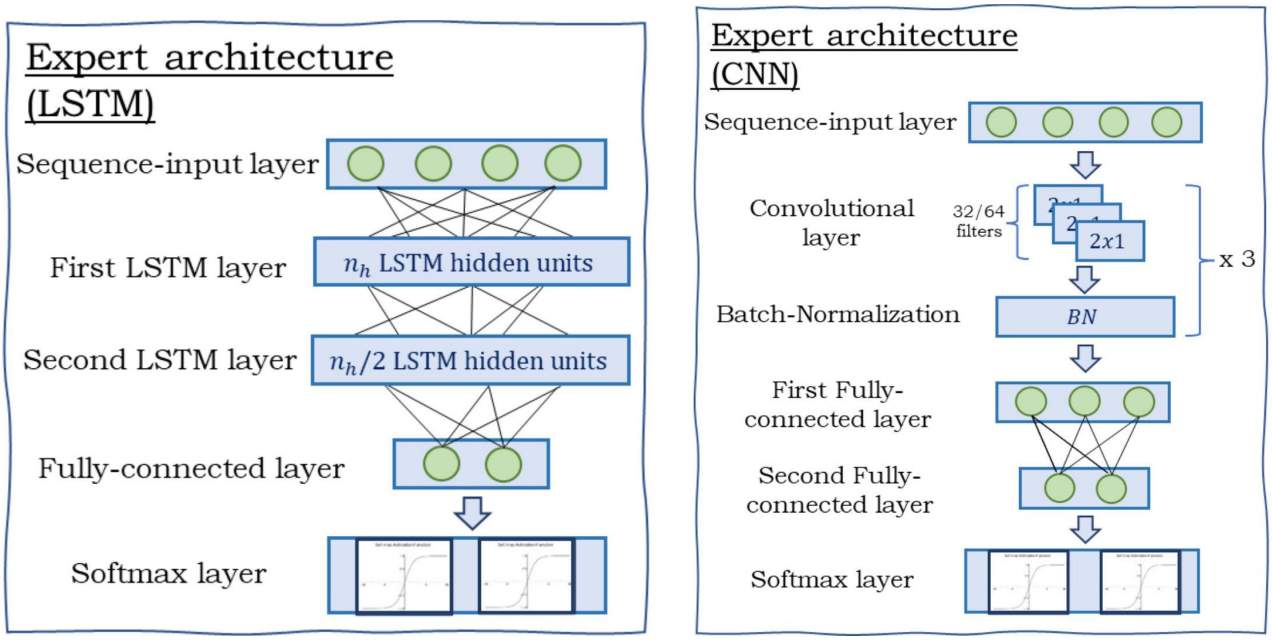


FIGURE 1. Schematic representation of the expert architectures. Left: the architecture based on the LSTM. network. Right: the architecture based on the CNN network.

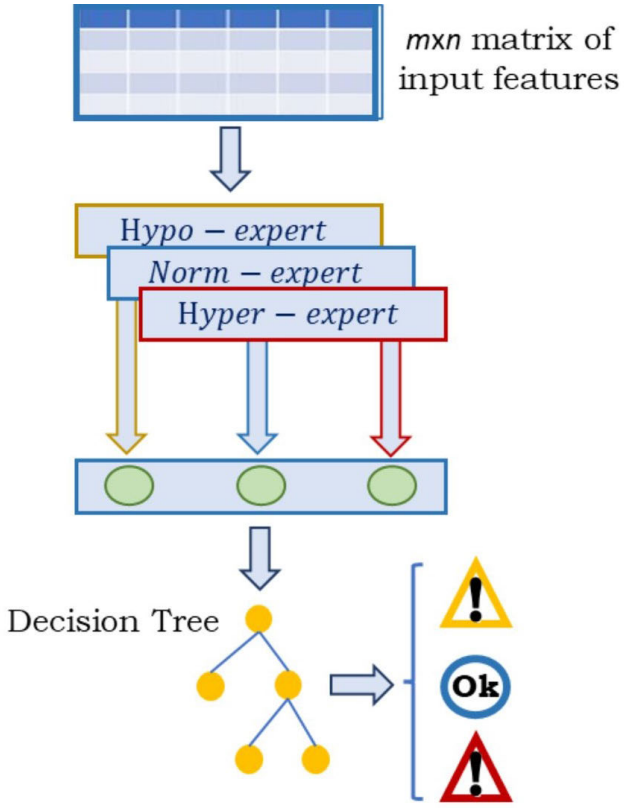


FIGURE 2. Schematic representation of the meta-learning algorithm and the single experts' architecture.

of basic tests (nodes), where a numeric feature is compared to a threshold value in each node [59]. Although it can be prone to overfitting, it is highly interpretable compared to artificial neural networks, and overfitting can be limited using pruning. It is characterized by hyperparameters such as the split criterion for nodes (we utilized the Gini diversity index as the split criterion) and a set of parameters optimized during training. The decision tree meta-learner automatically learns the optimal threshold from the outputs of the three experts. As will be discussed in the following sections, we proved that this meta-learner achieved better performance compared to other algorithms. We will refer to the complete systems (base-learner and meta-learner) as **ME-LSTM-DT** and **ME-CNN-DT** (Figure 2).

F. PARAMETER SEARCH

Before performing the tests, it is necessary to determine the optimal number of parameters of the base-learners, i.e., the number of hidden units n_h of the first LSTM layer of each expert (the number of hidden units of the second LSTM layer is always set equal to $n_h/2$), and the number of filters and kernel size for the CNN. With regard to the meta learner, we investigated whether or not using pruning or class weights would improve performance. In this phase, we use only the public dataset. Straightforwardly, taking apart data from one patient in each turn, we consider 12 different folds as the discovery set. Then, each discovery set is randomly split into a training (70%) and validation (30%) set.

About the LSTM, we investigate a variable number of hidden units n_h for each expert, ranging from 10 to 100, and evaluate the combination which guarantees the best performance through the medium of a grid search. For the CNN, we investigate the combinations with 32, 64, and 128 channels, considering all the parameter combinations by

performing a grid search. During this phase, we train each binary expert on each training set and evaluate its performance with a sample-based approach on the corresponding validation set. Then, we evaluate all the possible combinations of experts to determine the optimal configuration.

As mentioned, this work aims to develop a model capable of achieving high scores for both recall and precision per class, defined as: $Recall = TP/(TP + FN)$; $Precision = TP/(TP + FP)$, where TP, FP and FN are the total numbers of true positives, false positives, and false negatives per class. Straightforwardly, to maximize precision and recall per class at the same time, we considered as the evaluation metric the F1-Score: F1-Score=$2 \cdot Precision \cdot Recall/(Precision + Recall)$. In particular, we evaluated the quality of the predictions by measuring the geometric mean G of the F1-Scores per class: $G = \sqrt[K]{\prod_{i=1}^{K} F1\text{-}Score_i}$, considering $K = 3$ classes. The utilization of functions for the parameter selection that takes into account a combination of metrics, e.g., a combination of recall and specificity, has already proven to be effective for the prediction of nocturnal hypoglycemia, even for longer prediction horizons [60].

Since several combinations of parameters generate similar results for each validation set, we take the best 10 combinations from each fold and then check which of these was the most recurrent combination of parameters. Following this analysis, we select the triplet of 30-80-70 hidden units for the hypoglycemia-euglycemia-hyperglycemia experts for the ME-LSTM, and the triplet of 32-64-64 filters for the three subsequent convolutional layers for the ME-CNN. For the grid search routine, as well as for all the successive training phases described in the next sections, we set the mini-batch size equal to 1/10 of the size of the training set. To avoid overfitting, we set the maximum number of epochs to 1500 and stop the training phase by early stopping if the performance on the validation set does not improve for 10 consecutive checks. We check the validation performance every 25 training iterations and shuffle training and validation data after every epoch.

III. EXPERIMENTAL DESIGN

As widely mentioned in the previous sections, the sample-based approach presents several limitations. Consequently, we evaluate the performance using the event-based approach, as it provides a more realistic overview of the algorithm's capability to predict an adverse event compared to the sample-based approach. Nonetheless, taking into account the strong imbalance related to the event-based approach, we train the model with a sample-based approach. Then, we evaluate performance on event prediction in the aftermath according to the definition of event-based prediction reported in section I-A. We use this strategy as we assume that such training would improve performance because the model could see more samples belonging to the minority classes during the training and validation phase [61], [62].

A. EVENT DETECTION

Event-based performance evaluation requires preprocessing. According to the most widely used definition [18], we consider a true positive an event correctly predicted in advance, and a false positive an event predicted without an actual counterpart. We consider false negatives the events not predicted. Straightforwardly, we consider consecutive timestamps of hypo/hyperglycemia as a single event. In our approach, we use this definition for the events of classes hypoglycemia and hyperglycemia.

For the reasons reported in section II-C, we use a sample-based approach for class normoglycemia, instead. As a consequence, during the event-based performance evaluation, we follow a well-established strategy and consider consecutive misclassified samples as a single false-positive event when the actual observation is normoglycemia. Conversely, we consider each misclassified sample belonging to a minority class (either hypo- or hyperglycemia) a false negative for its class and a false positive for the wrongly assigned class.

Moreover, in order not to consider fluctuations in the read CGM signal nor in the predictions, we consider an event or a prediction as such if it lasts for at least 10 minutes, i.e., if it lasts for at least 3 consecutive timestamps. It is worth noting that our approach increases the imbalance of the dataset, making the classification task more difficult.

In most works, an event is considered correctly predicted if the prediction is supplied with any advance with respect to the actual event [18], [26]. Furthermore, fixed a prediction horizon PH, a parameter k is set so that a prediction is considered correct if performed from 1 to $PH + k$ minutes in advance. In the literature, values of k range from 10 to PH minutes. In this work, we considered $k = 10$ minutes. The standard approach provides no clue as to the actual advance of the prediction.

For this reason, here we introduce a parameter α ranging from 1 to 6 (i.e., from 5 to 30 minutes) to evaluate the number of correct predictions performed with a fixed advance in terms of timestamps. In particular, for classes hypoglycemia and hyperglycemia, we classify the events according to the following rules: **True Positive (TP)** if a correct prediction is performed in the time window $[-(PH + k),-\alpha]$ before the actual event; **False Positive (FP)** if an event is predicted and no actual counterpart is present in the $(k + PH)$ timestamps following the prediction; **False Negative (FN)** if an actual event is not predicted in the time window $[-(PH + k),-\alpha]$ before the actual event. It makes our approach differ from the standard approach, as it allows to evaluate how many events are effectively detected at least α timestamps in advance.

Figure 3 reports a graphical comparison between the proposed and the standard event prediction approaches and some examples of correct and wrong predictions. The figure refers to the prediction of adverse events, i.e., hypo- and hyperglycemia, whereas the prediction of class normoglycemia exploits a sample-based approach. In this example,

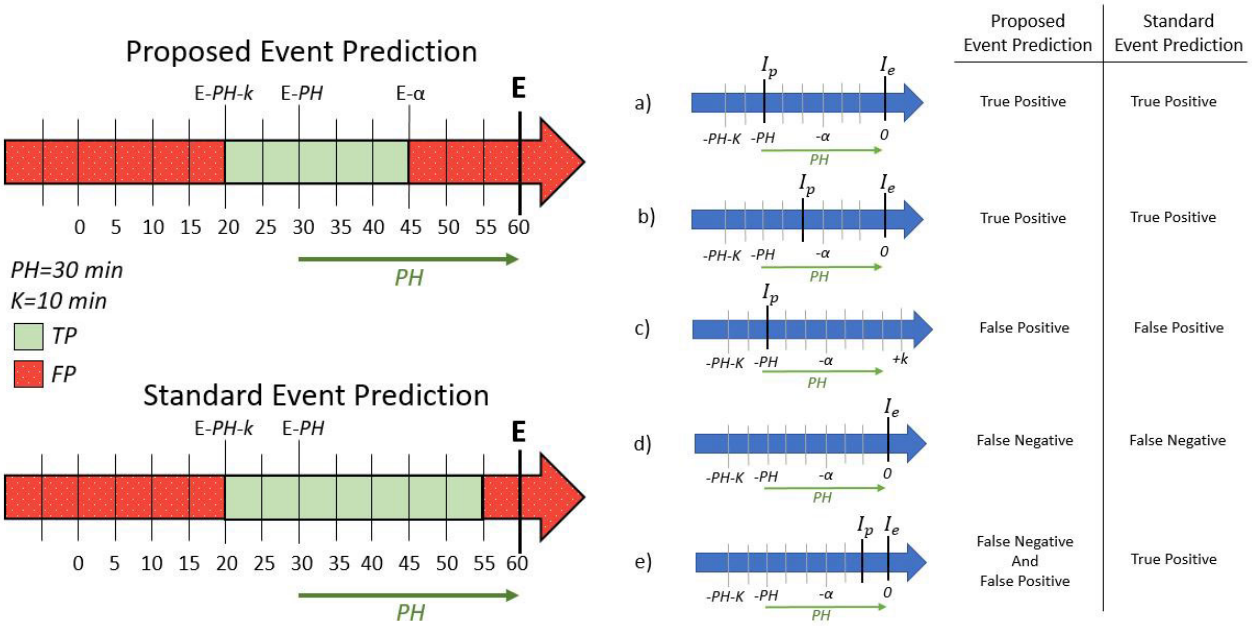


FIGURE 3. Comparison and differences between the proposed and the standard [18] event prediction approach. Left: example of how a predicted event is classified whether as a true positive or a false positive depending on the advance by which the prediction is performed. Given an actual event E beginning after 60 minutes, bright cells indicate when a prediction would produce a true positive, whereas dark cells indicate when a prediction would produce a false positive. Right: examples of predictions and relative classification with the proposed and the standard approach. a) An actual event I_e occurs at $t = 0$. The event is predicted (I_p) exactly PH timestamps in advance. Both approaches consider I_p as a true positive. b) The prediction is performed less than PH but more than α timestamps in advance. Both approaches consider I_p as a true positive. c) I_p is predicted without an actual counterpart. Both approaches consider I_p as a false positive. d) An actual event occurs, but it is not predicted at least ($PH + k$) minutes in advance. Both approaches consider I_e as a false negative. e) I_e occurs and it is predicted less than α timestamps in advance. The proposed approach considers I_p both as a false negative and a false positive, whereas the standard approach considers it as a true positive.

we consider α=3 for the proposed approach. In practice, the standard approach corresponds to our approach with $\alpha = 1$.

We performed three different tests, utilizing the public dataset, the private dataset, and by implementing the proposed architecture on an edge device. The tests are described below and a schematic representation is shown in Figure 4.

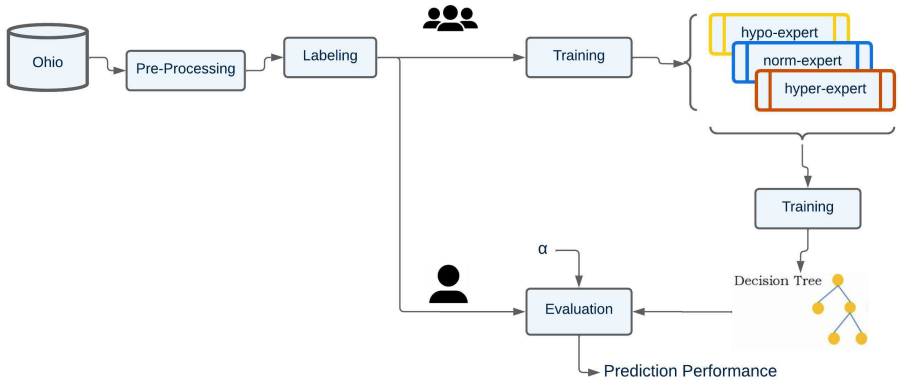
B. TEST 1: EVALUATION ON THE PUBLIC DATASET

We test the proposed approach on the Ohio T1DM dataset with a Leave-1-Patient-Out Cross-Validation (Fig. 4a). We fix, at each turn, data from one subject as the test set, and data from all the other subjects as the discovery set, randomly split into training (70%) and validation (30%) sets for the training of the base-learners. The outputs of the softmax layers of the three experts are passed as training data to the decision tree meta-learner, together with the corresponding target label. At inference time, we classify all the samples in the test set. We then compute for each subject the event detection performance and a confusion matrix; then, we derive the final results from the total confusion matrix calculated by summing all the confusion matrices of all subjects.

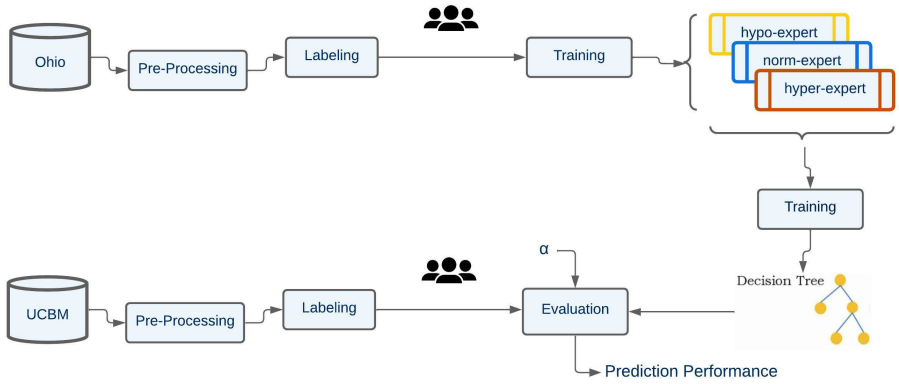
1) COMPARISON WITH OTHER METHODS

To further assess the proposed method, we compare the results we achieve on the public dataset to those of other state-of-the-art methods listed in section I-A. The list of competitors that we test on the Ohio T1DM includes:

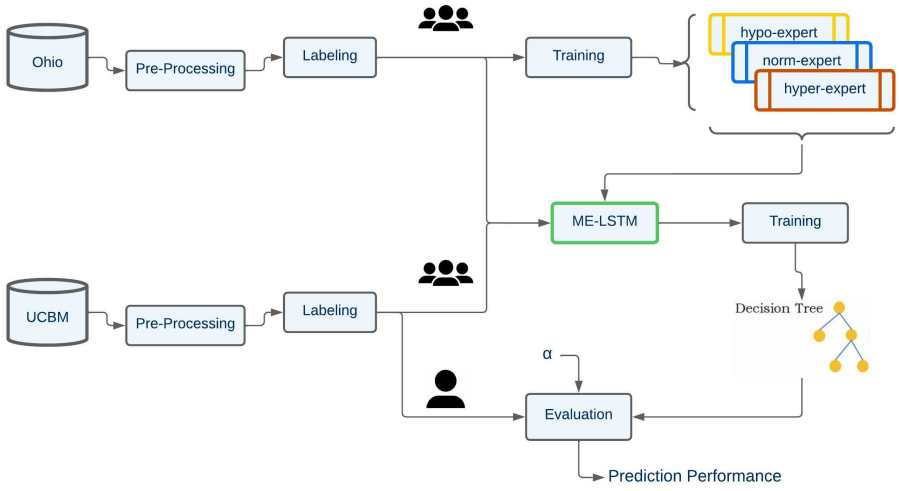
- A Support Vector Machine (SVM) with both polynomial (**SVM-poly**) and radial-basis-function (**SVM-rbf**) kernel. The latter model is the best classifier proposed by Gadaleta et al. [18]. Similar to our model, the learners were trained and tested with one-vs-all decomposition for the classification task.
- A Random Forest (**RF**), which was proposed by Seo et al. [20] and Dave et al. [21]. We performed a grid search on our data to detect the optimal number of learners, resulting in 100. We used the same weights as our proposed models to tackle the data imbalance. It is worth noting that this model consists of an ensemble of decision trees, i.e., the model utilized as a meta-learner in the proposed approach.
- Two different configurations of LSTM neural networks. We performed a grid search on our dataset to determine the optimal amount of LSTM hidden units for both models. The first presents a multi-expert architecture like the one proposed but includes simpler and lighter neural networks with only one hidden layer for each expert. The grid search returned a value of 10, 100, and 1 hidden units for the hypoglycemic, euglycemic, and hyperglycemic experts, respectively (**ME-LSTM 10/100/1**). The second setup consists of a single neural network that



(a) Test 1



(b) Test 2.1



(c) Test 2.2

FIGURE 4. Schematic representations of the experimental tests.

presents the same architecture as a proposed expert, performing a three-class classification task. The grid search returned an optimal value of 70 units in the first and 35 units in the second LSTM layers (**LSTM 3-class**).

- CNN as a three-class classifier (**CNN 3-class**). To keep the framework comparable with the multi-expert model, we implement an analogous architecture as in the ME-CNN system.

Furthermore, we optimized and tested additional meta learners following the optimal ME-LSTM and ME-CNN architecture already found as described in section II-F:

- A SVM (**ME-LSTM-SVM** and **ME-CNN-SVM**) whose optimal configuration resulted in a polynomial kernel with one-vs-one decomposition and no class weights.
- A Naive-Bayes classifier (**ME-LSTM-NB** and **ME-CNN-NB**) whose optimal configuration resulted in normal Kernel smoothing and class weights for each class.
- A feedforward neural network (**ME-LSTM-NN** and **ME-CNN-NN**) whose optimal configuration resulted in one hidden layer with 3 neurons, each having ReLU activation function, and a size of 256 for the mini batches.

We considered as additional competitors the **ME-LSTM** and the **ME-CNN**, i.e., the presented base-learners, in which the final decision on the label to assign to every sample is taken based on the greatest softmax output between the three experts. Finally, in order to assess if performance improves when including injected insulin and physiological features, we evaluated the proposed models, as well as every competitor, using all the four available input features (*Model*-**4F**).

C. TEST 2: EVALUATION ON THE PRIVATE DATASET

We further validate the proposed approach on a private (UCBM) dataset. To implement a realistic evaluation approach, we train the ME-LSTM using only data from the Ohio T1DM dataset, using data of all patients as a discovery set, and adopting a 70/30% split for training and validation set. Then, we perform tests on the five patients from the private dataset one by one. Before conducting these tests, we train the meta-learner following two different approaches:

1) utilizing only data from the public dataset (Fig. 4b). This approach consists of the application of a model trained using all the data available during test 1 to a different test set, consisting of patients that use different CGM sensors;
2) utilizing all the data from the public dataset and, at each turn, data from the four patients of the private dataset that are not the test patient (Fig. 4c). This approach is particularly suited for meta-learning because only the light meta-learner is updated with new data, while the base-learners remain unchanged.

D. TEST 3: EDGE IMPLEMENTATION

To date, there are many devices capable of improving the lives of people with T1D [63], but there are still no devices capable of predicting the onset of hypo- or hyperglycemic episodes without the aid of a doctor. To investigate the possibility of integrating our system on edge and evaluate the time performance due to the utilization of the proposed solution in real applications, we perform an edge implementation test on the edge devices presented in section II-D. We aim to obtain data on the training, transformation, and inference times of the proposed models and thus be able to discover their application scenarios and their possible limitation. We carry out the edge tests following a precise workflow. First, we train the classifiers, then we perform the transformation in *.tflite* to speed up the inference on the edge devices. Afterward, we run the classification process and feed the data to the decision trees downstream.

Regarding the number of operations accomplished:

- we train the base-learners 30 times each for each patient, for a total of 360 training for each classifier;
- we perform the transformations in *.tflite* 100 times for each classifier and each patient, for a total of 1200 transformations for each classifier;
- we calculate the inference times $100 \times N\_test\_samples$ times for each classifier and each patient, following the leave-1-patient-out approach.

Finally, for calculating the training and inference times of the decision trees downstream of the three base-learners, 1000 pieces of training were carried out and $1000 \times N\_test\_samples$ inference tests were calculated, always following the leave-1-patient-out approach. After that, we compute the mean and standard deviations for all the collected data.

IV. RESULTS AND DISCUSSION

In this section, we present and discuss the results achieved with the proposed meta-learning models. For compactness purposes, we use the abbreviations Hypo (hypoglycemia), Norm (normoglycemia), and Hyper (hyperglycemia) in the result tables.

A. TEST 1: RESULTS AND PERFORMANCE ANALYSIS

With regard to the event-based evaluation approach, we report the results achieved on the Ohio T1DM dataset with the proposed models in terms of recall per class, precision per class, and F1-Score per class. Table 2 reports the total results computed by summing all the confusion matrices of the patients, thus providing the performance on the whole dataset for the proposed models. The average results on the 12 patients are similar to the total results. We do not report them for brevity purposes.

Let us focus on the results achieved by the ME-LSTM-DT for different values of α. Recall, precision, and F1-Scores per class tend to become smaller as α increases. It indicates that the models are not fully capable of predicting adverse events with greater advance. The scores of class normoglycemia tend to remain high due to the strong imbalance of the dataset and the sample-based approach considered for this class. We can observe that more than half of the adverse events are predicted at least 30 minutes in advance; at the same time, the amount of FPs is very limited. In detail, the model can predict more than 81% hypoglycemic events and 83% hyperglycemic events at least 15 minutes in advance, while producing a small number of false alarms. Such a time advance could be sufficient to avoid or considerably mitigate

the complications [23]. More in detail, the average time gain, defined as the time between an alert and a real event (where the time gain is 0 in the case of an FN), is 22.8 minutes for hypoglycemia and 24.0 minutes for hyperglycemia. It is a good improvement compared with the literature, where a time gain of 15-20 minutes is usually achieved [17], [29].

It is worth noting that the decrease in the precision-per-class scores is due to the events predicted less than α timestamps in advance. In this case, they are considered false positives despite an actual event is going to occur; for this reason, the most appropriate precision scores to take into consideration are those obtained considering α=1, which express to what extent a wrongly predicted event is not going to occur. It is also interesting to focus on the number of false alarms produced per day by the proposed method. Indeed, a 79.3% precision for hypoglycemia means that, on average, only 2 out of 10 alarms generated by the model are false alarms; in total, the amount of FPs for this class is 201, corresponding to an average of 0.45 false alarms per day. Some of these false alarms might be due to hypoglycemic events which would have actually occurred without a patient intervention [64], or that have not been detected by the CGM sensor [22], [64]. Similarly, a total of 202 FPs is observed for hyperglycemia, corresponding to an average of 0.46 false alarms per day. Such values are small enough not to stress patients with constant alarms that would generate nuisance.

With regard to the results of the ME-CNN-DT, the F1-scores are always slightly greater than those achieved by the ME-LSTM-DT, with the exception of hypoglycemia for $\alpha \geq 5$. In particular, this model performs better on hyperglycemia prediction, as the recall scores are always slightly greater, while the precision scores are very similar. Taking into account hypoglycemia performance, this model presents greater precision (less false alarms) at the expense of a lower ability to detect events with greater advance, corresponding to values of $\alpha \geq 4$. It corresponds to an average time gain of 21.7 minutes for hypoglycemia and 25.0 minutes for hyperglycemia. The 87% precision achieved with $\alpha = 1$ corresponds to 1.3 false alarms every 10 alarms; in total, the amount of FPs for this class is 34, corresponding to an average of 0.087 false alarms per day. A total of 134 FPs are observed for hyperglycemia, corresponding to an average of 0.34 false alarms per day. Although the performance of the ME-CNN-DT model is better in general, the ME-LSTM-DT model would probably provide greater help to T1D patients, due to its improved ability to predict hypoglycemic events with greater advance while keeping small the number of false alarms. However, the ME-CNN-DT would be very helpful as well and would provide better performance in the prediction of hyperglycemia.

1) QUALITATIVE COMPARISON WITH THE LITERATURE

In this section, we provide a comparison with the results presented by other works. Straightforwardly, we focus on the total results we achieve considering α=1 because they correspond to the approach pursued in the literature [18]. The comparison is qualitative because works that performed event detection used different datasets.

For hypoglycemia, the best recall score is 95%, proving that almost all hypoglycemic events are predicted at least 5 minutes in advance, while precision is strictly greater than 79%. Of the models listed in section I-A, only our previous work [42] achieves a better precision (86.4%), which is lower than that of the ME-CNN-DT model, while achieving a sensitively lower recall (59.8%). The second best precision score is achieved by Zhu et al. [17] (65.6%) while achieving 84.1% recall. They proposed a bidirectional recurrent neural network refined with patient-specific model agnostic meta-learning for regression on three datasets (including the Ohio T1DM dataset), obtaining on average 0.48 false alarms per day. Similarly, the model proposed by Prendin et al. [29] achieves a good precision (64%), which also results in a smaller amount of 0.5 false alarms per day; however, the recall reported in that study is lower (82%). We outperform by more than 40% the remaining hypoglycemia precision scores. Daskalaki et al. [26] achieve 100% recall for both hypoglycemia and hyperglycemia; nonetheless, their work only aims at predicting events regardless of the precision per class. They report that their model generates on average 1.6 false alarms per day, but there is no clue on the number of events in the test set, so the computation of the precision per class is not possible. The same applies to the work from Yang et al. [19].

For hyperglycemia, the recall score is noteworthy as well, being about 92%, whereas precision is above 89%. It is worth pointing out that, although the prediction of hyperglycemia may seem of reduced practical impact because most patients experience hyperglycemia after a meal, the proposed models do not exploit carbohydrate information to perform such a prediction, in the view of a fully-automated system that does not require the patient to provide meal data manually. We outperform by more than 33% the only who reported hyperglycemia precision (Gadaleta et al. [18], 56%), although the same study outperforms our hyperglycemia recall (95%). Nonetheless, their proposed SVM model produces many false alarms (hypo/hyperglycemia precision equal to 36/57%). In general, the proposed meta-learning approaches outperform the previously presented ones. However, these comparisons are qualitative because tests are performed on different datasets.

The F1-Score per class, which can be interpreted as the ability of the model to perform accurate predictions while generating few false alarms, is greater than 86% for every class. It proves that the proposed approach could be reliable in a real-life application without stressing patients with many false alarms, which is rarely achieved in the literature. However, a value of α=1 means that predictions are performed at least 5 minutes in advance, which may not be a sufficient time to prevent adverse events. It is the reason why we investigated the performance with different values of α.

For sake of completeness, we report in Table 3 performance of the proposed meta-learning models with the sample-based

TABLE 2. Total results of the proposed meta-learning systems with the event-based approach, extracted from the total confusion matrix for Test 1. Results are reported in terms of recall [%], precision [%], and F1-Score [%] per class for the different values of α investigated.

Model	α	Hypoglycemia			Normoglycemia			Hyperglycemia		
		Recall	Precision	F1-Score	Recall	Precision	F1-Score	Recall	Precision	F1-Score
ME-LSTM-DT	1	95.0	79.3	86.4	92.5	99.6	95.9	91.9	89.2	90.5
	2	88.3	78.0	82.9	92.5	99.6	95.9	89.0	88.5	88.8
	3	81.0	76.6	78.8	92.5	99.6	95.9	83.9	86.7	85.3
	4	73.3	75.0	74.1	92.5	99.6	95.9	78.6	85.2	81.1
	5	65.9	73.1	69.3	92.5	99.6	95.9	72.1	82.6	77.0
	6	54.8	69.1	61.1	92.5	99.6	95.9	62.9	79.2	70.2
ME-CNN-DT	1	92.3	87.3	89.7	92.5	99.9	96.0	94.8	89.0	91.8
	2	83.9	86.0	85.0	92.5	99.8	96.0	91.1	87.9	89.5
	3	75.8	84.8	80.0	92.5	99.8	96.0	87.3	86.4	86.9
	4	67.5	83.2	74.5	92.5	99.8	96.0	83.2	84.9	84.0
	5	59.4	80.8	68.5	92.5	99.7	96.0	77.9	82.7	80.3
	6	48.5	77.8	59.7	92.5	99.6	95.9	66.8	79.5	72.6

TABLE 3. Results with a sample-based approach.

Model	Class	Recall [%]	Precision [%]	F1-Score [%]
ME-LSTM-DT	Hypo	90.6	71.2	79.7
	Norm	91.1	96.0	93.5
	Hyper	94.7	90.2	92.4
ME-CNN-DT	Hypo	78.2	77.6	77.9
	Norm	91.8	92.2	92.0
	Hyper	89.5	88.9	89.2

approach, albeit it is not fully indicative of a model's real performance, as widely discussed in the previous sections. The results achieved are highly competitive compared to those reported by the models listed in Table 1 that pursue a sample-based approach, since only the study from Dave et al. [21], who proposed a model composed of two Random Forests, one day-specific and one night-specific, achieves better hypoglycemia recall (93.7%) but at the expense of a considerably lower precision (15.1%). The opposite approach was pursued by Marcus et al. [28], who aimed to reduce as much as possible the number of false alarms per day, achieving a 4% false-positive rate; nonetheless, their recall is considerably lower than ours (64% and 61% for hypo- and hyperglycemia).

Finally, we report in Table 4 the results achieved by the proposed models when a longer PH of 60 or 120 minutes is considered. The performance worsens sensitively for both models. Although a longer PH would provide patients with more time to react to an incoming adverse event, a prediction over such a long temporal horizon necessarily increases the uncertainty in the predictions, for example, due to the attempt of the algorithm to maximize the performance for the minority classes, which leads to the generation of many false alarms, as demonstrated by the considerably lower recall scores for class normoglycemia. In the light of this analysis, a 30-minute PH seems appropriate for event detection. However, the results achieved by the proposed model are comparable to those of other recent studies that investigate a longer PH for the prediction of nocturnal hypo- or hyperglycemia [24], [65], which also suffer from a lower recall or precision score.

TABLE 4. Average percentage results over the 12 Ohio T1DM patients with the event-based approach of the two proposed models with a PH of 60 and 120 minutes.

Model	Class	Recall	Precision	F1-Score
ME-LSTM-DT	Hypo	29.1	44.7	35.2
PH = 60 min	Norm	80.1	97.6	87.9
	Hyper	41.6	47.4	42.9
ME-CNN-DT	Hypo	25.3	43.8	31.2
PH = 60 min	Norm	83.4	98.0	90.1
	Hyper	42.9	56.5	47.9
ME-LSTM-DT	Hypo	26.3	21.9	21.7
PH = 120 min	Norm	60.3	92.8	73.0
	Hyper	55.8	31.0	39.3
ME-CNN-DT	Hypo	24.6	24.8	24.7
PH = 120 min	Norm	65.7	92.9	76.9
	Hyper	55.3	37.2	43.9

2) RESULTS OF THE COMPARISON WITH OTHER METHODS ON THE OHIO T1DM DATASET

In this section, we compare our performance to the performance of the competitors listed in section III-B1. The results are referred to the event-based approach and are computed on the total confusion matrix with a Leave-1-Patient-Out Cross Validation approach. All the competitors have undergone a grid search to select the optimal model parameters. To provide a compact overview of the performance for different values of α, we report the results of each model in terms of the F1-Scores per class and of the geometric mean G of the F1-Scores per class, because they provide an overview of the model capability to achieve good performance for each class.

Table 5 reports the results of the comparison with the other methods when exploiting only CGM as an input feature. Results are reported in terms of the F1-Score for classes hypoglycemia (F_{Hypo}), normoglycemia (F_{Norm}), and hyperglycemia (F_{Hyper}), together with the geometric mean G of the F1-Scores per class. Considering all values of α, both the proposed models outperform all the competitors by a large margin, except for class normoglycemia for which the SVM with radial basis function always achieves better results.

Model	$\alpha=1$				$\alpha=2$				$\alpha=3$				$\alpha=4$				$\alpha=5$				$\alpha=6$			
	F_{Hypo}	F_{Norm}	F_{Hyper}	G	F_{Hypo}	F_{Norm}	F_{Hyper}	G	F_{Hypo}	F_{Norm}	F_{Hyper}	G	F_{Hypo}	F_{Norm}	F_{Hyper}	G	F_{Hypo}	F_{Norm}	F_{Hyper}	G	F_{Hypo}	F_{Norm}	F_{Hyper}	G
ME-LSTM-DT	86.4	95.9	90.5	90.9	82.9	95.9	**88.8**	89.0	78.8	95.9	85.3	86.4	74.1	95.9	81.8	83.4	**69.3**	95.8	77.0	80.0	**61.1**	95.8	70.2	74.3
ME-CNN-DT	**89.7**	96.0	**91.8**	**92.4**	**85.0**	96.0	85.9	**90.0**	**80.0**	96.0	**86.9**	**87.3**	**74.5**	96.0	**84.0**	**84.4**	68.5	96.0	**80.3**	**80.7**	59.7	95.9	**72.6**	**74.6**
ME-LSTM	74.0	95.9	90.2	86.2	72.1	95.9	86.1	84.1	70.1	95.9	82.1	82.0	64.8	95.9	79.0	78.9	58.8	95.9	75.9	75.4	52.8	95.8	71.1	71.1
ME-LSTM 10/100/1	64.3	92.5	86.2	80.1	63.4	92.4	81.8	78.3	62.0	92.4	77.9	76.4	60.3	92.4	74.9	74.7	57.7	92.4	72.2	72.7	52.3	92.4	64.7	67.9
LSTM 3-class	46.0	92.0	62.8	64.3	45.0	92.0	59.8	62.8	42.4	92.0	56.8	60.5	39.3	92.0	52.7	57.6	35.8	92.0	47.8	54.0	31.3	92.0	43.3	50.0
ME-CNN	87.6	97.9	90.3	91.8	76.2	97.9	85.6	86.1	64.8	97.9	81.9	80.3	57.7	97.9	78.5	76.2	48.9	97.9	74.2	70.8	41.5	97.9	66.5	64.6
CNN 3-class	86.0	97.7	90.2	91.2	77.3	97.7	86.6	86.8	69.5	97.7	82.4	82.4	61.2	97.7	78.6	77.2	53.8	97.7	73.9	72.9	43.8	97.6	66.2	65.6
SVM-rbf	80.6	**99.5**	85.0	88.0	68.6	**99.5**	80.0	81.8	59.8	**99.5**	77.2	77.2	51.0	**99.5**	73.8	72.1	46.5	**99.5**	71.7	69.2	37.9	**99.5**	63.6	62.2
SVM-poly	76.4	99.1	78.3	84.9	65.9	99.1	77.0	79.5	59.8	99.1	74.3	76.1	53.9	99.0	71.4	72.5	46.0	99.0	67.8	67.6	39.8	99.0	59.5	61.7
RF	61.5	96.3	81.6	78.4	58.8	96.3	78.1	76.0	55.2	96.3	74.2	73.4	50.0	96.3	70.2	69.7	45.8	96.3	66.3	66.4	38.3	96.3	59.4	60.3
ME-LSTM-SVM	88.1	99.2	94.5	93.8	63.4	99.2	78.4	79.0	46.2	99.2	62.2	65.8	34.2	99.2	51.3	55.8	27.9	99.2	43.9	49.5	24.2	99.1	39.1	45.4
ME-CNN-SVM	69.3	95.0	82.9	81.7	64.0	94.9	80.4	78.7	59.9	94.9	76.9	75.9	54.2	94.9	72.6	72.0	51.2	94.9	67.2	68.9	47.0	94.9	61.1	64.8
ME-LSTM-NB	60.9	93.1	86.3	78.8	60.7	93.1	84.6	78.2	59.2	93.1	78.4	75.6	56.0	93.1	69.0	71.1	52.3	93.1	61.4	66.9	44.6	93.1	52.1	60.0
ME-CNN-NB	50.7	82.9	86.2	71.3	50.6	82.9	85.6	71.0	50.2	82.9	83.1	70.2	50.0	82.9	80.2	69.3	49.4	82.9	77.1	68.1	48.6	82.9	72.4	66.3
ME-LSTM-NN	89.6	97.5	90.2	92.4	84.5	97.5	84.3	88.5	76.1	97.4	80.1	84.0	65.4	97.4	77.8	79.1	58.3	97.4	73.7	74.8	49.5	97.4	67.4	68.7
ME-CNN-NN	87.8	97.8	90.2	91.8	79.6	97.8	85.6	87.3	69.8	97.8	81.5	82.2	61.2	97.8	77.9	77.5	53.5	97.8	73.6	72.3	45.2	97.8	65.4	66.1

TABLE 5. Results of the proposed models and the competitors with the event-based approach, extracted from the total confusion matrix, for the different values of α investigated. All the competitors are tested using only CGM as input feature. Results are reported in terms of the F1-Score for classes hypoglycemia (F_{Hypo}), normoglycemia (F_{Norm}) and hyperglycemia (F_{Hyper}), together with the geometric mean G of the F1-Scores per class. We investigated the models listed in section III-B1, which are an SVM with radial-basis-function (SVM-rbf) [18] and with polynomial (SVM-poly) kernel, a Random Forest (RF) [20], [21], and variations of LSTM and CNN models. The results in the bottom panel refer to the proposed base learners followed by different meta learners. The best score of each column is highlighted in red.

Model	$\alpha=1$				$\alpha=2$				$\alpha=3$				$\alpha=4$				$\alpha=5$				$\alpha=6$			
	F_{Hypo}	F_{Norm}	F_{Hyper}	G	F_{Hypo}	F_{Norm}	F_{Hyper}	G	F_{Hypo}	F_{Norm}	F_{Hyper}	G	F_{Hypo}	F_{Norm}	F_{Hyper}	G	F_{Hypo}	F_{Norm}	F_{Hyper}	G	F_{Hypo}	F_{Norm}	F_{Hyper}	G
ME-LSTM-DT	86.4	95.9	90.5	90.9	82.9	95.9	88.8	89.0	78.8	95.9	85.3	86.4	74.1	95.9	81.8	83.4	**69.3**	95.8	77.0	80.0	**61.1**	95.8	70.2	74.3
ME-CNN-DT	89.7	96.0	91.8	**92.4**	**85.0**	96.0	85.9	**90.0**	**80.0**	96.0	86.9	**87.3**	**74.5**	96.0	84.0	**84.4**	68.5	96.0	80.3	**80.7**	59.7	95.9	72.6	**74.6**
ME-LSTM-DT-4F	88.5	95.6	91.9	91.8	82.4	95.6	88.5	88.3	73.4	95.6	83.0	83.5	67.6	95.5	79.1	78.2	55.2	95.5	70.0	70.5	46.0	95.4	62.0	63.9
ME-CNN-DT-4F	87.2	96.4	**93.7**	92.3	82.6	96.4	**91.6**	**90.0**	77.9	96.4	**88.6**	87.2	71.9	96.3	**85.0**	83.7	66.8	96.3	**80.9**	80.4	57.6	96.3	**72.9**	73.9
ME-LSTM-4F	73.7	95.9	89.9	86.0	70.7	95.9	84.6	83.1	64.0	95.9	74.4	77.0	56.3	95.9	64.2	70.2	46.9	95.8	57.1	63.5	40.3	95.8	49.9	57.8
LSTM 3-class-4F	44.1	92.2	61.2	62.9	41.9	92.2	58.4	60.9	39.5	92.1	54.7	58.4	36.1	92.1	51.0	55.4	31.8	92.1	46.8	51.6	28.1	92.1	43.1	48.1
ME-LSTM 10/100/1-4F	43.7	92.5	73.5	66.7	42.2	92.4	63.7	62.9	38.1	92.4	49.9	56.0	34.4	92.3	37.6	49.2	30.1	92.3	30.8	44.1	25.6	92.2	24.1	38.5
ME-CNN-4F	90.9	98.1	92.3	**93.6**	76.9	98.1	87.0	86.8	66.7	98.1	82.1	81.2	57.6	98.1	78.6	76.2	50.9	98.0	74.5	71.9	42.2	98.0	66.6	65.0
CNN 3-class-4F	**91.0**	97.9	91.0	93.2	79.3	97.9	86.4	87.5	69.2	97.9	82.6	82.4	61.7	97.8	79.0	78.1	54.3	97.8	74.2	73.3	46.7	97.8	66.2	67.1
SVM-rbf-4F	63.6	**99.2**	68.6	75.6	44.1	99.1	55.6	62.4	29.5	99.1	46.0	51.3	22.7	99.1	37.9	44.0	12.0	99.0	31.5	38.4	13.6	99.0	27.0	33.1
SVM-poly-4F	75.5	**99.2**	83.4	85.5	58.7	**99.2**	70.9	74.5	44.1	**99.2**	56.0	62.6	34.3	**99.2**	41.1	51.9	26.7	**99.1**	32.9	44.3	21.4	**99.1**	26.4	38.3
RF-4F	71.7	97.4	82.4	83.3	63.6	97.4	72.1	76.5	55.3	97.4	72.1	67.8	47.1	97.4	46.5	59.7	37.9	99.3	36.3	51.2	30.0	97.3	28.0	43.4
ME-CNN-SVM-4F	84.4	98.6	89.1	90.5	61.8	98.5	75.8	77.3	44.3	98.5	63.1	65.0	36.1	98.5	53.3	57.4	29.8	98.5	46.9	51.7	26.9	98.5	41.2	47.8
ME-CNN-SVM-4F	90.3	97.9	90.8	92.9	77.6	97.9	86.3	86.9	67.9	97.9	82.4	81.8	59.2	97.9	78.9	77.0								
ME-LSTM-NB-4F	57.1	91.5	86.1	76.6	56.4	91.5	82.8	75.3	54.1	91.5	75.7	72.0	50.3	91.5	66.1	67.2	46.0	91.5	58.6	62.7	41.5	91.4	51.1	57.9
ME-CNN-NB-4F	48.8	79.1	89.8	70.3	48.7	79.1	88.2	69.8	48.5	79.1	85.0	68.8	47.9	79.1	81.4	67.6	46.7	79.1	78.3	66.1	45.5	79.1	73.7	64.2
ME-LSTM-NN-4F	57.2	91.9	74.6	73.2	49.5	91.9	71.8	68.9	46.3	91.9	68.8	66.4	40.3	91.9	64.7	62.1	36.1	91.8	59.8	58.3	31.8	91.8	55.8	54.8
ME-CNN-NN-4F	82.7	98.1	91.7	90.6	75.8	98.1	86.2	86.2	64.1	98.1	81.7	80.0	55.7	98.1	78.0	75.2	48.6	98.1	73.9	70.6	42.6	98.0	65.9	65.0

TABLE 6. Results of the proposed models and of the competitors with the event-based approach, extracted from the total confusion matrix, for the different values of α investigated. The top panel reports the results of the proposed models with a univariate approach; results in the central panel refer to the proposed models and the competitors tested using all the 4 available features (Model-4F); results in the bottom panel refer to the proposed base learners followed by different meta learners. Results are reported in terms of the F1-Score for classes hypoglycemia (F_{Hypo}), normoglycemia (F_{Norm}) and hyperglycemia (F_{Hyper}), together with the geometric mean G of the F1-Scores per class. The best score of each column is highlighted in red.

However, this is the majority class and is less important to predict accurately. The best competitors are the other CNN-based models for $\alpha \leq 2$, and the ME-LSTM for greater values.

Let us focus on the comparison between the results achieved with and without resorting to meta-learning. With regard to hyperglycemia, a small improvement is observed for F1-scores, as the slight precision increase is balanced by the slight recall decrease. The major advantage of the meta-learning is observed with regard to hypoglycemia, where an increase of 10 to 15% is observed for all the F1-scores. In detail, although the recall is slightly decreased by 3 to 7%, a considerable improvement of about 20% is observed for the precision, resulting in a much lower amount of false alarms. We can conclude that using a meta-learner considerably improves the capability of predicting adverse events while producing few false alarms. We also tested two other meta-learners (Naive-Bayes classifier and Support Vector Machine) which returned very high recall scores (above 99%) for both hypo- and hyperglycemia, at the expense of very low precision (below 15%). We do not report these results for the sake of brevity. From a comparison with the ME-LSTM, the ME-CNN, and the Random Forest, it is clear that the utilization of the meta-learning approach as a whole guarantees sensitively better performance than any of the models it is composed of. It is also interesting to note that the multi-expert systems ME-LSTM and ME-CNN outperform the correspondent three-class model, suggesting that the ensemble strategy is more effective for this task.

Finally, we tested our models and the competitors using a multivariate approach, i.e., using all four available features as input (Model-4F); these results are reported in the bottom panel of Table 6, whereas the top panel reports the results of the proposed univariate approach. The reported results are referred to the total confusion matrix computed by adding the confusion matrices of all patients. In general, all the competitors perform better when using CGM alone as an input feature. The proposed models outperform all the competitors. The only exception concerns class normoglycemia, for which the SVM with polynomial kernel always achieves better results. The analysis is very similar to that provided for the models which exploit only CGM. An interesting behavior is observed for hyperglycemia prediction, for which the ME-CNN-DT-4F outperforms all the other models, including its univariate counterpart. This is probably due to the information concerning insulin boluses, which allows an easier prediction of postprandial hyperglycemia; however, such a feature complicates the data management, and the improvement compared to the univariate model is not very marked (3-4%).

In conclusion, by testing different models on the same dataset we observed that:

1) resorting to multi-expert systems with a majority-based decision policy provides better performance compared to utilizing a single model for a 3-class classification task;
2) using meta-learning considerably improves the performance of multi-expert base-learners.

B. TEST 2: RESULTS AND PERFORMANCE ANALYSIS

We tested a private dataset to evaluate the capability of the proposed approach to adapt to data of new patients. The UCBM dataset includes patients that utilize a different CGM sensor than the patients enrolled in the Ohio T1DM dataset, and that regularly perform physical activity. This test was performed twice: 1) by training the meta-learner only on the Ohio patients, and 2) by training the meta-learner on the Ohio dataset joined with the UCBM dataset with a leave-1-patient-out approach. Table 7 reports the results of these tests (we do not report the results for the normoglycemia class, which are all above 95%).

Let us focus on the results of the first implementation of the test, in which only the Ohio T1DM dataset was used to train the meta-learner. The performance worsens considerably, particularly for larger values of α. The main worsening concerns the hyperglycemia prediction of the ME-CNN-DT; however, also the ME-LSTM-DT model is able to predict only few more than half hyperglcemic events with any advance. This suggests that the different cohort of patients, with different habits and life style, joint with a different CGM sensor, presents completely different patterns preceding hyperglycemia. Conversely, the worsening for class hypoglycemia is less pronounced, suggesting that common patterns exist between the two datasets.

Let us now focus on the results achieved including part of the UCBM dataset in the training set. It is worth stressing out that data from the UCBM dataset were used only to train the meta-learners, whose training requires a very small amount of time; differently, only the public dataset was used (once) for the more onerous training of the base-learners. Again, the performance is considerably worse than Test 1; nonetheless, a pronounced improvement is observed for all classes and for all values of α, with the exception of class hypoglycemia of the ME-LSTM-DT model, which already achieved the best performance in the first configuration. The improvement is particularly noticeable for larger values of α and for the ME-CNN-DT, whose F1-scores increase by up to 4 times.

Although the results achieved with the second experimental setup are in line with those presented in previous works (e.g. an F1-score of 72% for hypoglycemia is presented in [29]), these results are considerably worse than those achieved in Test 1. This could be expected in the light of the huge difference between the two datasets under observation, and considering the limited size of the UCBM dataset for training. In addition, it has been widely investigated how the prediction of T1D events and glycemic levels is particularly challenging on patients that perform physical activity [48], [66]. In conclusion, the take-home message of this test is that the predictive performance of the proposed meta-learning approach can be considerably improved using a very limited amount of data from the new dataset. Such an improvement is achievable in the time required to train the meta-learner,

TABLE 7. Total results of the tests performed over the private dataset. Results are reported for the ME-LSTM-DT (left) and the ME-CNN-DT (right) in terms of recall [%], precision [%] and F1-Score [%] per class for the different values of α investigated. The top panel reports the results of the tests performed using only the Ohio dataset to train the meta-learner, whereas the results in the bottom panel are referred to the model in which the meta-learner is updated using data of the UCBM dataset using a leave-1-patient-out approach.

Training dataset	α	ME-LSTM-DT						ME-CNN-DT					
		Hypoglycemia			Hyperglycemia			Hypoglycemia			Hyperglycemia		
		Recall	Precision	F1-Score	Recall	Precision	F1-Score	Recall	Precision	F1-Score	Recall	Precision	F1-Score
Ohio	1	81.3	97.5	88.7	51.4	79.7	62.5	70.8	80.1	75.1	32.2	80.0	45.9
	2	67.9	96.7	79.8	39.9	74.0	51.9	38.8	55.8	45.7	25.9	76.7	38.8
	3	60.5	96.7	74.5	33.5	70.1	45.3	27.5	48.4	35.0	16.3	45.0	24.0
	4	46.8	95.0	62.7	21.5	65.0	32.3	19.5	41.4	26.5	14.3	45.0	21.7
	5	39.2	93.3	55.2	12.4	52.7	20.1	11.5	29.6	16.6	10.3	43.3	16.7
	6	34.6	90.0	50.0	8.4	46.0	14.2	11.5	29.6	16.6	7.3	23.3	11.2
Ohio + UCBM	1	91.8	90.9	91.3	84.6	91.2	87.8	88.4	66.3	75.7	63.3	73.6	68.1
	2	82.1	89.8	85.8	70.3	89.7	78.9	74.2	62.5	67.8	59.9	70.1	64.6
	3	64.6	87.7	74.4	47.1	87.8	61.3	74.2	62.5	67.8	52.8	65.3	58.4
	4	56.2	86.6	68.2	32.7	82.0	46.7	65.5	59.8	62.5	43.5	59.0	50.1
	5	40.6	77.1	53.2	23.8	78.0	36.5	53.0	50.2	51.5	39.1	56.9	46.4
	6	39.2	76.7	51.9	14.9	74.7	24.9	42.4	42.2	42.3	36.8	54.7	44.0

which is far less than a second, as discussed in the next subsection.

C. TEST 3: RESULTS OF THE EDGE IMPLEMENTATION

The tests on the edge system were carried out following the pipeline described in subsection III-D. The results concerning training, conversion and inference time are shown in Table 8.

From the data collected, on the one hand it can be observed that the training of CNNs is more onerous in terms of time required, when compared to that of LSTMs; on the other hand, the transformation times of the CNN models are less time consuming, by a factor of 5, with respect to the LSTM ones. This is due to the steps needed for the conversion into *.tflite*; in fact, in order to transform an LSTM, or in general an RNN, into *.tflite* it is necessary to build the graph of the model itself, an operation that can be performed through the use of the concrete functions of Tensor Flow. This operation, which is not required for the CNN transformation, results in a longer transformation time for this type of models. In all cases, no appreciable loss in performance was observed.

As far as inference times are concerned, it can be observed that, regardless of the model under consideration, they are around values of less than a tenth of a millisecond. We can therefore state that the time required to perform this operation has little or any influence on the total time count, thus allowing both the considered models to work effectively in real time when considering the 5-minute sampling window typical of CGM sensors. Moreover, the training and transformation times of the networks are in both cases greater than the single window required for prediction, but considerably shorter for LSTM. Therefore, in case of a possible implementation of an online learning system, i.e. a system capable of updating itself directly on the edge device using new incoming data, the use of multi-expert LSTMs would be preferable due to their speed in the training phase. The only data collected not shown in table 8 are those concerning the training and inference time of the decision trees. We made this choice because, for both the ME-LSTM-DT and the ME-CNN-DT, the results obtained are overlapping with a mean time for training the decision tree of $0.055 \pm 0.002\,s$ and inference time of $9.86 \cdot 10^{-8} \pm 1.86 \cdot 10^{-8}\,s$ and therefore, similarly to the inference times of the models, negligible for a real application scenario. This suggests that updating the meta-learners on the edge with new incoming data would have a very limited impact on the device in terms of computational time.

TABLE 8. Average time required with standard deviation for the edge implementation of the multi-expert architecture. The results for both individual experts and the two multi-expert approaches are reported.

Model	Training time (s)	Transformation time (s)	Inference time (s)
LSTM hypo	51.4 ± 19.4	55.2 ± 2.6	$1 \cdot 10^{-4} \pm 5 \cdot 10^{-5}$
LSTM norm	181.5 ± 62.9	55.2 ± 2.7	$2 \cdot 10^{-4} \pm 8 \cdot 10^{-5}$
LSTM hyper	147.7 ± 43.6	55.1 ± 2.8	$2 \cdot 10^{-4} \pm 6 \cdot 10^{-5}$
CNN hypo	1133.4 ± 415.2	10.6 ± 1.1	$3 \cdot 10^{-4} \pm 8 \cdot 10^{-4}$
CNN norm	1358.3 ± 610.0	10.6 ± 1.0	$2 \cdot 10^{-4} \pm 5 \cdot 10^{-5}$
CNN hyper	1467.3 ± 456.3	10.6 ± 0.9	$2 \cdot 10^{-4} \pm 5 \cdot 10^{-5}$
ME-LSTM	380.7 ± 125.9	165.5 ± 8.2	$5 \cdot 10^{-4} \pm 2 \cdot 10^{-4}$
ME-CNN	3958.9 ± 1481.6	31.8 ± 3.0	$6 \cdot 10^{-4} \pm 9 \cdot 10^{-4}$

V. CONCLUSION

Intensive management of T1D with multiple daily injections or continuous subcutaneous insulin infusion showed to reduce the risk of developing micro/macrovascular complications [67]. However, the insulin-intensive treatment exposes patients to frequent hyperglycemia and severe hypoglycemia events. Growing evidence indicates that significant glycemic variability significantly affects the onset and progression of diabetes complications [68]. However, these glycemic fluctuations are not easily predictable. The methods developed in this study exploits CGM data flow to predict hypo- and hyperglycemic events in the real world. Our approach would help subjects quickly adjust insulin therapy and nutrition, enabling them to speed up decision-making and improve personal self-management. Thus, it would reduce daily activities-related glycemic fluctuations and, consequently, the risk of acute and chronic complications.

We presented a layered meta-learning approach based on multi-expert models to predict adverse events in T1D. We approached a 3-class classification task, considering

classes hypoglycemia, normoglycemia, and hyperglycemia, and performed tests in 3 different experimental configurations. We pursued a univariate approach exploiting CGM data using a public dataset and further evaluated the performance on a private dataset. We also introduced a parameter α in order to evaluate the effective advance by which predictions are performed, and evaluated performance for α varying from 1 to 6 (i.e., from 5 to 30 minutes). Finally, we investigated the real-life feasibility of the proposed approach by implementing it on an edge device and observing training, transformation, and inference time.

The layered meta-learning model is composed of a base-learner and a meta-learner. The base-learner consists of three neural networks, which are either LSTM or CNN, each specialized in detecting one of the three classes. The softmax output of each expert is passed to a decision tree meta-learner, which is trained with the same data as the base-learners and provides the final model classification.

With regard to the test on the public dataset, the presented approach outperforms by a large margin all the works in the literature using any of the investigated base-learners. The ME-LSTM-DT model achieves better recall scores for hypoglycemia, with particular regard to larger values of α, whereas the ME-CNN-DT achieves better F1-scores in general. For what concerns hypoglycemia prediction, the average time gain of the two models is equal to 22.8 and 21.7 minutes, and the average number of false alarms per day is 0.45 and 0.087, respectively. Such time advances correspond to an $\alpha = 4$, for which an F1-Score greater than 74% is achieved by both the proposed models; this performance combined with such a time advance would be sufficient to avoid or at least mitigate these events. For what regards hyperglycemia, the average time gains of the two models, 24.0 minutes for the ME-LSTM-DT and 25.0 minutes for the ME-CNN-DT, combined with their low rate of false alarms (0.46 and 0.34 per day, respectively), as well as the corresponding F1-Scores above 77% for both models, would permit to avoid large excursions above the target glycemic range. Based on these results, both the proposed models could be very useful to T1D patients in the disease management; in particular, the ME-LSTM-DT model would probably provide greater help, due to its improved ability to predict hypoglycemic events with greater advance while keeping small the number of false alarms.

Based on the results obtained, the system shows robust performance to the use of the two possible therapies CSII (15 subjects) and MDI(2 subjects).

In addition, we tested some of the models presented in the literature and other well-established classification methods, proving the superiority of the layered meta-learning based on multi-expert systems. Tests on the private dataset suggest that the performance of the meta-learning approach can considerably improve by training only the meta-learner with a small amount of data from a different cohort of patients, without the need to train the entire model from scratch. Tests performed on an edge device confirm the real-life feasibility of the proposed approach.

Plus, while other works (e.g., [17]) proposed to adapt their systems to new patients (personalized fine-tuning approach), our model does not require any adaptation phase, being ready to be used by new users as-is. The main advantage of our (population-based) approach is to not require (several weeks of) new acquisitions, which, in a supervised classification task as in our case, would also require the involvement of physicians for labeling the data, which is mostly practically infeasible.

DATA AVAILABILITY

The Ohio T1DM database is publicly available[1] upon request via a Data Use Agreement. The UCBM dataset is available upon request by contacting the authors of this work.

AUTHORS' CONTRIBUTION

Federico D'Antoni: Conceptualization, methodology, software, validation, formal analysis, data curation, writing-original draft, visualization, and supervision; Lorenzo Petrosino: Methodology, software, validation, formal analysis, and writing-original draft; Alessandro Marchetti: Methodology, software, validation, formal analysis, and writing-original draft; Luca Bacco: Methodology, software, validation, visualization, writing-original draft, supervision; Silvia Pieralice: Data curation, validation, and writing-original draft; Luca Vollero: Methodology, writing-review and editing, and supervision; Paolo Pozzilli: Data curation and writing-review and editing; Vincenzo Piemonte: Writing-review and editing and funding acquisition; and Mario Merone: Conceptualization, methodology, writing-original draft, validation, supervision, and project administration.

REFERENCES

[1] J. A. Bluestone, K. Herold, and G. Eisenbarth, "Genetics, pathogenesis and clinical interventions in type 1 diabetes," *Nature*, vol. 464, no. 7293, pp. 1293–1300, Apr. 2010.

[2] I. M. Stratton, "Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): Prospective observational study," *BMJ*, vol. 321, no. 7258, pp. 405–412, Aug. 2000.

[3] T. Danne et al., "International consensus on use of continuous glucose monitoring," *Diabetes care*, vol. 40, no. 12, pp. 1631–1640, 2017.

[4] E. Cengiz and W. V. Tamborlane, "A tale of two compartments: Interstitial versus blood glucose monitoring," *Diabetes Technol. Therapeutics*, vol. 11, no. S1, pp. S-11–S-16, Jun. 2009.

[5] *Introduction: Standards of Medical Care in Diabetes—2022*, Amer. Diabetes Assoc., Arlington County, VA, USA, 2022.

[6] *FDA Expands Indication for Continuous Glucose Monitoring System, First to Replace Fingerstick Testing for Diabetes Treatment Decisions*, US Food, Drug Admin., Silver Spring, MD, USA, Dec. 2016.

[7] L. Heinemann, M. Schoemaker, G. Schmelzeisen-Redecker, R. Hinzmann, A. Kassab, G. Freckmann, F. Reiterer, and L. Del Re, "Benefits and limitations of MARD as a performance parameter for continuous glucose monitoring in the interstitial space," *J. Diabetes Sci. Technol.*, vol. 14, no. 1, pp. 135–150, Jan. 2020.

[1] http://smarthealth.cs.ohio.edu/OhioT1DM-dataset.html

[8] S. Pleus, A. Stuhr, M. Link, C. Haug, and G. Freckmann, "Variation of mean absolute relative differences of continuous glucose monitoring systems throughout the day," *J. Diabetes Sci. Technol.*, vol. 16, no. 3, pp. 649–658, May 2022.

[9] Y. Nakagawa, Y. Hirota, A. Yamamoto, T. Takayoshi, T. Takeuchi, T. Hamaguchi, A. Matsuoka, K. Sakaguchi, and W. Ogawa, "Accuracy of a professional continuous glucose monitoring device in individuals with type 2 diabetes mellitus," *Kobe J. Med. Sci.*, 68, no. 1, pp. E5–E10, 2022.

[10] O. Didyuk, N. Econom, A. Guardia, K. Livingston, and U. Klueh, "Continuous glucose monitoring devices: Past, present, and future focus on the history and evolution of technological innovation," *J. Diabetes Sci. Technol.*, vol. 15, no. 3, pp. 676–683, May 2021.

[11] W. H. Polonsky, D. Hessler, K. J. Ruedy, and R. W. Beck, "The impact of continuous glucose monitoring on markers of quality of life in adults with type 1 diabetes: Further findings from the DIAMOND randomized clinical trial," *Diabetes Care*, vol. 40, no. 6, pp. 736–741, Jun. 2017.

[12] C. G. Parkin, C. Graham, and J. Smolskis, "Continuous glucose monitoring use in type 1 diabetes: Longitudinal analysis demonstrates meaningful improvements in HbA1c and reductions in health care utilization," *J. Diabetes Sci. Technol.*, vol. 11, no. 3, pp. 522–528, May 2017.

[13] O. Diouri, M. Cigler, M. Vettoretti, J. K. Mader, P. Choudhary, E. Renard, and H. Consortium, "Hypoglycaemia detection and prediction techniques: A systematic review on the latest developments," *Diabetes/Metabolism Res. Rev.*, vol. 37, no. 7, Oct. 2021, Art. no. e3449.

[14] C. Marling and R. Bunescu, "The OhioT1DM dataset for blood glucose level prediction," in *Proc. 3rd Int. Workshop Knowl. Discovery Healthcare Data IJCAI-ECAI*, 2018, pp. 60–63.

[15] C. Marling and R. Bunescu, "The OhioT1DM dataset for blood glucose level prediction: Update 2020," *5th Int. Workshop Knowl. Discovery Healthcare Data (ECAI)*, 2020, p. 71.

[16] S. Oviedo, J. Vehí, R. Calm, and J. Armengol, "A review of personalized blood glucose prediction strategies for T1DM patients," *Int. J. Numer. Methods Biomed. Eng.*, vol. 33, no. 6, Jun. 2017, Art. no. e2833.

[17] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Personalized blood glucose prediction for type 1 diabetes using evidential deep learning and meta-learning," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 1, pp. 193–204, Jan. 2023.

[18] M. Gadaleta, A. Facchinetti, E. Grisan, and M. Rossi, "Prediction of adverse glycemic events from continuous glucose monitoring signal," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 650–659, Mar. 2019.

[19] M. Yang, D. Dave, M. Erraguntla, G. L. Cote, and R. Gutierrez-Osuna, "Joint hypoglycemia prediction and glucose forecasting via deep multitask learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 1136–1140.

[20] W. Seo, Y.-B. Lee, S. Lee, S.-M. Jin, and S.-M. Park, "A machine-learning approach to predict postprandial hypoglycemia," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, p. 210, Dec. 2019.

[21] D. Dave, D. J. DeSalvo, B. Haridas, S. McKay, A. Shenoy, C. J. Koh, M. Lawley, and M. Erraguntla, "Feature-based machine learning model for real-time hypoglycemia prediction," *J. Diabetes Sci. Technol.*, vol. 15, no. 4, pp. 842–855, 2021.

[22] S. L. Cichosz, J. Frystyk, O. K. Hejlesen, L. Tarnow, and J. Fleischer, "A novel algorithm for prediction and detection of hypoglycemia based on continuous glucose monitoring and heart rate variability in patients with type 1 diabetes," *J. Diabetes Sci. Technol.*, vol. 8, no. 4, pp. 731–737, Jul. 2014.

[23] O. Mujahid, I. Contreras, and J. Vehi, "Machine learning techniques for hypoglycemia prediction: Trends and challenges," *Sensors*, vol. 21, no. 2, p. 546, Jan. 2021.

[24] M. H. Jensen, C. Dethlefsen, P. Vestergaard, and O. Hejlesen, "Prediction of nocturnal hypoglycemia from continuous glucose monitoring data in people with type 1 diabetes: A proof-of-concept study," *J. Diabetes Sci. Technol.*, vol. 14, no. 2, pp. 250–256, Mar. 2020.

[25] V. Felizardo, D. Machado, N. M. Garcia, N. Pombo, and P. Brandao, "Hypoglycaemia prediction models with auto explanation," *IEEE Access*, vol. 10, pp. 57930–57941, 2022.

[26] E. Daskalaki, K. Nørgaard, T. Züger, A. Prountzou, P. Diem, and S. Mougiakakou, "An early warning system for hypoglycemic/hyperglycemic events based on fusion of adaptive prediction models," *J. Diabetes Sci. Technol.*, vol. 7, no. 3, pp. 689–698, May 2013.

[27] G. Cappon, A. Facchinetti, G. Sparacino, P. Georgiou, and P. Herrero, "Classification of postprandial glycemic status with application to insulin dosing in type 1 diabetes—An in silico proof-of-concept," *Sensors*, vol. 19, no. 14, p. 3168, Jul. 2019.

[28] Y. Marcus, R. Eldor, M. Yaron, S. Shaklai, M. Ish-Shalom, G. Shefer, N. Stern, N. Golan, A. Z. Dvir, O. Pele, and M. Gonen, "Improving blood glucose level predictability using machine learning," *Diabetes/Metabolism Res. Rev.*, vol. 36, no. 8, Nov. 2020, Art. no. e3348.

[29] F. Prendin, S. D. Favero, M. Vettoretti, G. Sparacino, and A. Facchinetti, "Forecasting of glucose levels and hypoglycemic events: Head-to-head comparison of linear and nonlinear data-driven algorithms based on continuous glucose monitoring data only," *Sensors*, vol. 21, no. 5, p. 1647, Feb. 2021.

[30] V. Felizardo, N. M. Garcia, N. Pombo, and I. Megdiche, "Data-based algorithms and models using diabetics real data for blood glucose and hypoglycaemia prediction—A systematic literature review," *Artif. Intell. Med.*, vol. 118, Aug. 2021, Art. no. 102120.

[31] H. Peng, "A comprehensive overview and survey of recent advances in meta-learning," 2020, *arXiv:2004.11149*.

[32] G. Sparacino, F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, and C. Cobelli, "Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 5, pp. 931–937, May 2007.

[33] W. P. T. M. van Doorn, Y. D. Foreman, N. C. Schaper, H. H. C. M. Savelberg, A. Koster, C. J. H. van der Kallen, A. Wesselius, M. T. Schram, R. M. A. Henry, P. C. Dagnelie, B. E. de Galan, O. Bekers, C. D. A. Stehouwer, S. J. R. Meex, and M. C. G. J. Brouwers, "Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The Maastricht study," *PLoS ONE*, vol. 16, no. 6, Jun. 2021, Art. no. e0253125.

[34] E. I. Georga, V. C. Protopappas, D. Polyzos, and D. I. Fotiadis, "A predictive model of subcutaneous glucose concentration in type 1 diabetes based on random forests," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2012, pp. 2889–2892.

[35] F. D'Antoni, L. Petrosino, F. Sgarro, A. Pagano, L. Vollero, V. Piemonte, and M. Merone, "Prediction of glucose concentration in children with type 1 diabetes using neural networks: An edge computing application," *Bioengineering*, vol. 9, no. 5, p. 183, Apr. 2022.

[36] N. P. Balakrishnan, G. P. Rangaiah, and L. Samavedham, "Review and analysis of blood glucose (BG) models for type 1 diabetic patients," *Ind. Eng. Chem. Res.*, vol. 50, no. 21, pp. 12041–12066, Nov. 2011.

[37] A. Z. Woldaregay, E. Årsand, T. Botsis, D. Albers, L. Mamykina, and G. Hartvigsen, "Data-driven blood glucose pattern classification and anomalies detection: Machine-learning applications in type 1 diabetes," *J. Med. Internet Res.*, vol. 21, no. 5, May 2019, Art. no. e11030.

[38] R. Bunescu, N. Struble, C. Marling, J. Shubrook, and F. Schwartz, "Blood glucose level prediction using physiological models and support vector regression," in *Proc. 12th Int. Conf. Mach. Learn. Appl.*, vol. 1, 2013, pp. 135–140.

[39] C. Midroni, P. J. Leimbigler, G. Baruah, M. Kolla, A. J. Whitehead, and Y. Fossat, "Predicting glycemia in type 1 diabetes patients: Experiments with XGBoost," in *Proc. 3rd Int. Workshop Knowl. Discovery Healthcare Data IJCAI-ECAI*, 2018, pp. 79–84.

[40] K. Li, C. Liu, T. Zhu, P. Herrero, and P. Georgiou, "GluNet: A deep learning framework for accurate glucose forecasting," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 414–423, Feb. 2020.

[41] Q. Wang, P. Molenaar, S. Harsh, K. Freeman, J. Xie, C. Gold, M. Rovine, and J. Ulbrecht, "Personalized state-space modeling of glucose dynamics for type 1 diabetes using continuously monitored glucose, insulin dose, and meal intake: An extended Kalman filter approach," *J. Diabetes Sci. Technol.*, vol. 8, no. 2, pp. 331–345, Mar. 2014.

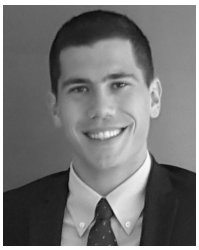
[42] F. D'Antoni, M. Merone, V. Piemonte, G. Iannello, and P. Soda, "Autoregressive time delayed jump neural network for blood glucose levels forecasting," *Knowl.-Based Syst.*, vol. 203, Sep. 2020, Art. no. 106134.

[43] F. Iacono, L. Magni, and C. Toffanin, "Personalized LSTM models for glucose prediction in type 1 diabetes subjects," in *Proc. 30th Medit. Conf. Control Autom. (MED)*, Jun. 2022, pp. 324–329.

[44] D. Kalita and K. B. Mirza, "LS-GRUNet: Glucose forecasting using deep learning for closed-loop diabetes management," in *Proc. IEEE 7th Int. Conf. for Converg. Technol. (I2CT)*, Apr. 2022, pp. 1–6.

[45] C. D. Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, and C. Cobelli, "The UVA/PADOVA type 1 diabetes simulator: New features," *J. Diabetes Sci. Technol.*, vol. 8, no. 1, pp. 26–34, 2014.

[46] O. Mujahid, I. Contreras, A. Beneyto, I. Conget, M. Giménez, and J. Vehi, "Conditional synthesis of blood glucose profiles for T1D patients using deep generative models," *Mathematics*, vol. 10, no. 20, p. 3741, Oct. 2022.
[47] S. Tsichlaki, L. Koumakis, and M. Tsiknakis, "Type 1 diabetes hypoglycemia prediction algorithms: Systematic review," *JMIR Diabetes*, vol. 7, no. 3, Jul. 2022, Art. no. e34699.
[48] B. D. Paoli, F. D'Antoni, M. Merone, S. Pieralice, V. Piemonte, and P. Pozzilli, "Blood glucose level forecasting on type-1-diabetes subjects during physical activity: A comparative analysis of different learning techniques," *Bioengineering*, vol. 8, no. 6, p. 72, May 2021.
[49] A. Fernández, S. García, M. J. D. Jesus, and F. Herrera, "A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets," *Fuzzy Sets Syst.*, vol. 159, no. 18, pp. 2378–2398, 2008.
[50] E. Li, L. Zeng, Z. Zhou, and X. Chen, "Edge AI: On-demand accelerating deep neural network inference via edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 447–457, Jan. 2020.
[51] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019.
[52] W. R. Hersh, M. Helfand, J. Wallace, D. Kraemer, P. Patterson, S. Shapiro, and M. Greenlick, "Clinical outcomes resulting from telemedicine interventions: A systematic review," *BMC Med. Informat. Decis. Making*, vol. 1, no. 1, pp. 1–8, Dec. 2001.
[53] C. Scott Kruse, P. Karem, K. Shifflett, L. Vegi, K. Ravi, and M. Brooks, "Evaluating barriers to adopting telemedicine worldwide: A systematic review," *J. Telemed. Telecare*, vol. 24, no. 1, pp. 4–12, Jan. 2018.
[54] M. Merone, A. Graziosi, V. Lapadula, L. Petrosino, O. d'Angelis, and L. Vollero, "A practical approach to the analysis and optimization of neural networks on embedded systems," *Sensors*, vol. 22, no. 20, p. 7807, Oct. 2022.
[55] T. Zhu, L. Kuang, J. Daniels, P. Herrero, K. Li, and P. Georgiou, "IoMT-enabled real-time blood glucose prediction with deep learning and edge computing," *IEEE Internet Things J.*, early access, Jan. 14, 2022, doi: 10.1109/JIOT.2022.3143375.
[56] J. Li, J. Hao, Q. Feng, X. Sun, and M. Liu, "Optimal selection of heterogeneous ensemble strategies of time series forecasting with multi-objective programming," *Expert Syst. Appl.*, vol. 166, Mar. 2021, Art. no. 114091.
[57] M. Bukhsh, M. S. Ali, M. U. Ashraf, K. Alsubhi, and W. Chen, "An interpretation of long short-term memory recurrent neural network for approximating roots of polynomials," *IEEE Access*, vol. 10, pp. 28194–28205, 2022.
[58] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[59] I. S. Damanik, A. P. Windarto, A. Wanto, Poningsih, S. R. Andani, and W. Saputra, "Decision tree optimization in C4.5 algorithm using genetic algorithm," *J. Phys., Conf. Ser.*, vol. 1255, no. 1, Aug. 2019, Art. no. 012012.
[60] A. Bertachi, C. Viñals, L. Biagi, I. Contreras, J. Vehí, I. Conget, and M. Giménez, "Prediction of nocturnal hypoglycemia in adults with type 1 diabetes under multiple daily injections using continuous glucose monitoring and physical activity monitor," *Sensors*, vol. 20, no. 6, p. 1705, Mar. 2020.
[61] Y.-C. Wang and C.-H. Cheng, "A multiple combined method for rebalancing medical data with class imbalances," *Comput. Biol. Med.*, vol. 134, Jul. 2021, Art. no. 104527.
[62] N. Nnamoko and I. Korkontzelos, "Efficient treatment of outliers and class imbalance for diabetes prediction," *Artif. Intell. Med.*, vol. 104, Apr. 2020, Art. no. 101815.
[63] S. El-Sappagh, F. Ali, S. El-Masri, K. Kim, A. Ali, and K.-S. Kwak, "Mobile health technologies for diabetes mellitus: Current state and future challenges," *IEEE Access*, vol. 7, pp. 21917–21947, 2019.
[64] Z. Mahmoudi, M. H. Jensen, M. D. Johansen, T. F. Christensen, L. Tarnow, J. S. Christiansen, and O. Hejlesen, "Accuracy evaluation of a new real-time continuous glucose monitoring algorithm in hypoglycemia," *Diabetes Technol. Therapeutics*, vol. 16, no. 10, pp. 667–678, Oct. 2014.
[65] A. Güemes, G. Cappon, B. Hernandez, M. Reddy, N. Oliver, P. Georgiou, and P. Herrero, "Predicting quality of overnight glycaemic control in type 1 diabetes using binary classifiers," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 5, pp. 1439–1446, May 2020.
[66] M. Sevil, M. Rashid, I. Hajizadeh, M. Park, L. Quinn, and A. Cinar, "Physical activity and psychological stress detection and assessment of their effects on glucose concentration predictions in diabetes management," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 7, pp. 2251–2260, Jul. 2021.
[67] Diabetes Control and Complications Trial Research Group, "The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus," *New England J. Med.*, vol. 329, no. 14, pp. 977–986, Sep. 1993.
[68] J. M. Lachin, S. Genuth, D. M. Nathan, B. Zinman, and B. N. Rutledge, "Effect of glycemic exposure on the risk of microvascular complications in the diabetes control and complications trial—Revisited," *Diabetes*, vol. 57, no. 4, pp. 995–1001, Apr. 2008.



FEDERICO D'ANTONI received the graduate degree (Hons.) in biomedical engineering from University Campus Bio-Medico di Roma, in 2019, where he is currently pursuing the Ph.D. degree in information technology. His current research interests include machine learning, deep learning, time-series analysis, computer vision, and the application of artificial intelligence to medicine.

LORENZO PETROSINO (Student Member, IEEE) recived the master's degree in biomedical engineering from University Campus Bio-Medico di Roma, in 2020, where he is currently pursuing the Ph.D. degree in science and engineering for humans and the environment. His current research interests include machine and deep learning, edge computing, DLT and its interaction with the IoT, and also decision support system for diabetes.

ALESSANDRO MARCHETTI received the M.Sc. degree (summa cum laude) in physics from University Sapienza, Rome, in 2003. He is currently pursuing the Ph.D. degree in AI for health and life sciences pillar with University Campus Bio-Medico di Roma. Since 2004, he has been a software analyst, a developer, and an IT consultant of information technology, carrying out over 40 projects for important Italian and foreign companies. His research interests include deep learning, reinforcement learning, and AI applications to diabetics and genomics.

LUCA BACCO received the M.Sc. degree in biomedical engineering from University Campus Bio-Medico di Roma, in 2019. He is currently pursuing the Ph.D. degree with the Department of Engineering, Unit of Computer Systems and Bioinformatics, in collaboration with the ItaliaNLP Laboratory, National Research Council, Pisa, Italy, and the Research and Development Laboratory, Webmonks S.r.l., Rome, Italy. His current research interests include (explainable) artificial intelligence and natural language processing solutions for the biomedical and healthcare fields.

SILVIA PIERALICE received the graduate degree in medicine and surgery with Sapienza University, Rome, in 2016, and the degree (cum laude) in endocrinology and metabolic disorders from University Campus Bio-Medico di Roma, in 2019, where she is currently pursuing the Ph.D. degree in integrated biomedical sciences and bioethics with a research project entitled ''Novel technologies to improve diabetes care and management.'' Currently, she works as a Consultant Endocrinologist with University Campus Bio-Medico di Roma. She has acted as study sub investigator in national and international clinical trials and has published in top-ranked peer-reviewed international medical journal. Her main research interests include advanced technologies and novel therapies for type 2 (T2D) and type 1 diabetes (T1D) from basic studies to clinical trials testing novel therapies to prevent beta-cell destruction in patients affected by autoimmune diabetes.

LUCA VOLLERO (Member, IEEE) is an Associate Professor of computer science with University Campus Bio-Medico di Roma. His research interests include of signal processing and digital imaging, AI, distributed systems, and embedded systems optimization. He is a member of the IEEE Computer Society, the IEEE Communication Society, ACM, and SIAM. He is a Co-Founder of the innovative startup Heremos, operating in the field of remote monitoring of frail persons and patients.

PAOLO POZZILLI trained in medicine and specialized in endocrinology and metabolic diseases at Sapienza University, Rome, Italy. He is a Professor of endocrinology and diabetes with University Campus Bio-Medico di Roma (UCBM) and a Professor of clinical research (diabetes) with the Centre of Immunobiology, Blizard Institute, Barts, and the London School of Medicine and Dentistry. He participated as the leader or a regional P.I. in relevant international clinical studies. He is the Deputy Leader of the Action LADA (Latent Autoimmune Diabetes in Adults) Project funded by the EU and the Coordinator of a Joint International Ph.D. Programme in integrated biomedical sciences and bioethics between UCBM and QMUL. His main research interests include pathogenesis and prevention of type 1 diabetes (T1D) from basic studies on the disease's autoimmune mechanisms to clinical trials testing novel therapies. He has received several awards and published more than 500 papers (Pubmed). He is the Editor-in-Chief of *Diabetes Metabolism Research and Reviews*.

VINCENZO PIEMONTE is a Full Professor with the University Campus Bio-Medico di Roma. He is the President of M.Sc. in Chemical Engineering for Sustainable Development and the Head of Laboratory of Chemical-Physics Fundamentals in Chemical Engineering. He is an Adjunct Professor at the Department of Chemical Engineering of the University La Sapienza of Rome. He has about 160 publications on chemical thermodynamics, kinetics, biomedical devices modeling, bioreactors, life cycle assessment (LCA) studies. His research interests include study of transport phenomena in the artificial and bioartificial organs; new biotreatment technology platform for the elimination of toxic pollutants from water; reuse of waste materials for the production of adsorbent materials for water treatment and zeolitic catalysts for the priolosis of waste plastic products; LCA of petroleum-based plastics and bio-based plastics; extraction of valuable substances (polyphenols, tannins) from natural matrices; hydrogen production by membrane reactors; and concentrated solar power plants integrated with membrane steam reforming reactor for the production of hydrogen and hydro-methane.

MARIO MERONE (Member, IEEE) received the Ph.D. degree in bioengineering and bioscience, in 2017. He is an Assistant Professor with the University Campus Bio-Medico di Roma. He has extensive experience in the field of artificial intelligence (machine learning and deep learning for time-series analysis). His research has led to 40 publications, all focused on AI for healthcare. He is currently the scientific leader of several projects involving the application of artificial intelligence in the medical field. He is the Founding Partner of an innovative start-up called BPCOMEDIA S.R.L., focused on technology transfer of research results. He is a member of the IEEE Computer Society, the IEEE Computational Life Sciences, and the IEEE Sensors Council.

Open Access funding provided by 'Università ''Campus Bio-Medico'' di Roma' within the CRUI CARE Agreement