IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 19, 2019, accepted May 20, 2019, date of publication May 27, 2019, date of current version June 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2919184

A Multi-Patient Data-Driven Approach to Blood Glucose Prediction

ALESSANDRO ALIBERTI[1], (Student Member, IEEE), IRENE PUPILLO[1], STEFANO TERNA[2], ENRICO MACII[3], (Fellow, IEEE), SANTA DI CATALDO[1], (Member, IEEE), EDOARDO PATTI[1], (Member, IEEE), AND ANDREA ACQUAVIVA[3], (Member, IEEE)

[1]Department of Control and Computer Engineering, Politecnico di Torino, 10129 Turin, Italy
[2]TomorrowData, 10100 Turin, Italy
[3]Interuniversity Department of Regional and Urban Studies and Planning, Politecnico di Torino, 10129 Turin, Italy

Corresponding author: Alessandro Aliberti (alessandro.aliberti@polito.it)

ABSTRACT Continuous glucose monitoring systems (CGMSs) allow measuring the blood glycaemic value of a diabetic patient at a high sampling rate, producing a considerable amount of data. These data can be effectively used by machine learning techniques to infer future values of the glycaemic concentration, allowing the early prevention of dangerous hyperglycaemic or hypoglycaemic states and better optimization of the diabetic treatment. Most of the approaches in the literature learn a prediction model from the past samples of the same patient, which needs extensive calibrations and limits the usability of the system. In this paper, we investigate the prediction models trained on glucose signals of a large and heterogeneous cohort of patients and then applied to infer future glucose-level values on a completely new patient. To achieve this purpose, we designed and compared two different types of solutions that were proved successful in many time-series prediction problems based respectively, on non-linear autoregressive (NAR) neural network and on long short-term memory (LSTM) networks. These solutions were experimentally compared with three literature approaches, respectively, based on feed-forward neural networks (FNNs), autoregressive (AR) models, and recurrent neural networks (RNN). While the NAR obtained good prediction accuracy only for short-term predictions (i.e., with prediction horizon within 30 min), the LSTM obtained extremely good performance both for short- and long-term glucose-level inference (60 min and more), overcoming all the other methods in terms of correlation between the measured and the predicted glucose signal and in terms of clinical outcome.

INDEX TERMS Continuous glucose monitoring, diabetes, non-linear autoregressive neural network, long short-term memory (LSTM) network, time-series analysis, machine learning.

I. INTRODUCTION

The human body, through the appropriate organs, breaks down carbohydrates into glucose. Insulin, a peptide hormone produced by beta cells of the pancreatic islets, plays a key role in this process, regulating the way the cells absorb glucose and use it as an energy source [1]. Defects in either insulin secretion or insulin sensitivity (or both) lead to diabetes, that is a chronic disease characterized by high glucose levels in the blood (i.e. hyperglycaemia) [2]. This is a severe condition, that is known to at least double a person's risk of early death.

Diabetes can be categorized into three main categories. Type I diabetes, that affects about 10% of the diabetic patients, usually develops since childhood due to a loss of beta cells, that are mistakenly attacked and killed by the immune system [3], [4], resulting in a loss of insulin. Type II diabetes occurs when the body does not make a proper use of the released insulin (i.e. insulin insensitivity) or does not produce enough insulin [5]. This is the most common form of diabetes (about 90% of the diabetic population [3]) and usually develops in adulthood. Gestational diabetes, a temporary condition occurring only during pregnancy, affects between 3 and 20% of mothers-to-be, with increased risks of developing future chronic diabetes for both mother and child [6].

Diabetes in all its forms is among the most widely diffused chronic disorders worldwide, with ever-increasing diffusion trends in both women and men. Hence, there are growing efforts directed towards the development of therapies as well

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Tucci.

as of better ways of keeping the effects of this disease under control. In this work, we focus especially on Type I diabetes. Nonetheless, our study can be easily generalized to the other forms of the pathology.

In a normal subject, glycaemia typically oscillates between $70 - 100$ mg/dl in a fasting state and does not exceed 140 mg/dl after a meal. Conversely, diabetic subjects have fasting values higher than 126 mg/dl, and may encounter either *hyperglycaemia* (when the glycaemic value in the blood exceeds $180 - 200$ mg/dl) or *hypoglycaemia* (when the glycaemic value is much lower than 70 mg/dl), that are both life-endangering situations. Such events need to be timely detected, in order to take all possible countermeasures to save the patient's life.

Traditionally, diabetic patients are monitored by measuring their glycaemic value multiple times in a day (four or more, for a Type I diabetes) using a fingerstick blood glucose meter, often accompanied by fixed insulin infusions. This type of monitoring has a major disadvantage, in that it cannot detect fluctuations of the glycaemia that may be caused by intense physical activity, sudden emotional stress or food assumption. As a result, insulin injections are often over- or under-dosage with respect to the actual need [7].

Today, thanks to the widespread use of increasingly intelligent and low-cost technological devices, the medical sector is moving more and more towards the concept of *smart healthcare* [8]. In this scenario, diabetic patients, especially Type I, are subjected to constant monitoring and appropriate and timely insulin treatment, by means of Continuous Glucose Monitoring Systems (CGMS) [9]. Using sensors applied to the skin, CGMS are able to measure the glycaemic value of a subject at a rate of up to one sample per minute. This generates a considerable amount of data that can be either stored or sent to a processing system, and used to infer the future values of glycaemic concentration within different prediction horizons, with a two-fold benefit: i) better prevention of potentially dangerous hyperglycaemic or hypoglycaemic states and ii) optimization of the insulin dose that needs to be injected [10]. On top of that, the patient can be subjected to a continuous remote monitoring by the primary care physicians, triggering automatic alert mechanisms and, whenever needed, faster hospitalization procedures [11], [12].

Since the introduction of CGMS, literature has proposed several approaches for short-time glucose prediction, that can be broadly categorized into two main groups: i) approaches based on a priori physiological models, that try to reproduce the metabolic response of a patient by means of equations that mathematically describe glucose kinetics [13], [14]; ii) data-driven approaches [15], [16], that infer the future values of glucose concentration by applying machine learning techniques trained on real glycaemic data (see [17] for a review of some recent methods). As they do not depend on fixed parameters, machine learning techniques promise higher flexibility and generalization capability of a fixed physiological model, especially in presence of unpredictable variability of glucose kinetics due to either internal (e.g. different device calibrations) or external factors (e.g. physical activities, food intake, sudden stress, etc.). Hence, thanks to the higher availability of data, the data-driven approaches are generally preferred in the last few years.

In this paper, we propose a blood glucose prediction system exploiting the data-driven approach.

In the following, we first review the literature of data-driven glucose prediction, highlighting the limitations of the available techniques. Then, we introduce the main contributions of our work.

A. RELATED WORK

The idea of predicting the future trend of glucose level by using its past values as the only input was first exploited in [18] and then refined in most recent years, taking advantage of more powerful and accurate data recording devices and more sophisticated machine learning models. The most frequent approaches provided by literature are either based on time-series autoregressive models (AR) or artificial neural networks (ANN) [17]. In [19], a first-order AR model is proposed and compared with a first-order polynomial model. In the same years, a method based on Kalman filtering is used to predict hypoglycaemic events [20] based on glucose monitoring signal. Nonetheless, these methods have still significant prediction errors and a very limited forecasting window (maximum 45 and 30 min, respectively). In [21], [22], ANN approaches were used to predict glycaemia up to 3 hours. Nonetheless, the accuracy of the prediction considerably decreases when the prediction horizon increases. Better accuracy values, albeit on different test sets, were shown by the most recent works, either based on ANN or on support vector regression techniques [23]–[26]. A common trait of these works is that they are usually calibrated on individual patients (i.e. the prediction for a certain patient is built on top of the past glucose level signal recorded from the same patient). While this approach has the obvious advantage of creating a personalized model that perfectly fits the characteristics of a specific patient and of a recording device, it has also multiple disadvantages: i) it limits the usability of the device, in that the system cannot be used on a patient until the calibration is fully performed, ii) it limits the generalization capabilities of the system and increases the risks of overfitting. Conversely, learning a model from a heterogeneous set of patients and recording devices increases in principle the robustness of the model to unpredictable and unseen changes of the input signal [27]. On top of that, the device can be used on a new patient immediately, without re-training.

Fewer studies in literature explored the idea of creating a generalizable glucose level prediction model from a multi-patient training cohort. In [28], [29] the authors propose an AR model with fixed coefficients (applying data filtering and Tikhonov regularization [30]) and compare three different configurations: respectively i) models trained on each individual subject, ii) a model trained on different subjects using the same CGMS device, and iii) a model trained on

different subjects using different devices. Their experimental results show comparable prediction errors for the three scenarios on a forecasting horizon of 30 min. Further developments of the same idea are presented by [31], this time using model based on feed-forward ANN, and by [32], using a recurrent neural network (RNN). Nonetheless, the forecasting accuracy obtained by these works is still modest, and the data used for the training is poor both in terms of number and type of patients (e.g. age categories), which intrinsically limits their generalization capability.

While the most consolidated works are generally based on shallow neural networks, few recent studies started proposing deep learning techniques [16], [33], [34] (e.g. Convolutional Neural Networks). Nonetheless, these methods are typically very demanding, both in terms of training data and computational resources. Hence, the proposed predictors suffer limitations due to the lack of annotated CGM signals used for training the networks, both in terms of number and type of patients considered in the analysis, as well as type of recording devices.

B. CONTRIBUTIONS

In this work, we continue on the path of learning a model from a multi-patient dataset and using this model to predict the future glucose level values of a new patient. By doing so, we try to improve the previous works in two ways: i) by refining the formulation of the neural network used for the prediction; and ii) by considerably enlarging the dataset used for learning the model. The aim is to improve the prediction accuracy, possibly on a much larger forecasting horizon, and to increase the generalization and robustness of the model.

As of point i), we designed and compared two different solutions, that were successfully applied to other time-series forecasting problems. The first solution exploits a Non-Linear Autoregressive Neural Network (NAR). This model extends and refines traditional linear AR architectures, in that it is not intrinsically limited by the assumption of linearity, and overcomes the stability problems of past formulations [35]. The second solution exploits a Long Short-Term Memory (LSTM) network, that is generally acknowledged as one of the best for time series prediction, thanks to its versatility and flexibility [36], [37]. This model is able to overcome the well-known problems of exploding and vanishing gradient that typically affect traditional RNN architectures, and to maintain long-term information over time [38], [39]. While, to the best of our knowledge, NAR was never applied before to the problem of glucose profile prediction, few recent works exploited LSTM, eventually embedded into deep learning frameworks [33], [34], [40]. These works show promising results of LSTM model compared to other approaches, but they are still limited by lack of training data.

As of point ii), we trained and tested our new solutions on a very large set of CGMS signals, with unprecedented variability both in terms of type of subjects and recording devices. This choice stems from the consideration that a higher variability of the training set is known to improve the generalization capabilities of the prediction model and to reduce the risks of over-fitting.

To evaluate the forecasting accuracy and robustness of our solutions, we assessed our results in comparison with other multi-patient techniques, exploiting traditional AR, feed-forward ANN and RNN formulations, respectively. For a fair comparison, we re-implemented the architectures proposed by previous works [29], [31], [32] and performed experiments on the same dataset as our proposed solutions, both for training and testing purposes.

II. DATASET

In the following, we describe in details the dataset that we used to train and test our prediction system.

As anticipated in Section I, our aim is a generalizable CGMS, that learns a prediction model from a fixed set of subjects and then is able to predict glucose level values on a new patient without needing any re-calibration. To do so, the system needs to be trained on a large set of CMGS signals, possibly representing a very wide range of possible outcomes. This inherently reduces the risks of learning a model that is too simple or unfit to deal with unseen data. More specifically, the training set should represent wide variations of glucose dynamics and reflect differences due to either different subject categories (e.g. adults and children, female and male, etc.) or different recording systems. On top of that, to ensure a good representation of glucose dynamics the training samples should be acquired on a continuous basis for a sufficiently long monitoring period, with a rate that is high enough to represent glycaemic fluctuations. Supported by past literature, we identified 5 min as the minimum sampling rate for the CGMS (lower frequency is acceptable only during sleeping time, when the signal is less subject to fluctuations), two days as the minimum monitoring time per subject and 30 samples as the minimum length of a monitoring sequence for it to be considered significant [41], [42].

Based on the considerations above, we decided to use the RT_CGM dataset [43], that is freely available for research purposes, in anonymized form in order to protect patients' privacy. This dataset includes glycaemia trends of an heterogeneous population of 451 patients affected by Type I diabetes, already randomized. Patients have different ethnic origins and gender (45% male and 55% female, respectively), and belong to three different age categories (respectively, adults > 25, adolescents and young adults $15 - 24$ and children $8 - 14$). The data consist in glucose level samples acquired every 5 min using three different CGMS devices (provided by Abbott Diabetes [44], DexCom [45] and Medtronic [46], respectively). On top of that, patients take insulin in two different ways, either by injections or using a micropump delivery system.

The original dataset was pre-processed to make it consistent for our analysis. More specifically, we removed sequences with too many gaps as well as sequences with less than 30 consecutive samples, due to device calibration or measurement errors. Then, we randomly split the resulting

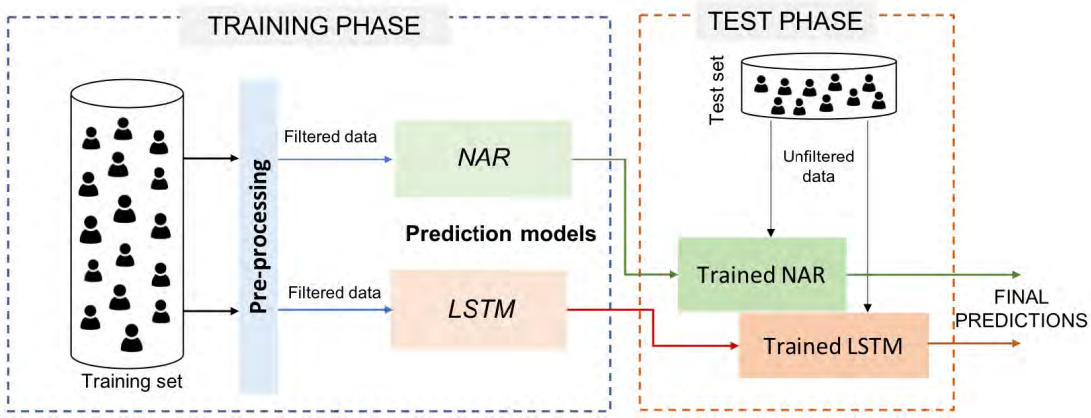


FIGURE 1. Main phases of the study.

data into two sub-sets for training and testing purposes, containing about 70% and 30% of the initial samples, respectively. The full characterization of these two sets is reported in Table 1. The training and test sets are completely independent in terms of patients (i.e. data from a certain patient can be either in training or on the test set).

TABLE 1. Dataset characterization.

	Age	# of patients	# of samples	# of hyperglyc. samples	# of hypoglyc. samples
Training set	>25	95	304450	142410	24827
	15-24	97			
	8-14	124			
Test set	>25	70	94128	34311	8253
	15-24	46			
	8-14	19			

The training set was used to train the prediction models and optimise their parameters, using a portion of the samples as validation set, whilst the test set was solely used for assessing the final prediction performance.

III. METHODOLOGY

In Figure 1, we show the main phases of our study. In the so-called training phase, training samples first undergo a pre-processing filtering step. The filtered data are used as input to build a prediction model, either based on NAR or LSTM neural networks. In the test phase, new unseen and unfiltered data are fed into the trained models to obtain the final predictions. All the implementations were done in Matlab (NNSYSID toolbox) and Python, using TensorFlow as back-end and exploiting Scikit-learn, pyrenn and pandas libraries.

In the following, we describe in details the main phases of the system.

A. PRE-PROCESSING

Generally, CGMS sensors introduce some amount of noise during signal sampling [42], [47], that needs to be attenuated or removed before the data can be used either to train a prediction system or to infer new values from unseen data [48].

As it is well known from literature, over-smoothing the glucose time series data translates into a higher risk of missing out hypoglycaemia and hyperglycaemia events. On the other hand, processing the data completely unfiltered might lead to false alarms caused by few unexpected sensing errors [48]. Hence, data filtering should be performed having as objective a reasonable compromise between these two opposite effects.

As a denoising pre-processing step, we applied to the training data Tikhonov regularization [49], that is widely used in time series analysis and in glucose level prediction system. As demonstrated by [28], [29], this method allows to obtain a filtered version of the signal without introducing significant delays into the time series.

Tikhonov is a filtering technique, where the N-dimensional filtered signal $\hat{y}$ ($N = 30$ in our case) is obtained as follows:

$$\hat{y} = U_d \omega, \tag{1}$$

where ω is the N-dimensional first derivative of the input signal and U_d is the $N \times N$ integral operator matrix:

$$U_d = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 1 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & \cdots & 1 & 0 & 0 \\ 1 & 1 & 1 & \cdots & 1 & 1 & 0 \\ 1 & 1 & 1 & \cdots & 1 & 1 & 1 \end{bmatrix} \tag{2}$$

To estimate ω, we need to minimize the functional $f(\omega)$:

$$f(\omega) = \| y - U_d \omega \|^2 + \lambda_d^2 \| L_d \omega \|^2, \tag{3}$$

where y is the N-dimensional input glucose time series, L_d is the second derivative operator matrix, as in [50], and λ_d is a regularization parameter, set to 3000 following the implementation of [28].

Then, we formulated Tikhonov regularization as an explicit functional minimization problem, as follows:

$$\hat{y} = (U_d^T U_d + \lambda_d L_d^T L_d)^{-1} U_d^T y \tag{4}$$

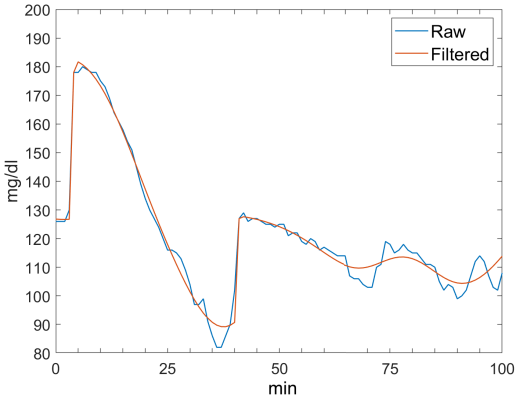


FIGURE 2. Effects of Tikhonov regularization on glucose level data: Example.

In Figure 2, we show an example of the effect of Tikhonov regularization applied to our CGMS training data. As it can be easily gathered from the plot, the filter attenuates sudden spikes in the signal, without altering the trend and without introducing delays, as already demonstrated by [28], [29].

B. PREDICTION MODEL BUILDING

In the following, we present our glucose prediction models based on NAR and LSTM neural networks, respectively.

More specifically, for both the solutions we describe i) how we selected and identified the final architecture and related parameters of the prediction models and ii) how we optimized such parameters in order to boost the performance and robustness of the models, preventing the risk of over-fitting.

1) NON-LINEAR AUTOREGRESSIVE (NAR) NEURAL NETWORK

As anticipated in Section I, Non-Linear Autoregressive (NAR) extends traditional linear autoregressive model [51], in that it is completely distribution-free. Hence, it can be applied even to time series with intrinsic non-linearities, such as sudden spikes and fleeting transient periods [52].

A NAR model computes the value of a signal y at time t using n past values of y as regressors (also called feedback delays), as follows:

$$y(t) = f(y(t-1), y(t-2), ..., y(t-n)) + e(t), \qquad (5)$$

where f is an unknown non-linear function and $e(t)$ is the model approximation error at the time t.

Function $f(\cdot)$ is computed by optimizing a multi-layered neural network, whose topology is depicted in Figure 3(a).

At the time t, the neural network is fed with the n past values of the signal y. Such inputs are transferred through multiple layers of neurons, where each neuron is a simple computational unit characterized by a set of weights W (one per each input connection j), a bias b and an activation function h. Hence, the output of a neuron i is computed as follows:

$$out_i = h_i \left(\sum_j w_{i,j} \cdot in_{i,j} + b_i \right) \qquad (6)$$

where the optimal values of $w_{i,j}$ and b_i are computed by back-propagation on the training set [52].

As anticipated in Section II, the minimum number of consecutive samples in our dataset is 30. Keeping this in mind, we initially designed a very simple fully-connected NAR as the one represented in Figure 3, with the following topology:

1) one input layer, with 30 units;
2) one hidden layer, again with 30 units;
3) one output layer, with one unit.

As reported in Figure 3(b), we used hyperbolic tangent activation functions for the hidden units and a linear activation function for the output unit.

As a learning paradigm for the NAR network, we implemented a Levenberg-Marquardt backpropagation procedure (LMBP), which is widely applied to NAR models. This technique approximates second-order derivatives leveraging a *trust region* approach [52], with no need to compute the Hessian matrix. This helps reducing the training speed compared to traditional backpropagation techniques.

As it is widely known, models with too many parameters (i.e. too many neurons and connections) are less fit for hardware implementation and may easily lead to over-fitting. To overcome this problem, we adopted a two-steps design procedure, as follows:

1) we applied an automated optimization strategy based on Lipschitz method [53] to possibly reduce the number of regressors of our model (i.e. the number of past glucose values to be given as input to the system).
2) we performed an automated pruning of the initial fully-connected structure, based on Optimal Brain Surgeon (OBS) method [54].

This two-steps procedure allowed us to design a more compact version of the network compared to the one of Figure 3(a), obtaining a non-redundant NAR model with optimal number of inputs and computational units.

As of point 1), we applied Lipschitz methodology for determining the optimum *lag-space*, that in our case is the number of delayed glucose signals to be used as regressors [53]. This method is successfully used for the analysis of Input-Output Models orders in Non-linear Dynamic Systems in many applications. As initially proposed by He and Asada [55], a reliable decision on the optimal order n of a non-linear model characterized by training input-output pairs $(\boldsymbol{x_i}, y_i)$ can be made based on so-called Lipschitz quotients:

$$q_{i,j}^{(n)} = \frac{|y_i - y_j|}{\|\boldsymbol{x_i} - \boldsymbol{x_j}\|}, \qquad (7)$$

where in our case $\boldsymbol{x_i}$ is a vector of inputs and y_i is the corresponding output of the system, with $i \neq j$ and $i, j = 1..N$ where N is the number of samples in the training set. Hence, the superscript n stands here for the number of regressors of the system.

Lipschitz order index $L^{(n)}$ is defined as the geometric mean of the m largest Lipschitz quotients, as follows:

$$L^{(n)} = \left[\prod_{k=1}^{m} \sqrt{n} \cdot l^{(n)}(k) \right]^{\frac{1}{m}}, \qquad (8)$$

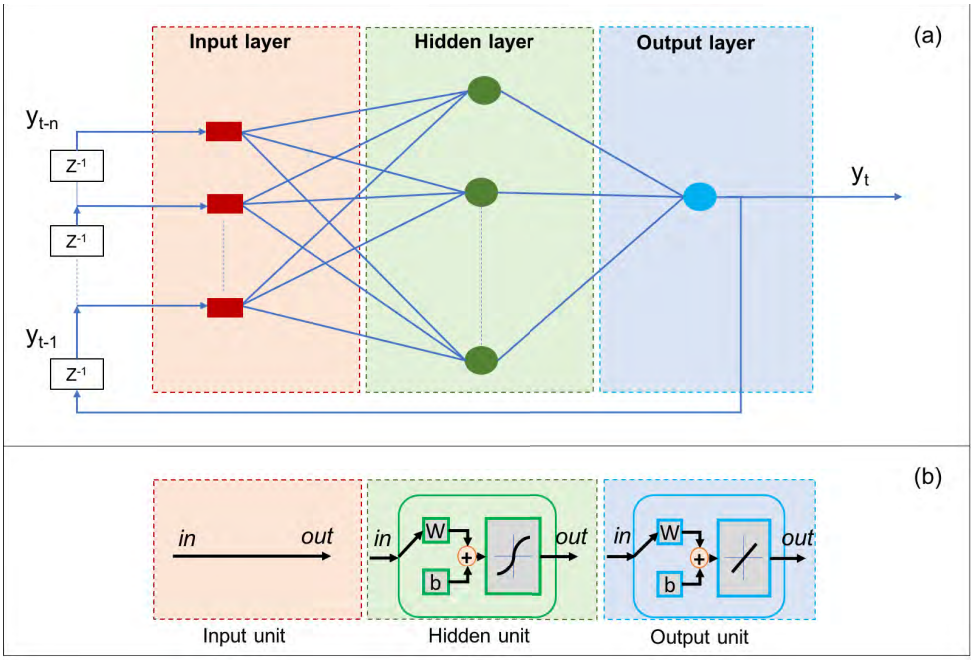


FIGURE 3. Non-linear autoregressive neural network. (a) Network topology. (b) Neuron models of the input, hidden and output layers, respectively.

where m is a positive number recommended to be $\sim 0.01N$ and $l^{(n)}(k)$ is the k-th largest Lipschitz quotient among all $q_{i,j}^{(n)}$.

Finally, as demonstrated by [55], the optimal number of regressors can be found by plotting Lipschitz order index at increasing values of n, in a forward sequential way, and selecting the knee-points of the obtained curve. As reported in Figure 4, by applying this procedure to our training data, we found $n = 8$ and $n = 17$ as the best candidates for the number of regressors of our system.

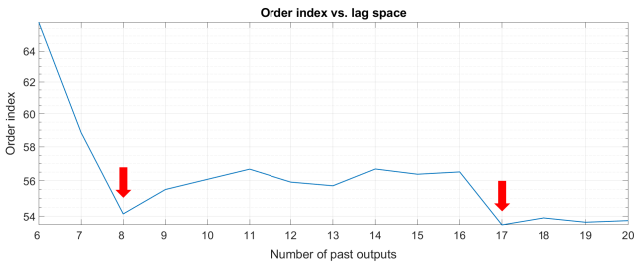


FIGURE 4. Evaluation of order index criterion for different lag-space.

Based on Lipschitz results, we implemented three fully-connected NAR models, respectively with 30 (that is the minimum number of consecutive samples of our dataset, and hence the upper-bound of the input size), 17 and 8 regressors. These three models were trained on the training set described in Section II, using LMBP as the learning algorithm. To assess the goodness of the training and the generalization capabilities of the three models after this first design phase, in the second column of Table 2 we report the normalized sum of squared errors (nSSE) obtained after training the 30, 17 and 8 regressors models, respectively. These values were computed on a portion of the training set solely used for validation and optimization purposes. (i.e. *validation set*). As it can be easily gathered from the table, the model with 8 regressors is the one obtaining the best performance.

As of point 2) of our two-steps design procedure, each of the fully-connected models obtained after Lipschitz underwent automated pruning. The objective of this second design phase is to eliminate redundant connections between the neurons, and hence obtain more efficient and compact models than the initial ones, possibly improving or at least not impacting on their prediction capability. For this purpose, we implemented an Optimal Brain Surgeon (OBS) methodology, as first introduced by [54]. The main idea behind this procedure is to estimate the increase in the training error when deleting weights, leveraging information in the second-order derivatives of the error surface. More specifically, the strategy works towards the minimization of the error variation, leveraging a recursive calculation of the inverse Hessian matrix from the training data to obtain better approximations of the error function (see [54] for details). This strategy has been demonstrated to eliminate more redundant neuron connections than other pruning techniques, yielding to models with improved generalization capabilities [56].

The three pruned models (respectively with 30, 17 and 8 regressors) were trained again with a Levenberg-Marquardt backpropagation procedure (LMBP). The nSSE values obtained after this second round of training are reported in the third column of Table 2.

From the analysis of this table, we can see that OBS pruning further reduced the validation error of the three models.

TABLE 2. Validation error (nSSE) before and after OBS pruning.

NAR Network Type	nSSE before pruning	nSSE after pruning
30 regressors	30.49	25.83
17 regressors	28.60	27.05
8 regressors	26.24	25.89

As it was reasonable to expect, the model that benefited the most from the pruning is the 30 regressors one. On the other hand, the 8 regressors NAR is the one that had the best compromise between generalization error (that is comparable with the one obtained by the 30 regressors after pruning) and simplicity of the model, which guarantees the lowest risks of over-fitting. Based on these considerations, we decided to select the pruned network with 8 regressors as the final NAR model for glucose level predictions. The architecture of this model is schematically represented in Figure 5.

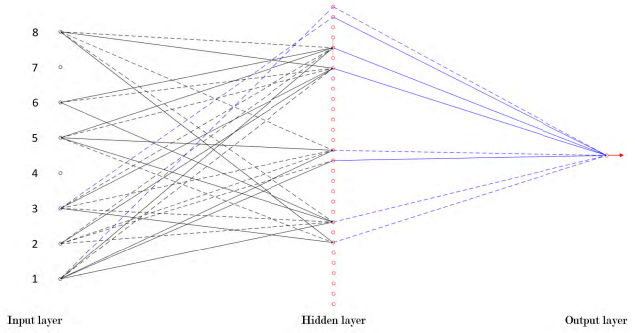


FIGURE 5. NAR final architecture.

2) LONG SHORT-TERM MEMORY NEURAL NETWORK

The Long Short Term Memory networks (LSTM) represent an evolution of the classic Recurrent Neural Networks (RNNs).

Unlike Feed-Forward models, where the information flows just in one direction (from the input to the output) and each layer is characterized by a different set of parameters, RNN models are characterized by multiple layers of recurrent units all sharing the same parameters, with loops allowing the information to propagate back to the same computational units (see the diagram in Figure 6). By doing so, each computational step takes into account not only the current input at time t, but also what was learnt from the previous inputs. Ultimately, this makes it particularly suitable for time-series predictions [57].

A well-known limitation of classic RNN architectures, where the parameters of a large number of layers are learnt by backpropagation, is the instability of long-term predictions due to either vanishing or exploding gradient problems [58]. Such problems arise during the training of a deep network, when the error gradients are propagated back in time to the initial layer, going through continuous matrix multiplications. As the gradients approach the earlier layers, if they have small

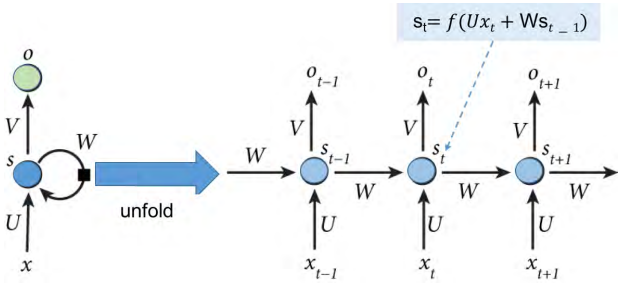


FIGURE 6. Unfolding of a recurrent neural network unit. x_t is the input at time step t, s_t the hidden state, also called *memory* of the network (initialized to 0), f is a non-linear activation function and o_t the output.

values (much lower than 1), they shrink exponentially until they vanish, making it impossible for the model to learn. Likewise, very large gradient values (much greater than 1) get larger and larger and eventually crash the model.

LSTM modules are specifically designed to overcome this limitation [59]. While the structure of the network is fundamentally very similar to RNN, a different function is used to compute the hidden state [60]. In the RNN, repeating modules consist of a single layer, typically with tangential activation function. The memory in LSTMs is instead implemented as *cells* (see Figure 7), where specific gating functions decide whether the information needs to be kept or erased from memory at each time step. The key to this process is the cell state (in the diagram of Figure 7, the green horizontal line), which conveys information to the next cell. The power of either removing or adding information to the cell state is regulated by gates (respectively, input, output and forget gates) consisting of sigmoidal activation functions coupled with pointwise multipliers. Each sigmoid outputs values in a 0 to 1 range, modulating how much of the corresponding signal should be let through (for more details, see [60]). No matter how deep the network is, the LSTM network will be able to remember values that are passed through gates all in 1 state. This makes the LSTM model intrinsically immune to vanishing and exploding gradient.

For our specific problem of glucose level prediction, we designed a LSTM network consisting of a layer of 30 LSTM units and a single output layer (dense), with a number of units equal to the future glucose samples that need to be predicted (i.e. 18, corresponding to a 90 min prediction horizon at a 5 min sampling rate). The architecture of the cells is the same that is depicted in Figure 7, consisting of input, forget and output gates with sigmoidal gating functions.

As we did for the NAR module, before deciding the final architecture of the LSTM model, we investigated the possibility of reducing the number of cells of the network as well as the number of past glucose levels to be used as regressors for the prediction. We run experiments with respectively 30 (our upper bound), 17 and 8 regressors, that represent the knee-points detected applying Lipschitz order index method to our training data (see Figure 4). In Figure 8, we show values of the validation error, in the form of Root Mean Square Error

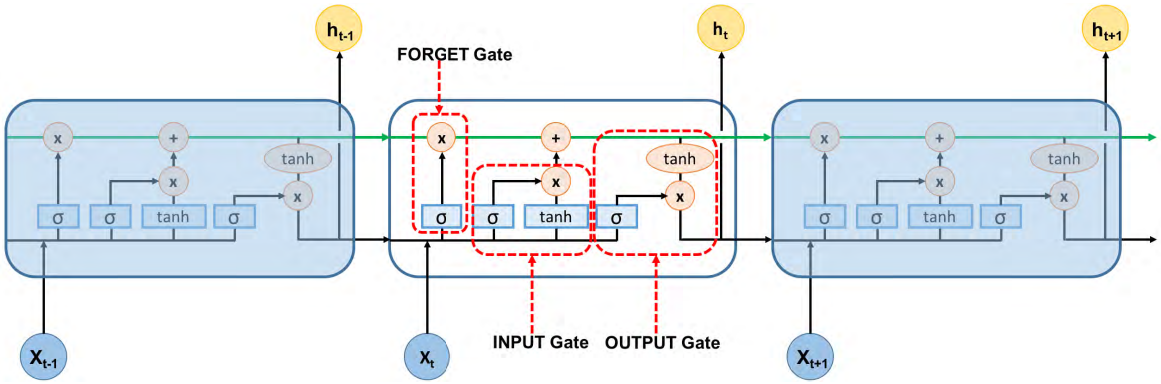


FIGURE 7. LSTM diagram, where x_t is the input and h_t is the output of a cell.

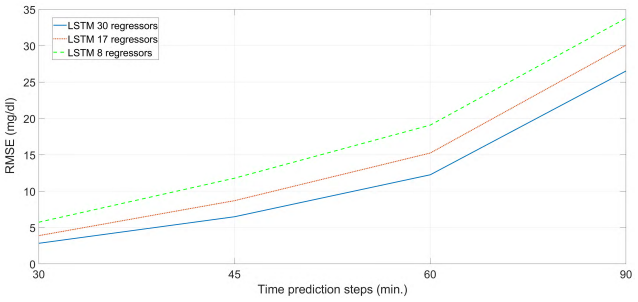


FIGURE 8. Validation error (RMSE) of LSTM models.

on the validation set, obtained by the 30, 17 and 8 regressors models, at increasing time prediction windows (respectively from a minimum of 30 up to a maximum of 90 min). As it is clear from the plot, the 30 regressors model is the one with the best performance. As expected, for all the models the prediction error increases exponentially with the prediction window. The 8 regressors model is consistently the one with the lowest prediction accuracy.

In general, the 30 regressors model is the one that provides the best performance for either short-time and long-time glucose level predictions.

To optimise the hyper-parameters of our network and prevent either under-fitting or over-fitting problems, we implemented a training procedure imposing an initial learning rate equal to 0.001, with dropout. More specifically, at each training stage individual nodes and corresponding links are randomly dropped out of the model, leaving a reduced network [61]. This regularization procedure helps reducing co-dependency of parameters, and hence prevents overfitting. Then, we chose *Adam* (Adaptive moment estimation) as the optimizer. This algorithm leverages adaptive learning rates methods to set individual learning rates per each parameter, combining this feature with the advantages of classical optimization techniques such as stochastic gradient descent and root mean square propagation. This technique was demonstrated to be particularly suitable for non-stationary problems with lot of noise [62].

The training steps for the learning procedure were set to 50000, with a batch size of 100. In order to keep the computational costs of the training under control without having to reduce the size of the training set, we implemented a *mini-batch* training paradigm. More specifically, the training set was split into a number of small sub-sets of size 100, that were used to compute the model error. Each mini-batch was computed in parallel during each iteration. Then, the average error over the mini-batches was used to update the model parameters at each iteration. This method, besides considerably reducing computational time and memory requirements, has been demonstrated to generally improve the model performance [62].

To find a good compromise between a learning rate too low (which may make the training too slow) or too high (which may lead to sub-optimal solutions), we imposed a step decay schedule, that dropped the initial learning rate by a 0.9 factor every 1000 iterations.

After running the learning algorithm on the training set, the so-obtained LSTM model was used as-is to predict glucose level values on an independent test set, with prediction horizons spanning from 30 up to 90 min.

IV. EXPERIMENTAL RESULTS

In the following, we present and discuss the experiments that were performed to assess the prediction models described in Section III. To obtain a thorough evaluation of our methods, our analysis was divided into two main parts:

1) *analytical assessment.* In this part, we assess the validity of the predictions from a regression analysis point of view, by computing a set of metrics that are widely used to quantify the similarity of a discrete time-series with a reference ground truth;
2) *clinical assessment.* In this part, we assess the validity of the predictions from a clinical point of view. To do so, we use metrics that are specifically designed to validate the clinical outcome of blood glucose measurements.

Both the assessments were performed on a test set (fully characterized in Section II) that is completely independent

from the one that was used to design, train and optimise the prediction models.

To evaluate the performance of our models, the results obtained by our solutions were compared with the results obtained by literature techniques. As already anticipated in Section I, the most prominent multi-patient data-driven prediction techniques are based either on autoregressive models (AR), dense feed-forward neural networks (FNN) or standard recurrent neural networks (RNN). Even though most of the works provide a quantitative assessment of these methods, a fair comparison requires that all the predictors are trained and tested on the same data. Hence, we re-implemented three state of the art glucose predictors based respectively on AR, FNN and RNN, strictly following the design reported in the respective publications [29], [31] and [32]. Then, we trained, optimized and tested the three models on our data, with the same procedure that was applied to our proposed solutions.

The experiments were run on a Linux computer equipped with an Intel® Core™ i7-8750H CPU with 2.20 GHz, 6 cores, 12 logical processors and 16,0 GB of installed Physical Memory.

A. ANALYTICAL ASSESSMENT

To evaluate the prediction accuracy of the models, we exploited a number of metrics that are widely used in descriptive statistics and in regression analysis to quantify the similarity between predicted and observed time-series. More specifically we focused on a list of metrics that are more often used by blood glucose level predictions literature [63]:

- *RMSE* is the *Root Mean Square Error*, defined as the standard deviation of the difference between the predicted and the observed values. It is the prediction error index that is most often used in literature.
- R^2 is the *Coefficient of Determination*, defined as square of the correlation (R) between predicted and observed values. Thus, it ranges from 0 (absence of correlation) to 1 (complete correlation).
- *Time lag* is the *Prediction Delay*, defined as minimum time-shift between the predicted and observed signals which provides the highest correlation coefficient between them.
- *MAD* is the *Mean Absolute Difference* between predicted and observed values.
- *FIT* is computed as the ratio of RMSE and the root mean square difference between the observed signal and its mean value, as reported in the following equation:

$$FIT = \left(1 - \frac{\sqrt{\frac{1}{N}\sum (Y - \hat{Y})^2}}{\sqrt{\frac{1}{N}\sum (Y - \bar{Y})^2}}\right) \cdot 100, \qquad (9)$$

where Y and $\hat{Y}$ are respectively the observed and predicted signals and $\bar{Y}$ is the mean value of the observed signal. FIT closer to 100% indicates better prediction accuracy.

Before assessing the prediction accuracy of the models in comparison with literature approaches, we run few preliminary experiments to demonstrate the goodness of the pre-processing stage based on Tikhonov regularization applied to the training data (see Section III-A).

In the plot of Figure 9, we report the results of these experiments. More specifically, we plot the RMSE values obtained by both the NAR and LSTM prediction models on the test dataset, with different prediction horizons (from 30 up to 90 min with steps of 15). The same experiments were performed training the models first on raw data and then on filtered data. In both cases, the testing was performed on the raw unfiltered data, as specified in the diagram of Figure 1.

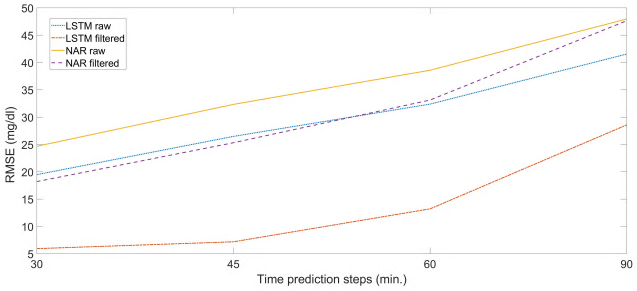


FIGURE 9. Impact of training data filtering on NAR and LSTM models predictions.

As it is clearly visible from the plot, the filtering significantly decreased the prediction error of both the models, with special benefit for LSTM model. On top of that, we can observe that the error trend against prediction horizon was more or less the same with or without filtering.

In Table 3, we report the values obtained running all the prediction models on the same test set, with our proposed NAR and LSTM networks highlighted in grey and light blue, respectively. Each sub-section of the table shows the values of all the figures of merit defined at the beginning of the section. Different columns show values obtained at different prediction horizons, starting from 30 min (short-term prediction) up to 90 min (very long-term prediction), at steps of 15 min. For completeness, we show on the left the performance of the models trained on raw data and on the right the values obtained by the same models trained on filtered data and validated on the unfiltered ones.

From the values of Table 3, we can draw the following considerations:

- All the models benefited from pre-processing the training data with Tikhonov. Hence, from now on we will mainly focus on the analysis of this set of experiments.
- the LSTM model is by far the one that obtained the best prediction performance. This is consistently confirmed by all the figures of merit, for both short-term and long-term predictions. Remarkably, the time-lag observed for this model was zero until a prediction horizon of 60 min. On the contrary, all the other models had a

TABLE 3. Prediction performance indicators for all the tested models.

		NOT FILTERED TRAINING SET prediction horizon (min)				FILTERED TRAINING SET prediction horizon (min)			
		30	**45**	**60**	**90**	**30**	**45**	**60**	**90**
AR	RMSE (mg/dl)	25.96	34.16	41.02	51.66	17.88	22.6	28.31	41.66
	R^2	0.84	0.72	0.60	0.37	0.92	0.88	0.81	0.59
	Time lag (samples)	5.00	8.00	11.00	17.00	3.00	3.00	4.00	9.00
	MAD (%)	9.80	13.90	17.40	22.97	3.64	5.19	8.57	15.64
	FIT (%)	60.05	47.43	36.88	20.52	72.48	65.22	56.44	35.9
FNN	RMSE (mg/dl)	26.75	35.18	42.39	53.95	21.1	29.6	38.51	54.95
	R^2	0.83	0.71	0.57	0.31	0.89	0.79	0.65	0.28
	Time lag (samples)	5.00	8.00	11.00	17.00	4.00	6.00	8.00	14.00
	MAD (%)	10.60	14.98	18.83	25.28	7.1	11.62	16.47	25.51
	FIT (%)	58.84	45.86	34.78	16.99	67.53	54.45	40.74	15.45
RNN	RMSE (mg/dl)	26.16	37.99	48.28	57.07	18.22	22.51	26.78	38.08
	R^2	0.84	0.66	0.45	0.23	0.92	0.88	0.83	0.66
	Time lag (samples)	5.00	8.00	11.00	17.00	3.00	3.00	3.00	3.00
	MAD (%)	9.41	13.45	16.88	22.10	4.02	4.60	5.80	13.06
	FIT (%)	59.75	41.55	25.72	12.19	71.97	65.36	58.80	41.41
NAR	RMSE (mg/dl)	24.66	32.33	38.58	47.96	18.2	25.31	33.12	47.64
	R^2	0.86	0.76	0.66	0.47	0.92	0.85	0.75	0.48
	Time lag (samples)	5.00	8.00	11.00	17.00	3.00	4.00	6.00	10.00
	MAD (%)	10.10	14.21	17.60	22.70	5.96	9.96	14.41	22.57
	FIT (%)	62.66	51.04	41.59	27.37	72.44	61.67	49.84	27.86
LSTM	RMSE (mg/dl)	19.47	26.47	32.38	41.54	5.93	7.18	13.21	28.57
	R^2	0.91	0.83	0.74	0.58	$\simeq$1.00	0.99	0.96	0.80
	Time lag (samples)	3.00	6.00	9.00	15.00	0.00	0.00	1.00	6.00
	MAD (%)	8.71	12.29	15.36	20.21	2.59	3.25	6.13	13.68
	FIT (%)	69.63	58.59	49.43	34.32	90.75	88.79	79.37	55.25

time-lag of 3 samples (15 min) at best for the short-term horizon.

- FNN is the model with the worst performance. On top of that, it is the one presenting the most rapid decrease of prediction accuracy when increasing the prediction horizon. The time lag is especially high, reaching 14 samples for long-term predictions.
- NAR, AR and RNN methods seem to have a comparable behavior, especially for short-term predictions.
- When comparing NAR with AR, the latter obtained slightly better results (e.g. RMSE was 17.88 against 18.2 mg/dl, at 30 min horizon). This difference is even more remarkable at 45 min horizons. On the other hand, if we observe the values obtained with unfiltered training data, we can see the opposite (e.g. 25.96 mg/dl against 24.66 mg/dl obtained by NAR at 30 min). Most reasonably, NAR is more robust to non-linearities in the training data when compared to AR. Nonetheless, if we consider the overall figures of merit, the performance of the two models in this specific application is almost equivalent.

To perform a more in-depth analysis of the results, in Figure 10, we show a plot of the RSME values obtained by all the tested models, this time in %, reporting the prediction horizon in the x-axis. To provide a better interpretation of the obtained results, we highlighted in green the area of the plot within a 20% RMSE range. In absence of clear indications by past literature, we used this value as an indicative threshold of acceptability for RMSE, consistent with the numerical criteria for accuracy described by International Organisation for Standardisation (ISO) [64], [65].

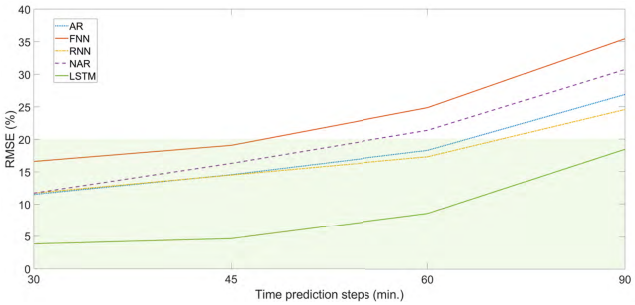


FIGURE 10. Overall RMSE comparison.

As it can be gathered from the plot, RSME consistently increases with the prediction time, more or less with the same slope for all models. The plot confirms the undisputed superiority of LSTM, which maintained RSME values at least 10% better that all the other models, throughout all the tested prediction horizons. On top of that, LSTM is the only one that provided RMSE always within the 20% range, and below 15% up to a 60 min prediction horizon. The behavior of NAR, AR and RNN models do not have significant differences for short-term predictions up to 30 min. The performance of NAR gets worse after that, still maintaining just within the acceptability range up to 60 min prediction horizon. AR and RNN have a similar behavior up until the longest prediction window, with slightly better performance for the AR model.

TABLE 4. Overall Clarke error grid results.

		NOT FILTERED TRAINING SET prediction horizon (min)				FILTERED TRAINING SET prediction horizon (min)			
		30	**45**	**60**	**90**	**30**	**45**	**60**	**90**
AR	A	87.01	76.99	68.55	56.30	97.74	96.33	91.8	71.68
	B	11.4	20.18	27.35	37.1	1.49	2.47	6.4	24.37
	C	0.29	0.50	0.78	1.38	0.25	0.4	0.55	0.99
	D	1.15	2.09	2.99	4.67	0.38	0.59	0.94	2.47
	E	0.15	0.24	0.33	0.57	0.14	0.22	0.3	0.49
FNN	A	85.05	74.14	64.62	50.66	94.41	83.4	70.82	51.62
	B	13.41	23.25	31.47	42.71	4.68	14.94	26.5	42.41
	C	0.27	0.51	0.96	2.19	0.26	0.42	0.72	2.4
	D	1.11	1.84	2.51	3.62	0.5	0.98	1.58	2.69
	E	0.16	0.27	0.44	0.82	0.14	0.25	0.38	0.88
RNN	A	88.06	77.96	69.44	57.67	97.51	96.43	95.14	82.04
	B	10.02	18.41	25.29	34.6	1.65	2.33	3.15	14.3
	C	0.25	0.38	0.57	0.99	0.26	0.44	0.62	0.99
	D	1.54	3.05	4.44	6.39	0.44	0.57	0.76	2.19
	E	0.13	0.21	0.26	0.34	0.14	0.23	0.32	0.48
NAR	A	84.86	72.96	64.26	52.92	95.03	84.44	73.06	54.36
	B	11.41	20.61	28.52	38.51	3.07	9.97	20.22	37.34
	C	0.18	0.32	0.46	0.72	0.16	0.28	0.44	1.02
	D	3.49	6.04	6.66	7.64	1.64	5.52	6.17	7.04
	E	0.07	0.07	0.11	0.21	0.09	0.09	0.11	0.24
LSTM	A	88.55	78.06	68.91	56.43	99.73	99.56	95.7	73.53
	B	9.88	18.73	26.28	36.73	0.21	0.36	3.76	22.7
	C	0.04	0.1	0.24	0.55	0.00	0.00	0.00	0.10
	D	1.52	3.1	4.55	6.2	0.06	0.09	0.54	3.66
	E	<0.01	0.01	0.02	0.09	0.00	0.00	0.00	0.01

As a qualitative confirmation of the good prediction accuracy of LSTM, in the following we show an example of glucose predictions performed by this model, respectively short-term (30 min) in Figure 11(a) and long-term (60 min) in Figure 11(b). In both the plots, the predicted signal is shown in blue, and the corresponding measured signal in red.

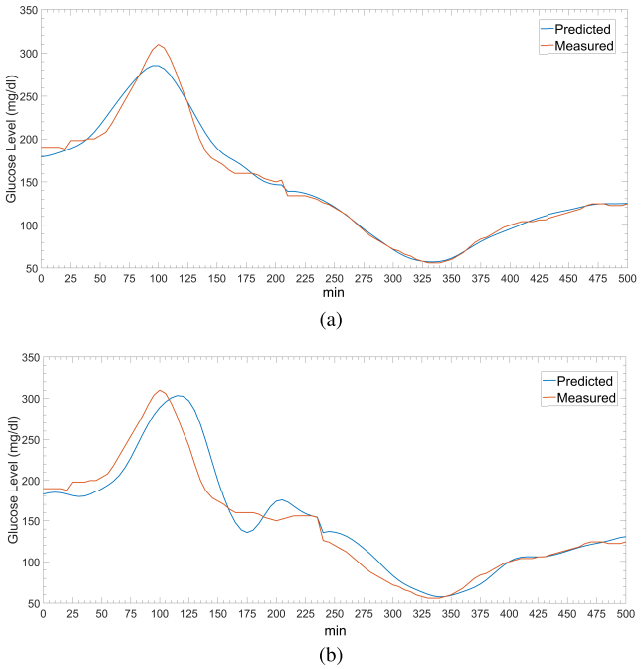


FIGURE 11. LSTM predictions at different forecasting horizons. (a) LSTM prediction (30 min horizon). (b) LSTM prediction (60 min horizon).

B. CLINICAL ASSESSMENT

Even though the metrics identified in Section IV-A are essential to understand the performance and prediction accuracy of the different models from a regression analysis point of view, they are not able to identify the most significant outliers, and they do not provide any information about the clinical impact of the prediction errors and of their consequences on medical treatment decisions. Then, to provide a more thorough picture of the models performance, we integrated our assessment with Clarke Error Grid analysis (EGA) [66].

EGA is a semi-quantitative methodology introduced in 1987 that is nowadays the most widely accepted tool for the analysis of clinical accuracy of blood glucose estimations. It provides a clinical interpretation of the mapping between predicted and measured blood glucose levels, that can be represented in a scatterplot with five main regions (see Figure 12):

- A: values within 20% of the reference;
- B: values that, in spite of being outside 20% of the reference, do not lead to inappropriate treatment of the patient;
- C: values leading to inappropriate treatment, but without dangerous consequences for the patient;
- D: values leading to potentially dangerous failure to detect hypoglycaemic or hyperglycaemic events;

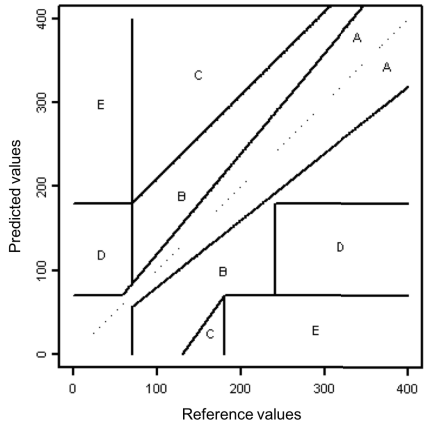


FIGURE 12. Clarke error grid analysis: Reference regions mapping.

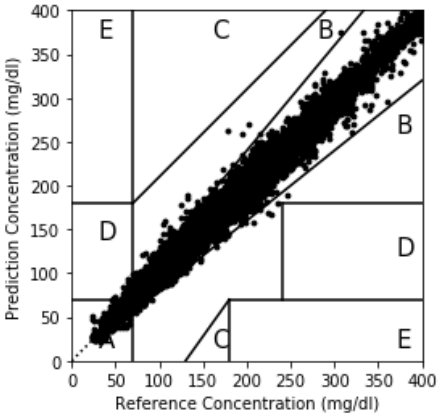
E: values leading to treat hypoglycaemia instead of hyperglycaemia and vice-versa.

Hence, zones A and B are the ones with full clinical acceptability. D and E, on the other hand, are the zones where prediction errors are mostly dangerous for a patient [67].

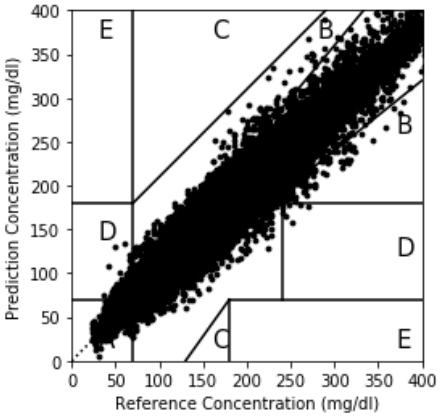
The overall results of our Clarke Error Grid analysis are reported in Table 4. More specifically, in each sub-section of this table we report the EGA results of a tested model, with our designed NAR and LSTM highlighted in grey and light blue, respectively. As for the analytical assessment, different columns of the table refer to different prediction horizons (30 to 90 min, with steps of 15). Values in different rows represent the percentage of values that fall within a specific zone of the EGA map (A to E, respectively). Again, we show for completeness the values obtained training the models on raw and on filtered data, respectively.

From the analysis of Table 4, we can draw the following considerations:

- The benefit of pre-processing the training data is again confirmed by Clarke EGA, for all the tested models. Hence, we will focus on these results.
- All the tested models had a satisfactory performance in short-term predictions. More than 94% of the data in the 30 min prediction horizon lay in the zone with best clinical outcome A. When considering a medium-term prediction horizon of 45 min, about 95% of the predictions fall within borders of the clinically acceptable zones A and B. Consistently with analytical assessment, the performance got worse when increasing further the prediction horizon.
- Partially contradicting the outcome of the analytical assessment, when considering the clinical outcome, NAR model was outperformed by the other methods. A possible interpretation of the discrepancy between the analytical and the clinical assessments is that the latter is much more affected by the presence of measurement spikes and outliers. Hence, while other prediction methods (and especially FNN) provide predictions that are on average not well-correlated with the observed measurements, they probably provided a lower number of outliers in comparison with NAR. Nonetheless, as previously observed, the short-term prediction performance of all the methods is quite comparable even from a clinical point of view.
- The undisputed superiority of LSTM model, both for short-term and long-term predictions, was confirmed even by EGA results. This model maintained 99% of the predicted values within the clinically acceptable zone up to 60 min horizon window and more than 99% in zone A up to 45 min forecasting. No other model showed comparable performance.



(a)



(b)

FIGURE 13. LSTM Clarke error grid analysis. (a) LSTM 30 min. (b) LSTM 60 min.

To have a better view of LSTM performance, in Figure 13, we show a graphical representation of LSTM EGA, respectively for the short-term (30 min) predictions and the long-term (60 min) predictions.

As it can be easily gathered from the plots, the data in Figure 13(a) distribute along the bisector, which is the region of highest correlation possible with the reference values. As expected, the data in Figure 13(b) have much higher dispersion. Nonetheless, we can observe that, even though the forecasting window was in this case extremely large (60 min), the wide majority of the values were still within the borders of best clinical acceptability.

V. CONCLUSIONS

In this work, we addressed the problem of automated glucose level prediction leveraging multi-patient CGMS data. Our specific aim is to learn a generalizable glucose level prediction model from a multi-patient training set, using this model to predict the future glucose values of a new patient. This allows improving the usability of models that are solely based on the past recordings of the same patient.

The main contribution of our work compared to past literature is two-fold: i) we learn the prediction models using a large set of CMGS data from a very heterogeneous set of diabetic patients. This possibly increases the generalization capability of the model and minimizes the risks of overfitting; ii) we design and compare different types of prediction models, analyzing the prediction outcome both from the analytical and from the clinical point of view. To address the limitations of literature approaches, we explored two types of models. The first solution exploits a Non-Linear Autoregressive Neural Network (NAR), that is supposed to extend the assumptions of linearity and overcome stability problems of traditional AR. The second solution exploits Long Short-Term Memory (LSTM), that addresses the exploding and vanishing gradient problems of classic RNN networks.

According to our experiments, the NAR network obtained satisfactory results only for short-term predictions, within 30 min. Nonetheless, if we take into account the model's simplicity (the NAR is based on just 8 regressors against 30 of all the other approaches), which makes it very convenient for hardware implementation, we can still consider it a good solution for systems not requiring a very large forecasting window.

Finally, as confirmed by both the analytical and clinical assessment, our LSTM network overcame by far the prediction accuracy of all the other models, for both short-term and long-term predictions. Hence, we can conclude that LSTM is the preferable approach for systems requiring a very long-term forecasting window.

In our future work, we will extend our multi-patient data-driven system by integrating real-time information. More specifically, we plan to perform a real-time fine-tuning of the model, leveraging the glucose level measurements of the patient that is currently using the system.

ACKNOWLEDGEMENT

Computational resources were provided by HPC@POLITO, a project of Academic Computing within the Department of Control and Computer Engineering at the Politecnico di Torino (http://www.hpc.polito.it)

REFERENCES

[1] A. R. Saltiel and C. R. Kahn, "Insulin signalling and the regulation of glucose and lipid metabolism," *Nature*, vol. 414, no. 6865, pp. 799–806, 2001.
[2] A. D. Association, "Diagnosis and classification of diabetes mellitus," *Diabetes Care*, vol. 33, no. 1, pp. S62–S69, 2010.
[3] E. J. Mayer-Davis, J. M. Lawrence, D. Dabelea, J. Divers, S. Isom, L. Dolan, G. Imperatore, B. Linder, S. Marcovina, and D. J. Pettitt, "Incidence trends of type 1 and type 2 diabetes among youths, 2002–2012," *New England J. Med.*, vol. 376, no. 15, pp. 1419–1429, 2017.
[4] M. A. Atkinson, G. S. Eisenbarth, and A. W. Michels, "Type 1 diabetes," *Lancet*, vol. 383, no. 9911, pp. 69–82, 2014.
[5] S. Chatterjee, K. Khunti, and M. J. Davies, "Type 2 diabetes," *Lancet*, vol. 389, no. 10085, pp. 2239–2251, 2017.
[6] M. W. Carpenter and D. R. Coustan, "Criteria for screening tests for gestational diabetes," *Amer. J. Obstetrics Gynecol.*, vol. 144, no. 7, pp. 768–773, 1982.
[7] M. Caswell, J. Frank, M. T. Viggiani, S. Pardo, N. Dunne, M. E. Warchal-Windham, and R. Morin, "Accuracy and user performance evaluation of a blood glucose monitoring system," *Diabetes Technol. Therapeutics*, vol. 17, no. 3, pp. 152–158, 2015.
[8] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2347–2376, 4th Quart., 2015.
[9] G. Freckmann, A. Baumstark, N. Jendrike, E. Zschornack, S. Kocher, J. Tshiananga, F. Heister, and C. Haug, "System accuracy evaluation of 27 blood glucose monitoring systems according to DIN EN ISO 15197," *Diabetes Technol. Therapeutics*, vol. 12, no. 3, pp. 221–231, 2010.
[10] K. Plis, R. Bunescu, C. Marling, J. Shubrook, and F. A. Schwartz, "A machine learning approach to predicting blood glucose levels for diabetes management," in *Proc. Workshops 28th AAAI Conf. Artif. Intell.*, Jun. 2014, pp. 35–39.
[11] G. Lanzola, E. Losiouk, S. D. Favero, A. Facchinetti, A. Galderisi, S. Quaglini, L. Magni, and C. Cobelli, "Remote blood glucose monitoring in mhealth scenarios: A review," *Sensors*, vol. 16, no. 12, p. 1983, 2016.
[12] E. Patti, M. Donatelli, E. Macii, and A. Acquaviva, "IoT software infrastructure for remote monitoring of patients with chronic metabolic disorders," in *Proc. IEEE 6th Int. Conf. Future Internet Things Cloud (FiCloud)*, Aug. 2018, pp. 311–317.
[13] C. D. Man, M. Camilleri, and C. Cobelli, "A system model of oral glucose absorption: Validation on gold standard data," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 12, pp. 2472–2478, Dec. 2006.
[14] 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.
[15] S. Dutta, T. Kushner, and S. Sankaranarayanan, "Robust data-driven control of artificial pancreas systems using neural networks," in *Computational Methods in Systems Biology—CMSB* (Lecture Notes in Computer Science), vol 11095, M. Èeška and D. Šafránek, Eds. Cham, Switzerland: Springer, 2018. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-99429-1_11#citeas
[16] H. N. Mhaskar, S. V. Pereverzyev, and M. D. van der Walt, "A deep learning approach to diabetic blood glucose prediction," *Frontiers Appl. Math. Statist.*, vol. 3, p. 14, Jul. 2017.
[17] S. Oviedo and 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, 2017, Art. no. e2833.
[18] T. Bremer and D. A. Gough, "Is blood glucose predictable from previous values? A solicitation for data," *Diabetes*, vol. 48, no. 3, pp. 445–451, 1999.
[19] 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.
[20] C. C. Palerm, J. P. Willis, J. Desemone, and B. W. Bequette, "Hypoglycemia prediction and detection using optimal estimation," *Diabetes Technol. Therapeutics*, vol. 7, no. 1, pp. 3–14, 2005.
[21] S. M. Pappada, B. D. Cameron, and P. M. Rosman, "Development of a neural network for prediction of glucose concentration in type 1 diabetes patients," *J. Diabetes Sci. Technol.*, vol. 2, no. 5, pp. 792–801, 2008.
[22] S. M. Pappada, B. D. Cameron, P. M. Rosman, R. E. Bourey, T. J. Papadimos, W. Olorunto, and M. J. Borst, "Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes," *Diabetes Technol. Therapeutics*, vol. 13, no. 2, pp. 135–141, 2011.
[23] E. I. Georga, V. C. Protopappas, D. Ardigò, M. Marina, I. Zavaroni, D. Polyzos, and D. I. Fotiadis, "Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 1, pp. 71–81, Jan. 2013.
[24] C. Zhao, E. Dassau, H. C. Zisser, L. Jovanovič, F. J. Doyle, III, and D. E. Seborg, "Online prediction of subcutaneous glucose concentration for type 1 diabetes using empirical models and frequency-band separation," *AIChE J.*, vol. 60, no. 2, pp. 574–584, 2014.
[25] J. B. Ali, T. Hamdi, N. Fnaiech, V. Di Costanzo, F. Fnaiech, and J.-M. Ginoux, "Continuous blood glucose level prediction of type 1 diabetes based on artificial neural network," *Biocybern. Biomed. Eng.*, vol. 38, no. 4, pp. 828–840, 2018.

[26] T. Hamdi, J. B. Ali, V. Di Costanzo, F. Fnaiech, E. Moreau, and J.-M. Ginoux, "Accurate prediction of continuous blood glucose based on support vector regression and differential evolution algorithm," *Biocybern. Biomed. Eng.*, vol. 38, no. 2, pp. 362–372, 2018.
[27] D. M. Hawkins, "The problem of overfitting," *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 1, pp. 1–12, 2004.
[28] A. Gani, A. V. Gribok, S. Rajaraman, W. K. Ward, and J. Reifman, "Predicting subcutaneous glucose concentration in humans: Data-driven glucose modeling," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 2, pp. 246–254, Feb. 2009.
[29] A. Gani, A. V. Gribok, Y. Lu, W. K. Ward, R. A. Vigersky, and J. Reifman, "Universal glucose models for predicting subcutaneous glucose concentration in humans," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 1, pp. 157–165, Jan. 2010.
[30] C. M. Bishop, "Training with noise is equivalent to Tikhonov regularization," *Neural Comput.*, vol. 7, no. 1, pp. 108–116, 1995.
[31] C. Pérez-Gandía, A. Facchinetti, G. Sparacino, C. Cobelli, E. J. Gómez, M. Rigla, A. de Leiva, and M. Hernando, "Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring," *Diabetes Technol. Therapeutics*, vol. 12, no. 1, pp. 81–88, 2010.
[32] F. Allam, Z. Nossai, H. Gomma, I. Ibrahim, and M. Abdelsalam, "A recurrent neural network approach for predicting glucose concentration in type-1 diabetic patients," in *Engineering Applications of Neural Networks—EANN* (IFIP Advances in Information and Communication Technology), vol. 363, L. Iliadis and C. Jayne, Eds. Berlin, Germany: Springer, 2011. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-23957-1_29#citeas
[33] T. Zhu, K. Li, P. Herrero, J. Chen, and P. Georgiou, "A deep learning algorithm for personalized blood glucose prediction," in *Proc. Int. Workshop Knowl. Discovery Healthcare Data*, 2018, pp. 1–5.
[34] K. Li, J. Daniels, P. Herrero, C. Liu, and P. Georgiou, "Convolutional recurrent neural networks for glucose prediction," Jul. 2018, *arXiv:1807.03043*. [Online]. Available: https://arxiv.org/abs/1807.03043
[35] S. A. Billings, *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains*. Hoboken, NJ, USA: Wiley, 2013.
[36] K. Greff, R. K. Srivastava, and J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.
[37] A. Sorjamaa, J. Hao, N. Reyhani, Y. Ji, and A. Lendasse, "Methodology for long-term prediction of time series," *Neurocomputing*, vol. 70, nos. 16–18, pp. 2861–2869, Oct. 2007.
[38] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. 30th Int. Conf. Int. Conf. Mach. Learn.*, Jun. 2013, pp. 1310–1318.
[39] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, Jul. 2015, pp. 2342–2350.
[40] Q. Sun, M. V. Jankovic, L. Bally, and S. G. Mougiakakou, "Predicting blood glucose with an LSTM and Bi-LSTM based deep neural network," in *Proc. 14th Symp. Neural Netw. Appl. (NEUREL)*, Nov. 2018, pp. 1–5.
[41] J. C. Pickup, M. F. Holloway, and K. Samsi, "Real-time continuous glucose monitoring in type 1 diabetes: A qualitative framework analysis of patient narratives," *Diabetes Care*, vol. 38, no. 4, pp. 544–550, 2015.
[42] D. Rodbard, "Continuous glucose monitoring: A review of successes, challenges, and opportunities," *Diabetes Technol. Therapeutics*, vol. 18, no. S2, p. S2-3, 2016.
[43] J. C. Health Research (JCHR). *Diabetes Research Studies*. Accessed: Oct. 2018. [Online]. Available: http://diabetes.jaeb.org/
[44] Abbott Diabetes Care Division. (2018). *WELCOME to the Forefront of Diabetes Care*. [Online]. Available: http://www.diabetescare.abbott/
[45] Dexcom, Inc. (2018). *DEXCOM Continuous Glucose Monitoring*. [Online]. Available: http://www.dexcom.com/
[46] M-PLC. (2018). *Medtronic*. [Online]. Available: https://www.medtronic.com/us-en/index.html
[47] X. Xie, J. C. Doloff, V. Yesilyurt, A. Sadraei, J. J. McGarrigle, M. Omami, O. Veiseh, S. Farah, D. Isa, and S. Ghani, "Reduction of measurement noise in a continuous glucose monitor by coating the sensor with a zwitterionic polymer," *Nature Biomed. Eng.*, vol. 2, pp. 894–906, Jul. 2018. [Online]. Available: https://www.nature.com/articles/s41551-018-0273-3
[48] B. W. Bequette, "Continuous glucose monitoring: Real-time algorithms for calibration, filtering, and alarms," *J. Diabetes Sci. Technol.*, vol. 4, no. 2, pp. 404–418, 2010.
[49] K. Ito and B. Jin, *Inverse Problems: Tikhonov Theory and Algorithms*. Singapore: World Scientific, 2015.
[50] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, vol. 4. Philadelphia, PA, USA: SIAM, 2005.
[51] L. Ljung, "System identification," in *Wiley Encyclopedia of Electrical and Electronics Engineering*. 1999, pp. 1–19. doi: 10.1002/047134608X.W1046.
[52] P. M. Nørgård, O. Ravn, N. K. Poulsen, and L. K. Hansen, *Neural Networks for Modelling and Control of Dynamic Systems—A Practitioner's Handbook*. London, U.K.: Springer, 2000. [Online]. Available: http://orbit.dtu.dk/en/publications/neural-networks-for-modelling-and-control-ofdynamic-systems–a-practitioners-handbook(e01355ec-56b4-4a72-a319-06ae7809eb34)/export.html
[53] R. Rajamani, "Observers for Lipschitz nonlinear systems," *IEEE Trans. Autom. Control*, vol. 43, no. 3, pp. 397–401, Mar. 1998.
[54] L. K. Hansen and M. W. Pedersen, "Controlled growth of cascade correlation nets," in *Proc. Int. Conf. Artif. Neural Netw.* Springer, May 1994, pp. 797–800.
[55] X. He and H. Asada, "A new method for identifying orders of input-output models for nonlinear dynamic systems," in *Proc. Amer. Control Conf.*, Jun. 1993, pp. 2520–2523.
[56] X. Dong, S. Chen, and S. J. Pan, "Learning to prune deep neural networks via layer-wise optimal brain surgeon," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Dec. 2017, pp. 4860–4874.
[57] G. Dorffner, "Neural networks for time series processing," *Neural Netw. World*, vol. 6, no. 4, pp. 447–468, 1996.
[58] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," in *Proc. 9th Int. Conf. Artif. Neural Netw.*, Sep. 1999, pp. 850–855.
[59] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[60] C. Olah, "Understanding LSTM networks," Tech. Rep., 2015. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/
[61] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing LSTM language models," Aug. 2017, *arXiv:1708.02182*. [Online]. Available: https://arxiv.org/abs/1708.02182
[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[63] C. A. Gueymard, "A review of validation methodologies and statistical performance indicators for modeled solar radiation data: Towards a better bankability of solar projects," *Renew. Sustain. Energy Rev.*, vol. 39, pp. 1024–1034, Nov. 2014.
[64] L. Ekhlaspour, D. Mondesir, N. Lautsch, C. Balliro, M. Hillard, K. Magyar, L. G. Radocchia, A. Esmaeili, M. Sinha, and S. J. Russell, "Comparative accuracy of 17 point-of-care glucose meters," *J. Diabetes Sci. Technol.*, vol. 11, no. 3, pp. 558–566, 2017.
[65] I. O. for Standardization, "ISO 15197:2013. *In vitro* diagnostic test systems–requirements for blood-glucose monitoring systems for self-testing in managing diabetes mellitus," Int. Org. Standardization, Geneva, Switzerland, Tech. Rep. 2nd Ed., 2013.
[66] W. L. Clarke, "The original Clarke error grid analysis (EGA)," *Diabetes Technol. Therapeutics*, vol. 7, no. 5, pp. 776–779, 2005.
[67] D. J. Cox, L. A. Gonder-Frederick, B. P. Kovatchev, D. M. Julian, and W. L. Clarke, "Understanding error grid analysis," *Diabetes Care*, vol. 20, no. 6, p. 911, Jun. 1997.

ALESSANDRO ALIBERTI received the B.S. and M.S. degrees in cinema and media engineering from the Politecnico di Torino, where he is currently pursuing the Ph.D. degree with the Department of Control and Computer Engineering. In 2016, he joined the Department of Control and Computer Engineering, Politecnico di Torino, as a Research Assistant. During his academic experience, he started working in a big Italian communication company and then in automation and control systems area. His research interests include the development of design methodologies, software tools, models, and sensor networks that can mitigate the temperature effects on digital electronics circuits in order to improve the knowledge in the fields of the IoT and smart systems and cities.

IRENE PUPILLO received the M.S. degree in biomedical engineering with specialization in IT medical engineering from the Politecnico di Torino, Turin, in 2017. During her academic career, she collaborated with NewAmuser srl in order to develop an e-health monitoring system.

STEFANO TERNA received the M.S. degree in physics from the University of Torino, in 1996, and the Ph.D. degree in theoretical physics from the Scuola Internazionale Superiore di Studi Avanzati, Trieste, in 2000. From 2000 to 2004, he was a Senior Technology Consultant at the Accenture Financial Services Business Unit. From 2004 to 2008, he was the Software Product Manager with SAET IS, Torino, developing distributed systems for the detection and remote control of industrial alarms. From 2008 to 2016, he led, as a freelancer, more than 30 software projects about field data acquisition, data analysis, and remote control in automotive, banking, pharma, energy, and telecommunication sectors. Since 2015, he has been the CTO and the Co-Founder of TomorrowData, a company specialized in developing and deploying AI algorithms on constrained hardware (the so-called edge computing). His research interests include recurrent neural networks and reservoir computing applied to the forecasting, and anomaly detection of multivariate time series. He was a member of the Advisory Board for the Master of Science in Data Analytics at the University of Central Florida, in 2016.

ENRICO MACII received the Laurea degree in electrical engineering from the Politecnico di Torino, in 1990, the Laurea degree in computer science from the Università di Torino, in 1991, and the Ph.D. degree in computer engineering from the Politecnico di Torino, in 1995. He was an Associate Professor, from 1998 to 2001, and an Assistant Professor, from 1993 to 1998, with the Politecnico di Torino, where he is currently a Full Professor of computer engineering. From 1991 to 1997, he was an Adjunct Faculty Member with the University of Colorado at Boulder. From 2009 to 2016, he was the Vice Rector for Research at the Politecnico di Torino. He was also the Vice Rector for European Affairs, from 2007 to 2009, the Vice Rector for Technology Transfer, from 2009 to 2015, and the Rector's Delegate for International Affairs, from 2012 to 2015. His research interest includes the design of electronic digital circuits and systems, with a particular emphasis on low power consumption aspects. In the last decade, he has extended his research activities to areas, such as bioinformatics, energy efficiency in buildings, districts and cities, sustainable urban mobility, and clean and intelligent manufacturing.

SANTA DI CATALDO received the Biomedical Engineering degree *(summa cum laude)* and the Ph.D. degree in systems and computer engineering from the Politecnico di Torino, Italy, in 2006 and 2011, respectively. In 2007, she joined the Department of Control and Computer Engineering, Politecnico di Torino, as a Research Assistant. She is currently an Assistant Professor (tenure track). Her main research interests include image processing, machine learning, and heterogeneous data integration, including techniques for pattern recognition, image segmentation, and feature quantification and classification.

EDOARDO PATTI received the M.Sc. and Ph.D. degrees in computer engineering from the Politecnico di Torino, in 2010 and 2014, respectively, where he is currently an Assistant Professor. From 2014 to 2015, he was Academic Visiting at The University of Manchester. His research interests include ubiquitous computing, the Internet of Things, smart systems and cities, software architectures with a particular emphasis on infrastructure for ambient intelligence, software solutions for simulating and optimizing energy systems, and software solutions for energy data visualization to increase user awareness. In these fields, he has authored over 70 scientific publications, from 2011 to 2018.

ANDREA ACQUAVIVA received the Ph.D. degree in electrical engineering from the University of Bologna, Italy, in 2003. In 2003, he became an Assistant Professor with the Computer Science Department, University of Urbino, Italy. From 2005 to 2007, he was a Visiting Researcher with the École Polytechnique Fédérale de Lausanne, Switzerland. In 2006, he joined the Department of Computer Science, University of Verona, Italy. He has been with the Interuniversity Department of Regional and Urban Studies and Planning, Politecnico di Torino, Italy. He is currently an Associate Professor with the Politecnico di Torino. His research interests include parallel computing for distributed embedded systems such as multicore and sensor networks, software solutions for smart cities, and simulation and analysis of biological systems using parallel architectures. In these fields, he has authored over 200 scientific publications, from 2000 to 2018.