Prediction of Personalized Blood Glucose Levels in Type 1 Diabetic Patients using a Neuroevolution Approach

Ivanoe De Falco
Institute for High Performance Computing and Networks
Naples, Italy
ivanoe.defalco@icar.cnr.it

Antonio Della Cioppa
University of Salerno
Salerno, Italy
adellacioppa@unisa.it

Angelo Marcelli
University of Salerno
Salerno, Italy
amarcelli@unisa.it

Umberto Scafuri
Institute for High Performance Computing and Networks
Naples, Italy
umberto.scafuri@icar.cnr.it

Luca Stellaccio
University of Salerno
Salerno, Italy
lucastellaccio@gmail.com

Ernesto Tarantino
Institute for High Performance Computing and Networks
Naples, Italy
ernesto.tarantino@icar.cnr.it

ABSTRACT

Diabetes mellitus is a lifelong disease in which either the pancreas fails to produce insulin or the produced amount is insufficient to control blood sugar levels. A way to tackle this malfunctioning is to devise an artificial pancreas endowed with a personalized control algorithm able to regulate the insulin dosage. A crucial step in realizing such a device is to effectively forecast future glucose levels starting from past glucose values, the knowledge of the food intake, and of the basal and the injected insulin. The increasing availability of medical diabetes data sets is providing unprecedented opportunities to identify correlations inside these data even harnessing innovative investigation methods, such as deep learning.

As an alternative to the deep learning methods successfully used as forecasting models, we exploit a neuroevolution algorithm to model and predict future personalized blood glucose levels. The discovered subjective regression model can represent the control algorithm of an artificial pancreas. This model is assessed through experiments performed on a real-world database containing data of six patients suffering from Type 1 diabetes. To further evaluate the effectiveness of the predictions derived from the proposed approach, the results are also compared against those obtained by other state–of–the–art recently proposed methods.

CCS CONCEPTS

• **Computing methodologies** → **Control methods**; • **Applied computing** → **Health care information systems**.

KEYWORDS

Blood glucose estimation, Regression models, Neuroevolution.

GECCO '21 Companion, July 10–14, 2021, Lille, France

ACM ISBN 978-1-4503-8351-6/21/07…$15.00
https://doi.org/10.1145/3449726.3463143

ACM Reference Format:
Ivanoe De Falco, Antonio Della Cioppa, Angelo Marcelli, Umberto Scafuri, Luca Stellaccio, and Ernesto Tarantino. 2021. Prediction of Personalized Blood Glucose Levels in Type 1 Diabetic Patients using a Neuroevolution Approach. In *2021 Genetic and Evolutionary Computation Conference Companion (GECCO '21 Companion), July 10–14, 2021, Lille, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3449726.3463143

1 INTRODUCTION

Diabetes mellitus (DM) is a metabolic disease provoked by the malfunction of the pancreas in releasing the insulin responsible for regulating the amount of the blood glucose (BG) in the body. Untreated high BG levels can damage nerves, eyes, kidneys, and other organs or sometimes involve fatal complications like heart attacks and strokes. Fortunately, managing diabetes is possible by changing the living habits and monitoring blood glucose levels, so reducing the risk of developing these medical complications. Thanks to monitoring, it is possible to establish whether the therapeutic plan adopted, consisting of a set of drugs, diet and exercise, is able to keep the blood glucose level under control. Being patients reluctant to control such a level by means of invasive devices because of the associated pain, the use of minimally invasive Continuous Glucose Monitoring System (CGMS) devices for the measurement of the Interstitial Glucose (IG), i.e., the glucose in the subcutaneous tissue, has spread. These devices allow continuous measurements that are gathered in medical data sets. These data can be exploited to derive control algorithms able to perform a glycemic control through the forecasting of future glucose levels.

In the last decade, to improve the quality of life of Type 1 DM (T1DM) subjects, researchers have focused on developing a closed-loop control system, the so-called Artificial Pancreas (AP). A fundamental part of the AP is represented by the control algorithm capable of automatically driving an insulin pump for providing the needed amount of insulin, so as to maintain the glucose level within a safe range.

Different models have been proposed in literature attempting to capture the future glucose trend, as reported in Section 2. However, despite the achieved progresses, glycemic control remains a very complex task, and there is a significant interest in developing innovative time-series forecasting methods for predicting blood glucose

levels. Such an interest has inspired the second Blood Glucose Level Prediction (BGLP) challenge helded in conjunction with the 5th International Workshop on Knowledge Discovery in Healthcare Data at ECAI in August 2020, in Santiago de Compostela, Spain [2]. During this challenge, the OhioT1DM data set, containing data from subjects suffering from T1DM [24], has been employed to evaluate the performance of different forecasting strategies. A deep learning approach has resulted to be the top-performing prediction model [30].

The contribution of this paper is to investigate neuroevolution for deriving personalized forecasting models. These models are used for estimating subjective glucose levels in future instants of time from past glucose measurements on the basis of the ingested carbohydrates, and of the basal and the administered insulin. The idea is to ascertain the ability of the neuroevolution to automatically obtain neural network models simpler than those typical of the deep learning structures [33], based on complex structure of artificial networks. In this way, the regression model discovered through the neuroevolution approach, characterized by low computational complexity, presents the advantage to be used for a personalized control algorithm in glucose regulating systems as the AP.

To perform a fair comparison, the experimental trials are performed over the Ohio T1DM data set, using a supervised learning approach to solve the problem of time-series glucose forecasting as in the BGLP challenge.

The paper is structured as follows. In Sect. 2 a brief overview of related works is outlined. Section 3 describes the neuroevolution methodology for solving the regression problem. The achieved results are discussed and compared with those from other models in Sect. 4. Conclusions and future work are exposed in Sect. 5.

2 RELATED WORKS

Numerous time-series forecasting methods for predicting blood glucose levels have been proposed [31]. These approaches can be roughly categorized in physiological, hybrid and data-based models. The physiological models are related to the development of glucose-insulin compartmental models each capable of emulating different human physiological behavior by describing the glucose dynamics during the insulin treatment [16, 28, 37].

The hybrid models combine physiological models with other soft computing methodologies techniques, such as swarm optimization [1], grammatical evolution [7], long short-term memory (LSTM) networks [36].

The approach here proposed falls within the data-based models, i.e. models that provide a prediction of glucose signals by learning evolution patterns from data. These data are represented by CGM measurements sometimes complemented with manual or computed recordings coming from other wearable sensors, and have been widely used for forecasting future glycemic trends for T1DM patients.

Sparacino et al. [32] exploited CGMS data within a first-order autoregressive model and compared the outcomes with a first-order polynomial model. For both methods, at each sampling time, a new set of model parameters is identified by means of weighted least squares techniques.

Peréz-Gandía [29] harnessed a fully connected artificial neural network model made up of three layers with a number of neurons equal to ten, five, and one, respectively. The transfer function is sigmoidal in the first two layers, while is linear in the third. This network exploits as inputs the current glucose measurement and its timestamps together with a limited number of previous glucose samples from the CGM system, and returns as output the glucose prediction at the chosen forecasting time horizon.

Zecchin et al. [40] proposed a jump neural network algorithm that, in addition to past CGMS data, also exploits carbohydrate intake information to derive a short-term (30-minute horizon) BG prediction model.

Zarkogiovanni et al. [39] presented a comparative evaluation of four models, a feedforward neural network, a self-organizing map, a neuro-fuzzy network with wavelets as activation functions, and a linear regression model for the glucose prediction of T1DM patients using BG concentration data extracted from sensors and physical activity information.

Li et al. [22] considered a deep learning approach relying on the combination of a Convolutional Neural Network for automatic feature extraction and an LSTM Recurrent Neural Networks (RNNs) for time series prediction. Such an approach leverages BG measurements from a CGM device along with insulin and carbohydrate intake information for estimating the BG levels at 30 and 60 minute prediction horizons.

Munoz-Organero [27] presented a new hybrid model that decomposes a deep machine learning model in order to mimic the metabolic behavior of physiological blood glucose methods. The inputs to the model are the current and past measurements from CGM devices, fast and slow acting insulin injections and food intake. The differential equations for carbohydrate and insulin absorption in the physiological models are modeled using an RNN implemented using LSTM cells.

Cappon et al. [5] proposed a new deep learning method relying on a personalized bidirectional LSTM, harnessed with an interpretability tool for future BG forecasting with time horizons of 30 and 60 minutes.

Zhu et al. [41] advanced a novel deep learning model to forecast future BG concentration by exploiting historical CGM data, ingested meal, and delivered insulin.

During the last years some attempts relying on evolutionary-based algorithms have been advanced to derive diabetes modeling by using as inputs CGMS values for extracting relationships under the form of explicit mathematical expressions [9–11, 13], even on the basis of previous and estimated future carbohydrate intakes and insulin injections [12, 18, 19].

3 THE NEUROEVOLUTION METHODOLOGY

Neuroevolution represents a sub-area of artificial intelligence and machine learning which harnesses evolutionary algorithms to optimize artificial neural networks by evolving not only the weights but also the architecture during the evolutionary process. This technique allows maintaining a population of solutions during the search, enabling extreme exploration and massive parallelization.

The neuroevolution methodology here proposed for the extraction of a regression model is the NeuroEvolution of Augmenting

Topologies (NEAT) algorithm developed by Stanley and Miikkulainen [34]. NEAT is a complex algorithm that evolves artificial neural networks by attempting to find a balance between the fitness of evolved solutions and their diversity. During evolution, NEAT can propose both feedforward and recurrent networks. As claimed by the same authors, three key techniques are the reason of the NEAT efficiency: (i) employing a principled method of crossover of different topologies, (ii) preserving structural innovation using speciation, and (iii) incrementally growing topologies from minimal and simple initial structures. In particular, differently from other neuroevolution approaches, NEAT starts with a population of networks each of which has no any hidden nodes, and mutation may incrementally add such nodes and the related links. Then, the use of fitness evaluation allows the survival of only those new, more complex, structures that turn out useful.

3.1 Prediction through NEAT

Given an objective function and a search space constituted by all the regression models representable as neural networks, the NEAT algorithm evolves a population of regression models, i.e., neural network structures, to be optimized by means of the objective function.

The problem to be solved is a time-series regression, represented by the glucose levels in future instants of time. With measurements taken every Δt minutes, we deal with a regression problem that aims at providing the predicted glucose value $\widehat{G}(t + h\Delta t)$ for the forecasting horizon $h\Delta t$. This regression is effected by considering available in input the values for the glucose levels $G(t)$, the injected insulin $I(t)$, and the ingested amounts of carbohydrates $C(t)$, in a time window of $k\Delta t$ before the current instant t. The time-series prediction task is transformed into the following symbolic regression problem:

$$\begin{aligned}\widehat{G}(t + h\Delta t) \quad &= f(G(t), G(t - \Delta t) \ldots, G(t - k\Delta t), I(t), I(t - \Delta t),\\ &\quad \ldots, I(t - k\Delta t), C(t), C(t - \Delta t) \ldots, C(t - k\Delta t))\end{aligned}$$

3.1.1 Objective function. Following the rules of the BGLP challenge, in this paper too the accuracy of the forecasting regression models found by NEAT is evaluated through a supervised learning process by means of two error functions to be minimized.

The first function is the weighted Root Mean Square Error (*RMSE*) between the actual glucose and the estimated glucose trends:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(G(i) - \widehat{G}(i))^2}$$

where $G(i)$ is the actual value of the glucose variable extracted from CGMS, $\widehat{G}(i)$ is its estimated value for the i-th item of the series, and n is the number of samples in the training phase.

The particularity of the RMSE is that it takes the unit of measure of the considered feature, in this case, *mg/dL*.

The second cost function is the Mean Absolute Error (MAE) that computes the mean of the error between the actual and the estimated glucose value:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|(G(i) - \widehat{G}(i))|$$

The reason why a check on the MAE is needed is that the metric based on RMSE is very susceptible to outliers, and even with a single outlier the RMSE may increase a lot. Given the rules of the BGLP challenge, during the evolution, the quality of each individual will be evaluated through RMSE only.

4 EXPERIMENTAL RESULTS

4.1 The database

The experiments are conducted on the Ohio T1DM data set that covers 8 weeks of daily living data for a total of 12 T1DM patients [24]. As mentioned in the introduction, this data set has been employed for evaluating the performance of several techniques during the BGLP challenge.

Within this paper the focus is only concentrated on the six patients added in 2020. All the involved people are subject to insulin pump therapy with CGMS. The data set includes: CGMS blood glucose estimates taken every $\Delta t = 5$ minutes; blood glucose measurements from periodic self-monitoring of blood glucose by finger sticks; insulin doses, both bolus and basal; self-reported meal times with carbohydrate estimates; self-reported times of sleep, work, and exercise, and 5-minute aggregations of heart rate, galvanic skin response, skin temperature, air temperature, and step count.

When dealing with machine learning and neural networks, the preprocessing phase is an important step, as the way the data are formatted and the selected features play an important role during the learning phase. Therefore a feature selection phase is essential. In fact, each feature is a potential candidate to be used to feed the forecasting model. We have made preliminary experiments by exploring all the variables in the data set, but we have only found useful the glucose level, the injected insulin (basal plus boluses), and the amount of carbohydrates ingested during the day. For the NEAT purpose, the data series of each patient is subdivided into a training set used to carry out the learning phase to build the model and a testing set employed to evaluate the quality of this model over unseen samples. The number of training and testing examples for each patient is reported in [24] together with information on how these items have been assigned to the two sets.

After the feature selection, an alignment phase is to be carried out on the Ohio T1DM data set [8] because the features are not reported at the same time. This involves that the raw data should be preprocessed to feed the neuroevolution algorithm. During this phase, the problem emerges of missing data in the data set.

To deal with these missing values, the padding preprocessing is used for the glucose values in the training and the testing sets. This preprocessing works by propagating the last available value until a new available data is encountered.

The preprocessing for insulin, instead, is performed on the basis of the Bergman model, which allows insulin to be better distributed over time, without having to extrapolate the last available value when there is a missing one. This model takes into account basal and bolus insulin and the absorption rate of carbohydrates [3].

As regards the carbohydrates intake, in presence of a meal the gut absorption rate is modeled according to [20] as:

$$C(t) = \frac{D_g \cdot A_g \cdot t \cdot e^{-t/t_{\max}}}{t_{\max}^2} \quad (1)$$

where $t_{\max} = 5$ is the time-of-maximum appearance rate of glucose in the accessible compartment, D_g is the amount of carbohydrates digested, and $A_g = 0.8$ is the carbohydrates bioavailability. This function which rapidly increases after the meal, and then decays to 0 in 2-3 hours. Outside such windows, the missing carbohydrate values are filled with zeroes.

4.2 The findings

To solve this regression problem, we have written an algorithm in Python that recalls a NEAT library in Python as well. This latter is freely downloadable from [25]. The library makes it available a configuration file in which different types of parameters are to be set [26]. Since the initialization of the algorithm can be crucial for the final results, a preliminary tuning is necessary to individuate the parameter configuration that leads to the best performance. A customized tuning for each patient and for all the possible configurations is impracticable. Therefore we only investigated the patient with identifier ID = 596 for some parameter configurations, and then the found 'sub-optimized' configuration is employed to test the algorithm performance on all the patients.

The experiments have been carried out by considering a value of $\Delta t = 5$ minutes, a forecasting horizon of $h\Delta t = 30$ minutes, and a time window $k\Delta t = 60$ minutes. For the experiments, ten independent executions have been made for each patient in order to have statistically meaningful results through the reduction of the randomness in the initialization of the NEAT algorithm.

Given the 60–minute width chosen for the time window, given the 5–minute sampling, and given that we consider the three parameters G, I, and C, the algorithm will take into account structures with at most 39 input nodes $(((60/5)+1)*3)$, and one output node representing the predicted $\widehat{G}(t+30)$.

During the tuning process, most of the parameters have been set to their default values in order to see how the setting of some specific parameters could influence the evaluation. Given the complexity of the problem, the population size and the number of generations have been set greater than the default values. By means of preliminary experiments, we found that the performance of the algorithm improves by using a wide set of activation functions and all the available aggregation functions with a small probability of mutation, as well as increasing the probability that, during a mutation, an existing connection and a node will be deleted or added so favoring the ability of the networks to mutate. The setting of the parameters modified with respect to the default values as result of the tuning process is reported in Table 1,.

Analogously to the BGLP challenge, as error metric the RMSE and MAE defined in Sect. 3.1.1 have been used. It is important to point out that the MAE is not used as a metric for the evaluation during the evolution. The model accuracy has been evaluated in terms of RMSE on the training and the outcomes on the testing set are reported in Table 2. In particular, for each patient, the table shows the average values, the standard deviation, and the best values of the error metrics on the testing set attained by using the ten discovered trained models with the lowest RMSE values over the ten runs performed. The last row outlines the respective average values achieved over all the patients.

Table 1: Parameter configuration for all the patients

Parameter	Value
generations	500
population size	500
activation functions	tanh, sigmoid, relu, sin, softplus, hat, gauss
activation mutation rate	0.3
aggregation functions	sum, product, max, min, median, mean
aggregation mutate rate	0.3
connection add prob.	0.6
connection delete prob.	0.6
node add prob.	0.6
node delete prob.	0.6

Table 2: Average, standard deviations, and best results obtained on the testing set.

	Average	**St. Dev.**	**Best**	
ID	RMSE	RMSE	RMSE	MAE
540	23.19	0.121	23.056	16,672
544	17.26	0.200	17.073	11.599
552	13.871	0.196	13.761	7.958
567	25.812	0.066	25.772	16.379
584	25.617	0.179	25.396	16.653
596	18.067	0.162	17.909	12.535
Average	20.636	0.154	20,494	13,633

In general, the results in the table show fairly good prediction accuracy for the personalized glucose predictions. In fact, the best personalized RMSE/MAE errors remain confined below 26/17 mg/dl. The model robustness is worth noting, given that the RMSE average value of 20.636 mg/dl over all the patients is very close to the global average of the best performance over all the patients that is 20.494 mg/dl. This aspect is also witnessed by the low values of the standard deviations.

4.3 Clarke Error Grid analysis

Clarke Error Grid (CEG) analysis [6] is an investigation typically performed in the diabetes field to estimate the clinical accuracy for a patient by using the predicted versus the actual observed BG values. The bisectrix represents a perfect prediction, while the 5 zones (A-E) in which the plot is split return an indication of the possible outcomes that may occur for a specific prediction. The predictions in zone A include values within 20% of the actual values, while those in zone B comprehend values outside of 20% but not involving dangerous treatments for the patient. The predictions in zone C could conduct to useless treatments. Prediction points in zone D denote a potentially dangerous failure to distinguish hypoglycemia or hyperglycemia. Predictions in zone E are extremely dangerous since they indicate a confusion of the treatment of hypoglycemia for hyperglycemia and vice versa. Predictions in the zones D and E should be avoided.

CEG analysis is used here to evaluate the performance over the testing set for the six patients. Specifically, the Table 3 reports the percentages of samples within the five grid areas for each patient denoted with the related ID. Specifically, the table shows the personalized best results on the testing set over the 10 runs obtained in correspondence of the forecasting model with the lowest RMSE in the training set. Furthermore, the last column of the table shows the best percentage of the points falling in the safe zone A+B, while the last row outlines the average values achieved over all the patients.

Table 3: Best percentages of the samples within the CEG areas

ID	A	B	C	D	E	A+B
540	82.42	16.01	0.03	1.51	0.03	98,43
544	92.87	6.88	0.00	0.25	0.00	99,75
552	92.87	6.50	0.00	0.63	0.00	99,37
567	85.81	12.61	0.00	1.55	0.03	98,42
584	86.14	12.73	0.43	0.70	0.00	98,87
596	89.10	9.50	0.00	1.37	0.03	98,60
Average	88,20	10,706	0,077	1,002	0,015	98,725

Some observations can be made on the basis of the CEG analysis reported in Table 3:

- an important issue to note is the reliability of the forecasting model. In fact, although 1 or 2 points are present in zone E for some patients, the percentages of the points predicted in the safe zone A+B for the best results are greater than 98%. The discovered prediction models show the best performance for the patients 544 and 552, who not only have the highest percentage of points in safe zones but also a number of points falling in zone E equal to 0.
- the percentages of samples predicted in the unsafe zones C, D, and E are very small, especially for the points falling in the very dangerous zones D, with less than 2%, and E, with less than 0,03%, so confirming the goodness of the neuroevolution forecasting;
- the difference between the actual and the estimated glucose values can be ascribed to the fact that the glycemic control was not performed under medical supervision, but also to the partial and imprecise information about the insulin bolus and carbohydrate intake provided by the patients and included in the data set. Moreover, even the imputation procedure of missing data can influence the prediction capability of the model, and so can the number of features considered.

In Fig. 1 the CEG and the related forecasting results on the testing set for the first three patients are reported.

4.4 Comparison with literature

The results achieved by our neuroevolution algorithm can be compared with those attained by other techniques on the same database during the BGLP challenge. In Table 4 the overall ranking of the most performing techniques of the challenge, denoted by the reference to the related paper, and the NEAT algorithm, is shown. Following the rules of the BGLP challenge, for each technique, its score is computed as the sum of the RMSE and MAE values averaged over the six patients, and the ranking is made in terms of increasing scores, the lower the better.

Table 4: Comparison with BGLP challenge 2020

Paper	RMSE	MAE	Overall
[30]	18.22	12.83	31.05
[41]	18.34	13.37	31.71
[17]	19.21	13.08	32.29
[38]	19.05	13.50	32.55
[4]	18.23	14.37	32.60
[35]	19.37	13.76	33.13
[21]	19.60	14.25	33.85
NEAT	20.49	13.63	34.12
[23]	20.03	14.52	34.55

Although a complete comparison is not possible because not all the participants to the challenge provided the CEGs corresponding to each patient, from this table it could seem that NEAT has worse performance when compared with those of the most performing algorithms. To better ascertain this aspect, in the following, we perform a further investigation.

4.4.1 Statistical analysis. With the aim to compare the results of NEAT in terms of RMSE against those of the other algorithms utilized in the BGLP competition, we have performed statistical analysis. It should be pointed out here that, out of the eight algorithms officially ranked in that competition, four reported average results whereas other four only reported results from one single run. Statistics is based on multiple executions and on average values, so that sound conclusions about the average behaviour of the participants can be drawn, therefore in this section we only consider the algorithms reporting average–based results, i.e., those in papers [4], [17], [38], and [41], and NEAT as well.

The average values obtained by each algorithm over each problem are reported in Table 5

Table 5: The average results in terms of RMSE obtained by the five algorithms over the six subjects.

ID	[17]	[41]	[38]	[4]	NEAT
540	19.55	20.14	21.00	21.03	23.19
544	16.56	16,28	16.69	16.14	17.26
552	15.04	16.08	16.92	15.82	13.87
567	23.07	20.00	21.93	20.29	25.81
584	25.19	20.91	21.88	20.39	25.62
596	15.85	16.63	15.87	15.70	18.07

As we wish to compare here five algorithms over six problems, we appeal to the use of non–parametrical statistical analysis. By following [14, 15], we have used Quade test, that takes account for the fact that not all the problems are equally difficult, or that the results obtained by the algorithms over them show large differences, as it is the case here. We have chosen this test because it has better discrimination ability than the classically used Friedman or Aligned Friedman, since it gives higher importance to more difficult tasks.

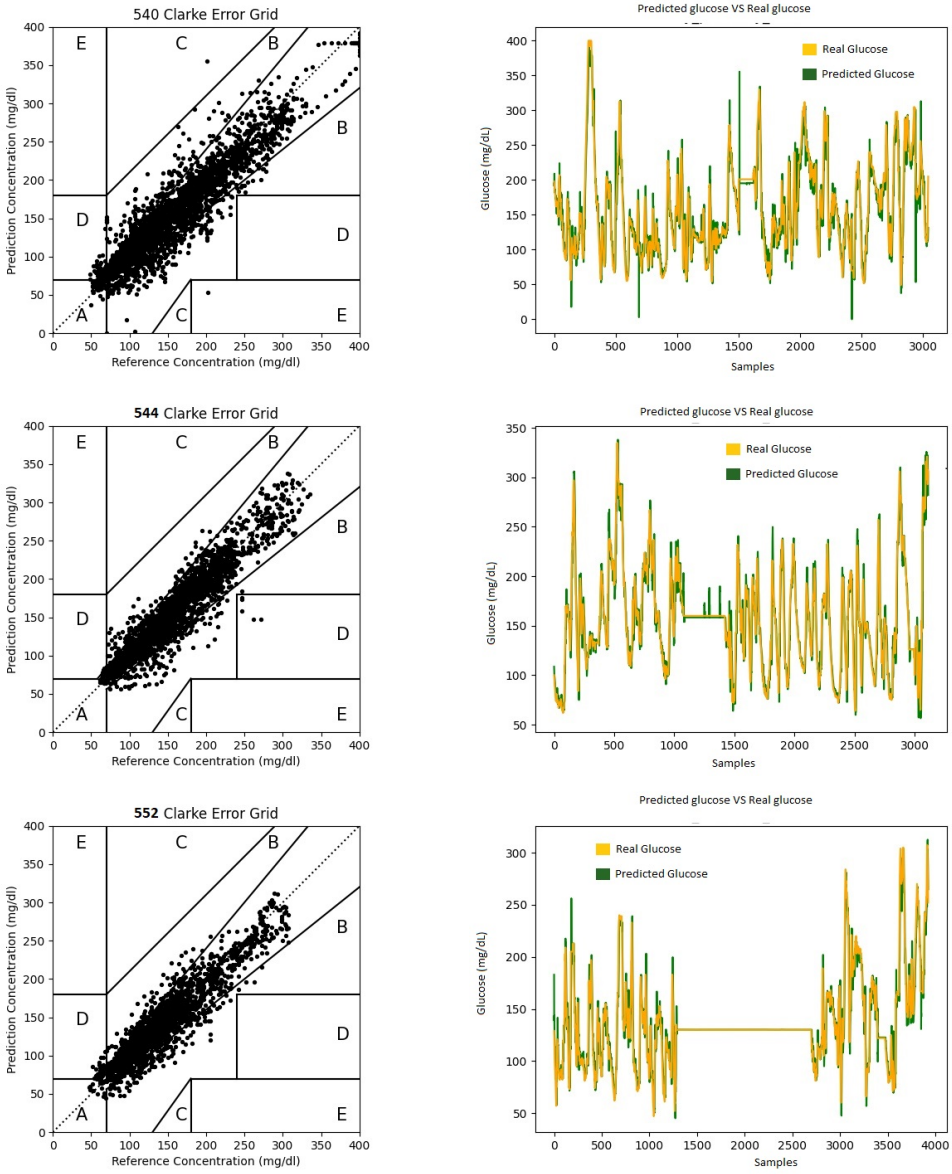

Figure 1: CEGs (left) and forecasting glucose results (right - in yellow and green the actual $G(t)$ and the estimated $\widehat{G}(t)$ respectively) on the testing set for the patients 540, 544 and 552.

We set as null hypothesis H_0 the fact that all the algorithms are statistically equivalent, i.e., that the results of any two algorithms are statistically the same. We run the tests at a significance level of 0.05, which means that there is a 5% of possibility of incorrectly rejecting the null hypothesis.

Table 6 shows the results of the execution of this test. Each column contains, for each algorithm, its rank value. For Quade test, the lower this value, the better the algorithm. Therefore, we can see that the best–performing algorithm is that in paper [41], the runner–up being that in paper [4].

Table 6: The results of the Quade test for the five algorithms.

Paper	**[41]**	**[4]**	**[17]**	**[38]**	**NEAT**
Score	2.067	2.200	3.000	3.267	4.467

statistic: 1.759 **p–value**: 0.186

It is important to note that the p–value is higher than the significance level, which means that the hypothesis H_0 is accepted, so we cannot exclude that all the algorithms are statistically equivalent.

From Table 6, we can see NEAT is in the last position, yet this by itself does not automatically imply that it is inferior to the other ones. Rather, this should be checked through the use of post–hoc procedures. Several such procedures exist, and we have used them all here.

Given that our aim is to investigate the quality of the NEAT algorithm, we choose it as the control method. This means that we contrast it against all the other algorithms.

Table 7 reports the outcome of the use of all these post–hoc procedures. In the table, each column from the third onward reports, for each procedure, the adjusted p–value for the comparison between the i–th algorithm and NEAT.

Table 7: Post–hoc procedure for Quade test over the five algorithms (NEAT being the control algorithm).

Algorithm	statistic	Bonferroni	Holm	Finner	Hochberg	Li
[41]	1.535	0.499	0.499	0.413	0.696	0.183
[4]	1.450	0.589	0.499	0.413	0.696	0.209
[17]	0.938	1.000	0.696	0.435	0.696	0.385
[38]	0.768	1.000	0.696	0.443	0.443	0.443

Each of the situations reported in Table 7 yields that the H_0 hypothesis is always accepted.

In conclusion, although the quantitative results provided by NEAT are a bit inferior to those from the other algorithms, there is no difference from a statistical viewpoint.

4.5 Discussion

An aspect that should be emphasized is the structure of the discovered network models. The neural networks proposed by the participants in the BGLP challenge have quite complex structures, although they perform really well. Instead, the NEAT algorithm presents the capability to maintain a low complexity of the individual structure. As an example, Fig. 2 depicts the most complex discovered structure, i.e. that related to best performing neural networks for the patient 584. The green lines represent a positive weight while the red lines represent a negative weight: the thicker the line, the bigger the value of the associated weight, as to specify the importance of that link. The hidden nodes are depicted in grey. From the figure it can be noted that a sort of feature selection of the input nodes used to feed the network is automatically performed. Moreover, it is easy to ascertain that even such an extracted structure is extremely simple. In fact, it is characterized by few hidden layers and few nodes per layer. This aspect represents an advantage with respect to more complex networks deriving from non-neuroevolution approaches. Just to better understand how simpler NEAT models are, suffice it to report here that the BGLP best–performing paper, i.e. [30], is based on a recurrent neural network architecture containing ten blocks with four layers each, with 512 hidden units output by each layer. Moreover, the BGLP runner–up paper, i.e., [41], relies on a three–layer recursive neural network with 32 hidden units together with LSTM units and gated recurrent units (GRUs). The difference in complexity between those architectures and ours is striking.

To further validate this low-complexity statement, in Table 8 the prediction features of neuroevolution models extracted for each patient are reported. From the table it is evident that the number of the features with an active rule in the prediction is limited. The most important prediction feature results to be the actual value of glucose level $G(t)$ that is present in all the models, the second being $G(t-5)$ that is present in four out of the six cases. On average, each model contains 6.83 items, of which 3.33 related to glucose, 1.33 to insulin, and 2.17 to carbohydrates.

In addition, the last column of Table 8 shows the number of the hidden nodes (hn) contained in the six best networks. The highest value is 6, the lowest being 0 in two cases. Their average number is 2.17, which proves that the inner structures of the proposed networks are characterized by a very low complexity.

Table 8: Structures of the personalized neuroevolution models.

ID	Function	hn
540	$\widehat{G}(t+30) = f(G(t), G(t-5), I(t-50), C(t-50))$	2
544	$\widehat{G}(t+30) = f(G(t), G(t-15), G(t-25), G(t-50), I(t-10), C(t), C(t-30), C(t-35), C(t-40), C(t-55))$	1
552	$\widehat{G}(t+30) = f(G(t), G(t-15), G(t-55), I(t-50))$	4
567	$\widehat{G}(t+30) = f(G(t), G(t-5), G(t-10), I(t-15), C(t-30))$	0
584	$\widehat{G}(t+30) = f(G(t), G(t-5), G(t-10), G(t-40), G(t-45), I(t-55), C(t-5), C(t-15), C(t-40), C(t-50), C(t-55))$	6
596	$\widehat{G}(t+30) = f(G(t), G(t-5), G(t-15), I(t), I(t-15), I(t-30), C(t-10))$	0

Another issue to underline is that the low complexity of the resulting personalized predictive models discovered through the neuroevolution approach gives the possibility to embed a control algorithm, in addition to monitoring services, as an application on a smart device. This control algorithm is responsible for the computation of the future amount of insulin to deliver, and for driving an insulin pump in the injection of the needed amount of insulin boluses for the patient. The possibility to have at disposal this control on a smart device can assure that the insulin administration is guaranteed even in conditions when a possible remote controller is either unreachable (absence of network) or not working (computing system power failure/unavailability). This allows patients to better manage blood glucose levels during the day. Furthermore, this low complexity, requiring a low computational power, permits saving the battery device charge, which is very important to avoid diabetic patients to remain without help.

5 CONCLUSIONS

In this work, we have presented a neuroevolutionary approach for the extraction of a forecasting model for the estimation of the future blood glucose values to be integrated in an AP as the control algorithm. The prediction accuracy has been evidenced by performing an evaluation of the discovered models over real data for a time horizon of 30 minutes.

Furthermore, the low computational complexity yields the personalized neuroevolution-extracted models suitable to be executed on low-power devices and also avoiding rapid consumption of their batteries.

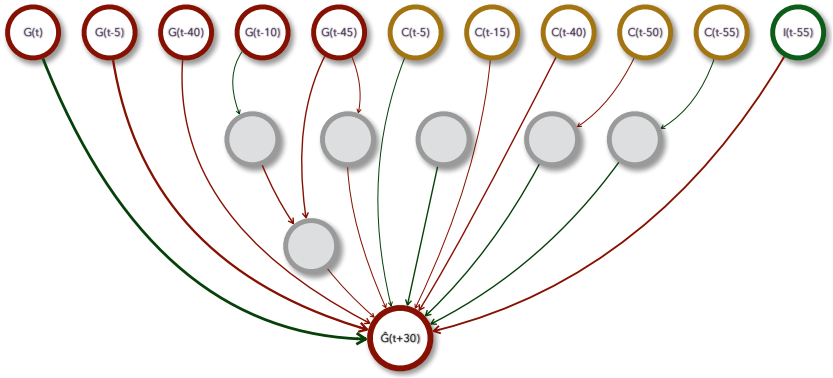

Figure 2: Best 30-minutes ahead predictive network model for the patient 584.

As future works, in addition to a further validation of the forecasting model over a broader set of clinical data, we plan to investigate multi-objective approaches that go beyond the simple measure of error metric but also rely on medical considerations related to the clinical accuracy.

As further objectives, we could consider issues as architecture simplicity, energy consumption, or other issues that could turn out important from the practical implementation point of view.

Also, another further step could consist in an optimization procedure not considering merely numerical prediction ability but also aiming at suggesting suited actions as well as, i.e., to compute minimal dosages and early interventions.

Besides, we aim to improve the model for capturing future glucose trends for diabetic patients over longer time horizons and to assess its performance against other competitive strategies.

REFERENCES

[1] Luis Acedo, Marta Botella, J. C. Cortés, and Ignacio Hidalgo. 2018. Swarm hybrid optimization for a piecewise model fitting applied to a glucose model. *Journal of Systems and Information Technology* 20, 4 (2018), 404–416.

[2] Kerstine Bach, R. Bunescu, C. Marling, and Nirmalie Wiratunga. 2020. Preface. In *5th International Workshop on Knowledge Discovery in Healthcare Data (KDH)*, Kerstin Bach, Razvan Bunescu, Cindy Marling, and Nirmalie Wiratunga (Eds.). 3–5.

[3] R.N. Begman, L.S. Phillips, and C. Cobelli. 1981. Physiologic evaluation of factors controlling glucose tolerance in man: measurement of insulin sensitivity and beta-cell glucose sensitivity from the response to intravenous glucose. *J. Clin Invest.* 68, 6 (1981), 1456–1467.

[4] Robert Bevan and Frans Coenen. 2020. Experiments in non-personalized future blood glucose level prediction. In *5th International Workshop on Knowledge Discovery in Healthcare Data (KDH)*, Kerstin Bach, Razvan Bunescu, Cindy Marling, and Nirmalie Wiratunga (Eds.). 106–110.

[5] Giacomo Cappon, Lorenzo Meneghetti, Francesco Prendin, Jacopo Pavan, Giovanni Sparacino, Simone Del Favero, and Andrea Facchinetti. 2020. A personalized and interpretable deep learning based approach to predict blood glucose concentration in Type 1 diabetes. In *Proc. 5th Int. Workshop Knowledge Discovery in Healthcare Data*, Kerstin Bach, Razvan Bunescu, Cindy Marling, and Nirmalie Wiratunga (Eds.). Santiago de Compostela, Spain, August 29-30, 81–85.

[6] W. L. Clarke, D. Cox, L.A. Gonder-Frederick, W. Carter, and S.L. Pohl. 1987. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care* 10 (1987), 622–628.

[7] Iván Contreras, Silvia Oviedo, Martina Vettoretti, Roberto Visentin, and Josep Vehí. 2017. Personalized blood glucose prediction: a hybrid approach using grammatical evolution and physiological models. *Plos One* 12, 11 (2017), e0187754.

[8] John Daniels, Pau Herrero, and Pantelis Georgiou. 2020. Personalised glucose prediction via deep multitask networks. In *Proc. 5th Int. Workshop Knowledge Discovery in Healthcare Data*, Kerstin Bach, Razvan Bunescu, Cindy Marling, and Nirmalie Wiratunga (Eds.). Santiago de Compostela, Spain, August 29-30, 116–120.

[9] I. De Falco, A. Della Cioppa, A. Giugliano, A. Marcelli, T. Koutny, M. Krcma, U. Scafuri, and E. Tarantino. 2019. A genetic programming-based regression for extrapolating a blood glucose-dynamics model from interstitial glucose measurements and their first derivatives. *Appl. Soft Comput.* 77 (2019), 316–328.

[10] I. De Falco, A. Della Cioppa, T. Koutny, M. Krcma, U. Scafuri, and E. Tarantino. 2018. An evolutionary approach for estimating the blood glucose by exploiting interstitial glucose measurements. In *Proc. 11th Int. Conf. Health Informatics, HEALTHINF - Part of the 11th Int. Joint Conf. Biomedical Engineering Systems and Technologies, BIOSTEC 2018*, Vol. 5. SCITEPRESS, Funchal, Madeira, Portugal, January 19-21, 625–632.

[11] I. De Falco, A. Della Cioppa, T. Koutny, M. Krcma, U. Scafuri, and E. Tarantino. 2018. Genetic programming-based induction of a glucose-dynamics model for telemedicine. *Journal of Network and Computer Applications* 119 (2018), 1–13.

[12] I. De Falco, A. Della Cioppa, T. Koutny, M. Krcma, U. Scafuri, and E. Tarantino. 2020. A Grammatical Evolution Approach for Estimating Blood Glucose Levels. In *Proceedings of the EEE Globecom Workshops*. IEEE Press, Taipei, Taiwan, 7-11 December, 1–6.

[13] I. De Falco, A. Della Cioppa, U. Scafuri, and E. Tarantino. 2017. Accurate estimate of blood glucose through interstitial glucose by genetic programming. In *Proc. Int. Symp. Computers and Communications*. IEEE Press, Heraklion, Greece July 3–7, 284–289.

[14] J Demsar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 7 (2006), 1–30.

[15] J Derrac, S Garcia, D Molina, and F Herrera. 2011. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation* 1 (2011), 3–18.

[16] S. Ghosh. 2015. A differential evolution based approach for estimating minimal model parameters from IVGTT data. *Comput. Biol. Med.* 46 (2015), 51–60.

[17] Hadia Hameed and Samantha Kleinberg. 2020. Investigating potentials and pitfalls of knowledge distillation across datasets for blood glucose forecasting. In *5th International Workshop on Knowledge Discovery in Healthcare Data (KDH)*, Kerstin Bach, Razvan Bunescu, Cindy Marling, and Nirmalie Wiratunga (Eds.). 91–95.

[18] J. I. Hidalgo, M. Botella, J. M. Velasco, O. Garnica, C. Cervigón, R. Martínez, A. Aramendi, E. Maqueda, and J. Lanchares. 2020. Glucose forecasting combining markov chain based enrichment of data, random grammatical evolution and bagging. *Appl. Soft Comput.* 88 (2020), 1–12.

[19] J. I. Hidalgo, J. M. Colmenar, G. Kronberger, S. M. Winkler, O. Garnica, and J. Lanchares. 2017. Data based prediction of blood glucose concentrations using evolutionary methods. *J. Med. Syst. 2017* 41, 9 (2017), 142.

[20] Roman Hovorka, Valentina Canonico, Ludovic J. Chassin, Ulrich Haueter, Massimo Massi-Benedetti, Marco Orsini Federici, Thomas R. Pieber, Helga C. Schaller, Lukas Schaupp, Thomas Vering, and Malgorzata E. Wilinska. 2004. Nonlinear model predictive control of glucose concentration in subjects with Type 1 diabetes. *Physiol. Meas.* 25 (2004), 905–920.

[21] David Joedicke, Oscar Garnica, Gabriel Kronberger, José Manuel Colmenar, Stephan Winkler, Jose Manuel Velasco, Sergio Contador, and Ignacio Hidalgo. 2020. Analysis of the performance of genetic programming on the blood glucose level prediction challenge 2020. In *5th International Workshop on Knowledge Discovery in Healthcare Data (KDH)*, Kerstin Bach, Razvan Bunescu, Cindy Marling, and Nirmalie Wiratunga (Eds.). 147–151.

[22] K. Li, J. Daniels, C. Liu, P. Herrero-Vinas, and P. Georgiou. 2019. Convolutional recurrent neural networks for glucose prediction. *IEEE J. Biomed. Health Inform.* 24 (2019), 603–613.

[23] Ning Ma, Yuhang Zhao, Shuang Wen, Tao Yang, Ruikun Wu, Rui Tao, Xia Yu, and Hongru Li. 2020. Online blood glucose prediction using autoregressive moving average model with residual compensation network. In *5th International Workshop on Knowledge Discovery in Healthcare Data (KDH)*, Kerstin Bach, Razvan Bunescu, Cindy Marling, and Nirmalie Wiratunga (Eds.). 152–156.

[24] C. Marling and R. Bunescu. 2020. The OhioT1DM dataset for blood glucose level prediction: Update 2020. In *5th International Workshop on Knowledge Discovery in Healthcare Data (KDH)*, Kerstin Bach, Razvan Bunescu, Cindy Marling, and Nirmalie Wiratunga (Eds.). 76–80.

[25] Alan McIntyre, Matt Kallada, Cesar G. Miguel, and Carolina Feher da Silva. 2019. NEAT-Python. https://github.com/CodeReclaimers/neat-python. (2019).

[26] Alan McIntyre, Matt Kallada, Cesar G. Miguel, and Carolina Feher da Silva. 2020. NEAT-Python Documentation: Configuration file description. (2020). https://neat-python.readthedocs.io/en/latest/config_file.html

[27] M. Munoz-Organero. 2020. Deep physiological model for blood glucose predictions in T1DM patients. *Sensors* 20, 14 (2020), 3896.

[28] A. Nath, S. Biradar, A. Balan, R. Dey, and R. Padh. 2018. Physiological models and control for Type 1 diabetes mellitus: a brief review. *IFAC Papers-OnLine* 51, 1 (2018), 289–294.

[29] C. Pérez-Gandía, A. Facchinetti, G. Sparacino, C. Cobelli, E.J. Gómez, M. Rigla, A. de Lieiva, and M. E. Hernando. 2010. Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. *Diabetes Technol. Ther.* 12, 1 (2010), 81–88.

[30] Harry Rubin-Falcone, Ian Fox, and Jenna Wiens. 2020. Deep residual time series forecasting: Application to blood glucose prediction. In *5th International Workshop on Knowledge Discovery in Healthcare Data (KDH)*, Kerstin Bach, Razvan Bunescu, Cindy Marling, and Nirmalie Wiratunga (Eds.). 111–115.

[31] Kyriaki Saiti, Martin Macaš, Kateřina Štechová, Pavlína Pithová, and Lenka Lhotská. 2017. A review of model prediction in diabetes and of designing glucose regulators based on model predictive control for the artificial pancreas. In *Lecture Notes in Computer Science*. Vol. 10433. Springer Nature Switzerland, 66–81.

[32] G. Sparacino, F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, and C. Cobelli. 2007. Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series. *IEEE Trans. Biomed. Eng.* 54, 5 (2007), 931–937.

[33] Kenneth O. Stanley, Jeff Clune, Joel Lehman, and Risto Miikkulainen. 2019. Designing neural networks through neuroevolution. *Nature Machine Intelligence* 1 (2019), 24–35.

[34] Kenneth O. Stanley and Risto Miikkulainen. 2002. Evolving Neural Networks Through Augmenting Topologies. *Evolutionary Computation* 10, 2 (2002), 99–127.

[35] Xiaoyu Sun, Mudassir Rashid, Mert Sevil, Nicole Hobbs, Rachel Brandt, Mohammad Reza Askari, Andrew Shahidehpour, and Ali Cinar. 2020. Prediction of blood glucose levels for people with type 1 diabetes using latent-variable-based model. In *5th International Workshop on Knowledge Discovery in Healthcare Data (KDH)*, Kerstin Bach, Razvan Bunescu, Cindy Marling, and Nirmalie Wiratunga (Eds.). 121–126.

[36] Jose M. Velasco, Sergio Contador, Oscar Garnica, and Ignacio Hidalgo. 2017. Forecasting glucose levels in patients with diabetes mellitus using semantic grammatical evolution and symbolic aggregate approximation. In *Proc. Genetic and Evolutionary Computation Conference Companion*. ACM, Berlin, Germany, July 15–19, 1387–1394.

[37] R. Visentin, E. Campos-Náñez, M. Schiavon, D. Lv, M. Vettoretti, M. Breton, B. P. Kovatchev, C. D. Man, and C. Cobelli. 2018. The UVA/PADOVA Type 1 diabetes simulator goes from single meal to single day. *J. Diabetes Sci. Technol.* 12, 2 (2018), 273–281.

[38] Tao Yang, Ruikun Wu, Rui Tao, Shuang Wen, Ning Ma, Yuhang Zhao, Xia Yu, and Hongru Li. 2020. Multi-scale long short-term memory network with multi-lag structure for blood glucose prediction. In *5th International Workshop on Knowledge Discovery in Healthcare Data (KDH)*, Kerstin Bach, Razvan Bunescu, Cindy Marling, and Nirmalie Wiratunga (Eds.). 142–146.

[39] K. Zarkogiovanni, K. Mitsis, E. Litsa, M.-T. Arredondo, G. Fico, A. Fioravanti, and K. S. Nikita. 2015. Comparative assessment of glucose prediction models for patient with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring. *Med. Bio. Eng. Comput.* 53 (2015), 1333–1343.

[40] C. Zecchin, A. Facchinetti, G. Sparacino, and C. Cobelli. 2014. Jump neural network for online short-time prediction of blood glucose from continuous monitoring sensors and meal information. *Comput. Meth. Prog. Bio.* 113, 1 (2014), 144–152.

[41] T. Zhu, X. Yao, K. Li, P. Herrero, and P. Georgiou. 2020. Blood glucose prediction for Type 1 diabetes using generative adversarial networks. In *Proc. 5th Int. Workshop Knowledge Discovery in Healthcare Data*, Kerstin Bach, Razvan Bunescu, Cindy Marling, and Nirmalie Wiratunga (Eds.). Santiago de Compostela, Spain, August 29-30, 96–1000.