A Personalized and Interpretable Deep Learning Based Approach to Predict Blood Glucose Concentration in Type 1 Diabetes

Giacomo Cappon[1], Lorenzo Meneghetti[1], Francesco Prendin[1], Jacopo Pavan[1], Giovanni Sparacino[1], Simone Del Favero[1], Andrea Facchinetti[1]

Abstract.

The management of type 1 diabetes mellitus (T1DM) is a burdensome life-long task. In fact, T1DM individuals are request to perform every day tens of actions to adapt the insulin therapy, aimed at maintaining the blood glucose (BG) concentration as much as possible into a safe range coping with the day-to-day variability of their life style. The recent availability of continuous glucose monitoring (CGM) devices and other low-cost wearable sensors to track important vital and activity signals, is stimulating the development of decision support systems to lower this burden. Modern deep learning models, trained using rich amount of information, are a suitable and effective instrument for such purpose, especially if used to predict future BG values. However, the high accuracy of deep learning approaches is often obtained at the expense of less interpretability.

To surpass this limit, in this work we propose a new deep learning method for BG prediction based on a personalized bidirectional long short-term memory (LSTM) equipped with a tool that enables its interpretability. The OhioT1DM Dataset was used to develop a model targeting future BG at 30 and 60 minute prediction horizons (PH). The accuracy of model predictions was evaluated in terms of root mean square error (RMSE), mean absolute error (MAE), and the time gained (TG) to anticipate the actual glucose concentration.

The obtained results show fairly good prediction accuracy (for PH = 30/60 min): RMSE = 20.20/34.19 mg/dl, MAE = 14.74/25.98 mg/dl, and TG = 9.17/18.33 min. Moreover, we showed, in a representative case, that our algorithm is able to preserve the physiological meaning of the considered inputs.

In conclusion, we built a model able to provide reliable glucose performance ensuring the interpretability of its output. Future work will assess model performance against other competitive strategies.

1 INTRODUCTION

Diabetes is a chronical metabolic disease in which patients are no longer able to effectively control blood glucose (BG) concentration [2]. In particular, type 1 diabetes mellitus (T1DM) is characterized by an autoimmune attack on the pancreatic β-cells resulting to impaired insulin production. As a consequence, people with T1DM are required to manage their glycemia to keep it within the safe range (i.e. BG $\in$ [70, 180] mg/dl) without incurring in dangerous complications induced by hypoglycemia (BG $<$ 70 mg/dl) and hyperglycemia (BG $>$ 180 mg/dl). Such a burdensome process can be eased by integrating in T1DM therapy newly developed decision support algorithms [15] [3]. Specifically, methodologies based on deep learning aimed to predict future BG levels [6] represent a unique way to equip people with T1DM with an effective tool to proactively tackle the shortcoming of adverse events.

The increasing amount of data that can be easily collected by sensors continuously monitoring BG levels (CGM), insulin infusion, and physical activity, just to mention a few, enables researchers to build new BG prediction algorithms that are effective, personalized, and able to empower T1DM management [4]. In particular, in 2020, Marling et al. [13] started the second edition of the Blood Glucose Level Prediction (BGLP) Challenge, i.e., an open competition aimed to promote and facilitate research in this field. Alongside with the competition, the second version of the so-called OhioT1DM Dataset was released. In particular, by including CGM recordings, insulin infusion logs, daily event reporting, and patient vital parameters' monitoring, this dataset represents a unique source of data that can be used for the purpose.

In this paper, we present a new BG level prediction method based on deep learning that we developed and submitted to the second BGLP Challenge. Specifically, given the complexity of the problem at hand and the "temporal" nature of the feature set, here we trained a long-short term memory (LSTM) [11] neural network targeting future BG levels. Even if recurrent neural networks such as LSTMs are known to achieve good performance for the specific task of BG prediction [14], they lack of interpretability. In fact, when developing models for T1DM decision support, there is the need of providing transparent models able to produce reliable but also interpretable predictions [1]. To the best of our knowledge, current state-of-the-art algorithms for BG prediction based on LSTMs have never been interpreted to explain the model "rationale" behind its outcomes. As such, the aim being equipping our model with this feature, we exploited SHapley Additive exPlanations (SHAP), i.e., a newly developed approach to interpret deep learning model predictions [12]. This represents a novelty in the field and offers useful insights on the use of recurrent neural networks for T1DM management.

2 DATASET PREPARATION

2.1 Dataset description and preprocessing

The model was trained and evaluated on data obtained from the updated OhioT1DM Dataset developed by Marling et al. [13]. In the specific, data from 6 people with T1DM were provided. These

[1] University of Padova, Department of Information Engineering, Padova Italy, email: {cappongi, meneghet, prendinf, pavanjac, gianni, sdelfave, facchine}@dei.unipd.it

anonymous people (numbered as 540, 544, 552, 567, 584, and 596) wore Medtronic 530G and 630G insulin pumps and Medtronic Enlite CGM sensors during an 8-week data-collection period. They reported their meals and other life-event data (time of exercise, sleep, work, stress, and illness) via a custom smartphone app. Furthermore, additional physiological data were collected by a Empatica fitness band, including galvanic skin response, skin temperature, and magnitude of acceleration.

In the training dataset, several intervals of missing values were observed. Such discontinuities reduce the number of training data available but also compromise the dynamical structure of the data, thus causing a bad impact on the training procedure. Because of this, a first order interpolation was performed, on the training set only, on the missing portions that were shorter than 30 minutes.

The data were re-sampled onto a uniform time grid with regular intervals of 5 minutes for training and testing the model. Each sample is placed in the new grid at the closest timestamp with respect to its original timestamp. The final prediction obtained was then realigned to the original timestamps by reassigning every predicted sample to the original timestamps, inverting the re-sampling procedure.

2.2 Feature extraction

Deep learning models, such as the one used in this work, are able to deal with raw data without resorting to manual feature engineering. However, this is in general true when large amount of data are used for their training. Therefore, given the limited size of the dataset at hand, we resorted to manual feature engineering. This is furtherly substantiated by several tests that we performed during our study (not reported here for the sake of simplicity), which confirmed that, using the extracted features described in the following, we were able to improve model performance.

An initial observation of the data revealed that the information registered by the fitness band were partial or incomplete in the majority of the people. Therefore, we decided to discard these signals. As such, along with the CGM measurements, we considered the following signals as input to our predictive algorithm: the injected insulin as reported by the pump, the reported meals and the self-reported physical exercise.

Since whenever a meal is consumed, an insulin bolus is injected to counter the post-prandial hyperglycemic excursion the two signals (meals and insulin) tend to be highly correlated. Therefore, to try to overcome this problem, we generated a new signal consisting of only the correction boluses (INS_C), determined as the injections of insulin that are administered at a time of minimum 90 minutes after a meal.

A consumed meal or an injected insulin bolus do not impact the BG levels immediately. Instead, their effect can only be observed after a minimum time of 30-60 minutes. Similarly, the impact of physical activity has a delayed effect on the BG levels [16]. Because of this, the signals of injected insulin (INS), INS_C, reported meals (MEA) and physical activity (PA) are transformed to better account for the underlying physiological dynamics. The transformation consisted of a 2^{nd} order low-pass filtering with impulse response $h(t) = \lambda t e^{-\lambda t}$, where we set λ=0.02. This procedure has been adopted in literature to produce feature sets for the development of ML algorithms for T1DM decision support [3] [15]. Additionally, a transformation of the CGM signal is obtained using the dynamic risk [7], which empowers the model with additional features that capture the dynamics of the CGM signal (e.g., glycemic variability).

In summary, the following features were considered: CGM, DR, INS, INS_C, MEA, and PA.

3 METHODS

3.1 A Bidirectional LSTM to Predict Future BG

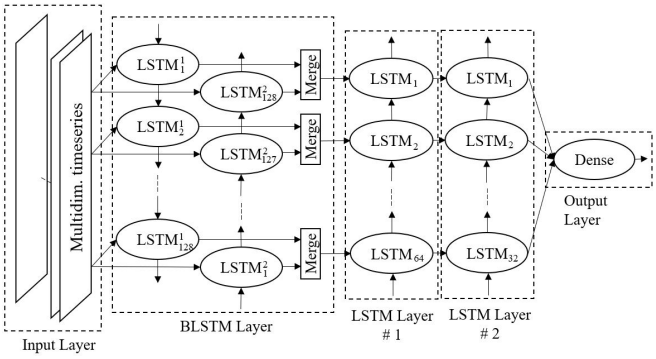

Figure 1. Scheme of the implemented bidirectional LSTM.

As introduced, BG level prediction is a very challenging and complex task. By analyzing the nature of our dataset, it is natural to think that proper modeling of the temporal between-feature dependencies is crucial to effectively solve the problem at hand. For this reason, in this work we decided to adopt an LSTM-based model architecture since LSTMs are well-known in the literature to be the ideal choice to build a predictive model for time series [9]. An LSTM consists of a set of recurrently connected blocks, known as LSTM memory cells. Each LSTM cell consists of an input gate, an output gate, and a forget gate. Each of the three gates can be thought of as a neuron, and each gate achieves a particular function in the cell. In particular, LSTMs are able to exploit learned temporal dependencies to predict the future output according to their previous states, thus well-fitting the purpose of this work. A common drawback of LSTM networks is that, by processing the input in a temporal order, they tend to produce as output, something that is strongly based on forwards dependencies only. To solve this issue, a bidirectional LSTM can be exploited [8]. Briefly, it consists of presenting, to two parallel LSTMs, each training sequence forwards and backwards and then merging the LSTMs outputs to obtain the resulting target estimate. As such, this allows to learn potentially richer representations and capture patterns that may have been missed by the chronological-order version alone. Moreover, the use of bidirectional LSTMs for BG level prediction allowed to obtained promising results in several seminal works [17][18].
The final model architecture, shown in Figure 1 and hereafter labeled as BLSTM, consists of a four-layer neural network: a bidirectional LSTM input layer composed of 128 cells having a look back period of 15 minutes (i.e. 3 samples), two LSTM layers respectively composed of 64 and 32 cells, and a fully connected layer consisting of a single neuron computing the BG level prediction at two different prediction horizons (PH), i.e. 30 and 60 min. BLSTM architecture, hyperparameters, and look back period have been chosen by trial-and-error to compromise between model complexity and accuracy. The BLSTM is implemented in Python using the Keras library [5].

3.2 Equipping BLSTM with interpretability

New algorithms for decision support in T1D management require to be interpretable [1] to avoid potentially adverse or even life-threatening consequences. Unlike traditional physiological-based strategies, deep learning models (such as LSTMs) are black-boxes,

meaning that their high accuracy is often achieved by learning complex relationships that even experts struggle to interpret. For black box models to be adopted in the field of T1D, it is thus desirable to understand whether or not they retain the physiological significance of the inputs they use.

In this work, we aim to overcome the issue of interpretability by analysing our BLSTM with a novel unified approach to interpret model predictions, SHAP [12]. SHAP is a newly developed game theoretical approach to explain how much a given feature impacts on model prediction (compared to if we made that prediction at some baseline value of that feature). By this method, we were able to fully interpret the BLSTM. Indeed, SHAP allowed to both visualize the feature importance and what is driving it.

3.3 Software framework

For each subject and considered PH we trained, thus personalized, a different BLSTM model. The training of each BLSTM has been performed through the gradient descent RMSprop algorithm applied in a mini-batch mode [10]. In particular, as schematized in Figure 2, we developed an *ad-hoc* software framework to automatically perform both model training and tuning. In details, in block A, the

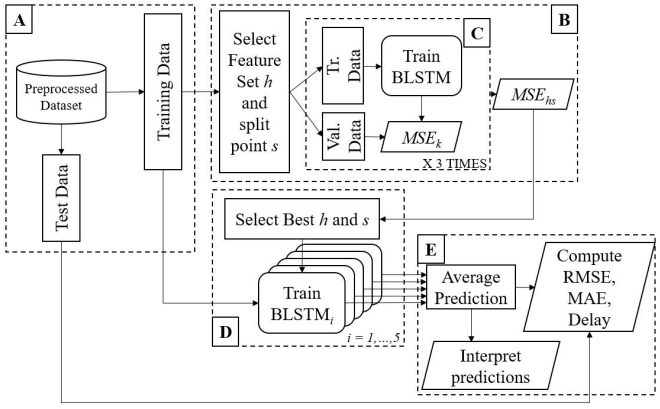

Figure 2. Scheme of the experimental framework.

preprocessed data have been divided into training and test data, respectively. Then, in block B, to optimally tune the BLSTM, feature selection is performed. To do so, we generated the power set of $S = \{\text{DR}, \text{INS}, \text{INS}_C, \text{CHO}, \text{PA}\}$, i.e. the set of all subsets of S, including the empty set and S itself. Then, given its obvious impact on model performance, we constrained each feature subset in the power set to also contain the CGM feature. As a result, we exhaustively examined all the possible sub-sets of features, each containing CGM and other, possibly useful, input features. Block B also splits data into training and validation set. Here, we explored multiple "split points" s, thus assigning $\{50, 60, 70, 80\}\%$ of the data to the training data and the remaining $\{50, 40, 30, 20\}\%$ to the validation data (used to early stop the training of BLSTM in block C to avoid overfitting). For each feature set h in the above-described power set and each considered split point s, the performance of the BLSTM is assessed in terms of mean squared error (MSE_{hs}). To prevent such evaluation from being affected by the random initialization of the BLSTM weights, the whole training and evaluation process is repeated, in block C, three times per feature set. In turn, for each feature set h and split point s we computed MSE_{hs} as:

$$MSE_{hs} = \frac{1}{3}\sum_{k=1}^{3} MSE_k \tag{1}$$

where subscript $k = 1, \ldots, 3$ refers to the repetition at hand. In block D, the best feature set h and split point s are selected as the h and s that obtained the minimum MSE_{hs}. Then, five BLSTMs, namely BLSTM$_i$ $i = 1, \ldots, 5$ are trained on the entire patient/prediction horizon-specific training set. Finally, in block E, we evaluated the model performance by comparing the true BG values in the test set against the respective predictions obtained by averaging each BLSTM$_i$ estimate and we interpret model predictions through SHAP.

4 ASSESSMENT OF BLSTM PERFORMANCE

For the BGLP challenge, the considered metrics for evaluating the accuracy of the obtained prediction are the Root Mean Squared Error, (RMSE) and the Mean Absolute Error (MAE). Considering the prediction error $e(n) = y(n) - \hat{y}(n)$, where $y(n)$ and $\hat{y}(n)$ are the CGM measurements and the computed prediction, respectively, the RMSE and MAE are obtained as follows:

$$RMSE = \sum_N \sqrt{e(t)^2}, \qquad MAE = \text{mean}(|e(t)|)$$

where N is the number of total points.

In this paper, we considered an additional performance metric: the Time Gain (TG), which quantifies the time gained thanks to the prediction. A measure of the average TG is obtained as:

$$TG(y, \hat{y}) = PH - \text{delay}(y, \hat{y})$$

where PH is the prediction horizon used to perform the prediction $\hat{y}$ and the delay$(y, \hat{y})$ between the original and the predicted profiles quantified by the temporal shift k that minimizes the distance between y and $\hat{y}$:

$$\text{delay}(y, \hat{y}) = \underset{k}{\text{argmin}} \sum_{i=1}^{N} \left(y(i) - \hat{y}(i-k)\right)^2$$

5 RESULTS

5.1 BLSTM performance in terms of prediction accuracy

In Figure 3, we present an example of the prediction obtained on a representative subject (544). In the top panel, we report in blue the actual CGM measurements and in red the prediction performed by the BLSTM; in the bottom panel, we report the consumed meals (in grams) as reported by the subject. Albeit affected by the CGM signal noise, the prediction is able to follow the CGM measurements during the post-prandial rises with minor delay. Predicting hypoglycemic episodes with high accuracy resulted to be one of the harder task (an example of inaccurate prediction can be seen at around 12:20). A possible explanation for this is that hypoglycemic episodes are sporadic events which do not happen often, therefore the BLSTM may not have enough training data to learn how to predict similar patterns occurring in the test set.

In Table 1, we report the optimal feature sets that were identified on the training set, in block C, for each subject and PH. The

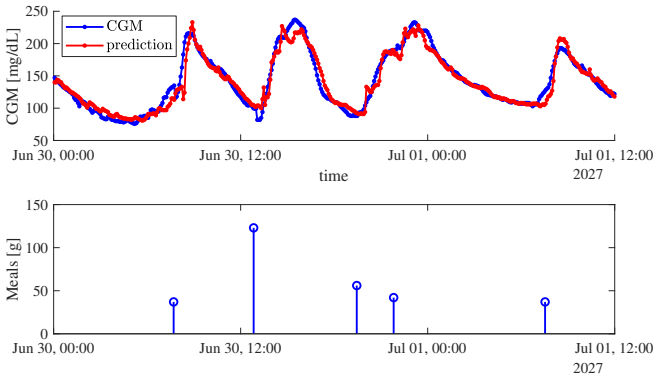

Figure 3. Example of prediction obtained on subject 596. In the top panel, the CGM measurements (blue) and the respective prediction (red). In the bottom panel, the consumed meals reported by the patient.

Table 1. Optimal feature set selected on the training set in block C.

ID	PH = 30 min	PH = 60 min
540	CGM, INS	CGM, DR, INS
544	CGM, DR, INS, MEA, INS_C	CGM, MEA, INS_C
552	CGM, INS, MEA	CGM, INS, INS_C
567	CGM, DR, INS	CGM, INS, PA
584	CGM, DR, INS_C	CGM, INS_C
596	CGM, INS, MEA	CGM, INS, MEA

CGM feature is included in every set by default as described earlier in Section 3.3. The feature INS is adopted in almost every case, expect some where it is replaced by the feature INS_C. The feature MEA is adopted less often, especially in patients where we observed a lower consistency in reporting meals. The feature PA was selected only once, denoting its limited effectiveness in improving the performance of the BLSTM. In general, different PH lead to different features sets for the same patient. This is due to the fact that some features, e.g. MEA, might be relevant, in a specific patient, for PH = 30 min and not for PH = 60 min given their impact on BG level in the very short-term.

In Table 2 we report the RMSE, the MAE and the TG obtained for each subject and PH. A mean RMSE = 20.20 mg/dl is obtained

Table 2. Results obtained on the test-set.

	PH = 30 min			PH = 60 min		
ID	RMSE	MAE	TG	RMSE	MAE	TG
540	23.19	17.33	10	41.41	31.77	20
544	18.88	13.23	15	31.06	22.54	30
552	17.97	13.50	10	31.20	24.48	20
567	21.18	15.20	10	37.40	28.50	20
584	21.91	16.38	5	35.95	27.59	5
596	18.09	12.81	5	28.13	20.99	15
mean	20.20	14.74	9.17	34.19	25.98	18.33

for PH = 30 min, together with a value of MAE = 14.74 mg/dl and TG = 9.17 min. For PH = 60 min, a mean RMSE = 34.19 mg/dl was obtained, together with MAE = 25.98 mg/dl and TG = 18.33 min.

Table 3 reports the number of samples predicted per patient and the percentage of predicted samples over the total CGM samples available. Except for one case, the BLSTM was able to compute a prediction for more than 90% of the samples.

Table 3. Predicted samples per subject.

		PH = 30 min		PH = 60 min	
ID	Total samples	predicted	%	predicted	%
540	2884	2820	97.78	2697	93.52
544	2704	2586	95.64	2638	97.56
552	2352	2275	96.73	2235	95.03
567	2377	2157	90.74	2232	93.90
584	2653	2354	88.73	2473	93.22
596	2731	2683	98.24	2647	96.92

5.2 Model interpretation

As discussed in Section 3.2, thanks to SHAP we are able to interpret each trained BLSTM. The plot in Figure 4 reports the application of SHAP to the BLSTM obtained for patient 596 for a PH of 60 min. This plot is made of many dots. Each dot has three characteristics:

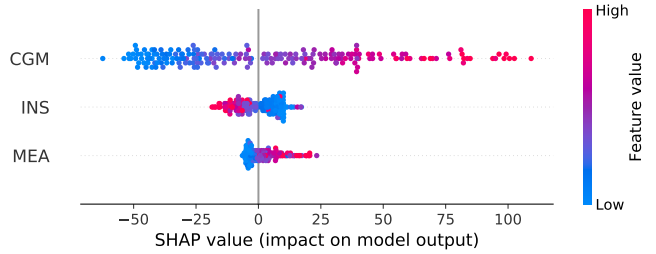

Figure 4. Impact of each input feature on model output obtained via SHAP in patient 596 with PH = 60 min.

vertical location shows what feature it is depicting, the color shows whether that feature assumed an high or low value for that row of the dataset, horizontal location shows whether the effect of that value caused a higher or lower prediction of future BG levels. Results show that high values of CGM translate in high predicted CGM values. On the other hand, high INS impacts negatively on model output mirroring the actual impact of insulin on BG dynamics. Parallely, high MEA induces an increase on predicted glucose values, correctly accounting for the effect of meal intakes on BG level. As such, the physiological meaning of all input features is preserved by the considered representative BLSTM.

For brevity, we do not report the results obtained on other patients, being very similar and consistent with that previously showed.

6 CONCLUSION

The possibility of collecting important vital and activity signals from low-cost wearable sensors in patients with T1DM is calling for the development of individualized proactive decision support systems to lower the daily burden in the application of BG control therapy. In this work, the aim being providing patients with reliable BG predictions, we leveraged the OhioT1DM Dataset to build a new deep learning-based approach for the scope that we submitted for the second edition of the BGLP Challenge. The novelty here is that, beside obtaining fairly good BG predictions considering both a 30 min and a 60 min-long PH, our algorithm is also interpretable. Indeed, the integration of SHAP in our procedure allowed to obtain a "transparent" model where the impact of each feature on model output is explicitly expressed.

The presented study has some limitations that need to be addressed in future work. In particular, we will concentrate on two main issues. First, to fully evaluate its performance, BLSTM will be assessed against other competing baseline and state-of-the-art BG prediction methodologies, e.g., neural networks, random forests, and vanilla LSTMs. Then, we will tackle the limitation represented by the dataset length. In fact, methodologies like LSTMs usually benefit from having more data to be used for their training and tuning. For this purpose, we will investigate the potential advantage of using longer datasets on BLSTM performance.

ACKNOWLEDGEMENTS

Part of this work was supported by MIUR (Italian Minister for Education) under the initiative "Departments of Excellence" (Law 232/2016).

CODE

A repository of the code used in this paper is available online [2].

REFERENCES

[1] M. A. Ahmad, C. Eckert, A. Teredesai, and G. McKelvey, 'Interpretable machine learning in healthcare', *2018 IEEE International Conference on Healthcare Informatics (ICHI), New York, NY*, 447, (2019).
[2] American Diabetes Association, 'Diagnosis and classification of diabetes mellitus: : Standards of medical care in diabetes', *Diabetes Care*, **43**, S14–S31, (2020).
[3] G. Cappon, A. Facchinetti, G. Sparacino, P. Georgiou, and P. Herrero, 'Classification of postprandial glycemic status with application to insulin dosing in type 1 diabetes—an in silico proof-of-concept', *Sensors*, **19**(14), 3168, (2019).
[4] G. Cappon, M. Vettoretti, G. Sparacino, and A. Facchinetti, 'Continuous glucose monitoring sensors for diabetes management: A review of technologies and applications', *Diabetes & metabolism journal*, **43**(4), 383–397, (2019).
[5] F. Chollet. Keras. `https://keras.io`, 2015.
[6] I. Contreras and J. Vehi, 'Artificial intelligence for diabetes management and decision support: Literature review.', *Journal of Medical Internet Research*, **20**, e10775, (2018).
[7] C. Fabris, S.D. Patek, and M.D. Breton, 'Exercise and glucose metabolism in persons with diabetes mellitus: perspectives on the role for continuous glucose monitoring', *Journal of Diabetes Science and Technology*, **14**(1), 50–59, (2015).
[8] A. Graves, S. Fernandez, and J. Schmidhuber, 'Bidirectional lstm networks for improved phoneme classification and recognition', *Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005*, 799–804, (2015).
[9] A. Graves, A.R. Mohamed, and G. Hinton, 'Speech recognition with deep recurrent neural networks', *IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649, (2013).
[10] G. Hinton, N. Srivastava, and K. Swersky, 'Overview of mini-batch gradient descent.', *Neural Network for Machine Learning. Coursera Course. Available at: http://www.cs.toronto.edu/ tijmen/csc321/ slides/lecture_slides_lec6.pdf*, (2017).
[11] S. Hochreiter and J. Schmidhuber, 'Long short term memory', *Neural Computation*, **9**, 1735–1780, (1997).
[12] S.M. Lundberg and S.I. Lee, 'A unified approach to interpreting model predictions', *Advances in neural information processing system.*, **30**, 4765–4774, (2017).
[13] C. Marling and R. Bunescu, 'The ohiot1dm dataset for blood glucose level prediction: Update 2020.', *in The 5th International Workshop on Knowledge Discovery in Healthcare Data*, (2020).
[14] J. Martinsson, A. Schliep, B. Eliasson, and O. Mogren, 'Blood glucose prediction with variance estimation using recurrent neural networks', *Journal of Healthcare Informatics Research*, **4**, 1–18, (2020).
[15] L. Meneghetti, G.A. Susto, and S. Del Favero, 'Detection of insulin pump malfunctioning to improve safety in artificial pancreas using unsupervised algorithms', *Journal of Diabetes Science and Technology*, **13**(6), 1065–1076, (2019).
[16] M. Riddell and B.A. Perkins, 'Exercise and glucose metabolism in persons with diabetes mellitus: perspectives on the role for continuous glucose monitoring', *Journal of Diabetes Science and Technology*, **3**(4), 914–923, (2009).
[17] 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 14th Symposium on Neural Networks and Applications (NEUREL)*, 1–5, (2018).
[18] T. Wang and W. Li, 'Blood glucose forecasting using lstm variants under the context of open source artificial pancreas system', *in 53rd Hawaii International Conference on System Sciences*, 3256–3563, (2020).

[2] https://github.com/meneghet/BGLP_challenge_2020