Blood Glucose Prediction for Type 1 Diabetes Using Generative Adversarial Networks

Taiyu Zhu[1], **Xi Yao**[2], **Kezhi Li**[3], **Pau Herrero**[4] and **Pantelis Georgiou**[5]

Abstract. Maintaining blood glucose in a target range is essential for people living with Type 1 diabetes in order to avoid excessive periods in hypoglycemia and hyperglycemia which can result in severe complications. Accurate blood glucose prediction can reduce this risk and enhance early interventions to improve diabetes management. However, due to the complex nature of glucose metabolism and the various lifestyle related factors which can disrupt this, diabetes management still remains challenging. In this work we propose a novel deep learning model to predict future BG levels based on the historical continuous glucose monitoring measurements, meal ingestion, and insulin delivery. We adopt a modified architecture of the generative adversarial network that comprises of a generator and a discriminator. The generator computes the BG predictions by a recurrent neural network with gated recurrent units, and the auxiliary discriminator employs a one-dimensional convolutional neural network to distinguish between the predictive and real BG values. Two modules are trained in an adversarial process with a combination of loss. The experiments were conducted using the OhioT1DM dataset that contains the data of six T1D contributors over 40 days. The proposed algorithm achieves an average root mean square error (RMSE) of 18.34 ± 0.17 mg/dL with a mean absolute error (MAE) of 13.37 ± 0.18 mg/dL for the 30-minute prediction horizon (PH) and an average RMSE of 32.31 ± 0.46 mg/dL with a MAE of 24.20 ± 0.42 for the 60-minute PH. The results are compared for clinical relevance using the Clarke error grid which confirms the promising performance of the proposed model.

1 INTRODUCTION

Diabetes is a chronic metabolic disorder that affects more than 400 million people worldwide with an increasing global prevalence [27]. Due to an absence of insulin production from the pancreatic β cells, people living with Type 1 diabetes (T1D) require long-term self-management through exogenous insulin delivery to maintain blood glucose (BG) levels in a normal range. In this regard, accurate glucose prediction has great potential to improve diabetes management, enabling proactive actions to reduce the occurrence of adverse glycemic events, including hypoglycemia and hyperglycemia.

In recent years, empowered by the advances in wearable devices and data-driven techniques, different BG prediction algorithms have been proposed and validated in clinical practice [29]. Among these, continuous glucose monitoring (CGM) is an essential technology that measures BG levels and provides readings in real-time. CGM has produced a vast amount of BG data with its increasing use in the diabetes population. Taking advantage of this, the emergence of deep learning algorithms for BG prediction has achieved recent success and outperformed several conventional machine learning approaches in terms of accuracy [1, 16, 17, 23, 28]. Generally, the major challenge of BG prediction lies in accounting for the intra- and inter-person variability that leads to various glucose responses under different conditions [25]. Furthermore, many external events and factors can influence glucose dynamics, such as meal ingestion, physical exercise, psychological stress, and illness. Deep learning is powerful at extracting hidden representations from large-scale raw data [15], making it suitable for accounting for the complexity of glucose dynamics in diabetes.

In this work, we propose a novel deep learning model for BG prediction using a modified generative adversarial network (GAN). As a recent breakthrough in the field of deep learning, GANs have shown promising performance on various tasks, such as generating realistic images [13], synthesizing electronic health records [4] and predicting financial time series [31]. Normally, a GAN framework is composed of two deep neural networks (DNNs) models as the generator and the discriminator, respectively. They are trained simultaneously through an adversarial process [10]. The proposed generator captures feature maps of the multi-variant physiological waveform data and generates predictive BG samples, while the discriminator is designed to distinguish the real data from generated ones. To model the temporal dynamics of BG data, we adopt a recurrent neural network (RNN) in the generator and a one-dimensional convolutional neural network (CNN) in the discriminator with dilation factors in each DNN layer to expand receptive fields, which have been verified as adequate network structures for BG prediction in our previous works [5, 17, 33].

2 METHODS

2.1 Dataset and Pre-processing

The data that we used to develop the model is the OhioT1DM dataset, provided by the Blood Glucose Level Prediction (BGLP) Challenge [20, 21]. It was produced by collecting BG-relevant data on 12 people with T1D over an eight-week period. The first half of the cohort released for the 2018 BGLP challenge was used for model pre-training, and we focus on the performance of the rest six individuals that numbered 540, 544, 552, 567, 584, and 596. The dataset contains BG levels collected by CGM readings every five minutes, insulin delivery from insulin pumps, self-reported events (such as meal, work, sleep, psychological stress, and physical exercise) via a smartphone app and physical activity by a sensor band. However,

[1] Imperial College London, UK, email: taiyu.zhu17@imperial.ac.uk
[2] Imperial College London, UK, email: x.yao19@imperial.ac.uk
[3] University College London, UK, email: ken.li@ucl.ac.uk
[4] Imperial College London, UK, email: pherrero@imperial.ac.uk
[5] Imperial College London, UK, email: pantelis@imperial.ac.uk

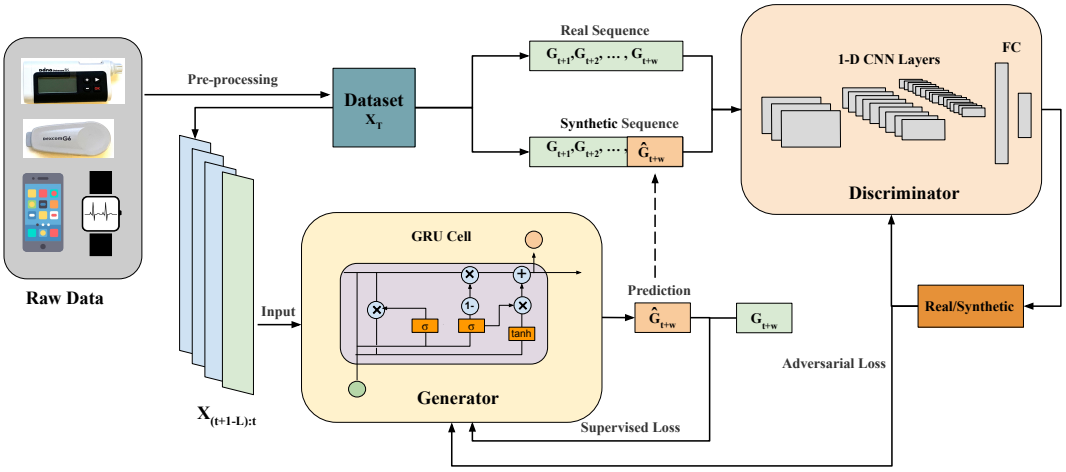


Figure 1: The system architecture of the proposed GAN framework to predict BG levels.

there are unavoidable differences between the collected data and actual physiological states. For example, the CGM sensor measures interstitial fluid glucose level and then estimate BG levels by applying signal processing techniques, such as filtering and calibration algorithms. The meal and insulin are discrete values manually input by users, instead of series of carbohydrates and insulin on board.

It should be noted that the dataset contains many missing gaps and outliers affecting BG levels, both in the training and testing sets, mainly due to CGM signal loss, sensor noise (e.g., compression artifacts), or some usage reasons, such as sensor replacement and calibration. To compensate for some of the missing data, we apply linear interpolation to fill the missing sequences in the training sets, while we only extrapolate missing values in the testing set to ensure that the future information is not involved as partial inputs in the prediction. We then align processed BG samples and other features, e.g. exogenous events, with the same resolution of CGM measurements, and normalize them to form a N-step time series: $\mathbf{X}_N = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{N \times d}$, where $\mathbf{x}$ is a d-dimensional vector mapping the multivariate data at each timestep.

2.2 Problem Formulation

Considering a target prediction horizon (PH) (e.g. 30 or 60 minutes), the goal of the predictor is to estimate the future BG levels G_{t+w} of individuals given past and current physiological states, where w is the number of timesteps determined by PH and CGM resolution (e.g. 5 minutes). Hence, the objective of predictor is consistent with that of GANs, aiming to learn the DNN approximator $\hat{p}$ from the pattern of glucose dynamics p measured in the human body, which can be expressed by the form of the Kullback-Leibler divergence [30]:

$$\min_{\hat{p}} \mathrm{D}((p(G_{t+w}|\mathbf{X}_{1:t})||(\hat{p}(G_{t+w}|\mathbf{X}_{1:t})) \tag{1}$$

where D is a measurement of the distance between distributions. Thus, we need to select highly-related data features to represent the physiological state. Referring to some previous work and hyper-parameter tuning [16, 17, 22, 23], we use $\mathbf{X} \triangleq [\mathbf{G}, \mathbf{M}, \mathbf{I}]$ as the physiological time series, where $\mathbf{G}$ is pre-processed CGM measurements (mg/dL); $\mathbf{M}$ denotes the carbohydrate amount of meal ingestion (g); and $\mathbf{I}$ is the bolus insulin delivery (U). In order to reduce the bias in the supervised learning, we set the changes of BG levels in PH as the training targets of the generator: $\Delta G_t = G_{t+w} - G_t$. Then the predictive BG level $\hat{G}_{t+w}$ from the generator is defined as follows:

$$\hat{G}_{t+w} = f_{\mathcal{G}}(\mathbf{X}_{t+1-L:t}) + G_t \tag{2}$$

where $f_{\mathcal{G}}$ represents the parameters of the generator. Instead of using the whole series, we divide $\mathbf{X}$ into small contiguous sequences with a length of L as a sliding window, then feed them into the deep generative model in a form of mini-batches, aiming at improving stability and generalization of the model [12]. According to the feature selection in [22] and the model validation, we empirically set $L = 18$ which indicates that the input contains 1.5 hours historical data.

2.3 System Architecture

The RNN-based algorithms performed well in BG level prediction in previous studies [1, 23, 28]. Thus, we instantiate a three-layer RNN with 32 hidden units to build the generator, which can be seen as a typical setup of time series GANs [9, 24, 30]. In general, vanilla RNN architecture faces the problem of gradient vanishing and exploding, making it difficult to capture long-term dependencies. Thus, the gated RNN units are proposed to meet this challenge using element-wise gating functions [7], including long short-term memory (LSTM) units [11] and gated recurrent units (GRUs) [6]. Compared to the vanilla RNN, the gated units are able to control the flow of information inside units by a set of gates, which allows an easier backward propagation process. Compared to the LSTM, the GRU was proposed more recently and removed the memory unit. This cell structure uses less parameters and computes the output more efficiently [32] . During the hyper-parameter tuning, GRU-based algorithms also achieved the best predictive outcomes, so we naturally adopt GRU cells in the RNNs .

As depicted in Figure 1, the multi-dimensional input is fed into a RNN with GRU cells given a state length of L. Then the data is processed by a set of hidden neurons to calculate the last cell state C_t. A fully connected (FC) layer with weights $\mathbf{W}_{FC}$ and a bias b_{FC} are used to model the final scalar output: $\Delta\hat{G}_t = \mathbf{W}_{FC}C_t + b_{FC}$. Finally, after adding the current BG level to predictive glucose change, we obtain the output $\hat{G}_{t+w}$.

In general, the prediction performance degrades with the increase of PH, due to the complicated physiological conditions of people with T1D and the uncertainties of exogenous events between t and $t+w$. For instance, if there was a meal intake with large carbohydrate

20-30 minutes before $t+w$, the BG level would raise fast and make the target ΔG_t suddenly increase. These cases occur frequently in the daytime with a large PH, which could affect a supervised learning model to achieve global optimum. This motivated us to make use of the information between t and $t+w$ during the training process to investigate the contiguous glucose change. Therefore, we append the predictive BG level to the end of series $G_{t+1:t+w-1}$ to form a synthetic sequence $\hat{\mathbf{y}}$ and use $G_{t+1:t+w}$ as the corresponding real sequence $\mathbf{y}$. Then we introduce a CNN-based discriminator to extract features and distinguish the real from synthetic sequences, benefiting from the good classification ability of CNNs [15]. There are three one-dimensional (1-D) causal CNN layers employed with rectified linear unit (ReLU) activation and 32 hidden units to compute the final binary output. The discriminator is expected to classify the real and synthetic sequences by 1 and 0, while the generator is pitted against the discriminator and aims to estimate a BG value that is close to the real BG distribution over the PH. Thus the loss of discriminator is computed by cross-entropy. Consequently, this adversarial training contains two loss functions $\mathcal{L}_\mathcal{G}$ and $\mathcal{L}_\mathcal{D}$ for the generator and the discriminator respectively, which are given by

$$\mathcal{L}_\mathcal{G} = \lambda_1 \mathcal{L}_{SL} + \lambda_2 m \sum_{i=1}^{m} \log(1 - f_\mathcal{D}(\hat{\mathbf{y}}^{(i)})), \tag{3}$$

$$\mathcal{L}_\mathcal{D} = \frac{1}{m} \sum_{i=1}^{m} [-\log f_\mathcal{D}(\mathbf{y}^{(i)}) - (\log(1 - f_\mathcal{D}(\hat{\mathbf{y}}^{(i)})))], \tag{4}$$

where $f_\mathcal{D}$ represents the calculation in the discriminator; $\mathcal{L}_{SL}$ is the means square error loss of supervised learning: $\mathcal{L}_{SL} = \sum_{i=1}^{m} (G_{t+w}^{(i)} - \hat{G}_{t+w}^{(i)})^2$; λ_1 and λ_2 are used to adjust the ratio between supervised loss and adversarial loss [31]; and m stands for the mini-batch size. In practice, we employ two separate Adam optimizer [14] to minimize $\mathcal{L}_\mathcal{G}$ and $\mathcal{L}_\mathcal{D}$ with batch size of 512 and learning rate of 0.0001.

Moreover, we introduce dilation to both the RNN and the CNN layers [3, 26], which has shown the promising performance of BG level prediction in previous work [5, 17, 32, 33]. By skipping certain number connections between neurons, the receptive field of the DNN layers can be exponentially increased, which is helpful to capture long-term temporal dependencies in the BG series. In particular, the dilation of layer l is set to $r^l = 2^{l-1}$, increasing from the bottom layer to the top layer. The computation of DNN layers are defined as follows:

$$h_t^{(l)} = f_N(h_{t-r^l}^{(*)}, in_t^{(l-1)}) \tag{5}$$

where $h_t^{(l)}$ and $in_t^{(l-1)}$ are the output and input of layer l at timestep t; f_N denotes the computation in hidden neurons, referring to convolution and cell operation in CNN and RNN layers, respectively. As a feed-forward neural network, the CNN hidden units fetch all the inputs from the layer at a lower level ($* = l-1$), whereas RNNs skip cell state by $r^l - 1$ timesteps to perform the recursive operation ($* = l$).

2.4 Training and Validation

The training and testing sets are separately provided by the BGLP challenge, which contains the data for around 40 and 10 days, respectively. To tune the hyper-parameters by grid search, we validated the models by the same range of hyper-parameters values as in our previous work [32]. We considered many validation methods, such as simple splitting, k-fold cross-validation, and blocked cross-validation [2]. Due to the temporal dependencies and limited size of the training set, we use the last 20% data of the training set to validate the models and guarantee that future information is not involved in current prediction. The early-stop technique is applied to avoid overfitting; we stop the training process when the validation loss keeps increasing. In particular, we set the maximum number of epochs to 3000 with stopping patience of 50. The data sufficiency and overfitting occurrences are further investigated by means of the learning curves.

2.5 Metrics

A set of metrics is applied to evaluate the performance of the GAN model, including root mean square error (RMSE) (mg/dL), mean absolute error (MAE) (mg/dL), which are denoted as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(G_k - \hat{G}_k)^2}, \quad \text{MAE} = \frac{1}{N}\sum_{k=1}^{N}|G_k - \hat{G}_k|, \tag{6}$$

In addition to the RMSE and MAE metrics, we also use the Clarke error grid (CEG) [8], which is a semi-quantitative tool from the clinical perspective. As shown in Figure 2, there are five zones labeled to intuitively reveal the medical consequence based on the prediction results. In general, the data points (BG pairs) in zone A and B are regarded as positive for medical treatment, while the rest (C, D and E) are considered undesirable.

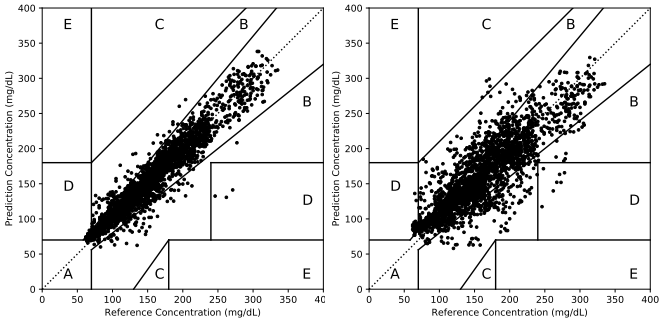
3 RESULTS

After tuning the hyper-parameters, we tested the model on the testing sets. Table 1 shows the RMSE and MAE results for the PH of 30 minutes and 60 minutes. Considering the randomness of the initial weights in DNNs, we conducted 10 simulations and reported results by Mean±SD, where SD is the standard deviation. The average (AVG) RMSE and MAE over all 6 contributors respectively achieve 18.34 ± 0.17 and 13.37 ± 0.18 mg/dL for 30-minute PH, and 32.21 ± 0.46 and 24.20 ± 0.42 for 60-minute PH. The best RMSE and MAE results in experiments are also presented in the last row, which are slightly smaller than the average results. It is noted the standard deviation of multiple simulations is small, which indicates the stability of the model.

Table 1: Prediction performance of the GAN model evaluated on 6 data contributors.

ID	Number	30-minute PH		60-minute PH	
	(#)	RMSE	MAE	RMSE	MAE
540	2884	20.14 ± 0.21	15.22 ± 0.17	38.54 ± 0.46	29.37 ± 0.21
544	2704	16.28 ± 0.11	11.62 ± 0.15	27.64 ± 0.43	20.09 ± 0.38
552	2352	16.08 ± 0.20	12.03 ± 0.22	29.03 ± 0.35	22.47 ± 0.34
567	2377	20.00 ± 0.14	14.17 ± 0.22	35.65 ± 0.41	26.68 ± 0.53
584	2653	20.91 ± 0.08	15.11 ± 0.11	34.31 ± 0.53	25.55 ± 0.52
596	2731	16.63 ± 0.25	12.12 ± 0.23	28.10 ± 0.57	21.06 ± 0.57
AVG		**18.34**	**13.37**	**32.21**	**24.20**
SD		0.17	0.18	0.46	0.42
Best		18.21	13.21	31.64	23.70

To visualize clinical significance between the reference and prediction outcomes, Figure 2 shows the CEG of the contributor 544 that obtains the best statistic performance in Table 1. The specific percentage of the distribution in five regions is presented in Table 2.



(a) 30-minute PH (b) 60-minute PH

Figure 2: The Clarke error grid plots for contributor 544

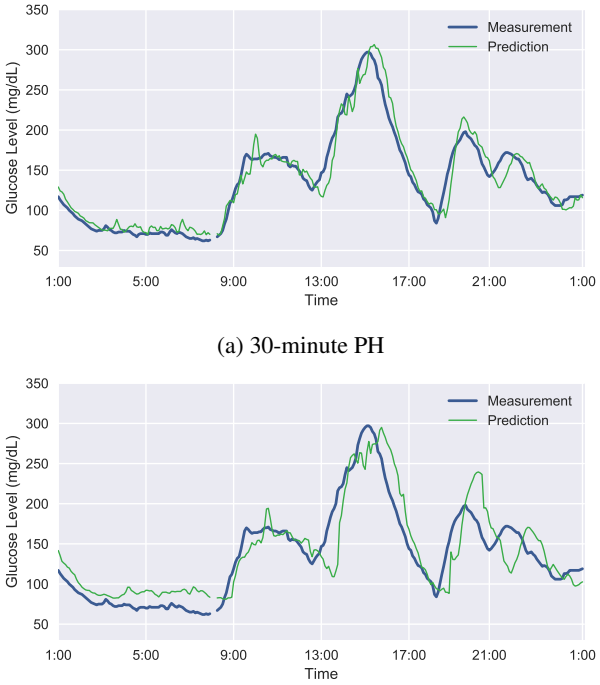
Table 2: The percentage distribution in Clarke error gird (%).

ID	540	544	552	567	584	596
30-minute PH						
CEG_A	86.15	93.91	89.41	89.01	86.75	91.03
CEG_B	12.18	5.76	8.80	10.06	12.26	7.57
CEG_C	0	0	0	0	0	0
CEG_D	1.67	0.33	1.79	0.93	0.98	1.40
CEG_E	0	0	0	0	0	0
60-minute PH						
CEG_A	60.22	79.38	68.01	60.81	69.46	76.60
CEG_B	33.37	19.20	28.91	30.80	28.34	20.78
CEG_C	0.14	0	0	0.25	0.18	0
CEG_D	6.27	1.38	3.08	8.14	2.01	2.62
CEG_E	0	0	0	0	0	0

4 DISCUSSION

As shown in Table 2, the majority of the CEG points are located in zones A and B. These zones signify that the data is within 20% value of the reference, where the treatment suggestions are appropriate regardless of the prediction error. It indicates the high clinical accuracy of the proposed model. The percentage of zone D is small for the 30-minute PH and increases for the 60-minute PH. The points in zone D mean the predictive model missed the hypoglycemia or hyperglycemia events and could lead to poor treatment. In Figure 2b, the most error points are concentrated on the bottom-right corner of the left panel of zone D. It reveals that the model outputs higher predictions when BG levels enter the hypoglycemia region, which is undesirable in the clinical setting. Figure 3 shows the corresponding BG curves for the contributor 544, where the findings from CEG analysis can be validated, and time lags between the predictions and measurements can be observed. The overestimation is observed in several BG regions with low BG levels or a sharp decrease. Aligning the error region with the timesteps, we find that some of the misestimation occurs in nocturnal hypoglycemia. Similar findings are identified by the CEG analysis and BG curves of the other contributors. Therefore, future work will include training and switching between different models for different glucose regions, evaluated by more advanced error grid analysis.

During the experiments, we explored Tikhonov regularization to filter out the outliers in training sets, as described in [1]. However, it was prone to degrade the validation performance but largely reduce the training loss. Then we used the 2018 OhioT1DM dataset [21] and the *in silico* datasets from UVA/Padova T1D simulator [19] for model pre-training. The simulator produced data of an average virtual adult subject with the scenarios defined in [32] over 360 simulated days. The population model was trained by 5 epochs and then fine-tuned



(a) 30-minute PH

(b) 60-minute PH

Figure 3: The comparison between the model predictions and the ground truth of CGM measurements during the first 24-hour period in the testing set of contributor 544. There are three missing BG values between 8:00 and 8:15.

by subject-specific data, but the average validation RMSE slightly increased by around 0.5 mg/dL, compared with the models without pre-training. As shown in Table 1, there are two groups: one including contributors 544, 552, and 596 with better RMSE and MAE performance, and the other including contributors 540, 567 and 584. We introduced the data from the former group to pre-train a population model for the latter group, but the RMSE almost remained unchanged. Thus, one explanation of the pre-training performance is that large inter-person variability exists. For example, in the testing set, contributor 552 has a gap of 1415 missing data points ($\sim$ 5 days), and contributor 567 did not record the meal ingestion, for which we reduced the dimension of the input data. To this end, multiple prepossessing methods are needed to mitigate these missing or incorrect inputs, such as the detection of unannounced meals. In addition, as future work, we consider incorporating personalized physiological and behavioral models [18], such as insulin and carbohydrate on board, to better explain the observed variability.

Compared with the RNN prediction model in our previous work [32], the GAN model achieved better validation performance and smaller RMSE for most of the data contributors in the training process, especially for the 60-minute PH. During the testing phase, the GAN model can output the predictions without using the discriminator. Hence, the complexity of the proposed model is similar to that of the conventional RNN models, which can be easily implemented on smartphone applications [16, 17] to provide real-time predictions and control insulin pump via Bluetooth connectivity. The code corresponding to this work is available at: https://bitbucket.org/deep-learning-healthcare/glugan.

5 CONCLUSION

In this work, a novel deep learning model using a modified GAN architecture is designed to predict BG levels for people with T1D. We developed the personalized models and conducted multiple evaluations for each data contributor in the OhioT1DM dataset. The proposed model achieves promising prediction performance for 30-minute and 60-minute PH in terms of average RMSE and MAE. The CEG analysis further indicates good clinical accuracy, but there are opportunities for enhancement. In particular the model falls short sometimes in capturing a small number of hypoglycemia events. Nevertheless, the model is able capture most of the individual glucose dynamics and has clear potential to be adopted in actual clinical applications.

ACKNOWLEDGEMENTS

The work is supported by EPSRC EP/P00993X/1 and the President's PhD Scholarship at Imperial College London.

REFERENCES

[1] Alessandro Aliberti, Irene Pupillo, Stefano Terna, Enrico Macii, Santa Di Cataldo, Edoardo Patti, and Andrea Acquaviva, 'A multi-patient data-driven approach to blood glucose prediction', *IEEE Access*, **7**, 69311–69325, (2019).

[2] Christoph Bergmeir and José M Benítez, 'On the use of cross-validation for time series predictor evaluation', *Information Sciences*, **191**, 192–213, (2012).

[3] Shiyu Chang, Yang Zhang, Wei Han, Mo Yu, Xiaoxiao Guo, Wei Tan, Xiaodong Cui, Michael Witbrock, Mark A Hasegawa-Johnson, and Thomas S Huang, 'Dilated recurrent neural networks', in *Advances in Neural Information Processing Systems*, pp. 77–87, (2017).

[4] Zhengping Che, Yu Cheng, Shuangfei Zhai, Zhaonan Sun, and Yan Liu, 'Boosting deep learning risk prediction with generative adversarial networks for electronic health records', in *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 787–792. IEEE, (2017).

[5] Jianwei Chen, Kezhi Li, Pau Herrero, Taiyu Zhu, and Pantelis Georgiou, 'Dilated recurrent neural network for short-time prediction of glucose concentration.', in *The 3rd KDH workshop, IJCAI-ECAI 2018*, pp. 69–73, (2018).

[6] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio, 'On the properties of neural machine translation: Encoder-decoder approaches', *arXiv preprint arXiv:1409.1259*, (2014).

[7] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio, 'Empirical evaluation of gated recurrent neural networks on sequence modeling', in *NIPS 2014 Workshop on Deep Learning, December 2014*, (2014).

[8] William L Clarke, Daniel Cox, Linda A Gonder-Frederick, William Carter, and Stephen L Pohl, 'Evaluating clinical accuracy of systems for self-monitoring of blood glucose', *Diabetes care*, **10**(5), 622–628, (1987).

[9] Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch, 'Real-valued (medical) time series generation with recurrent conditional gans', *arXiv preprint arXiv:1706.02633*, (2017).

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, 'Generative adversarial nets', in *Advances in neural information processing systems*, pp. 2672–2680, (2014).

[11] Sepp Hochreiter and Jürgen Schmidhuber, 'Long short-term memory', *Neural computation*, **9**(8), 1735–1780, (1997).

[12] Elad Hoffer, Itay Hubara, and Daniel Soudry, 'Train longer, generalize better: closing the generalization gap in large batch training of neural networks', in *Advances in Neural Information Processing Systems*, pp. 1731–1741, (2017).

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, 'Image-to-image translation with conditional adversarial networks', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, (2017).

[14] Diederik P Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization.', *International Conference on Learning Representations 2015*, 1–15, (2015).

[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, 'Deep learning', *nature*, **521**(7553), 436–444, (2015).

[16] Kezhi Li, John Daniels, Chengyuan Liu, Pau Herrero-Vinas, and Pantelis Georgiou, 'Convolutional recurrent neural networks for glucose prediction', *IEEE journal of biomedical and health informatics*, (2019).

[17] Kezhi Li, Chengyuan Liu, Taiyu Zhu, Pau Herrero, and Pantelis Georgiou, 'Glunet: A deep learning framework for accurate glucose forecasting', *IEEE journal of biomedical and health informatics*, (2019).

[18] Chengyuan Liu, Josep Vehí, Parizad Avari, Monika Reddy, Nick Oliver, Pantelis Georgiou, and Pau Herrero, 'Long-term glucose forecasting using a physiological model and deconvolution of the continuous glucose monitoring signal', *Sensors*, **19**(19), 4338, (2019).

[19] Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli, 'The uva/padova type 1 diabetes simulator: new features', *Journal of diabetes science and technology*, **8**(1), 26–34, (2014).

[20] C. Marling and R. Bunescu, 'The OhioT1DM dataset for blood glucose level prediction: Update 2020', in *The 5th KDH workshop, ECAI 2020*, (2020). CEUR proceedings in press, available at http://smarthealth.cs.ohio.edu/bglp/OhioT1DM-dataset-paper.pdf.

[21] Cindy Marling and Razvan C. Bunescu, 'The OhioT1DM dataset for blood glucose level prediction', in *The 3rd KDH workshop, IJCAI-ECAI 2018*, pp. 60–63, (2018).

[22] Cooper Midroni, Peter J Leimbigler, Gaurav Baruah, Maheedhar Kolla, Alfred J Whitehead, and Yan Fossat, 'Predicting glycemia in type 1 diabetes patients: experiments with XGBoost', in *The 3rd KDH workshop, IJCAI-ECAI 2018*, pp. 79–84, (2018).

[23] Sadegh Mirshekarian, Hui Shen, Razvan Bunescu, and Cindy Marling, 'Lstms and neural attention models for blood glucose prediction: Comparative experiments on real and synthetic data', in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 706–712. IEEE, (2019).

[24] Olof Mogren, 'C-rnn-gan: Continuous recurrent neural networks with adversarial training', *arXiv preprint arXiv:1611.09904*, (2016).

[25] Silvia Oviedo, Josep Vehí, Remei Calm, and Joaquim Armengol, 'A review of personalized blood glucose prediction strategies for t1dm patients', *International journal for numerical methods in biomedical engineering*, **33**(6), e2833, (2017).

[26] Tom Le Paine, Pooya Khorrami, Shiyu Chang, Yang Zhang, Prajit Ramachandran, Mark A Hasegawa-Johnson, and Thomas S Huang, 'Fast wavenet generation algorithm', *arXiv preprint arXiv:1611.09482*, (2016).

[27] Pouya Saeedi, Inga Petersohn, Paraskevi Salpea, Belma Malanda, Suvi Karuranga, Nigel Unwin, Stephen Colagiuri, Leonor Guariguata, Ayesha A Motala, Katherine Ogurtsova, et al., 'Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas', *Diabetes research and clinical practice*, **157**, 107843, (2019).

[28] Qingnan Sun, Marko V Jankovic, Lia Bally, and Stavroula G Mougiakakou, 'Predicting blood glucose with an LSTM and Bi-LSTM based deep neural network', in *2018 14th Symposium on Neural Networks and Applications (NEUREL)*, pp. 1–5. IEEE, (2018).

[29] Ashenafi Zebene Woldaregay, Eirik Årsand, Ståle Walderhaug, David Albers, Lena Mamykina, Taxiarchis Botsis, and Gunnar Hartvigsen, 'Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes', *Artificial intelligence in medicine*, (2019).

[30] Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar, 'Time-series generative adversarial networks', in *Advances in Neural Information Processing Systems*, pp. 5509–5519, (2019).

[31] Xingyu Zhou, Zhisong Pan, Guyu Hu, Siqi Tang, and Cheng Zhao, 'Stock market prediction on high-frequency data using generative adversarial nets', *Mathematical Problems in Engineering*, **2018**, (2018).

[32] Taiyu Zhu, Kezhi Li, Jianwei Chen, Pau Herrero, and Pantelis Georgiou, 'Dilated recurrent neural networks for glucose forecasting in type 1 diabetes', *Journal of Healthcare Informatics Research*, 1–17, (2020).

[33] Taiyu Zhu, Kezhi Li, Pau Herrero, Jianwei Chen, and Pantelis Georgiou, 'A deep learning algorithm for personalized blood glucose prediction.', in *The 3rd KDH workshop, IJCAI-ECAI 2018*, pp. 64–78, (2018).