A Deep Learning Algorithm For Personalized Blood Glucose Prediction

Taiyu Zhu*, **Kezhi Li***, **Pau Herrero, Jianwei Chen, Pantelis Georgiou**
Department of Electronic and Electrical Engineering, Imperial College London, London SW5 7AZ, UK
taiyu.zhu17@imperial.ac.uk, kezhi.li@imperial.ac.uk * †

Abstract

A convolutional neural network (CNN) model is presented to forecast the future glucose levels of the patients with type 1 diabetes. The model is a modified version of a recently proposed model called WaveNet, which becomes very useful in acoustic signal processing. By transferring the task into a classification problem, the model is mainly built by casual dilated CNN layers and employs fast WaveNet algorithms. The OhioT1DM dataset is the source of the four input fields: glucose levels, insulin events, carbohydrate intake and time index. The data is fed into the network along with the targets of the glucose change in 30 minutes. Several pre-processing approaches such as interpolation, combination and filtering are used to fill up the missing data in the training sets, and they improve the performance. Finally, we obtain the predictions of the testing dataset and evaluate the results by the root mean squared error (RMSE). The mean value of the best RMSE of six patients is 21.72.

1 Introduction

With an increasing incidence worldwide, type 1 diabetes is a severe chronic that requires long-term management of blood glucose relying on the glucose predictions [Daneman, 2006]. Aiming at improving the accuracy of the predictions, artificial intelligence researchers have been investigating machine learning approaches to develop efficient forecasting models.

In this paper, the main system is constructed by a deep convolutional neural network (CNN). It origins from a proposed model called WaveNet, which is firstly developed by the firm DeepMind to process raw audio signals [Van Den Oord *et al.*, 2016]. The glucose data of the patients are sequentially obtained by continuous glucose monitoring (CGM) in every five minutes. The change between the current glucose value and the future glucose value is quantised to 256 target categories. Under such circumstance, the prediction problem is converted to a classification task, which can be properly solved. After pre-processing datasets and building the modified WaveNet model, the prediction results for 30-minute prediction horizons (PH) are obtained.

2 Data Pre-processing

2.1 Database

The source of training and testing data is from the OhioT1DM dataset developed by [Marling and Bunescu, 2018]. There are six patients with type 1 diabetes wearing Medtronic 530G insulin pumps and Medtronic Enlite CGM sensors to collect the data during the 8-week period. Each of the patients reports the data on daily events via an app on a smartphone and a fitness band. In the OhioT1DM dataset, the patients are numbered as 559, 563, 570, 575, 588 and 591. Two of them are male with ID 563 and 570, while others are female. Three of the nineteen data fields, including previous CGM data by 'glucose_level', insulin value by 'bolus', carbohydrate intake by 'bwz_carb_input' and time index normalised to the unit for each day are used as the four channels of inputs. The meal information is obtained by the pump Bolus Wizard (BWZ), which is input by patients to calculate the bolus. Other fields of the patient data are also added to the input batch, such as heart rate and skin temperature. However, in our experiment, these fields slightly degrade the performance of classification and bring more variances to the model.

2.2 Interpolation and Extrapolation

By observing the glucose data, several intervals miss values in both training and testing sets. Since the targets of the CNN model rely on the differences between current and future data points, the discontinuities can cause negative influences. We fill up the missing values of the training dataset by the first-order interpolation. For the testing data, the first-order extrapolation is taken to ensure the future values are not involved. The predictions by extrapolated intervals are ignored to guarantee that the result has the same length as the CGM testing data when evaluating the performance.

*This work is submitted to the Blood Glucose Level Prediction Challenge, the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence (IJCAI-ECAI 2018), International Workshop on Knowledge Discovery in Healthcare Data.

†This work is supported by EPSRC, the ARISES project. T. Zhu and K. Li are the main contributors to the paper.

2.3 Combination

For the training dataset, it contains the data from six patients around 40 days and 115,200 CGM data points. Usually an effective machine learning model needs training data with much larger size. Moreover, the missing intervals appear frequently in the whole dataset. In our model, the data points with large missing gaps are discarded, and we interpolate the values only for short intervals. To have a longer training dataset and avoid the overfitting problem, we expand the training set and improve the generalisation. We introduce a part of the data with the longest continuous interval from other subjects and combine them into the current subject to form an extended training data. Notably, the strategy keeps half proportion data of the current subject, and the other five patients contribute to the other half of the training data with 10% each. The segments from other patients are selected by observation with the fewest missing values. Thus, the length of the dataset is expanded. This method significantly improves the performance of patient 591, who has long missing data interval (967-point) in the training set.

2.4 Filtering

After the interpolation and combination, it is found that there are many small spikes near the peaks or the turning points on the CGM data of the training dataset. Those spikes are a part of variances when the batches are used to train the model. To remove these variances, we use a median filter to filter out the noises at the cost of raising the bias slightly. The window size needs to be appropriately chosen, which is five-point in this work. The median filter is not used on the testing data, so the on-line prediction is still feasible. The outcome of data pre-processing is shown in Figure 1.

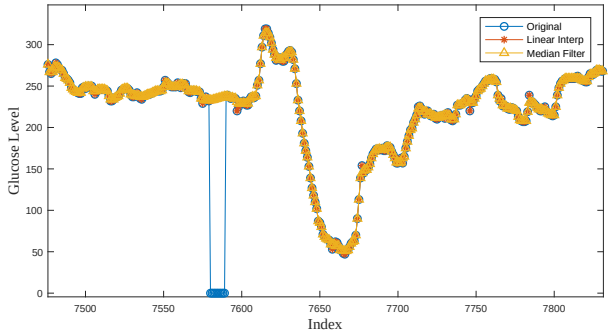

Figure 1: A part of CGM training set after pre-processing.

3 Methodology

Predicting future values for time series is one of the essential problems in data science. A conventional method is to use autoregressivemoving-average (ARMA) process to model the patterns. However, the performance of the ARMA model is not satisfactory in this task since it is incapable to capture non-linearities [Hamilton, 1994]. Feed-forward neural networks can overcome this difficulty and learn the patterns of multivariate time series well by feeding the data with extremely large size [Zhang *et al.*, 1998]. Thus it is a more suitable model with a four-channel feeding dictionary and non-linear glucose-insulin interaction. Moreover, our work is based on WaveNet that is more time efficient for training and testing with smaller weights compared with recurrent neural networks (RNNs) [Borovykh *et al.*, 2017]. It focuses on the long-term relationships between channels and is conditioned on all previous samples [Van Den Oord *et al.*, 2016], which is modelled as (1).

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t|x_1, ..., x_{t-1}) \tag{1}$$

where $\mathbf{x} = x_1, ..., x_T$ and $p(\mathbf{x})$ is the joint probability computed by the product of conditional probabilities. The output dimension is the same as the input because there are no pooling layers. Convolutional layers model the conditional probabilities, and a softmax layer is applied to maximise the log-likelihood probabilities.

3.1 The Causal CNN

The main components in WaveNet are causal convolutional layers. After shifting the outputs by several data points, the 1-D causal convolution layers can be implemented. The causality is essential for CNN to forecast time series since it guarantees that the model cannot use any information from future timesteps. Particularly, one special ingredient is causal dilated convolutional neural network (DCNN) layer. It largely increases the receptive field of the input signal. The structure is shown in Figure 2. Compared with regular causal convolutional layers, the dilated one involves a larger number of input nodes. This setting makes the system capable of learning long-term dependencies by skipping certain steps that determined by the dilation factor.

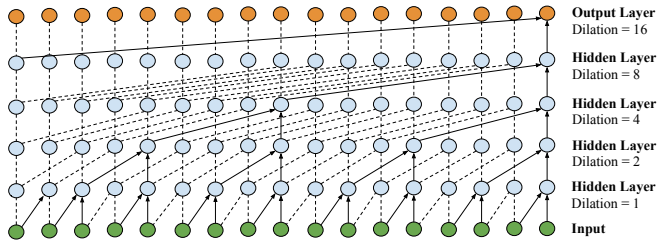

Figure 2: One Dilated causal convolution block with five layers.

In this work, there are three blocks, and the DCNN block contains five layers of dilated convolution. The dilation increases from one to a certain number in each block. The motivation behind this configuration is to exponentially grow the receptive field to cover more previous information and obtain a more efficient model with 1×32 convolution.

3.2 System Architecture

We adopt the fast approach to implement the WaveNet method, so it removes redundant convolutional operations and reduces the time complexity from $\mathcal{O}(2^L)$ to $\mathcal{O}(L)$ [Paine *et al.*, 2016], where L is the total number of the layers. The

fundamental technique is to create convolution queues that divide the operations into pop and push phases. The model functionally acts as a single step of RNNs. The system model for this project with fast WaveNet is shown in Figure 3.

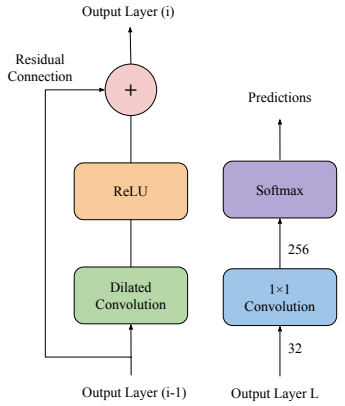

Figure 3: The system model. The output from the previous layer (i-1) is the input of the subsequent dilated convolutional layer (i). The process is repeated until obtaining the final output layer. Then the output is fed into a 1×1 convolutional layer, and a Softmax layer computes the predictions.

Compared with the work in [Van Den Oord *et al.*, 2016], we use a rectified linear unit (ReLU) as the activation function, instead of gated function, which is denoted as ReLU(x) := max(x, 0). How the model learns the non-linearities of the data series largely depends on the activation function. It is found in [Borovykh *et al.*, 2017] that the ReLU is very efficient for non-stationary or noisy time series. Moreover, it also reduces the training time and simplifies the model further. The output from layer i is written as (2).

$$f^i = [ReLU(w_1^i *_d f^{i-1}) + b, ..., ReLU(w_{T_i}^i *_d f^{i-1}) + b] \quad (2)$$

where f^i is the output of the CNN layer i after the dilated convolution $*_d$ with weight filters w_l^i, $l = 1, 2, ..., T_l$, and b stands for the bias. To model the conditional probabilities, Softmax is applied to calculate the training loss and output the predictions. The reason for this is because of its flexibility which has no requirement of the data shape. Thus it works well for continuous 1-D data [Oord *et al.*, 2016].

4 Training WaveNet

After pre-processing patient data and constructing the WaveNet system, the following step is to train the network; then the test data can be fed into the trained model.

4.1 Make Batches

The inputs of the neural network are four channels: CGM data, insulin event, carbohydrate intake, and time index. The batches of the testing phase have the same structure. The PH is 30 minutes, so it requires to forecast the CGM values 6 points in the future. Therefore, we calculate the glucose change between the current value and the future value in the PH. By quantisation, we put the change of glucose values in 256 classes/categories as targets with a difference of 1 mg/dl between each class. The number of classes is chosen carefully because a small number of classes cannot distinguish the difference while a large number of classes are not suitable for small training dataset. After investigating the training dataset, we think that the value of 256 is able to cover more than 95% of difference values, referring to the glucose change in the range of ± 128 mg/dl within 30 mins.

4.2 Weight Optimisation

The training process is to find the weights that minimise the cost function of the network. The cost function is one of the most important indicators in the training phase, which represents the error between targets and the network output. In the proposed system, sparse Softmax cross entropy is applied to optimise the model. Generally, the optimisation follows the gradient descent method that calculates weights through backward propagation, and the weights are updated after each iteration. Here we use adaptive moment estimation (Adam) Optimiser to adjust training steps by minimising the mean cost, and the learning rate is set to 0.0001. Adam optimiser has the promising performance on non-stationary time series [Kingma and Ba, 2014]. It uses first-order gradients and can be implemented with high computational efficiency. We set the number of total training iteration as 1,000 to avoid the underfitting or overfitting problem. The cost function loss versus the global steps is shown in Figure 4.

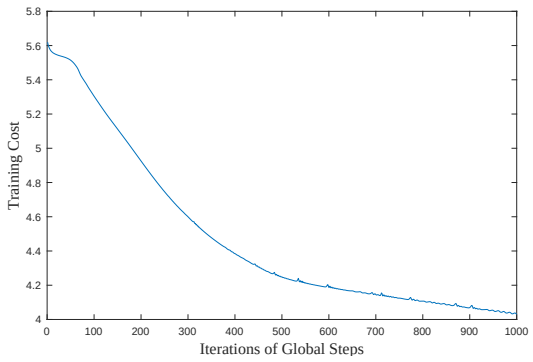

Figure 4: The training cost through the iterations

The input data has the same length as the outputs. The curve smoothly decreases in the first 500 iterations, and some spikes appear in iterations after 500, which is the consequence of mini-batch operations by Adam optimiser. It is noted that the cost is still converging after 1,000 global steps, but it causes over-fitting. The reason why over-fitting is likely to happen is the complexity of the model having limited training data.

5 Performance

5.1 Results

The unit of glucose level in this paper is mg/dl, and we use one of the essential evaluation metrics to evaluate the predic-

tion performance: root mean squared error (RMSE), which can be expressed as

$$\mathbf{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{x}_{(t|t-PH)} - x_t)^2}. \quad (3)$$

We mainly focus on RMSE for each patient and record the results for each patient after several runs, whose result are shown in Table1. The mean of the best results is 21.7267 with the standard deviation of 2.5237. The performance varies with subjects, and the method for subject 570 obtain the best result.

Table 1: The RMSE results of six patients in 30-minute PH.

Patient	P559	P563	P570	P575	P588	P591
Avg	22.48	20.35	18.26	25.65	21.69	24.59
Best	21.72	20.17	18.03	24.80	21.42	24.22
Point (#)	2514	2570	2745	2590	2791	2760
Best Avg	21.7267±2.5237					

5.2 Analysis

The forecasting curve and original CGM data are plotted in Figure 5. Notably, the predicted curve fits original CGM recording with similar trends in general. However, there still exists a degree of difference that can be seen in the detail view of one-day CGM data. First, it is noted that there is an obvious delay between predictions and raw data, especially for the turning points and peaks. In the bottom plot, the dashed line stands for the insulin events, and the red circle is the meal events. Intuitively, it is found the curve will change significantly after these events. The curve intensively fluctuates, and the error is high in these periods. The possible reason is that it is difficult for the model to learn the biological model explicitly, and those glucose level changes are not determined only by the input data. However, it is also found that the RMSE result improves by 0.8 mg/dl when feeding the four channels of data fields instead of solely CGM data.

Another finding is the effect of extrapolation. There are some missing values around 18:00 and we use the first-order extrapolation to fill up the time series. However, the error is still high for the data after these regions. Because the predictive curve is calculated by adding the predicted differences onto the previous values, it heavily depends on the data 6 timesteps before. We only extrapolate the CGM field, because the other three channels are discrete values. As shown in Figure 5, the insulin and carbohydrate intake have significant impacts on the future values, so it would cause more error if these data points are missing. Several different interpolation methods are also tested in the training phase, such as cubic and spline. The first-order interpolation performs best by reducing mean RMSE by 2.1 mg/dl because it captures linearities of 1-D signal well.

For the first six data points, the predictions are calculated on concatenating the last part of the training data at the front of the testing data. As long as the trend of concatenated data is the same as the subsequent timesteps, it will not affect the

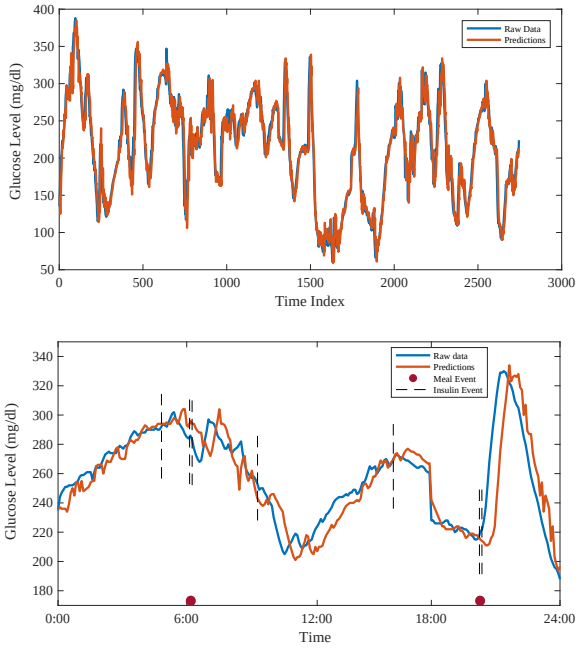

Figure 5: The curves of forecasting CGM series and original CGM recordings. Top: The whole test data set for patient 570, RMSE:18.027, MARD: 6.545. Bottom: The predictions and test data for patient 570 on 21/01/2022.

RMSE results much. The length of the predictions is the same as the testing dataset, as the point number shown in Tabel 1.

The predictions for subject 575 and 591 performs much worse than other subjects. There are two reasons. On the one hand, it is the large gap in the training dataset, as in subject 591. Although we use data combination approach to compensate for this loss and successfully reduce the mean RMSE by around 0.2 mg/dl, the RMSE is still quite high compared with others. On the other hand, it is the condition of the testing dataset. For subject 591, the CGM data of testing sets fluctuates a lot with plenty of spikes, and the error is high near the turning point of the curve. However, those fluctuations are determined by the biological model and the conditions of patient health, such as the plasma insulin model in [Lehmann and Deutsch, 1992]. Moreover, subject 570 and 563 use "Humalog" insulin while other four subject use "Novalog". We found that predictions are more accurate for the subjects with "Humalog" insulin. Another possible feature of these two patients is they have the same gender. The data from a larger group of subjects is required to prove and explore the correlations.

Compared with existing models, this paper presents a novel deep learning model based on CNN layers. The performance outperforms the simple autoregressive models using only the same CGM data, which follows the structure in [Sparacino *et al.*, 2007]. As for the results from other deep learning models, the RMSE results cannot be compared directly, due to different subjects and datasets, such as [Mougiakakou *et al.*, 2006]. Nevertheless, the major advantages of this models are higher efficiency with less global training steps and small weights,

and it has the fast algorithmic implementation with the low time complexity $\mathcal{O}(L)$.

6 Conclusion

In this paper, the task to predict the glucose level in 30-minute PH is converted into a classification problem, and a new model based on modified WaveNet is developed. With the pre-processed dataset and causal DCNN system architecture, the network is trained to obtain the network weights. Four channels are selected to input the network because we find they have the strong correlations with glucose levels.

The mean value of the best RMSE for six subjects is 21.7267 with standard deviation equals to 2.5237. The model is different from existing RNN models and outperforms many current algorithms. The prediction performances mainly affected by the missing CGM values and the length of the training sets. By integrating other data fields with biological models is a potential approach to improve the prediction accuracy in the future work.

References

[Borovykh *et al.*, 2017] Anastasia Borovykh, Sander Bohte, and Cornelis W Oosterlee. Conditional time series forecasting with convolutional neural networks. *arXiv preprint arXiv:1703.04691*, 2017.

[Daneman, 2006] Denis Daneman. Type 1 diabetes. *The Lancet*, 367(9513):847–858, 2006.

[Hamilton, 1994] James Douglas Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, 1994.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Lehmann and Deutsch, 1992] ED Lehmann and T Deutsch. A physiological model of glucose-insulin interaction in type 1 diabetes mellitus. *Journal of biomedical engineering*, 14(3):235–242, 1992.

[Marling and Bunescu, 2018] Cindy Marling and Razvan Bunescu. The OhioT1DM dataset for blood glucose level prediction. In The 3rd International Workshop on Knowledge Discovery in Healthcare Data, Stockholm, Sweden, July 2018. CEUR proceedings in press, available at `http://smarthealth.cs.ohio.edu/bglp/OhioT1DM-dataset-paper.pdf`, 2018.

[Mougiakakou *et al.*, 2006] Stavroula G Mougiakakou, Aikaterini Prountzou, Dimitra Iliopoulou, Konstantina S Nikita, Andriani Vazeou, and Christos S Bartsocas. Neural network based glucose-insulin metabolism models for children with type 1 diabetes. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pages 3545–3548. IEEE, 2006.

[Oord *et al.*, 2016] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

[Paine *et al.*, 2016] 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.

[Sparacino *et al.*, 2007] Giovanni Sparacino, Francesca Zanderigo, Stefano Corazza, Alberto Maran, Andrea Facchinetti, and Claudio Cobelli. Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series. *IEEE Transactions on biomedical engineering*, 54(5):931–937, 2007.

[Van Den Oord *et al.*, 2016] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[Zhang *et al.*, 1998] Guoqiang Zhang, B Eddy Patuwo, and Michael Y Hu. Forecasting with artificial neural networks:: The state of the art. *International journal of forecasting*, 14(1):35–62, 1998.