Experiments in non-personalized future blood glucose level prediction

Robert Bevan [1] and **Frans Coenen** [2]

Abstract. In this study we investigate the need for training future blood glucose level prediction models at the individual level (i.e. per patient). Specifically, we train various model classes: linear models, feed-forward neural networks, recurrent neural networks, and recurrent neural networks incorporating attention mechanisms, to predict future blood glucose levels using varying time series history lengths and data sources. We also compare methods of handling missing time series data during training. We found that relatively short history lengths provided the best results: a 30 minute history length proved optimal in our experiments. We observed long short-term memory (LSTM) networks performed better than linear and feed-forward neural networks, and that including an attention mechanism in the LSTM model further improved performance, even when processing sequences with relatively short length. We observed models trained using all of the available data outperformed those trained at the individual level. We also observed models trained using all of the available data, except for the data contributed by a given patient, were as effective at predicting the patient's future blood glucose levels as models trained using all of the available data. These models also significantly outperformed models trained using the patient's data only. Finally, we found that including sequences with missing values during training produced models that were more robust to missing values.

1 Introduction

Accurate future blood glucose level prediction systems could play an important role in future type-I diabetes condition management practices. Such a system could prove particularly useful in avoiding hypo/hyper-glycemic events. Future blood glucose level prediction is difficult - blood glucose levels are influenced by many variables, including food consumption, physical activity, mental stress, and fatigue. The Blood Glucose Level Prediction Challenge 2020 tasked entrants with building systems to predict future blood glucose levels at 30 minutes, and 60 minutes into the future. Challenge participants were given access to the OhioT1DM dataset [8], which comprises 8 weeks worth of data collected for 12 type-I diabetes patients. The data include periodic blood glucose level readings, administered insulin information, various bio-metric data, and self-reported information regarding meals and exercise.

In the previous iteration of the challenge, several researchers demonstrated both that it is possible to predict future blood glucose levels using previous blood glucose levels only [9], and that past blood glucose levels are the most important features for future blood glucose level prediction [10]. In this study, we aimed to extend this research into future glucose level prediction from historical glucose levels only. Most previous work involved training personalized models designed to predict future blood glucose level data for a single patient [9, 10, 2]. Others used schemes coupling pre-training using adjacent patient's data with a final training phase using the patient of interest's data only [3, 12].

In this work, we investigate the possibility of building a single model that is able to predict future blood glucose levels for all 12 patients in the OhioT1DM data set, and the effectiveness of applying such a model to completely unseen data (i.e. blood glucose series from an unseen patient). We also investigate the impact of history length on future blood glucose level prediction. We experiment with various model types: linear models, feed-forward neural networks, and recurrent neural networks. Furthermore, inspired by advances in leveraging long distance temporal patterns for time series prediction [6], we attempt to build a long short-term memory (LSTM) model that is able to use information from very far in the past (up to 24 hours) by incorporating an attention mechanism. Finally, we compare the effectiveness of two methods for handling missing data during training.

2 Method

2.1 Datasets

As stated above, one of our primary aims was to build a single model that is able to predict future blood glucose levels for each patient in the data set. To this end, we constructed a combined data set containing data provided by each of the patients. Specifically, we created a data set composed of all of the data points contributed by the six patients included in the previous iteration of the challenge (both training and test sets), as well as the training data points provided by the new cohort of patients. This combined data set was split into training and validation sets: the final 20% of the data points provided by each patient were chosen for the validation set. The test data sets for the 6 new patients were ignored during development to avoid bias in the result. For experiments in building patient specific models, training and validation sets were constructed using the patient in question's data only (again with an 80/20 split).

2.2 Data preprocessing

Prior to model training, the data were standardized according to:

$$x = \frac{x - \mu_{train}}{\sigma_{train}} \quad (1)$$

[1] University of Liverpool, UK, email: robert.e.bevan@gmail.com
[2] University of Liverpool, UK, email: coenen@liverpool.ac.uk

Model	Hyper-parameters
Feed-forward	# hidden units $\in \{16, 32, 64, 128, 256, 512, 1024\}$
	# layers* $\in \{1, 2\}$
	activation function $\in \{ReLU\}$
Recurrent	# hidden units $\in \{16, 32, 64, 128, 256, 512, 1024\}$
	recurrent cell* $\in \{LSTM, GRU\}$
	# layers* $\in \{1, 2\}$
	output dropout* $\in \{0, 0.1, 0.2, 0.5\}$
LSTM + Attention	# $\boldsymbol{\alpha_t}$ hidden units $\in \{4, 8, 16, 32, 64, 128\}$

Table 1. Lists of hyper-parameters tuned when training feed-forward, recurrent neural networks, and LSTM with attention networks. Note that not all possible combinations were tried - parameters marked with an asterisk were tuned after the optimal number of hidden units was chosen.

where μ_{train}, and σ_{train}, are the mean and standard deviation of the training data, respectively. There is a non-negligible amount of data missing from the training set, which needed to be considered when preprocessing the data. We investigated two approaches to handling missing data: discarding any training sequences with one or more missing data points; and replacing missing values with zeros following standardization. It was hypothesized that the second approach may help the system learn to be robust to missing data.

2.3 Model training

We experimented with linear models, feed-forward neural networks, and recurrent neural networks. Each model was trained to minimize the root mean square error (RMSE):

$$RMSE = \sqrt{\sum_{i=1}^{n} \left(\frac{\hat{y_i} - y_i}{n}\right)^2} \tag{2}$$

where $\hat{y_i}$ is the predicted value, y_i is the true value, and n is the total number of points in the evaluation set. Various hyper-parameters were tuned when training the feed-forward and recurrent neural networks; Table 1 provides a summary. During development, each model was trained for 50 epochs using the Adam optimizer (α=0.001, β_1=0.9, β_2=0.999) [5] and a batch size of 256. The final model was trained with early stopping using the same optimizer settings, and a batch size of 32, for a maximum period of 500 epochs with early stopping and a patience value of 30. Each model was trained 5 times in order to get an estimate of the influence of the random initialization and stochastic training process on the result.

Model selection and hyper-parameter tuning were performed for the 30 minute prediction horizon task. The best performing model was then trained for 60 minute prediction. Experiments were repeated for blood glucose history lengths of 30 minutes, 1 hour, 2 hours, and 24 hours.

2.4 Improving long distance pattern learning with Attention

It can be difficult for recurrent neural networks to learn long distance patterns. The LSTM network was introduced to address this problem

	RMSE		MAE	
Patient ID	**PH=30**	**PH=60**	**PH=30**	**PH=60**
540	21.03 (0.07)	37.37 (0.09)	16.64 (0.1)	30.8 (0.13)
544	16.14 (0.12)	28.4 (0.14)	12.85 (0.11)	23.57 (0.16)
552	15.82 (0.06)	27.6 (0.15)	12.43 (0.12)	22.78 (0.16)
567	20.29 (0.08)	34.28 (0.18)	15.9 (0.12)	28.95 (0.13)
584	20.39 (0.07)	32.97 (0.09)	15.99 (0.03)	27.04 (0.07)
596	15.7 (0.03)	25.99 (0.12)	12.4 (0.04)	21.33 (0.13)
AVG	18.23 (2.36)	31.1 (4.05)	14.37 (1.83)	25.75 (3.43)

Table 2. Root mean square error and mean absolute error (mg/dl) computed using the test points for each patient, at different prediction horizons (30 minutes, and 60 minutes) for a single layer LSTM with 128 hidden units.

[4]. Even so, LSTM networks can struggle to learn very long range patterns. Attention mechanisms - initially introduced in the context of neural machine translation - have been shown to improve LSTM networks' capacity for learning very long range patterns [7, 1]. Attention mechanisms have also been applied to time series data, and have proven to be effective in instances where the data exhibit long range periodicity - for example, in electricity consumption prediction [6]. We hypothesised that blood glucose level prediction using a very long history, coupled with an attention mechanism, could lead to improved performance, due to periodic human behaviours (e.g. eating meals at similar times each day; walking to and from work e.t.c). In order to test this hypothesis, we chose the best performing LSTM configuration trained with a history length of 24 hours, without attention, and added an attention mechanism as per [7]:

$$score(\boldsymbol{h_t}, \boldsymbol{h_i}) = \boldsymbol{h_t^T W h_i} \tag{3}$$

$$\alpha_{ti} = \frac{exp(score(\boldsymbol{h_t}, \boldsymbol{h_i})}{\Sigma_{j=1}^{t} exp(score(\boldsymbol{h_t}, \boldsymbol{h_j}))} \tag{4}$$

$$\boldsymbol{c_t} = \sum_i \alpha_{ti} \boldsymbol{h_i} \tag{5}$$

$$\boldsymbol{a_t} = f(\boldsymbol{c_t}, \boldsymbol{h_t}) = tanh(\boldsymbol{W_c}[\boldsymbol{c_t}; \boldsymbol{h_t}]) \tag{6}$$

where $score(\boldsymbol{h_t}, \boldsymbol{h_i})$ is an alignment model, $\boldsymbol{c_t}$ is the weighted context vector, and $\boldsymbol{a_t}$ is the attention vector, which is fed into the classification layer (without an attention mechanism, the hidden state $\boldsymbol{h_t}$ is fed into the classification layer). The dimensionality of $\boldsymbol{\alpha_t}$ was chosen using the validation set - see Table 1 for details. We also experimented with attention mechanisms in LSTM networks designed to process shorter sequence lengths: we chose the optimal LSTM model architecture for each history length (30 minutes, 60 minutes, 2 hours), added an attention mechanism, and re-trained the model (again, the optimal $\boldsymbol{\alpha_t}$ dimensionality was chosen using the validation set).

2.5 Investigating the need for personal data during training

In order to investigate the need for an individual's data when training a model to predict their future blood glucose levels, we trained 6 different models - each with one patient's training data excluded from

Patient ID	Patient only	All patients	Patient excluded
540	21.68 (0.04)	**21.03 (0.07)**	21.16 (0.11)
544	17.28 (0.1)	**16.14 (0.12)**	16.22 (0.09)
552	16.87 (0.12)	**15.82 (0.06)**	15.87 (0.09)
567	21.15 (0.3)	**20.29 (0.08)**	20.5 (0.11)
584	22.11 (0.13)	**20.39 (0.07)**	20.46 (0.06)
596	16.16 (0.11)	**15.7 (0.03)**	15.71 (0.02)
AVG	19.21 (2.48)	**18.23 (2.36)**	18.32 (2.4)

Table 3. Root mean square error (mg/dl) computed using the test points for each patient, with a prediction horizon of 30 minutes for a single layer LSTM with 128 hidden units, trained using different data sets. Values listed in the first column correspond to models trained using the individual patient data only; values in the middle column correspond to models trained using data from all patients; values in the final column correspond to models trained using data from all patients except for the patient for which the evaluation is performed.

the training set - using the optimal LSTM architecture determined in previous experiments. The models were then evaluated using the test data for the patient that was excluded from the training set. We also trained 6 patient specific models, each trained using the patient's training data only. We again used the optimal architecture determined in previous experiments, but tuned the number of hidden units using the validation set in order to avoid over-fitting due to the significantly reduced size of the training set (compared with the set with which the optimal architecture was chosen). Each model was trained using the early-stopping procedure outlined in 2.3.

3 Results and Discussion

Our evaluation showed that recurrent models performed significantly better than both linear and feed-forward neural network models, for each history length we experimented with (p=0.05, corrected paired t-test [11]). We also found that feed-forward networks generally outperformed linear models, likely due to their ability to model non-linear relationships. The optimal feed-forward network contained 512 hidden units. We found no difference between LSTM and gated recurrent unit (GRU) networks - remaining evaluations will be performed for LSTM recurrent networks only for simplicity. Figure 1 compares the performance of the different model types as a function of history length. For each model class we observed that performance decreased linearly with increasing history length. The LSTM appeared better able to deal with longer history lengths - the performance degradation was less severe than for the other model classes. We found a history length of 30 minutes to be optimal for each model class. The best performing LSTM model contained a single layer with 128 hidden units, and was trained without dropout. The test set results for this model are listed in Table 2.

Table 3 compares LSTM models (with the same architecture as above) trained with the following data sources: the individual patient's data only, data from all of the available patients, and data from every other patient (excluding data contributed by the patient in question). We observed that models trained using a large amount of data, but excluding the patient's data, outperformed models trained using the patient's data only (p=0.05). We also found no significant difference in performance between models trained using all of the available training data (i.e. including the patient's data) and those trained excluding the patient's data, highlighting the general nature of the models. We found that including sequences with values in the training set produced models that were more robust to missing data, as evidenced by the improved RMSE scores listed in Table 4: RMSE scores were significantly improved for each patient using this approach to training (p=0.05).

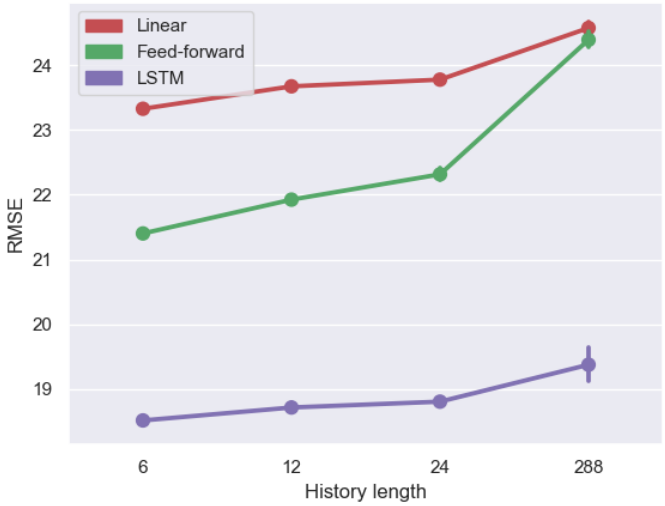

Figure 1. Comparison of validation set scores for linear, feed-forward, and LSTM neural networks as a function of history length.

Incorporating an attention mechanism further improved performance in most instances: we observed significant improvements for history lengths of 30 minutes, 60 minutes, and 2 hours (p=0.05), but not for history lengths of 24 hours. Figure 3 compares the regular LSTM and LSTM with attention performance as a function of history length. Figure 2 shows partial auto-correlation plots for 4 different patients. Interestingly, two of the patients' blood glucose data - patient 540, and patient 544 - don't show any significant long term correlation, whereas the other two - patient 552, and patient 567 - both exhibit significant correlation at time lags of approximately 6 and 12 hours. We observed this behaviour in half of the patients. We

Patient ID	Exclude missing data	Include missing data
540	21.45 (0.06)	**21.03 (0.07)**
544	16.79 (0.06)	**16.14 (0.12)**
552	16.27 (0.13)	**15.82 (0.06)**
567	21.19 (0.1)	**20.29 (0.08)**
584	21.16 (0.06)	**20.39 (0.07)**
596	16.08 (0.07)	**15.7 (0.03)**
AVG	18.82 (2.45)	**18.23 (2.36)**

Table 4. Root mean square error (mg/dl) computed using the test points for each patient, with a prediction horizon of 30 minutes for a single layer LSTM with 128 hidden units, trained using different methods of handling missing data. Values in the first column correspond to a model trained with full sequences only (any sequences with missing values were discarded). Values in the second column correspond to a model trained with sequences including missing values - missing values were replaced with zeros following standardization.

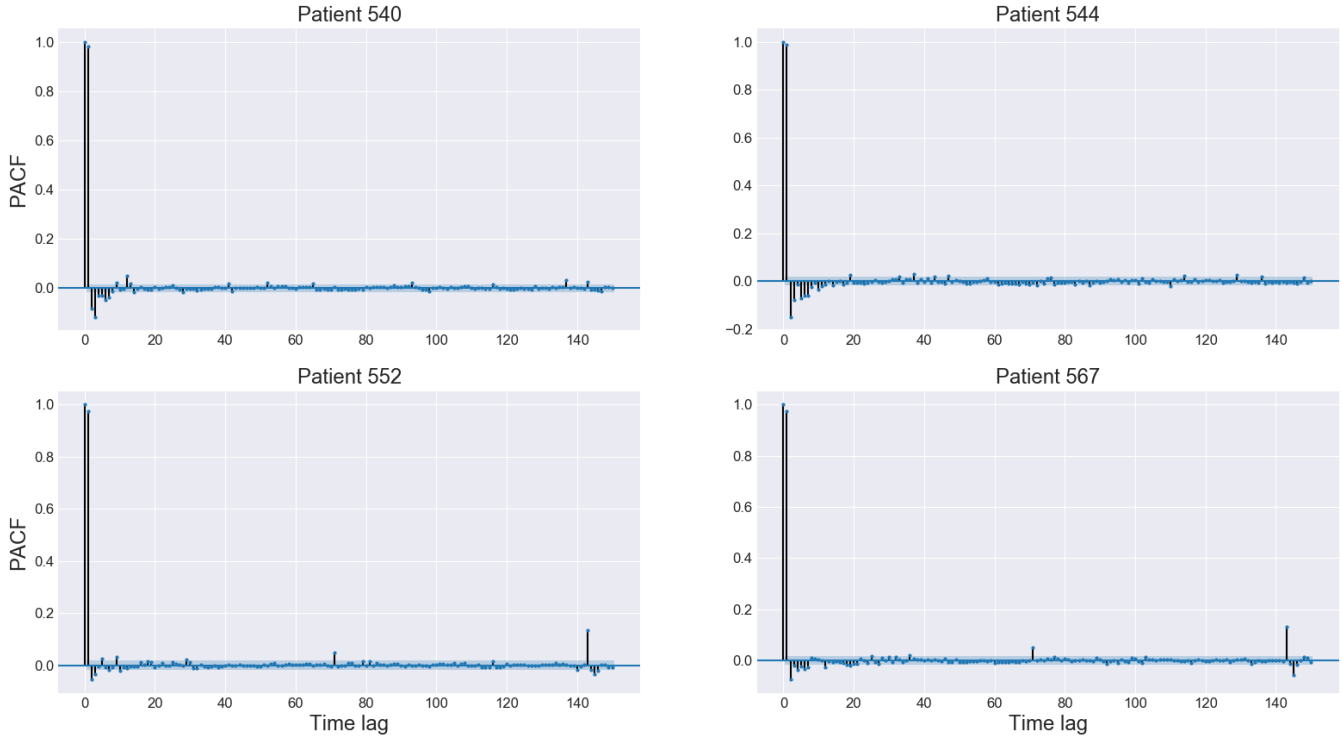

Figure 2. Partial auto-correlation plots for 4 different patients. The patients in the top row exhibit short term patterns only, but those in the bottom row show significant correlations at time lags of approximately 6 and 12 hours.

also observed correlations at even greater time lags, corresponding to multiples of 6 hours. The difference in the patients' partial auto-correlation plots suggests it may be sub-optimal to train an attention mechanism using each patient's data at once, and that training at the patient-level may enable the model to learn very long range patterns. Furthermore, while we were able to train an LSTM model with a history length of 30 minutes that generalized across all patients, it may be the case that short range blood glucose patterns are quite general, and long range patterns are more personalized, and tuning the history length per patient could improve prediction performance. All of the results presented in this section can be reproduced using publicly available code [3].

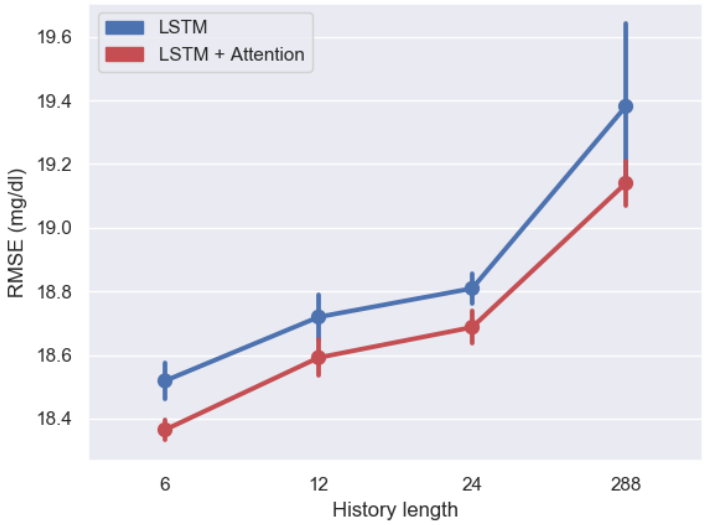

Figure 3. Comparison of validation set RMSE scores (prediction horizon = 30 minutes) for LSTMs and LSTMs incorporating an attention mechanism as a function of history length.

4 Conclusion

In this study we showed that it is possible to train a single LSTM model that is able to predict future blood glucose levels for each of the different patients whose data are included in the OhioT1DM data set. We also demonstrated that an individual patient's data is not required during the training process in order for our model to effectively predict the patient's future blood glucose levels. Furthermore we showed that incorporating an attention mechanism in the LSTM improved performance, and that including sequences with missing values during training produced models that were more robust to missing data.

[3] https://github.com/robert-bevan/bglp

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, 'Neural machine translation by jointly learning to align and translate', in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, (2015).

[2] Arthur Bertachi, Lyvia Biagi, Iván Contreras, Ningsu Luo, and Josep Vehí, 'Prediction of blood glucose levels and nocturnal hypoglycemia using physiological models and artificial neural networks', in *Proceedings of the 3rd International Workshop on Knowledge Discovery in Healthcare Data*, pp. 85–90.

[3] Jianwei Chen, Kezhi Li, Pau Herrero, Taiyu Zhu, and Pantelis Georgiou, 'Dilated recurrent neural network for short-time prediction of glucose concentration', in *Proceedings of the 3rd International Workshop on Knowledge Discovery in Healthcare Data*, pp. 69–73.

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

[5] Diederik P. Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, (2015).

[6] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu, 'Modeling long- and short-term temporal patterns with deep neural networks', *CoRR*, **abs/1703.07015**, (2017).

[7] Thang Luong, Hieu Pham, and Christopher D. Manning, 'Effective approaches to attention-based neural machine translation', in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pp. 1412–1421, (2015).

[8] Cynthia R. Marling and Razvan C. Bunescu, 'The OhioT1DM dataset for blood glucose level prediction', in *Proceedings of the 3rd International Workshop on Knowledge Discovery in Healthcare Data*, pp. 60–63, (2018).

[9] John Martinsson, Alexander Schliep, Bjorn Eliasson, Christian Meijner, Simon Persson, and Olof Mogren, 'Automatic blood glucose prediction with confidence using recurrent neural networks', in *Proceedings of the 3rd International Workshop on Knowledge Discovery in Healthcare Data*, pp. 64–68.

[10] 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 *Proceedings of the 3rd International Workshop on Knowledge Discovery in Healthcare Data*, pp. 79–84.

[11] C. Nadeau and Y. Bengio, 'Inference for the generalization error', *Mach. Learn.*, **52**(3), 239–281, (September 2003).

[12] Taiyu Zhu, Kezhi Li, Pau Herrero, Jianwei Chen, and Pantelis Georgiou, 'A deep learning algorithm for personalized blood glucose prediction', in *Proceedings of the 3rd International Workshop on Knowledge Discovery in Healthcare Data*, pp. 74–78.