Multi-lag Stacking for Blood Glucose Level Prediction

Heydar Khadem[1] and **Hoda Nemat**[1] and **Jackie Elliott**[2] and **Mohammed Benaissa**[1]

Abstract. This work investigates blood glucose level prediction for type 1 diabetes in two horizons of 30 and 60 minutes. Initially, three conventional regression tools—partial least square regression (PLSR), multilayer perceptron, and long short-term memory—are deployed to create predictive models. They are trained once on 30 minutes and once on 60 minutes of historical data resulting in six basic models for each prediction horizon. A collection of these models are then set as base-learners to develop three stacking systems; two uni-lag and one multi-lag. One of the uni-lag systems uses the three basic models trained on 30 minutes of lag data; the other uses those trained on 60 minutes. The multi-lag system, on the other hand, leverages the basic models trained on both lags. All three stacking systems deploy a PLSR as meta-learner. The results obtained show: i) the stacking systems outperform the basic models, ii) among the stacking systems, the multi-lag shows the best predictive performance with a root mean square error of 19.01 mg/dl and 33.37 mg/dl for the prediction horizon of 30 and 60 minutes, respectively.

1 INTRODUCTION

Diabetes mellitus is a metabolic disorder and a significant cause of morbidity and mortality worldwide [1]. As yet, there is no cure developed for diabetes; and management of the corresponding life-impeding conditions is recommended as the most successful way to control the disease [6]. In fact, the occurrence of the associated complications can be suspended or even prevented by effective management of the disease [11].

Among different types of diabetes, the importance of the self-management for type 1 diabetes mellitus (T1DM) is accentuated [8, 19]. The key factor in T1DM management is to control the blood glucose level (BGL) within the normal range [2]. BGL predictive models could contribute to achieving this goal. They can help avert adverse glycaemic events by forecasting them and giving patients the chance to take corrective actions ahead of time [2].

The importance of the development of BGL predictive models in T1DM management has spurred research into this field [16, 22]. According to the knowledge requirement, predictive models can be classified as; physiological, data-driven, and hybrid models [21]. Data-driven models interpret trends in sequences of data to make estimations of future BGLs. Machine learning approaches are broadly adopted in this area [21].

Mirshekarian et al. [17] developed a model to predict blood glucose in 30-minute and 60-minute horizons using a recursive neural network (RNN) with long short- term memory (LSTM) units. The model explored BGL, insulin, food, and activity information as inputs. For the same prediction horizons, Bertachi et al. [4] and Georga et al. [9], in separate studies, proposed predictive models. Bertachi et al. applied an artificial neural network contemplating glucose, insulin, carbohydrate and physical activity as inputs for their system. BGL profile, insulin, carbohydrate intake and physical activity were inputs for a support vector regression (SVR) in the model developed by Georga et al. Investigating continuous glucose monitoring (CGM) data by recursive and direct deep learning approaches, Xie et al. [22] recommended a model for BGL prediction. Martinsson et al. [15] proposed an automatic forecast model for a prediction horizon of up to 60 minutes using RNN. The model used only the information from past BGLs as input. Bunescu et al. [7] created descriptive features to train a SVR using a physiological model of blood glucose dynamics. Carbohydrate intake, insulin administration, and the current and past BGLs were inputs of their model. Despite extensive research devoted to the development of predictive models, the performance of the proposed models remains a challenge [3].

In this work, we contributed to the improvement of BGL prediction for T1DM by applying a multi-lag stacking methodology. Initially, three conventional regression tools—partial least squares, multilayer perceptron, and long-short term memory—were applied to forecast BGLs in horizons of 30 and 60 minutes. Each tool was trained twice; once on a lag of 30 minutes and once on a lag of 60 minutes of CGM data. Therefore, six basic models were created for each prediction horizon. For each horizon, three stacking systems were then developed where predictions from a selection of the basic models were used as features to train a new regression. The first two stacking systems followed a uni-lag approach. They used predictions from the three base models trained on a history of 30 minutes and 60 minutes, respectively. The third system was multi-lag and used predictions from all six base models. The stacking systems resulted in appreciable improvements in predictive accuracy as compared to the basic predictive models. The third stacking system showed a predictive performance better than the other systems.

This is the first paper, to our knowledge, that has combined models with different time-lags to generate a multi-lag BGL prediction system.

2 DATASET

The Ohio T1DM dataset comprises several features collected from 12 individuals with type 1 diabetes in 8 weeks [14, 13]. The last ten days' worth of data for each contributor was considered as the test set. Data for a cohort of six subjects was released in 2018 for the first BGL prediction challenge [14]; data for another six subjects was released in 2020 for the second challenge [13].

In this work, the 2020's data was investigated for developing and evaluating predictive models. Among the collected features were

[1] Department of Electronic and Electrical Engineering, University of Sheffield, UK, email addresses: h.khadem@sheffield.ac.uk, hoda.nemat@sheffield.ac.uk, m.benaissa@sheffiels.ac.uk

[2] Department of Oncology and Metabolism, University of Sheffield, UK, email address: j.elliott@sheffield.ac.uk

CGM data every 5 minutes, which was the only feature explored in this work. A brief description of the CGM data in the Ohio T1DM dataset released for 2020 BGL prediction challenge is displayed in Table 1.

Table 1. Number of test and training examples of each participant in Ohio T1DM dataset released in 2020 [13].

Patient ID	Number of Training Examples	Number of Test Examples
540	11947	2896
544	10623	2716
552	9080	2364
567	10858	2389
584	12150	2665
596	10877	2743

3 METHODS

As mentioned earlier, this work proposes methodologies to predict BGL in horizons of 30 and 60 minutes. The detail of the pursued methodologies is presented in this section

3.1 Pre-processing

The first pre-processing task was taking care of missing data. Missing data in the training set was imputed applying a simple linear interpolation. Alternatively, for the test set, a linear extrapolation was employed. This was to ensure the model is not contaminated by observing future data in its pre-processing stage.

The next pre-processing step was transferring the time series forecasting problem to a supervised learning task. To this end, a rolling window consisting of a lag and future data was used as explanatory and dependent variables respectively. To give an illustration, for forecasting BGL of 30 minutes later using a history of 60 minutes, for example, we used a window with the length of 18. As a consequence of the 5-minute interval between data points, it therefore follows that the first 12 data points in the window were explanatory variables, and the rest were dependent variables.

3.2 Prediction methods

First, six basic predictive *models* were created by means of three conventional regression tools. Subsequently, employing stacking learning, three more advanced predictive *systems* were developed where a collection of the basic models were considered as base-learners and a partial least squares regression as meta-learner. All proposed models/systems were personalised to individuals.

3.2.1 Basic models

Initially, for each prediction horizon of 30 and 60 minutes, the following three conventional regressions tools were employed to generate six basic predictive models—two models by each tool. For this purpose, these tools were trained once on a history of 30 and once on a history of 60 minutes.

- *Partial least squares regression (PLSR)*
  PLSR, as a basic linear regression, holds substantial popularity in different applications due to its easy-to-apply nature and minimal computation time requirement. In a previous work, we applied PLSR for glucose quantification which provided promising results [12].
  In this work, PLSR was used as one of the regression tools. For the number of components, different values ranging from 1 to the length of the input variable were tried. Each time, the predicted residual sum of squares ($PRESS$) was calculated as follows. The number of components (A) resulting in the minimum value for $PRESS/(N-A-1)$ was then selected [20].

$$PRESS = \sum_{i=1}^{N}(y_i - \hat{y_i})^2 \tag{1}$$

  where, N is the size of the evaluation set , y_i is reference value, and $\hat{y_i}$ is predicted value.

- *Multilayer perceptron (MLP)*
  An MLP [18] with an architecture of one hidden layer including 100 nodes and an output layer was implemented. ReLU was used as the activation function for he hidden layer, Adam as the optimiser, and mean absolute error as the loss function. Learning rate was 0.01, and the training process was based on 100 epochs.

- *Long short-term memory (LSTM)*
  We used a Vanila LSTM [10] composed of a single hidden LSTM layer with 200 nodes, a fully connected layer with 100 nodes, and an output layer. ReLU was the activation function for both hidden layers, mean squared error was the loss function, and Adam was the optimizer. The model trained on 100 epochs with a learning rate of 0.01.

3.2.2 Stacking systems

Ensemble learning is a machine learning technique that combines decisions from several models to create a new model. Stacking (Figure 1) is an ensemble approach that uses predictions from multiple base-learners (first level models) as features to train a meta-learner (second level model). This meta-learner then makes the final predictions on the test set [23].

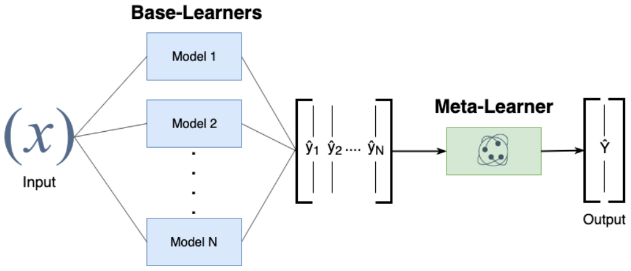

Figure 1. A stacking system uses predictions from multiple base-learners as features to train a meta-learner[5] .

In this paper, for each prediction horizon of 30 and 60 minutes, three stacking systems comprised of two uni-lag and one multi-lag were developed.

- *System 1*
  The three basic models trained on a history of 30 minutes were the base-learners of this uni-lag system and a PLSR was its meta-learner.

- *System 2*
  This system was also uni-lag. It was similar to system 1, except it used the three basic models trained on a history of 60 minutes in place of 30 minutes as base-learners.

- *System 3*
  In this multi-lag system, all the six basic models were considered as the base-learners and again a PLSR was the meta-learner. By performing a multi-lag approach the idea was to help capture a broader frequency range of BGL dynamics.

3.3 Evaluation

The test set was held out, and the train set was used to create the predictive models/systems. The developed models/systems were then utilised to predict the test data. The set of evaluation points starts 60 minutes after the beginning of the test set. First evaluation points would be otherwise similar to the training data, and it can affect the reliability of the results. Hence, the number of evaluated points for each patient is 12 less than the number of test examples mentioned in Table 1. Root mean square error (RMSE) and mean absolute error (MAE) were calculated as follows and then used as evaluation metrics.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{2}$$

$$MAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{N} \tag{3}$$

where, N, y_i , and $\hat{y}_i$ carry the same definition as in (1).

4 RESULTS AND DISCUSSION

This section presents the evaluation results for both the basic models and stacking systems. Models/systems with a performance depended on random initialization ran five times, and corresponding results have been reported in the form of mean and standard deviation. Extrapolated points were excluded when calculating the evaluation metrics. All models were built to predict future BGLs up to the end of the intended prediction horizon, but only the evaluation results for the horizon of interest are reported.

4.1 Prediction horizon of 30 minutes

4.1.1 Basic models

The results of the RMSE and MAE of the basic predictive models for the prediction horizon of 30 minutes are displayed in Table 2.

Based on the average of RMSE and MAE for all patients, *LSTM* trained on a history of 30 minutes showed the best performance among the basic models. *PLSR* with 60-minute lag was the second-best model. All models had satisfactory standard deviations.

LSTM yielded the best overall predictive accuracy among the three regression tools. However, the results of the other two tools were also comparable to that of *LSTM*. It is worth remarking that *PLSR*, as a linear regression tool, was able to generate results comparable to that of *LSTM* and even better than that of *MLP*.

Among all patients, patient 552 had the best overall evaluation results. The worst results, on the other hand, belonged to patients 584 and 540.

Table 2. Evaluation results of the basic predictive models for a 30-minute prediction horizon.

Patient ID	Basic Model	History (min)	RMSE (mg/dl)	MAE (mg/dl)
540	PLSR	30	22.11	16.58
		60	22.07	16.56
	MLP	30	21.98 ± 0.48	16.52 ± 0.33
		60	22.52 ± 0.78	16.76 ± 0.62
	LSTM	30	21.65 ± 0.28	16.06 ± 0.12
		60	**21.58 ±0.67**	**16.20 ± 0.61**
544	PLSR	30	18.08	13.34
		60	18.09	13.33
	MLP	30	18.22 ± 0.18	13.38 ± 0.37
		60	18.25 ± 0.28	13.21 ± 0.35
	LSTM	30	**17.63 ± 0.15**	**12.63 ± 0.10**
		60	18.42 ± 0.60	13.36 ± 0.44
552	PLSR	30	16.76	12.76
		60	16.79	12.78
	MLP	30	17.08 ± 0.36	12.91 ± 0.40
		60	17.03 ± 0.34	12.77 ± 0.17
	LSTM	30	**16.49 ± 0.10**	**12.29 ± 0.24**
		60	17.06 ± 0.70	12.88 ± 0.51
567	PLSR	30	20.98	15.12
		60	21.00	15.07
	MLP	30	21.24 ± 0.70	15.42 ± 0.76
		60	21.10 ± 0.46	15.13 ± 0.58
	LSTM	30	**20.66 ± 0.16**	**14.79 ± 0.25**
		60	20.77 ± 0.36	14.72 ± 0.40
584	PLSR	30	22.00	16.15
		60	21.97	16.12
	MLP	30	**21.67 ± 0.18**	**15.63 ± 0.16**
		60	22.43 ± 0.48	16.35 ± 0.61
	LSTM	30	22.23 ± 0.70	16.33 ± 0.67
		60	22.04 ± 0.22	16.11 ± 0.28
596	PLSR	30	17.79	12.77
		60	**17.62**	**12.67**
	MLP	30	17.74 ± 0.04	12.55 ± 0.05
		60	18.44 ± 0.26	13.49 ± 0.42
	LSTM	30	17.76 ± 0.67	12.74 ± 0.55
		60	17.71 ± 0.28	12.50 ± 0.33
Average	PLSR	30	19.62	14.45
		60	19.59	14.42
	MLP	30	19.65 ± 0.32	14.40 ± 0.35
		60	19.96 ± 0.43	14.62 ± 0.46
	LSTM	30	**19.40 ± 0.34**	**14.14 ± 0.32**
		60	19.60 ± 0.47	14.30 ± 0.43

4.1.2 Stacking systems

Table 3 shows the evaluation results of the stacking systems for a prediction horizon of 30 minutes. For all patients, the performance of the stacking systems surpassed that of the basic models. *System 3* proposed the best predictions overall based on average RMSE and MAE values. This system resulted in the best predictive accuracy for all patients except patient 544 and 584. All systems possessed small standard deviation values. The best result among all patients belonged to patient 552. The worst results, on the other hand, were those of patients 584, 540, and 567.

Table 3. Evaluation results of the stacking systems for a 30-minute prediction horizon.

Patient ID	Stacking System	RMSE (mg/dl)	MAE (mg/dl)
540	System 1	21.13 ± 0.08	15.72 ± 0.10
	System 2	21.11 ± 0.18	15.69 ± 0.14
	System 3	**20.93 ± 0.11**	**15.52 ± 0.13**
544	System 1	**17.47 ± 0.05**	**12.50 ± 0.05**
	System 2	17.92 ± 0.10	12.93 ± 0.08
	System 3	17.52 ± 0.05	12.50 ± 0.07
552	System 1	16.29 ± 0.06	12.13 ± 0.06
	System 2	16.43 ± 0.12	12.33 ± 0.16
	System 3	**16.21 ± 0.09**	**12.08 ± 0.08**
567	System 1	20.43 ± 0.07	14.47 ± 0.06
	System 2	20.51 ± 0.14	14.51 ± 0.16
	System 3	**20.43 ± 0.06**	**14.41 ± 0.06**
584	System 1	**21.61 ± 0.06**	**15.68 ± 0.04**
	System 2	21.83 ± 0.14	15.86 ± 0.08
	System 3	21.75 ± 0.08	15.76 ± 0.07
596	System 1	17.26 ± 0.03	12.19 ± 0.03
	System 2	17.47 ± 0.15	12.25 ± 0.11
	System 3	**17.22 ± 0.10**	**12.09 ± 0.04**
Average	System 1	19.03 ± 0.06	13.78 ± 0.06
	System 2	19.21 ± 0.14	13.93 ± 0.12
	System 3	**19.01 ± 0.08**	**13.73 ± 0.07**

4.2 Prediction horizon of 60 minutes

4.2.1 Basic models

Table 4 lists RMSE and MAE of the basic models for 60-minute prediction horizon. Among all models, *LSTM* trained on a lag of 30 minutes showed the best performance. *MLP* trained on 300 minutes was the second high-performance model. The value of standard deviation for all models were satisfactory. Among the implemented regression tools, *LSTM* resulted in the highest overall prediction accuracy. *PLSR* produced acceptable results in this case too. Data for patients 596 and 552 showed the highest overall predictability. In, contrast, patients 540, 567, and 584 had the lowest predictable data.

4.2.2 Stacking systems

Evaluation results of the stacking systems for a prediction horizon of 60 minutes are displayed in Table 5. *System 3* proposed the best overall predictions based on average RMSE and MAE values. The best result among all patients belonged to patient 596. All systems had low values of standard deviation.

Table 4. Evaluation results of the basic predictive models for a 60-minute prediction horizon.

Patient ID	Basic Model	History (min)	RMSE (mg/dl)	MAE (mg/dl)
540	PLSR	30	41.03	31.68
		60	41.03	31.70
	MLP	30	40.20 ± 0.38	30.90 ± 0.21
		60	41.94 ± 2.18	32.14 ± 1.53
	LSTM	30	40.36 ± 0.91	30.80 ± 0.64
		60	**39.65 ± 1.16**	**30.28 ± 0.84**
544	PLSR	30	31.80	24.71
		60	31.83	24.71
	MLP	30	31.58 ± 0.53	24.19 ± 0.99
		60	32.15 ± 0.63	24.13 ± 0.83
	LSTM	30	**30.61 ± 0.19**	**22.97 ± 0.26**
		60	31.79 ± 0.31	24.57 ± 0.73
552	PLSR	30	30.23	23.67
		60	30.24	23.68
	MLP	30	30.14 ± 0.09	23.27 ± 0.24
		60	30.59 ± 1.01	23.65 ± 0.63
	LSTM	30	**29.84 ± 0.25**	**22.52 ± 0.29**
		60	31.36 ± 1.43	23.72 ± 1.77
567	PLSR	30	37.47	28.28
		60	37.53	28.24
	MLP	30	36.81 ± 0.28	27.52 ± 0.50
		60	37.73 ± 1.28	28.57 ± 1.35
	LSTM	30	**36.56 ± 0.17**	**27.58 ± 0.28**
		60	37.17 ± 0.58	27.90 ± 0.72
584	PLSR	30	36.71	27.65
		60	36.84	27.75
	MLP	30	3**6.32 ± 0.59**	**26.95 ± 0.66**
		60	37.35 ± 0.82	27.82 ± 0.92
	LSTM	30	37.14 ± 0.98	28.03 ± 1.14
		60	37.03 ± 0.99	27.42 ± 0.54
596	PLSR	30	29.63	22.05
		60	29.48	21.97
	MLP	30	29.68 ± 0.27	21.87 ± 0.31
		60	29.97 ± 0.39	22.08 ± 0.39
	LSTM	30	**28.98 ± 0.29**	**21.14 ± 0.19**
		60	29.71 ± 0.72	22.09 ± 0.80
Average	PLSR	30	34.48	26.43
		60	34.55	26.34
	MLP	30	34.12 ± 0.36	25.78 ± 0.49
		60	34.95 ± 1.05	26.40 ± 0.94
	LSTM	30	**33.92 ± 0.47**	**25.51 ± 0.47**
		60	34.45 ± 0.86	26.00 ± 0.90

Table 5. Evaluation results of the stacking systems for a 60-minute prediction horizon.

Patient ID	Stacking System	RMSE (mg/dl)	MAE (mg/dl)
540	System 1	39.47 ± 0.17	30.10 ± 0.17
	System 2	39.14 ± 0.28	29.76 ± 0.20
	System 3	**39.00 ± 0.20**	**29.65 ± 0.12**
544	System 1	**30.47 ± 0.10**	**22.92 ± 0.13**
	System 2	31.12 ± 0.12	23.72 ± 0.14
	System 3	30.54 ± 0.09	22.95 ± 0.17
552	System 1	29.39 ± 0.15	22.39 ± 0.13
	System 2	29.38 ± 0.20	22.46 ± 0.20
	System 3	**29.10 ± 0.13**	**22.10 ± 0.14**
567	System 1	**36.11 ± 0.11**	**27.08 ± 0.15**
	System 2	36.54 ± 0.14	27.36 ± 0.14
	System 3	36.31 ± 0.14	27.09 ± 0.08
584	System 1	**36.15 ± 0.16**	**27.04 ± 0.18**
	System 2	36.68 ± 0.19	27.43 ± 0.19
	System 3	36.52 ± 0.10	27.30 ± 0.14
596	System 1	**28.74 ± 0.16**	20.84 ± 0.12
	System 2	29.06 ± 0.21	21.13 ± 0.27
	System 3	28.75 ± 0.10	**20.78 ± 0.05**
Average	System 1	33.39 ± 0.14	25.06 ± 0.15
	System 2	33.65 ± 0.19	25.31 ± 0.19
	System 3	**33.37 ± 0.13**	**24.98 ± 0.12**

5 CONCLUSION

BGL prediction improved using stacking learning concepts. Initially, a time series problem was translated into a supervised learning task. Three conventional regression tools were trained with on different history length of 30 and 60 minutes, resulting in six basic predictive models. Predictions from the basic models trained with a history of 30 minutes were fed as features to a regression to build a combined learner. The learner was then used to make final predictions on the test set. The same scenario was repeated using the basic models trained on 60-minute lag observations. In both cases, the combined learner was able to make more accurate predictions on the test set. The overall performance further improved when predictions from all basic models—trained on both histories of 30 and 60 minutes—were considered as features to train a new learner.

6 SOFTWARE AND CODE

For data analysis we used Python 3.6, TensorFlow 1.15.0 and Keras 2.2.5. Pandas, NumPy and Sklearn packages of python were used. The codes are available at: `https://gitlab.com/Heydar-Khadem/multi-lag-stacking.git`

REFERENCES

[1] Florencia Aguiree, Alex Brown, Nam Ho Cho, Gisela Dahlquist, Sheree Dodd, Trisha Dunning, Michael Hirst, Christopher Hwang, Dianna Magliano, Chris Patterson, et al., ‘Idf diabetes atlas’, (2013).

[2] Ramzi Ajjan, David Slattery, and Eugene Wright, ‘Continuous glucose monitoring: A brief review for primary care practitioners’, *Advances in therapy*, **36**(3), 579–596, (2019).

[3] Muhammad Asad and Usman Qamar, ‘A review of continuous blood glucose monitoring and prediction of blood glucose level for diabetes type 1 patient in different prediction horizons (ph) using artificial neural network (ann)’, in *Proceedings of SAI Intelligent Systems Conference*, pp. 684–695. Springer, (2019).

[4] 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 *KHD@ IJCAI*, pp. 85–90, (2018).

[5] Julio Borges, *The Power of Ensembles in Deep Learning*, 2019. https://towardsdatascience.com/the-power-of-ensembles-in-deep-learning-a8900ff42be9.

[6] Danielle Bruen, Colm Delaney, Larisa Florea, and Dermot Diamond, ‘Glucose sensing for diabetes monitoring: recent developments’, *Sensors*, **17**(8), 1866, (2017).

[7] Razvan Bunescu, Nigel Struble, Cindy Marling, Jay Shubrook, and Frank Schwartz, ‘Blood glucose level prediction using physiological models and support vector regression’, in *2013 12th International Conference on Machine Learning and Applications*, volume 1, pp. 135–140. IEEE, (2013).

[8] Mol Ecol, ‘HHS Public Access’, **25**(5), 1032–1057, (2017).

[9] Eleni I Georga, Vasilios C Protopappas, Diego Ardigò, Demosthenes Polyzos, and Dimitrios I Fotiadis, ‘A glucose model based on support vector regression for the prediction of hypoglycemic events under free-living conditions’, *Diabetes technology & therapeutics*, **15**(8), 634–643, (2013).

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

[11] George S Jeha, Lefkothea P Karaviti, Barbara Anderson, EO’Brian Smith, Susan Donaldson, Toniean S McGirk, and Morey W Haymond, ‘Continuous glucose monitoring and the reality of metabolic control in preschool children with type 1 diabetes’, *Diabetes Care*, **27**(12), 2881–2886, (2004).

[12] Heydar Khadem, Mohammad R Eissa, Hoda Nemat, Osamah Alrezj, and Mohammed Benaissa, ‘Classification before regression for improving the accuracy of glucose quantification using absorption spectroscopy’, *Talanta*, **211**, 120740, (2020).

[13] Cindy Marling and Razvan Bunescu, ‘The ohiot1dm dataset for blood glucose level prediction: Update 2020’.

[14] Cindy Marling and Razvan C Bunescu, ‘The OhioT1DM Dataset For Blood Glucose Level Prediction.’, in *3rd International Workshop on Knowledge Discovery in Healthcare Data*, pp. 60–63, (2018).

[15] John Martinsson, Alexander Schliep, Björn Eliasson, Christian Meijner, Simon Persson, and Olof Mogren, ‘Automatic blood glucose prediction with confidence using recurrent neural networks’, in *3rd International Workshop on Knowledge Discovery in Healthcare Data, KDH@ IJCAI-ECAI 2018, 13 July 2018*, pp. 64–68, (2018).

[16] 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’, *CEUR Workshop Proceedings*, **2148**, 79–84, (2018).

[17] Sadegh Mirshekarian, Razvan Bunescu, Cindy Marling, and Frank Schwartz, ‘Using lstms to learn physiological models of blood glucose behavior’, in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2887–2891. IEEE, (2017).

[18] Fionn Murtagh, ‘Multilayer perceptrons for classification and regression’, *Neurocomputing*, **2**(5-6), 183–197, (1991).

[19] Shauna S Roberts, ‘Type 1 diabetes’, *Diabetes Forecast*, **55**, 19, (2002).

[20] Svante Wold, Michael Sjöström, and Lennart Eriksson, ‘Pls-regression: a basic tool of chemometrics’, *Chemometrics and intelligent laboratory systems*, **58**(2), 109–130, (2001).

[21] Ashenafi Zebene Woldaregay, Eirik Årsand, Taxiarchis Botsis, David Albers, Lena Mamykina, and Gunnar Hartvigsen, ‘Data-driven blood glucose pattern classification and anomalies detection: machine-learning applications in type 1 diabetes’, *Journal of medical Internet research*, **21**(5), e11030, (2019).

[22] Jinyu Xie and Qian Wang, ‘Benchmark machine learning approaches with classical time series approaches on the blood glucose level prediction challenge.’, in *KHD@ IJCAI*, pp. 97–102, (2018).

[23] Zhi-Hua Zhou, *Ensemble methods: foundations and algorithms*, CRC press, 2012.