Introduction

Diabetes Mellitus (DM), commonly known as Diabetes, comprises a group of metabolic disorders characterized by elevated blood glucose levels (hyperglycemia). This condition results from defects in insulin secretion, insulin action, or both, affecting carbohydrate, protein, and fat metabolism [1]. Patients with diabetes are categorized into Type 1 diabetes (T1DM) and Type 2 diabetes (T2DM), with T1DM individuals relying on exogenous insulin due to insufficient natural production.

Effective diabetes management involves lifestyle adjustments, continuous glucose monitoring, and vigilant follow-up actions to enhance patient well-being and reduce medical costs [1]. This entails a meticulous regimen, including monitoring blood glucose levels, tracking carbohydrate intake, and administering precise insulin doses. Technological advancements, such as Continuous Glucose Monitoring Systems (CGM), play a crucial role in automating the monitoring process, especially when coupled with insulin pumps [1]. Despite these advances, predicting blood glucose values accurately remains a complex task, necessitating the development of dynamic models.

In the realm of predictive modeling, machine learning techniques are applied, particularly neural networks (NN), for blood glucose prediction. Addressed here, is a meta-study exploring the effectiveness of long short-term memory neural (LSTM) networks glucose prediction. This study utilizes the OhioT1DMDataset common, features, sampling rates, and metrics, offering a comprehensive analysis to guide informed decision-making in patient care.

Experimental setup

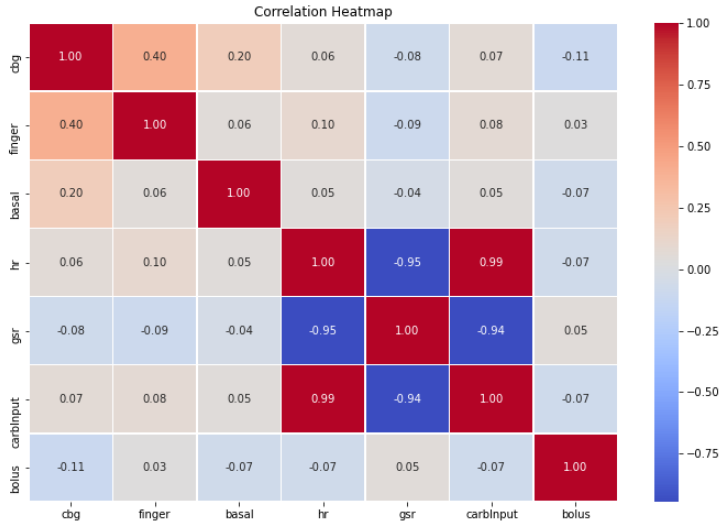
For this study, a model based on LSTMs was developed using Python 3.7, Tensorflow 2.2.0 and Keras 2.4.3. These models use the OhioT1DM dataset [2] for training and testing each model.
It contains eight weeks' worth of continuous glucose monitoring, insulin, physiological sensor, and self-reported life event data for each of 12 people with type 1 diabetes. The OhioT1DM Dataset was first released in 2018 for the first Blood Glucose Level Prediction (BGLP) Challenge. At that time, the dataset was half its current size, containing data for only six people with type 1 diabetes. Data for an additional six people was released in 2020 for the second BGLP Challenge.

Methods

Data Processing:

In this first step, a search for missing values was done. There are several ways to deal with missing values, but each of them has its ups and downsides. Removing them would be the more straightforward way, but the signal temporality would be lost this way, so interpolation was opted for. For the training set the cubic splines [1] were used to complete missing values and create a time series compatible with the models. Since the cubic splines fit a polynomial function to fill in the missing values, it ends up using future points. That way to the testing set the forward fill interpolation was used. This method fills the missing values with the first non-null value before them. After dealing with the missing values, a cross-correlation analysis function was used to explore the relationships between the different signals and the target variable (blood glucose levels) (see Figure 1).

Figure 1: Correlation analysis for patient 559.



For this dataset, the input features are:

- ***cbg*** - Blood glucose levels.
- ***finger*** - Blood glucose values obtained through self-monitoring by the patient with a finger stick.
- ***basal*** - The rate at which basal insulin is continuously infused. The basal rate begins at the specified timestamp ts, and it continues until another basal rate is set.
- ***hr*** - Heart rate, aggregated every 5 minutes. This data is only available for people who wore the Basis Peak sensor band (2018).
- ***gsr*** - Galvanic skin response, also known as skin conductance or electrodermal activity. For those who wore the Basis Peak, the data was aggregated every 5 minutes. Despite this attribute's name, it is also available for those who wore the Empatica Embrace. For these individuals, the data is aggregated every 1 minute.
- ***carbInput*** - the patient's carbohydrate estimate for the meal.
- ***bolus*** - Insulin delivered to the patient, typically before a meal or when the patient is hyperglycemic.

The input features and the target variable were normalised between 0 and 1 with the MinMaxScaler. Normalising the data is a common practice in machine learning, especially when using neural networks like LSTMs. Because neural networks perform better when all input features and the target variable are on a similar scale, by preventing certain features from dominating the learning process solely due to their scale. It also reduces the convergence speed, increases the stability of the learning process, and finally improves the model generalization.

Feature selection

Feature selection is a crucial step in building an effective predictive model, especially when dealing with datasets that contain many features, which was not the case as there were only seven. Even though there is a panoply of methods at our disposal, the Correlation Analysis was the one used, to calculate the correlation between each feature and the blood glucose levels. Because it is simple to implement, provides insights into linear relationships, but is limited to linear relationships. Other methods like Recursive Feature Elimination, Mutual Information, and Tree-Based Methods can capture non-linear relations but are more computationally expensive.

Model description

Table 1 summarises the architecture parameters. It consists of just one hidden layer with five units, followed by a dense layer with one unit acting as the output layer. The recurrent activation function for the LSTM layer is the sigmoid function, and the activation function is tanh. The activation function of the output neuron is the linear function.

Table 1: Model's architecture hyperparameters.

Layer	Hyperparameter	Value
Hidden	Type	LSTM
	Layers	1
	Units per layer	10
	Recurrent activation function	sigmoid
	Activation function	tanh
Out	Units	1
	Activation function	Linear
	Number of parameters	691

Training

The training and test data are split as provided by the OhioT1DM database, and training is performed by 85% to training and 15% for validation. Table 2 presents the number of samples each patient has for training and testing procedures. Due to a shortage of time and computational power, the training was not done using the 80/20 5-fold cross-validation approach. In the model validation, the mean absolute error, would have been used as the metric function.

Table 2: Number of training and testing samples per patient.

Number of Samples					
2018			2020		
Patient	Training	Testing	Patient	Training	Testing
559	12080	2876	540	13109	3065
563	13097	2691	544	12671	3136
570	11611	2879	552	11097	3950
575	13103	2718	567	not used	
588	13105	2880	584	13247	2995
591	12755	2847	596	13629	3002

The Adam algorithm [1] with a learning rate of 0.001 and the mean squared error, as loss function is used in the training. The training consists of 10 epochs with an early stopping of 3 epochs' patience. Ideally, training would have consisted of 100 epochs with an early stopping of 10 epochs' patience [1].

Results and Discussion

Blood glucose was forecasted at three prediction horizons, PH = {30; 60; 120} min; ph = 30 min was taken as the starting short-term prediction horizon and doubled progressively to define the medium- and long-term prediction horizons. The predictions were evaluated on a per-patient basis using the most common error metrics: mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), R-squared (R^2), correlation coefficient (CC), fit (Fit), and mean absolute relative difference (MARD). Tables 2, 3, and 4 show the results for the two groups of patients over the different prediction horizons. The results are the average of the metrics over the patients with the standard error of the mean deviation.

The Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) are common metrics used in forecasting and regression analysis to evaluate the performance of a predictive model. Each of these metrics provides different insights into the accuracy and precision of the model's predictions. The MSE and RMSE, emphasize larger errors due to squaring, penalize larger errors more heavily, and are useful when large errors are critical or when the distribution of errors is not symmetric. On the other hand, the MAE, provides a more balanced view of errors since it does not square the differences, is less sensitive to outliers, and is useful when all errors, regardless of size, contribute equally to the overall assessment. Given that a simple LSTM with a very small number of training epochs was used, large values of MSE, RMSE, and MAE are to be expected as a potential result of underfitting. These metrics point to a better performance obtained with a prediction horizon of 30 min followed by 120 min. On average these two prediction horizons triumph over the 60-minute one, on the one hand, maybe because it is easier to forecast less time into the future. On the other hand, having a bigger sequence with 120 minutes to analyse might help the model capture better hidden patterns in the time series.

The R^2 score is a metric used to evaluate the performance of a regression model. It can be interpreted as the explainability of the model, i.e., how much of the data can be explained by each of the models. For all the prediction horizons the R^2 score was negative indicating that the model is performing worse than a simple mean.

The Correlation Coefficient (CC), Fit (FIT), and Mean Absolute Relative Difference (MARD) are additional forecasting metrics that provide insights into different aspects of model performance. A high correlation coefficient indicates a strong linear relationship between predicted and actual values. On obtained results, the CCs between the predicted and actual values were consistently below 0.5, indicating a relatively weak linear relationship between the predictions and the actual blood glucose levels. The highest value was obtained, for a PH of 30 minutes for the 2020 group of patients.

For all the prediction horizons the FIT was negative indicating that the model is performing worse than a simple mean, this might be happening because the model may not be capturing the patterns in the data, leading to predictions that are worse than a simple mean baseline.

Finally, MARD is the most common metric used to analyse the accuracy of Continuous Glucose Monitoring systems [1]. It measures the difference between the actual values and the predicted ones. So, the lower the MARD is the more accurate the predictions are. For ph = 30 min the average values are 0.20 and 0.27, the predictions deviate by 20% from the actual values. For ph = 60 min the difference is greater, as expected, 0.29 and -5.06. And, for ph = 120 min MARD values are 0:32 and -0.08. Negative values are unusual for MARD and might indicate an issue with the calculations or data. MARD is typically non-negative as it involves taking the absolute value of the relative differences. A further analysis of the calculations is necessary.

Overall, the results achieved were better with a prediction horizon of 30 minutes, but the performance metrics also suggest the need for a careful investigation and adjustments, which might lead to a forecasting performance enhancement.

Table 3: Average results for the LSTM model for each performance metric for ph = 30 min.

Cohort	Metrics						
	MSE	**RMSE**	**MAE**	**R^2**	**CC**	**FIT**	**MARD**
2018	32478,10 ± 65294,06	118,03 ± 136,19	58,78 ± 33,70	-11,78 ± 26,41	0,33 ± 0,17	-1,33 ± 1,01	0,20 ± 0,14
2020	16316,99 ± 8006,14	122,35 ± 36,69	74,72 ± 17,50	-0,38 ± 1,09	0,48 ± 0,22	-0,71 ± 0,83	0,27 ± 0,11

Table 4: Average results for the LSTM model for each performance metric for ph = 60 min.

Cohort	Metrics						
	MSE	**RMSE**	**MAE**	**R^2**	**CC**	**FIT**	**MARD**
2018	46787,02 ± 96680,40	134,95 ± 169,04	64,23 ± 36,29	-17,22 ± 38,41	0,15 ± 0,17	-5,036 ± 0,19	0,29 ± 0,04
2020	24727,63 ± 15524,67	146,38 ± 57,439	83,14 ± 22,24	-0,62 ± 0,92	0,20 ± 0,31	-1,82 ± 1,99	-5,06 ± 11,16

Table 5: Average results for the LSTM model for each performance metric for ph = 120 min.

Cohort	Metrics						
	MSE	**RMSE**	**MAE**	**R^2**	**CC**	**FIT**	**MARD**
2018	3144,41 ± 1843,60	63,92 ± 8,12	51,08 ± 0,77	-0,19 ± 0,16	0,02 ± 0,09	-5,52 ± 2,25	0,32 ± 0,03
2020	25525,05 ± 22672,09	142,41 ± 72,41	88,41 ± 35,66	-0,26 ± 0,24	0,13 ± 0,28	-4,22 ± 4,27	-0,08 ± 0,84

Figure 2: Blood Gluose Prediction vs actual for patient 559 for ph = 30 min.

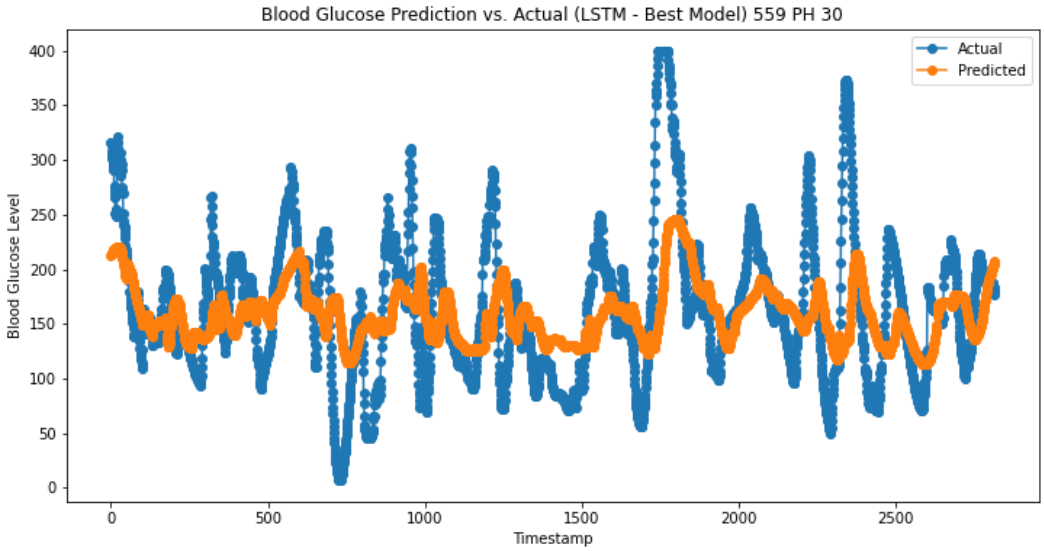


Figure 3: Blood Gluose Prediction vs actual for patient 559 for ph = 60 min.

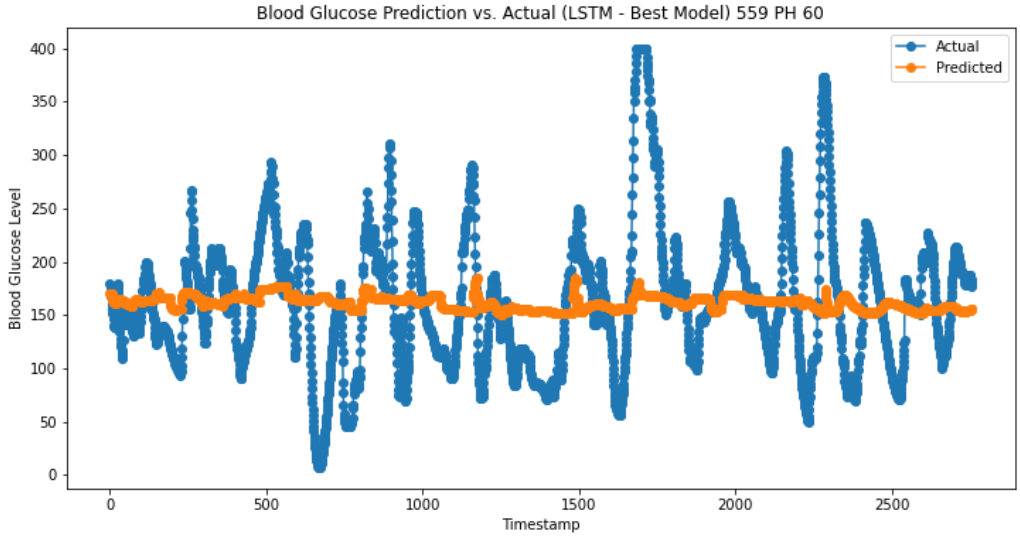
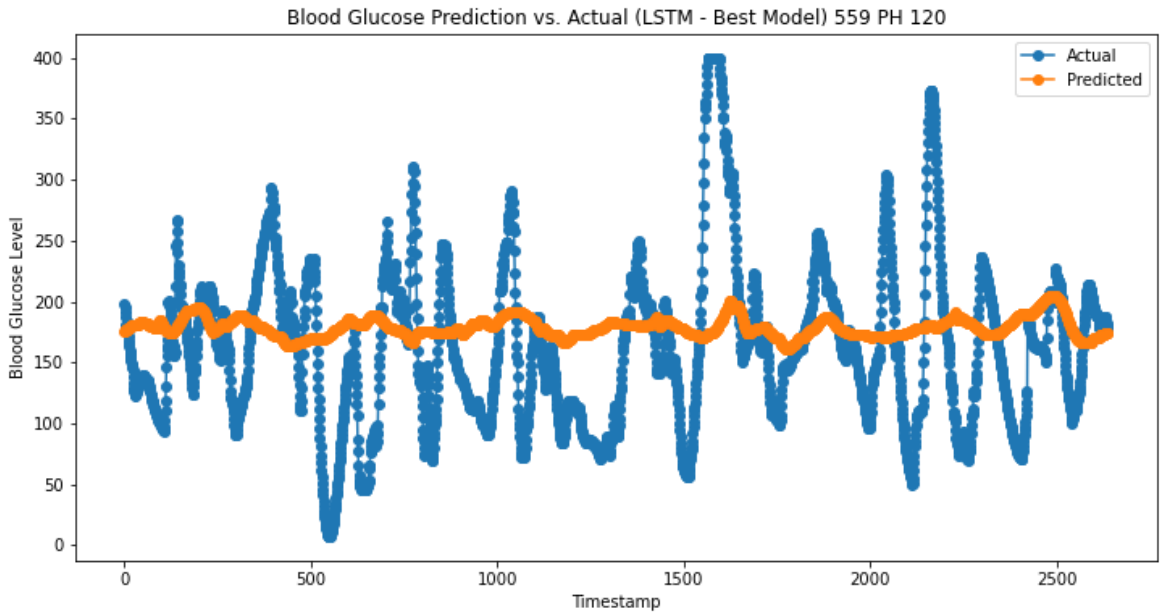


Figure 4: Blood Gluose Prediction vs actual for patient 559 for ph = 120 min.



Conclusion

In this challenge, it was compared three different prediction horizons: 30, 60, and 120 minutes. The performance metrics point to better results with a 30-minute PH. But only with a more robust training and model will it be possible to assess the veracity of this hypothesis.

References

[1] TENA, Felix, et al. A Critical Review of the state-of-the-art on Deep Neural Networks for Blood Glucose Prediction in Patients with Diabetes. *arXiv preprint arXiv:2109.02178*, 2021.

[2] MARLING, Cindy; BUNESCU, Razvan. The OhioT1DM dataset for blood glucose level prediction: Update 2020. In: *CEUR workshop proceedings*. NIH Public Access, 2020. p. 71.