engineering proceedings

MDPI

Proceeding Paper

Glucose Prediction with Long Short-Term Memory (LSTM) Models on Three Distinct Populations †

Cleber Carvalho [1] and Zilu Liang [1,2]

[1] Ubiquitous and Personal Computing Lab, Faculty of Engineering, Kyoto University of Advanced Science (KUAS), Kyoto 615-8577, Japan; email1@emai.com (C.C.); email2@email.com (Z.L.)
[2] Institute of Industrial Science, The University of Tokyo, Tokyo 153-8505, Japan
* Correspondence:
† Presented at The 11th International Electronic Conference on Sensors and Applications (ECSA-11), 26–28 November 2024; Available online: https://sciforum.net/event/ecsa-11.

Abstract: Diabetes mellitus is a chronic metabolic disorder characterized by dysregulation of blood glucose, which can lead to a range of serious health complications if not properly managed. Continuous glucose monitoring (CGM) is a cutting-edge technology that tracks glucose levels in real-time, providing continuous and detailed information about glucose fluctuations throughout the days. The CGM data can be leveraged to train deep learning models forecasting blood glucose levels. Several deep learning based glucose prediction models have been developed for diabetes populations, but their generalizability to other populations such as prediabetic individuals remains largely unknown. Prediabetes is a condition where blood glucose levels are higher than normal but not yet high enough to be classified as diabetes. It is a critical stage where intervention can prevent the progression to type 2 diabetes. To fill in the knowledge gap, we developed Long Short-Term Memory (LSTM) glucose prediction models tailored for three distinct populations: type 1 diabetes (T1D), type 2 diabetes (T2D), and prediabetic (PRED) individuals. We evaluated the internal and external validity of these models. The results showed that the model constructed with the prediabetic dataset demonstrated the best internal and external validity in predicting glucose levels across all three test sets, achieving a normalized RMSE (NRMSE) of 0.21 mg/dL, 0.11 mg/dL, 0.25 mg/dL when tested on the prediabetic, T1D, and T2D test sets, respectively.

Keywords: Continuous glucose monitoring; glucose prediction; machine learning; deep learning; LSTM; prediabetes; T1D; T2D

Citation: Carvalho, C.; Liang, Z. Glucose Prediction with Long Short-Term Memory (LSTM) Models on Three Distinct Populations. *Eng. Proc.* **2024**, *6*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s):

Published: 26 November 2024

1. Introduction

Diabetes Mellitus is a disease characterized by elevated blood glucose levels due to impaired insulin production or insulin resistance. It is broadly classified into two primary types: Type 1 diabetes (T1D), an autoimmune condition resulting in the destruction of insulin producing β-cells in the pancreas and Type 2 diabetes (T2D), which is predominantly associated with insulin resistance. Another increasingly common condition is prediabetes (PRED), characterized by higher-than-normal blood glucose levels that are not yet high enough to be diagnosed as diabetes. Often a precursor to T2D and cardiovascular diseases, PRED serves as a critical warning and a time point for intervention. It is usually diagnosed through measurements of glycated hemoglobin (HbA1c) when it is at a value of 5.7–6.4% [1].

Continuous glucose monitoring (CGM) is a method for tracking blood glucose levels in real-time, providing a view of glucose trends and fluctuations throughout the day and generating a large amount of data. This data can be utilized to uncover insights into glycemic dynamics and their relationship to other aspects of human physiology and behavior [2,3]. Deep learning (DL) has emerged as a powerful technique to predict glucose based

on CGM data in individuals with diabetes and thus be able to generate more accurate predictions. Long Short-Term Memory (LSTM) is a popular DL algorithm that can work with sequential and long CGM data, identifying complex patterns and relationships within the data and provide real-time glucose predictions [4]. These models have been widely used in subjects with T1D to predict glucose levels. On the other hand, there are few studies related to glucose prediction in prediabetes population, highlighting the need for more studies targeting this population. Acknowledging the differences in glycemic dynamics across T1D, T2D and prediabetes, we developed three LSTM models on three distinct populations and investigated the internal and external validation of the developed models on datasets of T1D and T2D subjects.

2. Materials and Methods

2.1. Dataset

Three databases were used in this study. For individuals with T1D, the Ohio dataset contains data from 12 individuals, of whom 58.3% were women and the mean age was 50 years. They were undergoing treatment with Medtronic Enlite 530G or 630G insulin pump [8]. There is no information about HbA1c values of this T1D cohort. Data collection spanned over 8 weeks using Medtronic Enlite CGM sensors. Blood glucose levels were measured every 5 min. The second dataset has time series of blood glucose readings from 100 individuals, of which 44% are women and the average age is 60.1 years, with T2D, who wore a FreeStyle Libre sensor for 3 to 14 days. Glucose data was automatically stored in the sensor every 15 min [4]. After removing duplicated data, 92 people were included in the T2D dataset for model training and test. The last dataset contains information collected from 16 prediabetic subjects, which 56.2% was female using Dexcom 6 for 10 days [1]. Glucose readings were recorded every 5 min. The average of HbA1c of the prediabetic cohort was 75.9 mmol/mol.

2.2. Model Development

The LSTM models were developed to predict the glucose level at time t + 1 based on the glucose level at time t. Before fitting the dataset, the glucose data were scaled using MinMaxScaler from the scikit-learn Python module. The LSTM models were built using the Keras platform and have 128 LSTM units, followed by a dense layer (150 units), dropout layer (0.20), dense layer (100 units), dropout layer (0.15), dense layer (50 units), dense layer (20 units) and a final layer with one unit (for prediction). ReLU was used as the activation function with Adam optimizer. The loss was calculated in mean squared error (MSE) and later converted into root mean squared error (RMSE). They were trained for 200 epochs with a batch size of 32. The models were tuned using 5-fold cross validation.

The three models were trained with data from the subjects with the greatest number of glucose records in each dataset separated T1D (3066), T2D (1328) and PRED (2846). For internal validation, each model was tested on the rest of the subjects in the correspondent dataset. For external validation, each model was tested on the entirely different datasets.

2.3. Evaluation Measures

Mean absolute error (MAE), root mean squared error (RMSE) and normalized RMSE (NRMSE) were used to evaluate the models. Prior studies recommended NRMSE as a better metric over the other two, as it is useful for comparing models of different scales [9]. In this study, NRMSE was calculated as the RMSE normalized over the standard deviation of true values of each dataset.

This study employed the Bland-Altman (B&A) analysis to compare the differences between the true values and the model predictions. It plots the difference between the two measurements against their average. The B&A method quantifies the agreement between two quantitative measurements by analyzing the mean difference and establishing limits of agreement (LoA). These statistical limits are calculated by using the mean and the

standard deviation of the differences between two values. The plot offers a straightforward approach to detect any bias in the mean differences and to estimate an agreement interval, within which 95% of the differences between the two methods are expected to fall [12].

The Continuous Glucose-Error Grid Analysis (CG-EGA) is a method of evaluation the accuracy of continuous glucose-monitoring. It serves as an evolution of the original Error Grid Analysis (EGA), which was created to assess the clinical accuracy of blood glucose readings obtained through estimation or self-monitoring systems. The CG-EGA operates on the principle that the data generated by a monitoring system should be reliable enough to support clinically accurate decision-making by the user. The CG-EGA graphic plots sensor blood glucose values (SBG) versus reference blood glucose (RBG) divided into A, B, C, D and E zones. Zones A and B are usually considered to be clinically acceptable, otherwise zones C to E are considered to be clinically significant errors.

3. Results

Table 1 shows a summary of the mean and standard deviation of the performance achieved by the three LSTM models in internal and external tests. The results showed that the model LSTM_pred obtained better MAE and RMSE metrics compared to the others. We observed that MAE and RMSE metrics are higher when the model was tested on the T2D dataset. This is due to the fact that the prediction window of this dataset is 15 min, which is longer than the 5-min prediction window in the other two test sets thus causing larger errors. The NRMSE metric demonstrates lower values for the T1D test set, which can be attributed to the greater variability in glucose levels within this group. Our findings highlight significant differences in model performance when evaluated using different metrics.

Table 1. Summary of metrics of all models.

Model	Dataset	MAE	RMSE	NRMSE
LSTM_pred	PRED	2.66 ± 0.54	4.00 ± 0.83	0.21 ± 0.04
	T1D	4.07 ± 0.43	6.39 ± 1.25	0.11 ± 0.02
	T2D	6.83 ± 1.48	9.55 ± 2.06	0.25 ± 0.06
LSTM_t1d	PRED	2.70 ± 0.60	4.07 ± 0.88	0.22 ± 0.05
	T1D	4.24 ± 0.38	6.59 ± 1.17	0.12 ± 0.02
	T2D	6.70 ± 1.44	9.54 ± 2.04	0.25 ± 0.06
LSTM_t2d	PRED	3.11 ± 0.65	4.55 ± 0.96	0.26 ± 0.05
	T1D	5.97 ± 0.69	8.77 ± 1.18	0.17 ± 0.01
	T2D	7.45 ± 1.74	10.42 ± 2.27	0.29 ± 0.05

Figure 1 illustrates the box-plots of all models metrics when tested on the three different datasets. The results corroborate those found in Table 1, observing the lowest values of the MAE and RMSE metrics for the prediabetic population, while NRSME presents a lower value for the T1D dataset. Analyzing the graphs, we observe the large dispersion of errors for the T2D dataset, which is likely due to its large number of subjects.

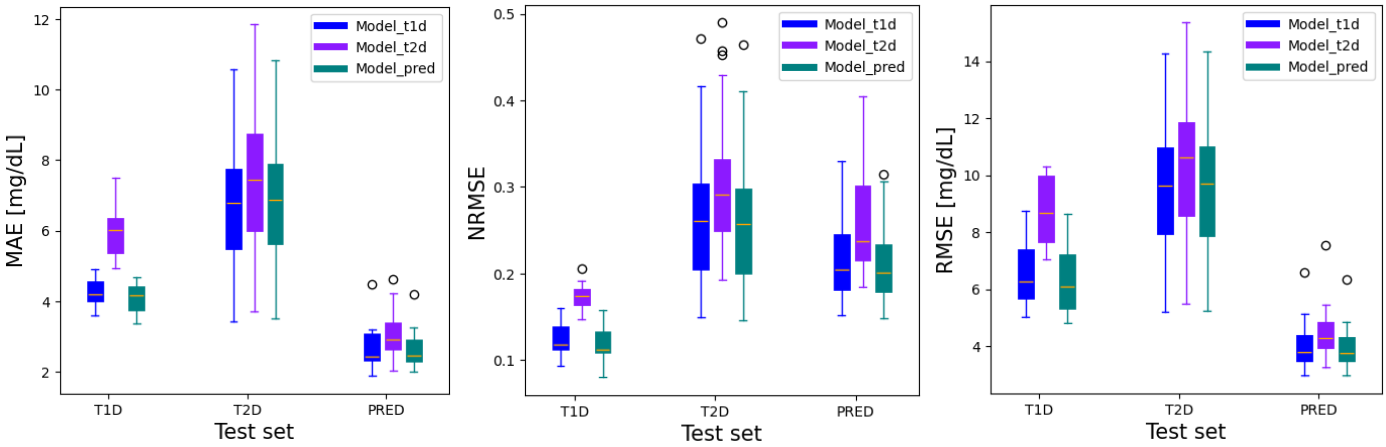


Figure 1. Box-plot of the metrics. (**a**) MAE, (**b**) RMSE, (**c**) NRMSE.

3.1. Bland-Altman Plots

Figures 2–4 represent the B&A plots for the models applied in all test sets. We observed that the majority of data points are within the upper and lower limits of agreements (LoA) and are centered around the average in all B&A plots. No systematic bias was detected in any of the models. In addition, the graphs related to the T2D subjects demonstrate a larger standard deviation and consequently a larger LoA compared to all others, which can be explained by this data set being recorded at 15-min interval, while the other data sets are recorded at 5-min. Besides, this data set has the largest number of individuals (n = 92). The Table 3 shows a summary of all values of the models and the Model_pred obtained the best results as indicated by the narrowest LoA especially when the test set were subjects with prediabetes (+8 and −8 mg/dL).

Table 2. Summary of Bland-Altman results of all models.

Model	Dataset	Mean Differences	+1.96 DP	−1.96 DP
LSTM_pred	PRED	0.02	8.0	−8.1
	T1D	−0.72	12.0	−13.0
	T2D	−0.72	18.0	−18.1
LSTM_t1d	PRED	0.87	8.9	−7.2
	T1D	0.95	14.0	−12.0
	T2D	0.32	19.0	−19.0
LSTM_t2d	PRED	1.43	10.0	−7.3
	T1D	1.81	19.0	−15.0
	T2D	0.50	21.0	−21.0

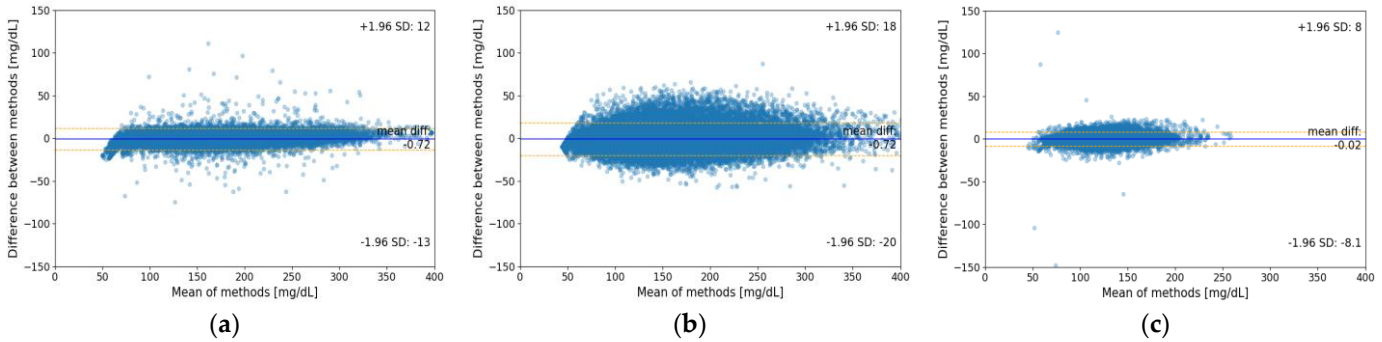
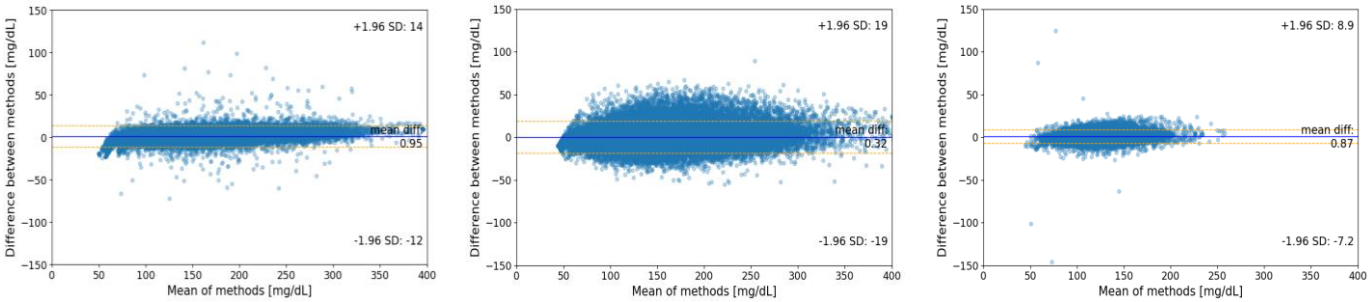
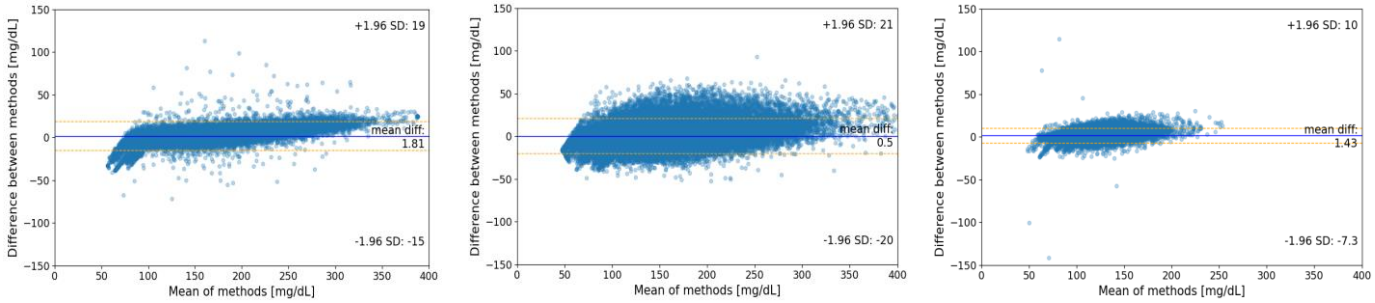


Figure 2. Bland-Altman plots of Model_pred on three test sets. (**a**) T1D, (**b**) T2D, (**c**) PRED.



(**a**) (**b**) (**c**)

Figure 3. Bland-Altman plots of Model_t1d on three test sets. (**a**) T1D, (**b**) T2D, (**c**) PRED.



(**a**) (**b**) (**c**)

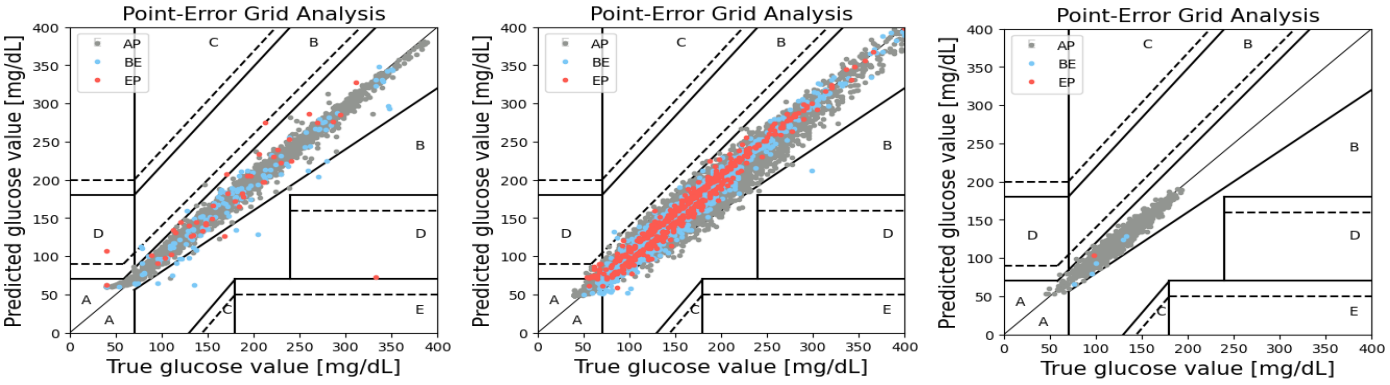
Figure 4. Bland-Altman plots of Model_t2d on three test sets. (**a**) T1D, (**b**) T2D, (**c**) PRED.

3.2. Continuous Glucose Error Grid Analysis—CG-EGA

Table 3 shows information from the analysis of the Continuous Glucose Error Grid Analysis (CG-EGA) graphs of all models when validated on the three test sets. Figures 5–7 present the EGA plots for the models applied in the three datasets.

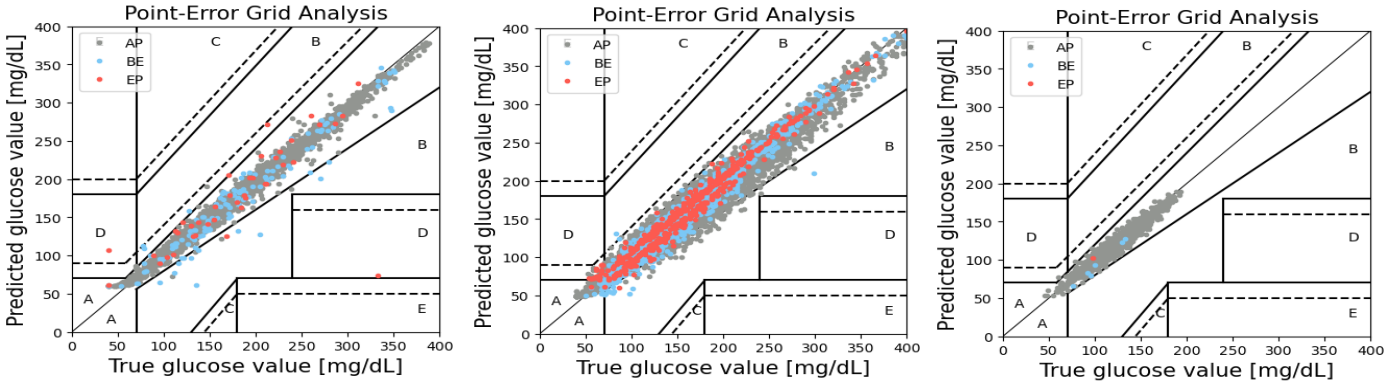
Table 3. Summary of CG-EGA of all models.

Model	Dataset	AP	BE	EP
LSTM_pred	PRED	99.8%	0.15%	0.15%
	T1D	99.6%	2.50%	0.90%
	T2D	89.9%	7.20%	2.80%
LSTM_t1d	PRED	99.8%	0.14%	0.06%
	T1D	96.8%	2.39%	0.81%
	T2D	89.9%	7.22%	2.81%
LSTM_t2d	PRED	99.7%	0.15%	0.10%
	T1D	95.3%	2.34%	2.36%
	T2D	89.9%	7.52%	2.50%



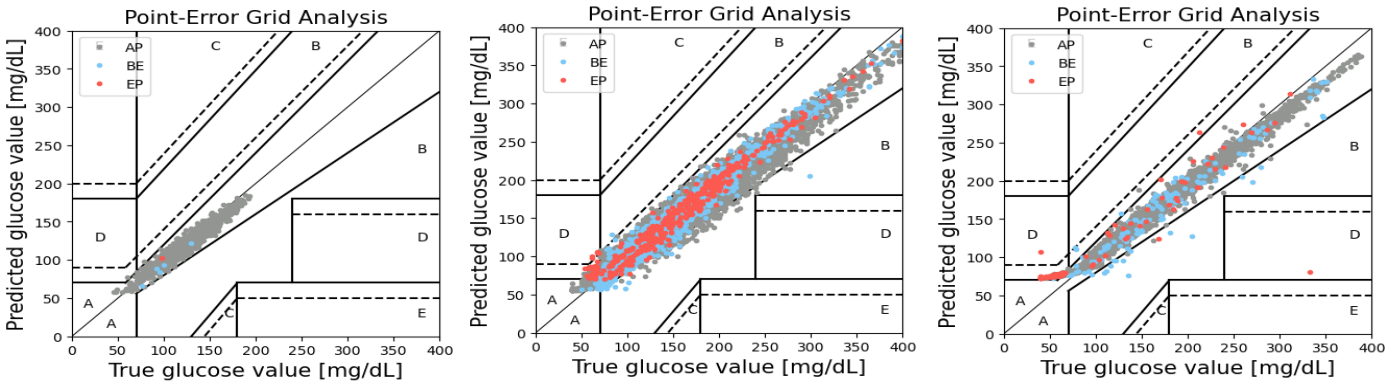
(**a**) (**b**) (**c**)

Figure 5. CG-EGA plots of Model_pred on three test sets. (**a**) T1D, (**b**) T2D, (**c**) PRED.



(**a**) (**b**) (**c**)

Figure 6. CG-EGA plots of Model_t1d on three test sets. (**a**) T1D, (**b**) T2D, (**c**) PRED.



(**a**) (**b**) (**c**)

Figure 7. CG-EGA plots of Model_t2d on three test sets. (**a**) T1D, (**b**) T2D, (**c**) PRED.

The CG-EGA results indicate that none of the three models achieved good clinical accuracy on the T2D dataset. The Model_pred achieved the highest values in the AP region when validated on the prediabetic test set and reasonably good AP when validated on the T1D and T2D test sets. As shown in the Table 3 this pattern manifests again for Model_t1d and Model_t2d. The compromised performance of all three models may be explained by the longer records interval of the T2D dataset (i.e., 15 min) compared to the other two datasets. The T2D dataset also has a much larger number of individuals, potentially leading to more prediction errors, which may in turn results in wrong clinical decisions.

4. Discussion

4.1. Principal Findings

Most existing deep learning based glucose prediction models only underwent an internal validation, leaving the generalizability of the models to other populations unclear. In this study, we trained three different models with different types of diabetes conditions (T1D, T2D, PRED) and performed both internal and external validation.

4.2. Comparison with Prior Studies

Numerous studies have endeavored to build deep learning models for predicting glucose levels in diabetic population. Models such as LSTM were used in the BGLP Challenge in 2020, using the OhioT1DM dataset to forecast glucose values. Bhimireddy et al. used DL models to forecast glucose level at 30 min horizon, the model LSTM achieved average RMSE value of 25.0 mg/dL [5]. LSTM type models were also used by [6], being tested on T1D subjects. They achieved an MAE of 19.4 mg/dL and an average RMSE of 14.1 mg/dL. A novel multivariate predictor with a multi-scale LSTM was used to predict glucose of T1D subjects through 30-min and 60-min [7]. The RMSE values of 30-min and 60-min predictions were 19.0 and 32.0 mg/dL, respectively, and the MAE values of 30-min and 60-min predictions were 13.5 and 23.8 mg/dL. In addition, Butt et al. [10] developed a multi-layered LSTM based recurrent neural network for forecasting horizon of 30 and 60-min. The model achieved the lowest RMSE score of 14.76 mg/dL and 25.48 mg/dL for prediction horizons of 30 and 60-min, respectively. Research effort has also been made to improve the interpretability of LSTM-type glucose prediction models. Researchers of [11] developed personalized bidirectional LSTM equipped with a tool that enables its interpretability. Their algorithm was able to preserve the physiological meaning of the considered inputs and achieved a RMSE of 20.20 mg/dL and a MAE of 14.74 mg/dL for 30-min prediction. In comparison to those existing models, all our three models obtained better results regarding MAE and RMSE. However, it is worth noting that our work differs from prior studies regarding the prediction horizon. While most existing models were developed for 30 or 60-min horizon, our models were developed for a much shorter horizon of 5 or 15 min depending on the dataset adopted.

4.3. Limitations and Future Work

The present study has several limitations. The first limitation is the sizes of the datasets. For example, the T1D dataset consists of only 12 participants, while the PRED dataset only has 16 participants. The granularity of demographic information also varies across the datasets used, with the T1D and PRED datasets only contain age range rather than specific age of the participants. The data quality imposes another potential limitation, particularly regarding the sampling rate of the glucose levels. While the T2D data were recorded every 15 min, the T1D and PRED groups had their records every 5 min. This difference may have produced distorted performance when the models were tested on the T2D dataset. In our future work, we will attempt other types of deep learning algorithms, as well as utilizing larger datasets with diverse demographics. We also plan to incorporate other types of signals in addition to historical glucose data into model construction.

Author Contributions:.

Funding:

Institutional Review Board Statement:

Informed Consent Statement:

Data Availability Statement:

Conflicts of Interest:

References

1. Goldberger, A.L.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, 101, e215–e220.
2. Liang, Z. Mining associations between glycemic variability in awake-time and in-sleep among non-diabetic adults. *Front. Med. Technol.* **2022**, *4*, 1026830.
3. Bertrand, L.; Cleyet-Marrel, N.; Liang, Z. The Role of Continuous Glucose Monitoring in Automatic Detection of Eating Activities. In Proceedings of the 2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech), Nara, Japan, 9–11 March 2021; pp. 313–314.
4. Zhao, Q.; Zhu, J.; Shen, X.; Lin, C.; Zhang, Y.; Liang, Y.; Cao, B.; Li, J.; Liu, X.; Rao, W.; Wang, C. Chinese diabetes datasets for data-driven machine learning. *Sci. Data* **2023**, *10*, 35.
5. Bhimireddy, A; Priyanshu, S.; Bolu, O.; Judy, W.G.; Saptarshi, P. Blood Glucose Level Prediction as Time-Series Modeling using Sequence-to-Sequence Neural Networks. In Proceedings of the KDH@ ECAI 2020, Santiago de Compostela, Spain, 29–30 August 2020; pp. 125–130.
6. Heydar, K.; Hoda, N.; Jackie, E.; Mohammed, B. Multi-lag Stacking for Blood Glucose Level Prediction. In *CEUR-Workshop Proceedings, Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data colocated with 24th European Conference on Artificial Intelligence, Santiago de Compostela, Spain, 29–30 August 2020*; CEUR-WS Team: Aachen, Germany, 2020; pp. 146–150.
7. Yang, T.; Wu, R.; Tao, R.; Wen, S.; Ma, N.; Zhao, Y.; Yu, X.; Li, H. Multi-Scale Long Short-Term Memory Network with Multi-Lag Structure for Blood Glucose Prediction. In Proceedings of the KDH@ ECAI 2020, Santiago de Compostela, Spain, 29–30 August 2020; pp. 136–140.
8. Marling, C.; Bunescu, R. The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020. *CEUR Workshop Proc.* **2020**, *2675*, 71–74.
9. Jacobs, P.G.; Herrero, P.; Facchinetti, A.; Vehi, J.; Kovatchev, B.; Breton, M.D.; Cinar, A.; Nikita, K.S.; Doyle, F.J.; Bondia, J.; et al. Artificial Intelligence and Machine Learning for Improving Glycemic Control in Diabetes: Best Practices, Pitfalls, and Opportunities. *IEEE Rev. Biomed. Eng.* **2024**, *17*, 19–41.
10. Butt, H.; Khosa, I.; Iftikhar, M.A. Feature Transformation for Efficient Blood Glucose Prediction in Type 1 Diabetes Mellitus Patients. *Diagnostics* **2023**, 13, 340.
11. Cappon, G.; Meneghetti, L.; Prendin, F.; Pavan, J.; Sparacino, G.; Del Favero, S.; Facchinetti, A. A Personalized and Interpretable Deep Learning Based Approach to Predict Blood Glucose Concentration in Type 1 Diabetes. In Proceedings of the KDH@ ECAI 2020, Santiago de Compostela, Spain, 29–30 August 2020; pp. 81–85.
12. Giavarina, D. Understanding Bland Altman analysis. *Biochem. Med.* **2015**, *25*, 141–151.