Journal of Data Science and Modeling, Vol. 2, No. 1, 115-131, December 2023

Research Manuscript

Blood glucose range, direction, and abnormal value prediction based on regular and NPH insulin doses using machine learning and statistical methods

Morteza Amini *, Kiana Ghasemifard

[1]Dpt. of Statistics, School of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, Iran.

Received: 21/04/2024 Accepted: 20/09/2024

Abstract:

The diabetes data set gathered by Michael Kahn at Washington University, St. Louis, MO, which is available online at the UCI Machine Learning Repository, is one of the rarely used data sets, especially for glucose prediction purposes in diabetic patients. In this paper, we study the problem of blood glucose range prediction, rather than raw glucose prediction, along with two other important tasks: the detection of glucose increments or decrements and the prediction of abnormal values, based on regular and NPH insulin doses, using this data set. Two commonly used machine learning approaches for time series data, namely LSTM and CNN, are used along with a promising statistical regression approach, specifically the non-parametric multivariate Gaussian additive mixed model, for the prediction task. It is observed that although LSTM and CNN models are preferable in terms of prediction error, the statistical method performs significantly better for detecting abnormal values, which is a critical task for diabetic patients.

Keywords: Long-Short-Term-Memory recurrent neural network, Convolutional neural network, multivariate time series, sequential data, generalized additive mixed models.

Mathematics Subject Classification (2010): 62J12, 68T05.

*Corresponding Author: morteza.amini@ut.ac.ir

1. Introduction

Diabetes is known as a disease characterized by high blood sugar levels, which is either the result of a lack of insulin production or a disorder in its effectiveness. In the absence of insulin, glucose is not absorbed by the cells, and blood glucose levels increase (Makroum et al., 2022). The International Diabetes Federation (IDF) estimates that 578 million adults will have diabetes by 2030, and 700 million by 2045 (IDF Diabetes Atlas, 2021).

Over the last decade, many advanced diabetes control and detection technologies have been developed using artificial intelligence (Ameen et al., 2021). Many predictive analytics methods, including machine learning algorithms, data mining techniques, and statistical approaches, are used to predict future events and risks for diabetic patients. One of the most important tasks in aiding diabetic patients is blood glucose prediction. Various methods, including classical time series approaches, regression models, and machine learning techniques, have been applied to different data sets to predict blood glucose based on various input variables, such as insulin, carbohydrate intake, and physical activity (see Gani et al., 2008; Eren-Oruklu et al., 2009; Sparacino et al., 2007; Turksoy et al., 2013; Wang et al., 2014; Xie and Wang, 2017, among others). In recent years, there has been a growing trend of applying machine learning algorithms to predict blood glucose levels (see Zecchin et al., 2012; Plis et al., 2014; Mirshekarian et al., 2017; Mhaskar et al., 2017; Fox et al., 2018; Xie and Wang, 2020, and references therein). For a systematic review of machine learning and smart devices for diabetes management, see Makroum et al. (2022).

In this paper, we analyze one of the less frequently considered diabetes data sets, gathered by Michael Kahn at Washington University, St. Louis, MO, which is available online at UCI machine learning repository (2017). This data set has only been used by several authors (Xu et al., 2019; Backurs et al., 2019; Yang and Tan, 2021; Gosiewska et al., 2019; Atamturk and Gomez, 2019), mainly as a benchmark to examine the performance of regression methods. Although many other papers mentioned on the data set page in UCI machine learning repository (2017) claim to have used this data set, they have actually used another diabetes data set, called the Pima Indians diabetes data set, also available online at UCI machine learning repository (2017). The Pima Indians data set is used for classifying subjects as diabetic or non-diabetic (see Zhou and Jiang, 2004; Melville and Mooney, 2004; Eggermont et al., 2004, among many others), while the diabetes data set gathered by Michael Kahn at Washington University, St. Louis, MO, is more suitable for regression tasks since the response variables are blood glucose measurements.

Another novel aspect of our work is related to the prediction task. Many of the works mentioned above focus on predicting the raw blood glucose levels of patients,

while in this paper, we aim to predict the blood glucose range — the minimum and maximum glucose levels during a day — along with two other important events. The first event is the increment or decrement of the maximum or minimum blood glucose, and the second event is an abnormal situation, defined as high glucose levels where the maximum blood glucose exceeds 180 mg/dL. We use two machine learning (ML) methods and one statistical method (SM) for predicting the current and future blood glucose range, as well as the two criteria mentioned above.

The rest of the paper is organized as follows. The diabetes data set is introduced in Section 2, along with the pre-processing approach used in this paper. The ML and SM approaches are described in Section 3. Finally, the results of the data analysis are presented in Section 4. The R and Python codes are available online at https://github.com/mortamini/diabetes/.

2. Diabetes data set and pre-processing

The diabetes data set, gathered by Michael Kahn at Washington University, St. Louis, MO, and available online at UCI machine learning repository (2017), includes patient records obtained from two sources: an automatic electronic recording device and paper records. The automatic device had an internal clock to timestamp events, whereas the paper records only provided "logical time" slots (breakfast, lunch, dinner, bedtime). For paper records, fixed times were assigned to breakfast (08:00), lunch (12:00), dinner (18:00), and bedtime (22:00). Thus, paper records have fictitious uniform recording times, whereas electronic records have more realistic time stamps.

The raw diabetes data set contains files for 69 cases (subjects), and for each case, it includes the following features:

(1) Date in MM-DD-YYYY format

(2) Time in XX:YY format

(3) Code

(4) Value

The Code field is deciphered in Table 1

There are a large number of missing values in this data set since most subjects reported only a few codes during the day. Therefore, we only considered codes 33 (regular insulin dose), 34 (NPH insulin dose), and 58-64 (blood glucose measurements).

The first step in pre-processing was to impute the missing values. We used the Multivariate Imputation by Chained Equations (MICE) method (Van Buuren

Table 1: Description of the codes in the diabetes data set.

code	description
33	Regular insulin dose
34	NPH insulin dose
35	UltraLente insulin dose
48	Unspecified blood glucose measurement
57	Unspecified blood glucose measurement
58	Pre-breakfast blood glucose measurement
59	Post-breakfast blood glucose measurement
60	Pre-lunch blood glucose measurement
61	Post-lunch blood glucose measurement
62	Pre-supper blood glucose measurement
63	Post-supper blood glucose measurement
64	Pre-snack blood glucose measurement
65	Hypoglycemic symptoms
66	Typical meal ingestion
67	More-than-usual meal ingestion
68	Less-than-usual meal ingestion
69	Typical exercise activity
70	More-than-usual exercise activity
71	Less-than-usual exercise activity
72	Unspecified special event

and Groothuis-Oudshoorn, 2011) to impute the missing values. Since our aim was to predict daily information, we computed the average regular and NPH insulin doses for each day (as the input variables). The blood glucose range was then obtained by computing the minimum and maximum daily blood glucose based on codes 58-64 (as the output variables). Therefore, the dimension of the input and output variables is 2, considering the independent time series for each case.

Figure 1 shows the plot of the number of days recorded for each case in the diabetes data set.

The resulting pre-processed data set is suitable for predicting the current blood glucose range based on the average regular and NPH insulin doses. However, a more important task is to predict the future blood glucose range based on the current values of the variables. Thus, a one-day lagged data set was also constructed, where the next day's blood glucose range is the output variable (case-independent values of dimension 2) and the current values of average regular and NPH insulin doses, the current values of the blood glucose range, and the pre-determined val-

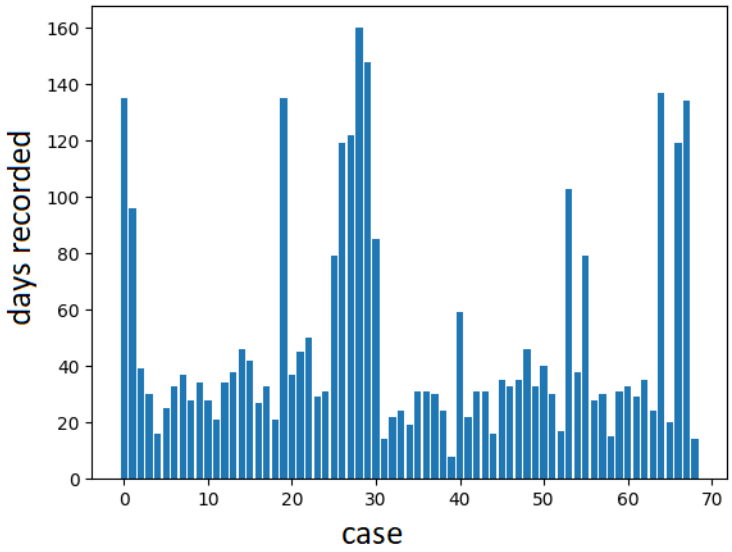

Figure 1: Number of days recorded for each case in diabetes data-set.

ues of the next-day average regular and NPH insulin doses are the input variables (case-independent time series of dimension 6).

3. Methods

Two ML and one SM approaches are considered for blood glucose range prediction based on time series data. The Long-Short-Term-Memory recurrent neural network (LSTM) and Convolutional Neural Network (CNN) are the most commonly used ML methods for sequential data sets. The SM method, the Multivariate Gaussian Additive Mixed Model (MGAMM), is also used as a statistically promising method for regression tasks.

3.1 LSTM

The LSTM (Hochreiter and Schmidhuber, 1997) is a generalization of the recurrent neural network (RNN) that is used for sequential data sets, including a sequence of inputs $x_t t = 1^T$ and an output (target) variable, which can be either a sequence $y_t t = 1^T$ or a single output y. RNNs have three common types: many-to-many, many-to-one, and one-to-many. A single-hidden-layer, many-to-many RNN has

the following structure:

$$\begin{aligned} x_t &= \sigma(w_x^\top x_{t-1} + b_x) \\ h_t &= \sigma(w_{h1}^\top x_t + w_{h2}^\top h_{t-1} + b_h) \\ o_t &= \sigma(w_o^\top h_t + b_o), \end{aligned} \tag{3.1}$$

where $\sigma(\cdot)$ is the activation function, and w_x, w_{h1}, w_{h2}, and w_o are unknown weight vectors, while b_x, b_h, and b_o are unknown biases that should be learned from the data by minimizing an appropriate loss function. The common loss function used for regression tasks is the mean squared error loss, which is also used in this study.

A single-hidden-layer, many-to-one RNN has the same structure as in (3.2), except that the single output is obtained as follows:

$$o = \sigma(w_o^\top h_T + b_o), \tag{3.2}$$

where h_T stands for the last value of the hidden states in the sequence.

The LSTM network is an RNN that overcomes the vanishing gradient problem (Hochreiter, 1991) encountered in traditional RNNs by flowing the information through input, output, and forget gates (Hochreiter and Schmidhuber, 1996). These gates are shown inside a cell of the LSTM network in Figure 2.

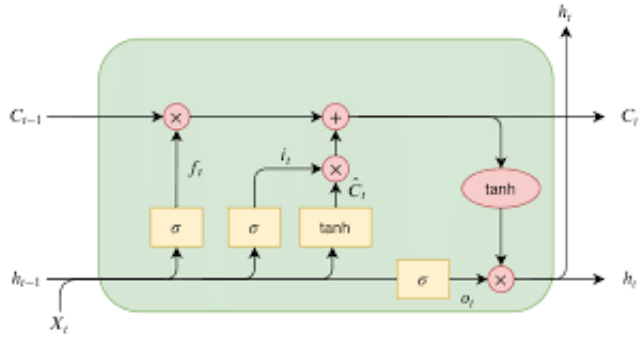

Figure 2: Architecture of LSTM model.

The LSTM has been frequently used for classification and regression tasks based on sequential data (for papers that have used LSTM for raw glucose prediction, see, for instance Makroum et al., 2022; Xie and Wang, 2020, and references therein).

In this study, we used four modules of many-to-many LSTMs for the current-time prediction of the glucose range, with an architecture presented in Figure 3 (top). A combination of two many-to-many and two many-to-one LSTM modules were also used for future prediction, with the architecture shown in Figure 3 (below).

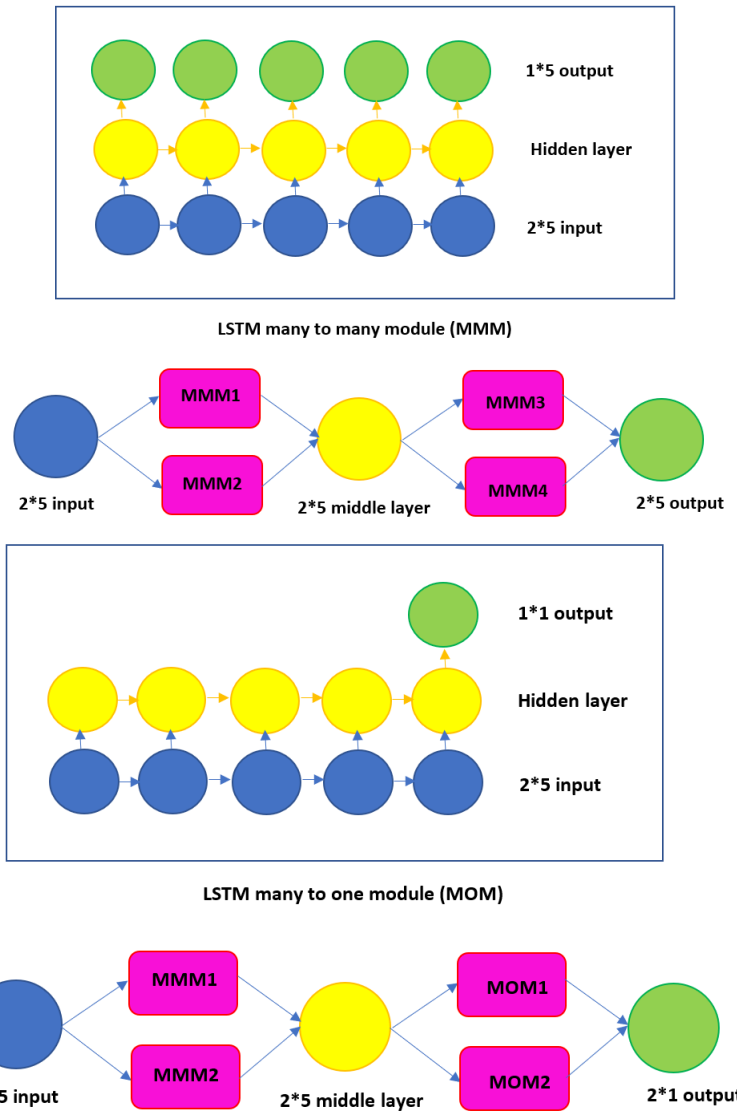

Figure 3: Architecture of LSTM for current-time (top) and future (below) Glucose-range prediction.

3.2 CNN

Another common ML method that can be used for regression based on sequential data sets is CNN (see, e.g., Venkatesan and Li, 2017). CNN is a special type of network in which the matrix multiplication operator is replaced by convolution operators. In a 2D CNN layer with activation function $\sigma(\cdot)$, a $d_1 \times d_2$ filter (kernel)

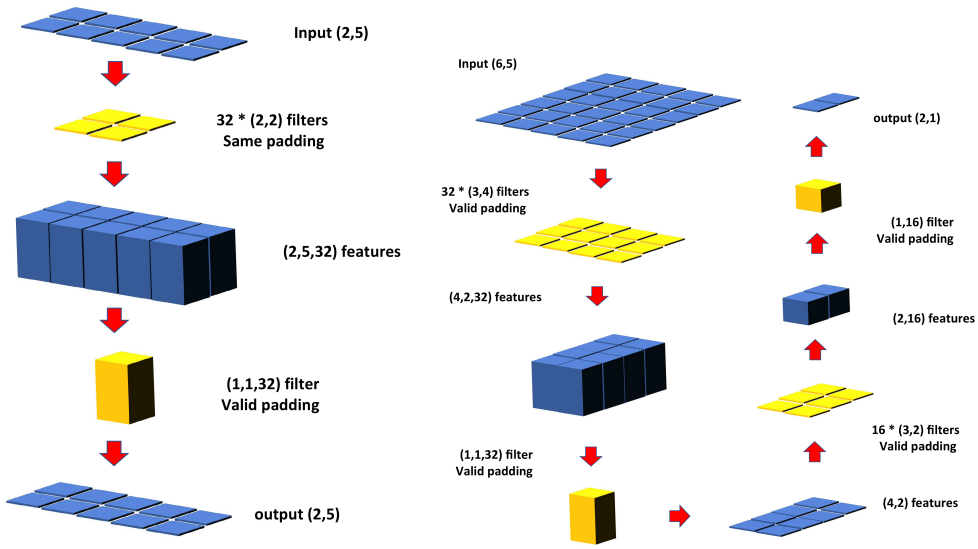

Figure 4: Architecture of CNN for current-time (left) and future (right) glucose-range prediction.

K, and stride s, the output of the layer is obtained by the following convolution:

$$Z_{ij} = \sigma\left(\sum_{l=1}^{d_1}\sum_{m=1}^{d_2} K_{lm} V_{(i-1)*s+l,(j-1)*s+m}\right) \tag{3.3}$$

The architecture of CNN for current-time glucose-range prediction is shown in Figure 4 (left), and for future prediction, the architecture is shown in Figure 4 (right).

3.3 MGAMM

Generalized Additive Models (GAMs) (Hastie and Tibshirani, 1987) are non-parametric regression models for modeling the relationship between a set of inputs and a target variable, with a distribution belonging to the exponential family of distributions. A common case is the Gaussian (or Multivariate Gaussian for multivariate targets) additive model. For our study, where the target sequence is a two-dimensional time series, the Multivariate Gaussian Additive Model (MGAM) is expressed as follows:

$$y_t^{(i)} \sim \mathcal{N}_2\left(\mu_0 + \sum_{j=1}^{p} f_j(x_{j,t}^{(i)}), \Sigma\right), \tag{3.4}$$

where $\mathcal{N}_2$ stands for the bivariate normal distribution, μ_0 is the vector of intercepts, Σ is a 2×2 unknown covariance matrix, and f_js are unknown functions (centered to

zero), which are learned from the data. Typically, one assumes that f_js belong to the space spanned by spline bases of a specified degree and are estimated through the coefficients of their expansion over this space (see Hastie and Tibshirani, 1987, and references therein).

The MGAMM is a generalization of MGAM by adding a random effect term to the mean of the multivariate Gaussian distribution. For our case, it is as follows

$$y_t^{(i)} \sim \mathcal{N}_2 \left(\mu_0 + \sum_{j=1}^{p} f(x_{j,t}^{(i)}) + a_i, \Sigma \right), \tag{3.5}$$

where $a_i \sim N(0, \sigma^2)$, is the random effect term. The random effect term is common across the sequence (i.e. for all values of t). This helps the MGAMM to model the dependency across the time series.

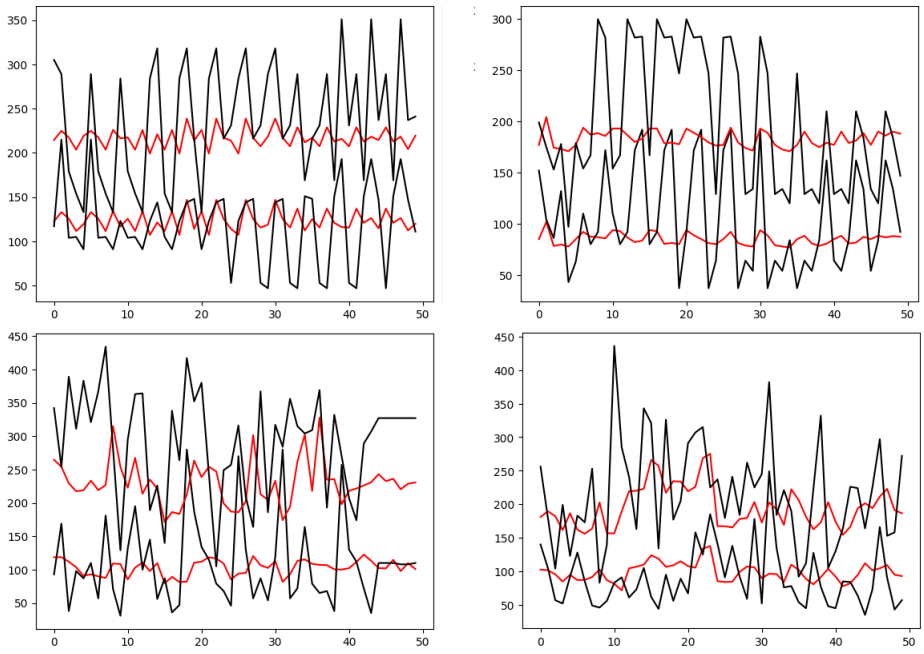

Figure 5: A sample of Glucose-range prediction results using CNN (left) and LSTM (right) for current-time (top) and future (below) prediction.

4. Results

All three models (LSTM, CNN, and MGAMM) were applied to both current-time and future glucose-range predictions. The out-of-sample prediction errors for all

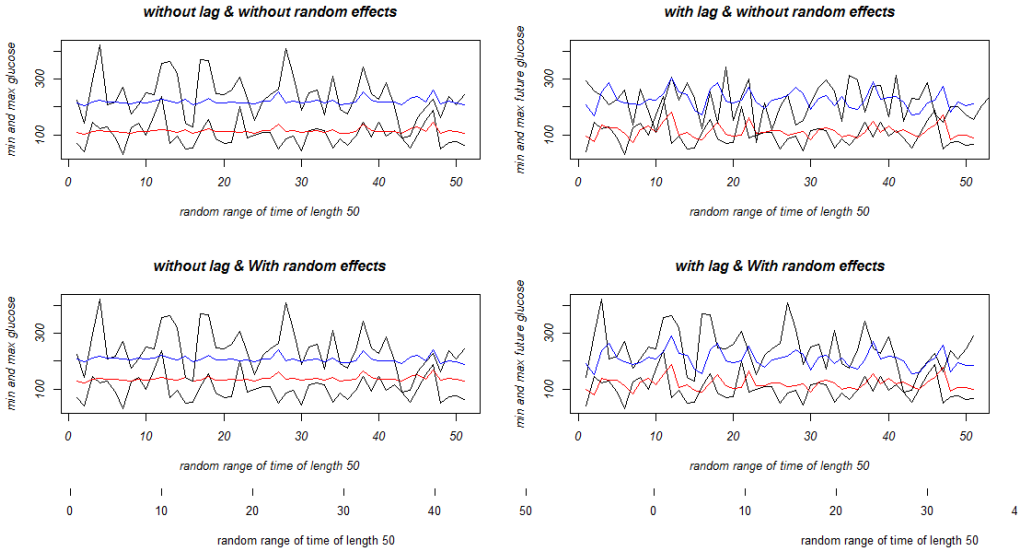

Figure 6: A sample of Glucose-range prediction results using MGAMM without (top) and with (below) random-effects for current-time (left) and future (right) prediction.

methods were computed using 3-fold Root Mean Square Error (RMSE):

(4.6)

where TS_f represents the test set of the fth fold, and N_i is the number of days reported by case i in TS_f.

The 3-fold Percent of Correct Direction (PCD) was also calculated to compare the models' ability to correctly identify the increase or decrease in glucose levels over time, as follows:

$$\text{3-fold PCD} = \frac{1}{3}\sum_{f=1}^{3}\sum_{i\in TS_f}\frac{1}{N_i}\sum_{t=2}^{N_i}\mathrm{I}\left(\text{sign}(y_t^{(i)} - y_{t-1}^{(i)}) = \text{sign}(\widehat{y}_t^{(i)} - \widehat{y}_{t-1}^{(i)})\right) \quad (4.7)$$

One critical task in glucose prediction is detecting abnormal situations where glucose levels exceed 180 mg/dL. The 3-fold Percent of Correct Abnormal Detection (CAD) criterion was defined and calculated to compare the models on this task:

$$\text{3-fold CAD} = \frac{1}{3}\sum_{f=1}^{3}\sum_{i\in TS_f}\frac{\sum_{t=1}^{N_i}\mathrm{I}\left(y_{t,\max}^{(i)} > 180 \;\;\&\;\; \widehat{y}_{t,\max}^{(i)} > 180\right)}{\sum_{t=1}^{N_i}\mathrm{I}\left(y_{t,\max}^{(i)} > 180\right)} \quad (4.8)$$

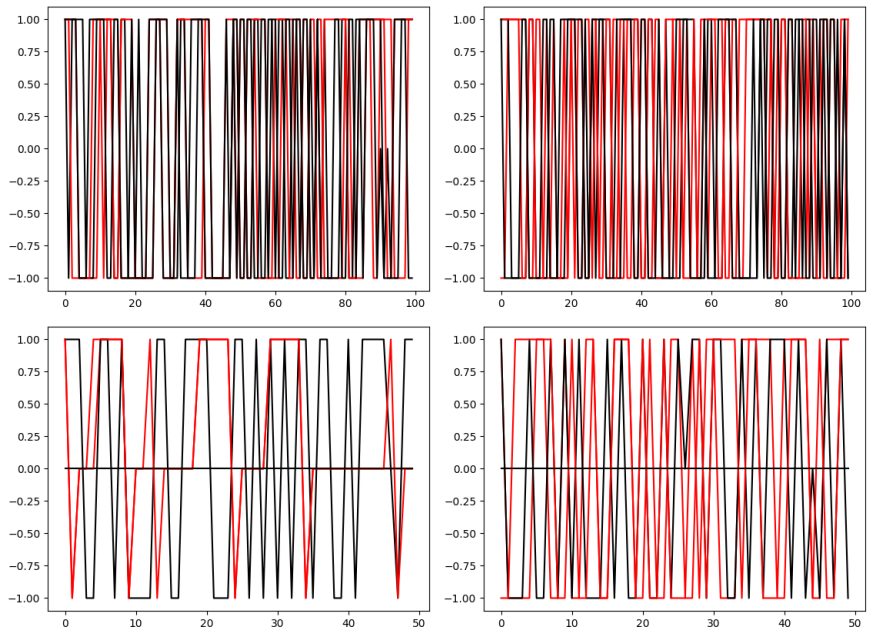

Figure 7: Direction prediction for LSTM (left) and CNN (right) for future (below) and current time (top) lag for random ranges of lengths 100 (top) and 50 (below). The +1 values are for increment, the -1 values show the decrement and zero values show no changes in time.

Table 2: Results of different methods for diabetes data set.

	criteria		
method	3-fold RMSE	3-fold PCD (%)	3-fold CAD (%)
LSTM	139.5	**42.73**	45.20
CNN	2013.8	37.57	68.59
lagged LSTM	**109.0**	12.90	48.20
lagged CNN	**109.1**	43.60	73.70
MGAM	**125.3**	**47.20**	**88.12**
lagged MGAM	201.1	35.99	**95.10**
MGAMM	128.9	**55.63**	72.30
lagged MGAMM	195.3	36.63	**84.72**

Different samples of glucose-range predictions for each method are shown in Figures 5 and 6. Additionally, the values of the three proposed criteria are presented in Table 2 for all competing models. While the ML models (LSTM and CNN) performed better based on the 3-fold RMSE and PCD, the SM model

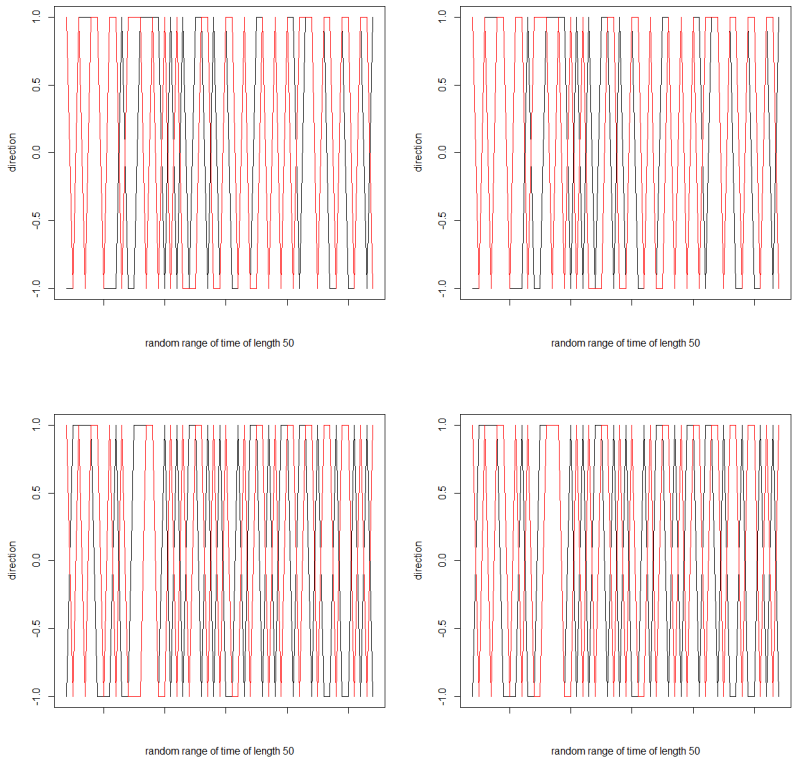

Figure 8: Direction prediction for MGAMM with (right) and without (left) random-effect for the current time (top) and future (below) for a random range of length 50. The +1 values are for increment, the -1 values show the decrement and zero values show no changes in time.

(MGAMM) achieved significantly higher values for the 3-fold CAD. Moreover, future glucose prediction is generally more useful (lagged models in Table 2), leading us to conclude that the lagged MGAMM is the most reliable technique for glucose control in diabetic patients. However, combining the lagged MGAMM with a lagged LSTM (or even CNN) could be beneficial for multi-purpose next-day glucose-range prediction.

Direction prediction samples are illustrated in Figures 7 and 8 for CNN-LSTM and MGAMM models, respectively. Furthermore, random samples of abnormal glucose detection results are presented in Figures 9 and 10.

It is important to note that RMSE, PCD, and CAD are different criteria. RMSE measures the distance error between predicted and real values, while PCD and CAD measure correct direction prediction and abnormal detection, respec-

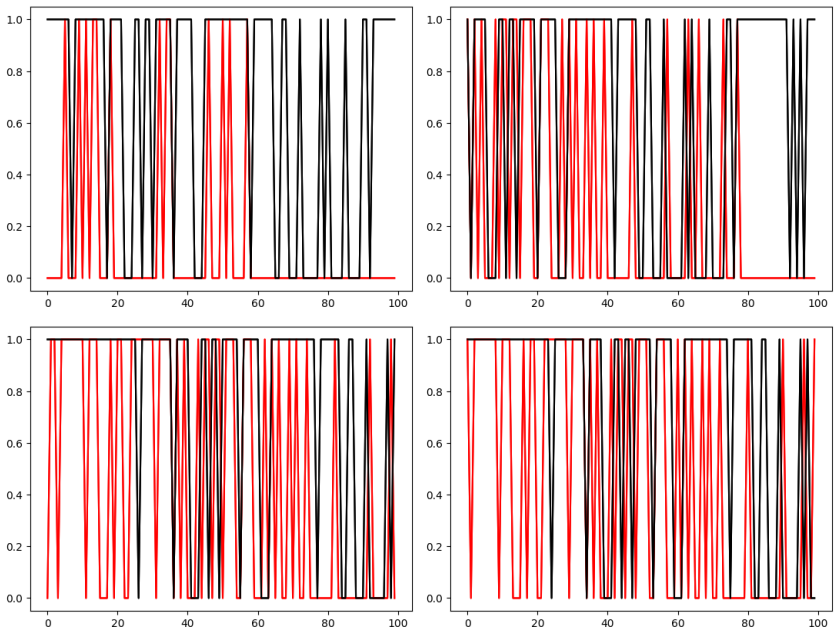

Figure 9: Abnormal glucose detection for LSTM (left) and CNN (right) for future (below) and current time (top) lag for random ranges of length 100. The 1 values are for abnormal glucose, and zero values show no abnormality.

tively. A model may achieve a low RMSE across the entire test sequence but still predict the direction (increase or decrease) incorrectly for many steps and detect only a few abnormal glucose values. This is possible since the percentage of abnormal situations is low compared to the sequence length, and a small prediction error in abnormal values (e.g., a real value of 183 mg/dL versus a predicted 178 mg/dL) may lead to incorrect detection. Furthermore, while the predicted values may closely follow real values, the direction of two successive points may be predicted in reverse.

5. Concluding remarks

It remains an open question how best to combine models to develop a multi-purpose, high-performance solution for blood glucose range and abnormal situation prediction and control. This could be explored in future research. Another area of interest is utilizing additional available information (such as codes 35, 48, 57, and 65-72) to improve prediction accuracy. Lastly, how to predict glucose ranges with incomplete inputs (with or without imputation) is another challenge for future studies.

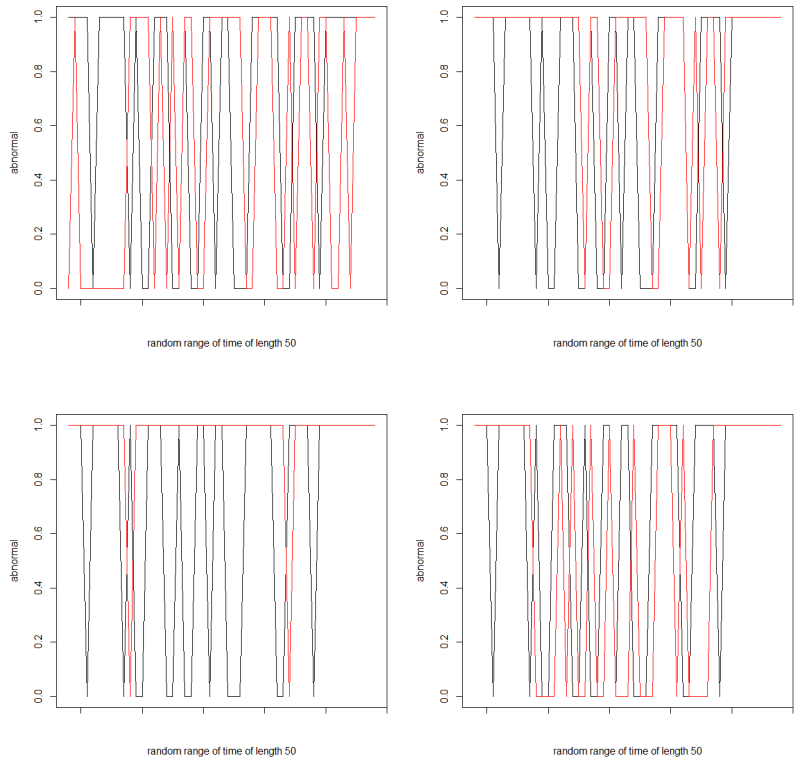

Figure 10: Abnormal glucose detection for MGAMM with (right) and without (left) random-effect for the current time (top) and future (below) for a random range of length 50. The 1 values are for abnormal glucose, and zero values show no abnormality.

References

Ameen, N., Tarhini, A., Reppel, A., and Anand, A. (2021). Customer experiences in the age of artificial intelligence. *Computers in human behavior*, **114**, 106548.

Atamturk, A., and Gomez, A. (2019). Rank-one convexification for sparse regression. *arXiv preprint arXiv:1901.10334*.

Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., and Wagner, T. (2019, May). Scalable fair clustering. In *International Conference on Machine Learning*, 405-413. PMLR.

Dua, D., and Graff, C. (2017). UCI machine learning repository. URL http://archive.ics.uci.edu/ml.

Eggermont, J., Kok, J. N., and Kosters, W. A. (2004). Genetic programming for data classification: Partitioning the search space. In *Proceedings of the 2004 ACM symposium on Applied computing*, 1001-1005.

Eren-Oruklu, M., Cinar, A., Quinn, L., and Smith, D. (2009). Estimation of future glucose concentrations with subject-specific recursive linear models. *Diabetes technology and therapeutics*, **11(4)**, 243-253.

Fox, I., Ang, L., Jaiswal, M., Pop-Busui, R., and Wiens, J. (2018, July). Deep multi-output forecasting: Learning to accurately predict blood glucose trajectories. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining*, 1387-1395.

Gani, A., Gribok, A. V., Rajaraman, S., Ward, W. K., and Reifman, J. (2008). Predicting subcutaneous glucose concentration in humans: data-driven glucose modeling. *IEEE Transactions on Biomedical Engineering*, **56(2)**, 246-254.

Gosiewska, A., Gacek, A., Lubon, P., and Biecek, P. (2019). SAFE ML: Surrogate Assisted Feature Extraction for Model Learning. *arXiv preprint arXiv:1902.11035.*

Hastie, T., and Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, **82(398)**, 371-386.

Hochreiter, S. (1991). *Untersuchungen zu dynamischen neuronalen Netzen.* Diploma, Technische Universität München, **91(1)**, 31.

Hochreiter, S., and Schmidhuber, J. (1996). LSTM can solve hard long time lag problems. *Advances in neural information processing systems*, 9.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, **9(8)**, 1735-1780.

IDF Diabetes Atlas (2021). Key global findings 2021. Available online: https://web.archive.org/web/20211208190021/https://diabetesatlas.org/

Makroum, M. A., Adda, M., Bouzouane, A., and Ibrahim, H. (2022). Machine learning and smart devices for diabetes management: Systematic review. *Sensors*, **22(5)**, 1843.

Melville, P., and Mooney, R. J. (2004). Diverse ensembles for active learning. In *Proceedings of the twenty-first international conference on Machine learning*, 74.

Mhaskar, H. N., Pereverzyev, S. V., and Van der Walt, M. D. (2017). A deep learning approach to diabetic blood glucose prediction. *Frontiers in Applied Mathematics and Statistics*, **3**, 14.

Mirshekarian, S., Bunescu, R., Marling, C., and Schwartz, F. (2017, July). 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)*, 2887-2891, IEEE.

Plis, K., Bunescu, R., Marling, C., Shubrook, J., and Schwartz, F. (2014, June). A machine learning approach to predicting blood glucose levels for diabetes management. In *Workshops at the Twenty-Eighth AAAI conference on artificial intelligence.*

Sparacino, G., Zanderigo, F., Corazza, S., Maran, A., Facchinetti, A., and Cobelli, C. (2007). Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series. *IEEE Transactions on biomedical engineering*, **54(5)**, 931-937.

Turksoy, K., Bayrak, E. S., Quinn, L., Littlejohn, E., and Cinar, A. (2013). Adaptive multivariable closed-loop control of blood glucose concentration in patients with type 1 diabetes. In *2013 american control conference*, 2905-2910, IEEE.

Van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, **45**, 1-67.

Venkatesan, R., and Li, B. (2017). *Convolutional neural networks in visual computing: a concise guide*. CRC Press.

Wang, Q., Molenaar, P., Harsh, S., Freeman, K., Xie, J., Gold, C., ... and Ulbrecht, J. (2014). Personalized state-space modeling of glucose dynamics for type 1 diabetes using continuously monitored glucose, insulin dose, and meal intake: an extended Kalman filter approach. *Journal of diabetes science and technology*, **8(2)**, 331-345.

Xie, J., and Wang, Q. (2017). A personalized diet and exercise recommender system in minimizing clinical risk for type 1 diabetes: An in silico study. In *Dynamic Systems and Control Conference*, **58271**, V001T08A003, American Society of Mechanical Engineers.

Xie, J., and Wang, Q. (2020). Benchmarking machine learning algorithms on blood glucose prediction for type I diabetes in comparison with classical time-series models. *IEEE Transactions on Biomedical Engineering*, **67(11)**, 3101-3124.

Xu, L., Honda, J., Niu, G., and Sugiyama, M. (2019). Uncoupled regression from pairwise comparison data. *Advances in Neural Information Processing Systems*, 32.

Yang, T., and Tan, Z. (2021). Hierarchical total variations and doubly penalized ANOVA modeling for multivariate nonparametric regression. *Journal of Computational and Graphical Statistics*, **30(4)**, 848-862.

Zecchin, C., Facchinetti, A., Sparacino, G., De Nicolao, G., and Cobelli, C. (2012). Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration. *IEEE transactions on biomedical engineering*, **59(6)**, 1550-1560.

Zhou, Z. H., and Jiang, Y. (2004). NeC4. 5: Neural ensemble based C4. 5. *IEEE Transactions on knowledge and data engineering*, **16(6)**, 770-773.