Comprehensive Data Processing and Predictive Modeling of Blood Glucose Levels Using Shanghai_T1DM and Shanghai_T2DM Datasets*

Coralluzzo Gabriele Vittorio[1] and Rossi Mariarosaria[2]
[1] *Università degli Studi di Salerno, Via Giovanni Paolo II, 132, Fisciano, 84084 SA, Italia*

height 0.5pt

ARTICLE INFO

- Diabetes
- Long short-term memory
- Glucose prediction

ABSTRACT

Type 1 and 2 diabetes demands continuous, personalized care. Precision medicine is vital, adapting treatments to individual needs based on age, lifestyle, and genetics. Accurate blood glucose prediction is crucial for effective management of treatments. This research introduces a deep learning model of LSTM networks for sequence-based prediction, aimed at extending blood glucose prediction horizons. Using the Shanghai_T1DM and Shanghai_T2DM datasets, the model shows improved performance in predicting blood glucose levels, supporting proactive diabetes management and enhancing patient well-being.

1 Introduction

Given the complexity of human health, type 1 and type 2 diabetes mellitus pose a unique challenge as autoimmune disorders that disrupt insulin production by the pancreas. Managing type 1 and type 2 diabetes is critical due to the risk of severe complications such as cardiovascular problems, kidney diseases, and nerve damage. Regulating blood sugar levels introduces two contrasting conditions: hypoglycemia, characterized by dangerously low blood sugar levels, and hyperglycemia, where levels reach risky peaks, both posing significant health risks. Hypoglycemia manifests as an energy deficiency, causing symptoms like dizziness and confusion. Conversely, hyperglycemia resembles a sugar overload, leading to excessive thirst and long-term complications such as cardiovascular diseases. Fortunately, modern diabetes management leverages technologies like continuous glucose monitoring (CGM) and insulin pumps, serving as digital guardians for those navigating the complexities of both types of diabetes. These innovations herald a new era of personalized care, allowing anticipation of glucose fluctuations and addressing challenges with greater security. In the intricate orchestration of diabetes management, predicting future blood glucose levels is akin to possessing a superpower. This capability not only enhances control but also liberates individuals from the unpredictable nature of the condition, enabling people with diabetes to live life to the fullest without the constant burden of uncertainty. In the context of type 1 and type 2 diabetes management, this study presents an innovative approach aimed at improving glucose prediction. The proposed model is a multi-step prediction method for blood glucose levels designed for both types of diabetes. It employs a Long Short-Term Memory (LSTM) network for sequence-based prediction. The LSTM model utilizes features extracted from input data to generate multi-step predictions of blood glucose levels. The extended prediction horizon of the model allows anticipation of glucose fluctuations, with prediction horizons of 15, 30, 60, and 75 minutes, offering valuable insights for proactive diabetes management

2 Related Works

This literature review explores the complex field of blood glucose prediction, analyzing various deep learning architectures, data sources, and prediction horizons. By examining pioneering works and addressing key trends and challenges, the review highlights the significant impact of deep learning on glycemic control. This journey offers an overview of the progress made and the promising prospects for the future of blood glucose prediction.

Recent advancements in deep learning have revolutionized healthcare applications, particularly in diabetes management through blood glucose level prediction. A significant contribution is the CRNN model developed by Kezhi Li and colleagues at Imperial College London in 2019. This approach integrates convolutional and recurrent neural networks to capture spatial and temporal dependencies in multi-dimensional time-series data, such as continuous glucose monitoring, carbohydrate intake, and insulin administration. The model has demonstrated superior performance with RMSE values of 9.38±0.71 mg/dL at 30 minutes and 18.87±2.25 mg/dL at 60 minutes for simulated data, and 21.07±2.35 mg/dL at 30 minutes and 33.27±4.79 mg/dL at 60 minutes for real-world data. Implemented on Android, it provides precise real-time predictions with an execution time of 6 milliseconds, highlighting the potential of deep learning to enhance diabetes management ([5]). The combination of CNN and LSTM models was thoroughly studied in 2020 by Touria El Idrissi and Ali Idri, aiming to identify their strengths and weaknesses ([2]).

In 2022, Jaloli, Lipscomb, and Cescon investigated the impact of stress and physical activity on blood glucose prediction for type 1 diabetes patients. Leveraging raw data from wearable devices, they utilized LSTM models and a CNN-LSTM hybrid to forecast glucose levels at 30 and 60 minute intervals. Their study yielded promising results, with MAE values of 9.13, 17.75, and 31.85, and RMSE values of 12.35, 24.71, and 41.64, highlighting the effectiveness in capturing glucose variations influenced by behavioral and physiological factors ([3]).

Tao et al.'s 2023 project on Type 2 diabetes mellitus (T2DM) utilized real-world medical data from Sichuan Province, China spanning 2015 to 2020. Their focus was on developing personalized predictive models for blood glucose variations post three-month therapy. Employing machine learning algorithms, they identified five high-performance models incorporating ensemble learning and modified random forest inputs. These models demonstrated significant precision, achieving AUC values of 0.819 for fasting blood glucose (FBG) and 0.970 for glycosylated hemoglobin (HbA1c). Factors influencing FBG levels include therapy compliance, follow-up outcomes, dietary practices, BMI, and waist circumference. The critical relation ship between HbA1c and FBG as key predictors underscores the therapeutic value of these precise models in enhancing blood glucose management for T2DM patients ([4]).

In 2023, Yang et al. introduced the "PBGTAM" method for precise blood glucose prediction in diabetic patients. This approach utilizes a multi-head temporal attention model, an algorithm for anomalous blood glucose detection, and a missing value imputation technique to analyze sequential data and extract spatiotemporal characteristics. PBGTAM achieved a root mean square error of 20.57, prediction accuracy of 84.35%, and clinical consistency of 85.18%, surpassing six advanced techniques ([7]).

3 Methods and Materials

After a preliminary analysis of existing literature from 2015 onwards concerning the use of models for predicting blood glucose levels in patients with type 1 diabetes mellitus, our research study focused on employing a Long Short-Term Memory (LSTM)-based neural network. Specifically, we extended the application of this model not only to predict glucose levels in type 1 diabetes patients but also in those with type 2 diabetes. Through meticulous data preprocessing, training, and feature extraction, our model demonstrates remarkable predictive accuracy

3.1 Dataset

In our study, we utilized the Shanghai_T1DM dataset, a valuable resource that includes Continuous Glucose Monitoring (CGM) data from 12 individuals facing the challenges of Type-1 Diabetes Mellitus (T1DM) in Shanghai, China. This meticulously curated dataset documents real-life scenarios over a temporal span ranging from 3 to 14 days, providing detailed information such as the date and time of each CGM measurement, corresponding blood glucose levels, insulin administration, carbohydrate intake, and demographic data, including age and gender.

Since our model has also been employed for predicting blood glucose levels in patients with Type 2 Diabetes Mellitus (T2DM), in addition to those with Type 1 Diabetes Mellitus, we utilized the Shanghai_T2DM dataset in our study. This dataset is a valuable resource, collecting detailed data related to T2DM from a population of patients. Shanghai_T2DM includes a variety of clinical information such as blood glucose measurements through Continuous Glucose Monitoring (CGM), recorded with specific dates and times. Additionally, it provides details on insulin administration, carbohydrate intake, and other antidiabetic medications, as well as laboratory data such as glycated hemoglobin (HbA1c) levels and lipid profiles.

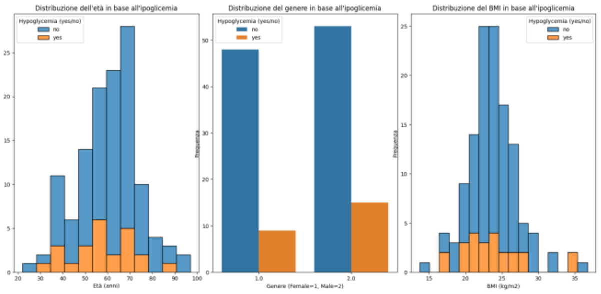
The dataset also contains demographic information including age, sex, weight, and height of the participants. Behavioral and lifestyle data, including physical activity, smoking habits, and alcohol consumption, are also documented. This dataset supports clinical and epidemiological research on T2DM, facilitating studies on the efficacy of therapies, cardiovascular risk, genetic predisposition, and the impact of lifestyle modifications ([8]Zhao et al., 2023).

3.2 Data Pre-Processing

The data pre-processing operation begins with defining the directories for input and output files, creating the output folder if it does not exist. Subsequently, subfolders are created for different time intervals (30, 60, and 75 minutes) to organize the processed data. The structure of the expected data columns is established, including various relevant parameters such as continuous and capillary glucose measurements, dietary intake, and insulin doses administered subcutaneously and intravenously. Date strings in the data are converted to datetime objects, attempting various common date formats.In case the conversion fails, the errors are handled through an imposed fix. Additionally, numerical values are extracted from text strings, treating missing values as zeros and using regular expressions to identify numbers. During file processing, the data are read and the number of columns is checked to match expectations. Column names are correctly assigned. Dates are converted and the elapsed time between successive recordings is calculated. The basal insulin amount is calculated by multiplying the administration rate by the elapsed time.

3.3 Feature Extraction

Following the initial pre-processing, our project team utilized several features derived from the raw data, including: BMI (Body Mass Index), Age, Gender (Female:1 - Male:2). Below are boxplots summarizing the aforementioned demographic information:



During the feature extraction process, we considered certain insulin-related. These data include:

- *CSII - basal insulin*
- *CSII - bolus insulin*
- *Insulin dose s.c.*
- *Insulin dose i.v.*
- *Non-insulin hypoglycemic agents*

Due to the discontinuities in insulin values and administration methods for each patient, it was decided to associate the data with a flag indicating whether the drug was administered or not. The flag assumes the following values:

- 0 - Drug not administered
- 1 - Drug administered

Additionally, it was decided to associate another flag with the Dietary intake column, indicating whether the patient consumed a meal. The flag assumes the following values:

- 0 - Meal not consumed
- 1 - Meal consumed

Furthermore, we included the Carbs column by manually calculating the number of carbohydrates consumed by each patient in their dietary regimen. The calculation procedure is as follows:

Step 1. The Dietary intake column is extracted;

Step 2. The Dietary intake column is renamed to Ingredient;

Step 3. Any redundant values (meals) are removed;

Step 4. A new column, titled Carbohydrates (g), is created;

Step 5. The carbohydrate values in the meals are calculated

using the website Prospre.io by manually entering the name and quantity in grams of the meals.

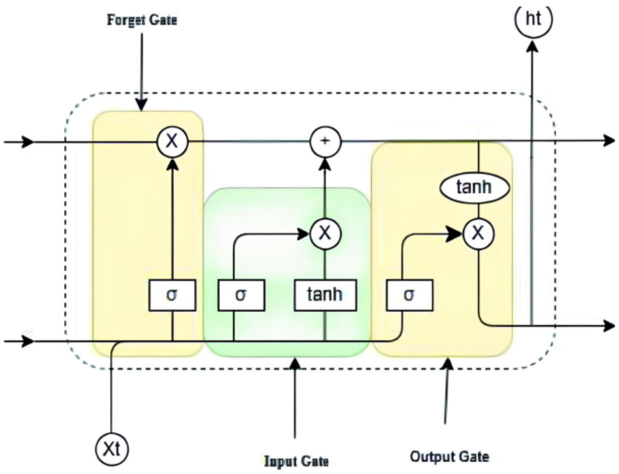


Figure 1: *Representation of a LSTM cell.*

Some of this data was grouped into 30 and 60 minute intervals. Specifically:

- BMI, Age, and Gender are static values and are not calculated over time intervals. These values were ex-tracted from the Shanghai_T1T2DM_Summary.xlsx file based on the patient ID and were associated with the files grouped into 30 and 60 intervals.

- Dietary intake-Flag, Insulin-Flag, and the Carbs column were recorded based on events rather than fixed intervals. The 30 and 60 minute intervals are used to create temporal sequences with sliding windows, which implies an overlap between the sequences. These aggregated files were converted to Excel format.

Finally, the files generated for each interval were merged into a single file, combining data from various files into a unified dataset. The entire process concluded with confirmation that all pre-processing operations were successfully completed.

4 Model Description

4.1 LSTM

Long Short-Term Memory (LSTM) networks are specialized recurrent neural networks (RNN) designed for modeling sequential data while maintaining a robust memory of past inputs. An LSTM network consists of interconnected memory cells with three main gates: input, forget, and out-put. These gates regulate the flow of information, allowing the network to learn and maintain long-range dependencies.

(1) graphically shows the structure of the LSTM model just described.

LSTMs are particularly effective in recognizing patterns in sequential data, integrating new information, and discarding irrelevant data. This capability makes them suitable for handling long-term dependencies in applications such as natural language processing and time series analysis. In blood glucose prediction, LSTMs are advantageous due to their ability to assess the importance of inputs, manage memory, and determine the output, making them ideal for modeling temporal sequences ([6]).

4.2 LSTM-Based Predictive Modeling Process

The modeling process begins with the use of processed data to ensure the quality and integrity of the information utilized. The data input into the Long Short-Term Memory (LSTM) model includes continuous glucose monitoring (CGM) measurements and additional features, such as *BMI, Age, Gender, Insulin-Flag, Dietary Intake-Flag, and Carbs.* Each input sequence has a length determined by the ‘prediction_step‘ parameter. For example, if ‘prediction_step‘ is 30, the input sequence will consist of the feature data collected over the previous 30 time steps. The shape of each sequence is ‘(prediction_step, number_of_features + 1)‘, where the additional ‘1‘ accounts for the CGM measurement itself. For instance, if ‘prediction_step‘ is 30 and there are 6 additional features (BMI, Age, Gender, Insulin-Flag, Dietary Intake-Flag, and Carbs), the sequence will have the shape ‘(30, 7)‘. Demographic and clinical information about the patients is integrated into the model. The model learns from these sequences to predict future glucose levels, and the accuracy of these predictions depends on the representation and normalization of the input data. A detailed overview of the data provided to the model is outlined below:

- **Input Data (x):**
  - *Shape: (number of sequences, prediction step, number of features + 1)*
  - Content:
    - Each sequence of ‘prediction step‘ time steps contains columns for BMI, Age, Gender, Dietary intake - Flag, Insulin-Flag, Carbs and CGM.
    - For each sequence, the values of these features are used to predict the glucose level in the subsequent time step.
- **Target (y):**
  - Shape: (number of sequences)
  - Content: Glucose values (CGM (mg/dl)) for the time step following each sequence.

Data from Excel files are loaded and concatenated into a single dataset, with patient identifiers extracted from the filenames. This data is then merged with a summary file **(Shanghai T1_T2DMSummary.xlsx)** containing additional relevant information for each patient, such as ***BMI, Gender,and Age***.

A crucial phase of the process involves normalizing the continuous variables (*BMI, Age, and Carbs)* using the StandardScaler to ensure that all variables are on a common scale, thereby reducing biases due to different measurement units; meanwhile, the categorical/binary variables (Gender, Insulin-Flag, and Dietary intake-Flag) are not normalized. Temporal sequences of a predefined length are created, with features and the target variable to be predicted separated. This procedure is applied to both training and test data, with careful verification to ensure there are no missing values, as they could compromise the accuracy of the model.

To ensure robust and reliable model training and evaluation, we employed an approach that dividing each subject’s about their Id data into training and testing sets individually, rather than combining data from multiple subjects. For example, if patient 1001 has 600 glucose readings, 480 readings are used for training and the remaining 120 for testing, with the chronological order of the readings being respected. By using the technique we explain first, we ensure that the model learns and is tested on data specific to each individual, which aids in capturing the unique physiological patterns and variations in blood glucose levels inherent to each subject. This method enhances the model’s ability to provide personalized predictions and improves its generalizability across different subjects ([1]).

The predictive model utilizes a recurrent neural network with Long Short-Term Memory (LSTM) units, which are well-suited for capturing the intrinsic temporal dependencies in sequential data. The network architecture includes an LSTM layer configured with a specific number of units, followed by a dropout layer to prevent overfitting, and finally a dense layer that produces the final prediction.

During the training phase, the model is optimized using a mean squared error (MSE) loss function and an advanced optimization algorithm. The model is trained for a predefined number of epochs, using both training and validation data to monitor and improve its performance. After training, the model is evaluated on test data to assess its ability to generalize to new data.

5 Results and Discussion

Our study used a sophisticated integrated pipeline as the implementation basis for comprehensive data processing and predictive modeling of blood glucose levels using the Shanghai_T1DM and Shanghai_T2DM datasets. In fact, modifications to the insulin management mechanisms were made to the pipeline.

Implemented in Python, the pipeline leverages core libraries such as Pandas, Scikit-Learn, TensorFlow, and Matplotlib to preprocess and analyze continuous glucose monitoring (CGM) data. Initially, the pipeline validates Excel files containing patient data to ensure format compatibility. Patient IDs are extracted from the file names and summary data, facilitating the integration of individual records with corresponding demographic and clinical information, including age, gender, and body mass index (BMI).

The data loading process is performed on a patient-by-patient basis. For each patient, the ID is read from the corresponding file, and using this ID, the information of interest, such as age, gender, and other characteristics, is extracted from the summary file. Then, the specific data for that patient is loaded from the corresponding Excel file (e.g., for patient 1001), which is then split into training and testing data: 80% of the data is used for training, while the remaining 20% is used for testing. The feature set includes BMI, age, gender, insulin-Flag, dietary intake-Flag, and carbohydrates, with CGM (mg/dL) as the target variable. For example, if there are 100 observations for a patient, we use 80 of them to train the model and 20 to test it. We decided to set a fixed number of 30 epochs for the Long Short-Term Memory (LSTM) model, training the model on the first 30 observations to try to predict the 31st. This process is repeated for each patient, ensuring that training occurs independently for each individual.

The LSTM model was built using TensorFlow's Keras API, comprising two LSTM layers followed by dropout layers to prevent overfitting and a dense layer for final prediction. The data was normalized using StandardScaler to improve model performance. The training process involved splitting the data as previously mentioned, creating sequences, and checking for NaN or infs post-normalization. Evaluation of model performance includes calculating the mean squared error (MSE) and coefficient of determination (R**²**), providing measures of accuracy and goodness of fit. To calculate the accuracy of the model within a specific tolerance, we use the following formula:

$$\text{Accuracy} = \frac{\sum_{i=1}^{n} \left(1 - \left|\frac{y_{\text{pred},i} - y_{\text{test},i}}{y_{\text{test},i}}\right| < \text{tolerance}\right)}{n} \times 100 \quad (1)$$

where:

- $y_{\text{pred},i}$ are the predicted values from the model
- $y_{\text{test},i}$ are the actual values
- n is the total number of observations
- tolerance is the accuracy threshold(es. 0.1 for 10

The Mean Squared Error (MSE) is calculated as:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(y_{\text{test},i} - y_{\text{pred},i}\right)^2 \quad (2)$$

where:

- $y_{\text{test},i}$ are the actual values
- $y_{\text{pred},i}$ are the predicated values from the model
- n is the total number of observations

The coefficient of determination is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_{\text{test},i} - y_{\text{pred},i})^2}{\sum_{i=1}^{n} (y_{\text{test},i} - \bar{y}_{\text{test}})^2} \quad (3)$$

- $\bar{y}_{\text{test}}$ is the mean of the actual value $y_{\text{test},i}$
- $y_{\text{test},i}$ are the actual values
- $y_{\text{pred},i}$ are the predicated values from the model
- n is the total number of observations

The Model Performance, in percentage terms, is calculated as:

$$\text{Model Performance} = R^2 \times 100 \quad (4)$$

Model predictions are compared with observed values through graphical representations illustrating correspondence between predicted and actual values, as well as loss trends during training and validation epochs. Additionally, model accuracy within a 10% tolerance is determined by evaluating the percentage of predictions falling within this interval compared to actual values.

This systematic approach enables raw data processing, LSTM model training for accurate blood glucose level prediction, and comprehensive evaluation of model performance using quantitative metrics and visual representations for detailed and comprehensive predictive analysis.

This approach has effectively preprocessed, trained, and evaluated a deep learning model for CGM value prediction, showcasing the potential of machine learning in managing and analyzing temporal health data. The implemented pipeline signifies significant progress in leveraging clinical data to enhance diabetes management and monitoring.

We conducted a comprehensive evaluation of the blood glucose prediction model using time intervals of 30 ands 60 minutes. The evaluation metrics for each time interval are presented below:

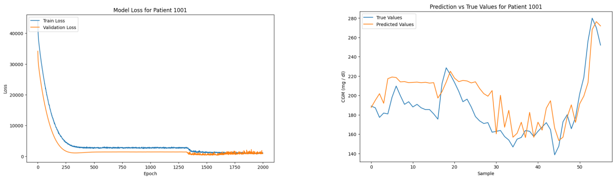


Figure 2: *Loss Rate & prediction rate for 1001 pazient in 30 minute.*

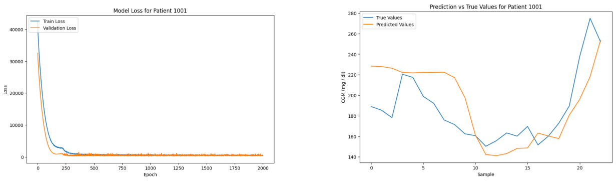


Figure 3: *Loss Rate & prediction rate for 1001 pazient in 60 minutes.*

Each figure consists of two graphs: one plots the glucose value predictions and the other plots the model loss during the training and validation phases for the corresponding time interval.

The full evaluation of the model performance at 30- and 60-minute forecast horizons using parameters such as MSE, R-squared and accuracy within a tolerance of 10.0% is documented in 1.

The performance of the proposed model was evaluated at two different forecast horizons: 30 and 60 minutes. The results demonstrate a clear trade-off between forecast accuracy and the length of the forecast horizon.

For a 30-minute forecast horizon, the model achieved a high R^2 score of 46.86%, indicating that the model can explain a substantial part of the variance in blood glucose levels at shorter periods. The accuracy within a 10% tolerance was 51.79%, which suggests that the model is reliable for short-term forecasts and can potentially be useful in real-time glucose monitoring applications where immediate decisions are required.

When the forecast horizon is extended to 60 minutes, the R^2 score drops to 21.45% and the accuracy decreases to 47.83%. While these results are still reasonably good, they indicate that the model's ability to predict blood glucose levels decreases as the forecast horizon increases.

Prediction Horizon	R^2 Score (%)	Accuracy (%)	MSE
30 minutes	46.86	51.79	600.402
60 minutes	21.45	47.83	838.5034

Table 1: *Evaluation of the model's performance.*

This decline is expected due to increased uncertainty and variability in physiological responses over longer periods. For the 60-minute forecast horizon, the model performance further declines with an R^2 score. While the model retains a fair degree of predictive power, its reduced accuracy suggests that it may be less effective for long-term forecasts, where more complex or additional features may be needed to capture the underlying dynamics of glucose fluctuations. Overall, the results indicate that the model performs better at shorter forecast horizons, making it a suitable tool for short-term glucose forecasting. However, for extended horizons, the model predictions become less reliable, highlighting the need for further refinement or integration with other predictive methodologies to maintain high levels of accuracy over longer periods.

CRediT - Author Contribution Statement

- **Gabriele Vittorio Coralluzzo**: *Metodology, Software, Testing and writing of the Paper;*
- **Mariarosaria Rossi**: *Metodology, Software, Testing and writing of the Paper;*

References

[1] H. Butt, I. Khosa, and M.A. Iftikhar. Feature transformation for efficient blood glucose prediction in type 1 diabetes mellitus patients. *Diagnostics*, 13, 2023.

[2] T. El Idrissi and A. Idri. Deep learning for blood glucose prediction: Cnn vs lstm. In O. Gervasi, B. Murgante, S. Misra, C. Garau, I. Blecic, D. Taniar, B.O. Apduhan, A.M.A. Rocha, E. Tarantino, C.M. Torre, and Y. Karaca, editors, *Computational Science and Its Applications ICCSA 2020*, pages 379–393. Springer International Publishing, Cham, 2020.

[3] M. Jaloli, W. Lipscomb, and M. Cescon. Incorporating the effect of behavioral data in multi-step ahead deep learning based multivariate predictors for blood glucose forecasting in type 1 diabetes. *BioMed Informatics*, 2:715–726, 2022.

[4] S. Langarica, M. Rodriguez-Fernandez, F. Núñez, and F.J. Doyle. A meta-learning approach to personalized

blood glucose prediction in type 1 diabetes. *Control Engineering Practice*, 135, 2023.

[5] K. Li, J. Damians, C. Liu, P. Herrero, and P. Georgiou. Convolutional recurrent neural networks for glucose prediction. *IEEE Journal of Biomedical and Health Informatics*, 24:603–613, 2020.

[6] S. Rajagopal and N. Thangarasu. A novel hybrid deep learning model for blood glucose prediction with extended prediction horizons in type-1 diabetic patients. In *2023 International Conference on Emerging Research in Computational Science (ICERCS)*, pages 1–7, 2023.

[7] G. Yang, S. Liu, Y. Li, and L. He. Short-term prediction method of blood glucose based on temporal multi-head attention mechanism for diabetic patients. *Biomedical Signal Processing and Control*, 82, 2023.

[8] Q. Zhao, J. Zhu, X. Shen, C. Lin, Y. Zhang, Y. Liang, B. Cao, J. Li, X. Liu, W. Rao, and C. Wang. Chinese diabetes datasets for data-driven machine learning. *Scientific Data*, 10:35, 2023.