Blood Glucose Prediction Model for Type 1 Diabetes based on Extreme Gradient Boosting

Ganjar Alfian¹, Muhammad Syafrudin², Jongtae Rhee², Muhammad Anshari³, M. Mustakim⁴, Imam Fahrurrozi⁵

¹Nano Information Technology Academy, Dongguk University, Seoul, Korea
²Department of Industrial and Systems Engineering, Dongguk University, Seoul, Korea
³School of Business & Economics, Universiti Brunei Darussalam, Brunei Darussalam
⁴Teknik Informatika, Universitas Islam Negeri Sunan Kalijaga, Indonesia
⁵Departemen Teknik Elektro dan Informatika, Sekolah Vokasi, Universitas Gadjah Mada, Indonesia

E-mail: ganjar@dongguk.edu, udin@dongguk.edu, jtrhee@dongguk.edu, anshari.ali@ubd.edu.bn, taqiem235@yahoo.com, imam.fahrurrozi@ugm.ac.id

Abstract. Predicting future blood glucose (BG) level for diabetic patients will help them to avoid critical conditions in the future. This study proposed Extreme Gradient Boosting (XGBoost), an ensemble learning model to predict the future blood glucose value of diabetic patients. The clinical dataset of Type 1 Diabetes (T1D) patients was utilized and the prediction models were generated to predict future BG of 30 and 60 minutes ahead of time. The prediction models have been tested to five children who develop T1D and showed that BG prediction model based on XGBoost outperformed other models, with average of Root Mean Square Error (RMSE) are 23.219 mg/dL and 35.800 mg/dL for prediction horizon (PH) 30 and 60 minutes respectively. In addition, the result showed that by utilizing statistical-based features as additional attributes, most of the performance of predictions model were increased.

1. Introduction
Diabetes is chronic disease when the body cannot provide enough insulin (T1D, Type 1 diabetes) or the body cannot effectively use its produced insulin (T2D, Type 2 diabetes) [1]. It is expected for the patient to manage their BG level in the normal value between 70–180 mg/dL. High blood glucose or hyperglycemia is the condition where BG > 180 mg/dL and can affect several complications such as cardiovascular disease and retinopathy. Other condition that occurs to the diabetic patients is low blood sugar or hypoglycemia (BG < 70 mg/dL) and can cause death [2]. To monitor blood glucose level, the diabetes patient can utilize continuous glucose monitoring (CGM) device. CGM device is an advanced wearable medical devices that can monitor and record glucose levels concentration every 1–5 min and has shown improvement for diabetes management [3,4].

Several studies have demonstrated BG prediction model, so that future value of BG can be obtained earlier. Blood glucose prediction models have been presented in previous studies by considering glucose level as single input. Perez-Gandia et al [5] proposed neural network (NN) to predict future blood glucose by using the last 20 minutes of BG level as input. Ben Ali et al [6] proposed neural network-based model to predict BG level of Type 1 Diabetes (T1D) using only CGM data as inputs.
Finally, Hamdi et al [7] proposed prediction model by utilizing vector regression and differential evolution algorithms to forecast blood glucose. The proposed method was tested on CGM data and showed that the proposed model achieved high prediction accuracy.

For the case of internet of things (IoT) sensor, the previous studies have shown that statistical based features can be utilized as input for prediction model. Park et al [8] proposed fault detection system from IoT sensor in the manufacturing. The statistical features were extracted from sensor data to be utilized as input for prediction model. Finally, Alfian et al [9] utilized statistical features from the received signal strength (RSS) of radio frequency identification (RFID) data as input for prediction model. The prediction model showed significant result on predicting the movement of RFID tag.

Previous studies showed that the blood glucose can be utilized as single input for blood glucose prediction model with high accuracy. In addition, the statistical features also showed significant result on improving the performance of prediction model for IoT sensor. Therefore, in present study the prediction model by considering blood glucose as input and statistical features as additional attributes are presented. The proposed prediction model based on XGBoost will be utilized to predict future blood glucose level of 30 and 60 minutes ahead of time.

2. Methodology
The history of blood glucose of diabetic patients needs to be gathered by CGM device, so that forecasting from this time-series data could be executed. For predicting future blood glucose, multi-step forecasting can be utilized where the task is predicting next \( H \) values \([y_{N+1}, \ldots, y_{N+H}]\) of a historical time series \([y_1, \ldots, y_N]\) composed of \( N \) observations and \( H > 1 \) denotes the forecasting horizon [10]. In multi-step forecasting \((t + h)\), given \( n \) previous values and \( N \) total data, the training set can be derived by creating the input data matrix \( X \) and output vector \( Y \).

Figure 1 shows the illustration of sliding window approach to convert time-series data into input matrix \( X \) and output vector \( Y \). In this scenario, the time series dataset was split into several windows with the size \( n = 6 \) as previous data for learning and is used to predict the blood glucose for the next \( h = 6 \). In our study, the blood glucose dataset was measured for every 5 minutes, therefore our prediction model (as presented in Figure 1) used the last 30 minutes of historical data to predict the next 30 minutes of blood glucose. In Figure 1, the windows (presented in red lines) were collected and presented as input matrix \( X \), while the collection of next 30 minutes value (presented in blue lines) were used as output vector \( Y \).

![Figure 1. Sliding window approach for timeseries blood glucose data.](image)

We propose statistical features as attributes addition for the \( n \) previous values which is generated by sliding window approach. Most of the previous studies, only considered \( n \) historical data as the input. Table 1 shows the proposed eight relevant statistical features extracted from each window (historical data) to improve forecasting performance.
Table 1. Statistical attributes extracted from window.

| Attribute name | Description |
|----------------|-------------|
| Min            | Minimum value of BG in a window. |
| Max            | Maximum value of BG in a window. |
| Mean           | Average value of BG in a window. |
| Std            | Standard deviation value of BG in a window. |
| Diff           | Difference between highest and lowest BG value in a window. |
| Median         | Middle value of BG in a window. |
| Kurtosis       | Describing tails of BG distribution differ from a normal distribution. |
| skew           | Distribution asymmetry of BG in a window. |

Next step, the additional statistical features are combined with previous values generated by sliding window approach. Finally, given $n$ previous values or window size, $h$ forecasting horizon, 8 (eight) number of statistical features and total data, the input can be derived by creating the $[ (N-n-h+1) \times (n+8) ]$

\[
X = \begin{bmatrix}
    y_1 & \ldots & y_5 & y_h & \min(y_1, \ldots, y_h) & \ldots & \text{skew}(y_1, \ldots, y_h) \\
    \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
    y_{N-n-h} & \ldots & y_{N-h-2} & y_{N-h-1} & \ldots & \text{skew}(y_{N-n-h}, \ldots, y_{N-h-1}) \\
    y_{N-n-h+1} & \ldots & y_{N-h} & y_h & \ldots & \text{skew}(y_{N-n-h+1}, \ldots, y_{N-h})
\end{bmatrix} \tag{1}
\]

and the $[ (N-n-h+1) \times 1 ]$ output vector $Y$.

\[
Y = \begin{bmatrix}
    y_{h+1} \\
    \vdots \\
    y_{N-1} \\
    y_N
\end{bmatrix} \tag{2}
\]

In order to investigate how the proposed prediction model can predict future blood glucose, the dataset from DirecNet was utilized for this study [11]. The subjects are children, age 3 to less than 7 and 12 to 17 with diabetes type 1. The subjects were asked to use the Guardian-RT (i.e., continuous glucose monitoring device) to record their blood glucose for approximately 6 days. In this study, we selected blood glucose measurement from 5 patients, and the detail dataset distribution can be seen in Table 2.

Table 2. Distribution of BG dataset from type 1 diabetes patients.

| Patient ID | #Data Point | Min | Max  | Mean  | STD   |
|------------|-------------|-----|------|-------|-------|
| 1          | 1496        | 41.17 | 400.0 | 200.351 | 68.353 |
| 15         | 2145        | 66.37 | 353.0 | 179.361 | 58.381 |
| 4          | 2212        | 53.74 | 303.72 | 151.202 | 52.796 |
| 21         | 2076        | 53.25 | 400.0 | 158.012 | 72.418 |
| 22         | 2173        | 40.0  | 232.33 | 121.561 | 37.766 |

In this study, we utilized XGBoost model to predict future blood glucose level of type 1 diabetes patient. XGBoost is decision-tree-based ensemble model that utilizes gradient boosting [12]. The objective function contains training loss and regularization. The objective function (loss function and regularization) can be presented as follows.
\[
\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)
\]

where \(\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2\)

Here, \(l\) is a differentiable loss function that shows the difference between the target \(y_i\) and the prediction \(\hat{y}_i\). The regularized term \(\Omega\) penalizes the complexity of the model and \(T\) represents the total number of leaves in tree. Each \(f_k\) represents tree \(q\) and weight of leaf \(w\). The term \(\gamma\) corresponds to the threshold and pre-pruning is performed while optimizing to limit the growth of the tree and \(\lambda\) is used to smooth the final learned weights to prevent overfitting.

The prediction models were implemented in Python V3.6.6 and Scikit-learn V0.19.1[13]. To simplify implementation of machine learning models, we used default parameters provided by Scikit-learn. Datasets were split into two parts: approximately 50:50 for training and testing. To present the performance of prediction models, the correlation coefficient \((r)\), RMSE and MAPE will be used. The prediction model fits the data when the RMSE and MAPE generates small values. The model correctly fits the data when the \(r\) generates higher value. The detailed formula of \(r\), RMSE and MAPE for a dataset of size \(n\), predicted value \(\hat{y}\) and actual value \(y\) are presented in Table 3.

### Table 3. The performance metrics for the forecasting models.

| Performance Metric                      | Formula                                      |
|----------------------------------------|----------------------------------------------|
| Root Mean Square Error (RMSE)           | \(\sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}\) |
| Pearson correlation coefficient \((r)\) | \(\frac{\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}}\) |
| Mean Absolute Percentage Error (MAPE)   | \(100\% \times \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i}\) |

### 3. Result and Discussion

Table 4 showed the performance of prediction models to predict the future blood glucose for different prediction horizon (PH). Multi-Layer Perceptron (MLP), K-Nearest Neighbour (KNN), Support Vector Regression (SVR), Random Forest (RF), Decision Tree (DT) and AdaBoost have been compared with proposed XGBoost model for predicting multi-step ahead of blood glucose. In this experiment, window size or previous values \(n = 6\) (last 30 minutes) has been utilized for machine learning models to predict future blood glucose (BG). Each prediction model was generated for each patient and the performance average was computed over 5 models (generated from 5 patients). The statistical features were utilized by proposed XGBoost model as additional attributes (see formula 1 and 2), while other models only utilized last 30 minutes values generated by sliding window approach (Figure 1). The result showed that as increasing the prediction horizon, the prediction models generated higher RMSE and MAPE, and lower \(r\). The proposed XGBoost model showed highest performance by generating lowest RMSE and MAPE as compared to other prediction models. The RMSE of proposed XGBoost are 23.219 mg/dL and 35.800 mg/dL for prediction horizon (PH) 30 and 60 minutes respectively, while the MAPE of proposed model are 10.896 and 17.681 for PH 30 and 60 minutes respectively. Furthermore, the proposed model also generated highest \(r\) (correlation coefficient) compared to other models, they are 0.900 and 0.750 for PH 30 and 60 minutes respectively.
Table 4. Performance of blood glucose prediction models.

| Method  | PH = 30 min | PH = 60 min |
|---------|-------------|-------------|
|         | RMSE        | r           | MAPE        | RMSE        | r           | MAPE        |
| MLP     | 30.997      | 0.854       | 13.683      | 42.277      | 0.705       | 19.046      |
| SVR     | 31.839      | 0.855       | 17.138      | 40.553      | 0.707       | 22.665      |
| KNN     | 26.439      | 0.872       | 12.562      | 41.302      | 0.698       | 20.291      |
| DT      | 33.788      | 0.824       | 16.446      | 52.638      | 0.580       | 25.967      |
| RF      | 26.791      | 0.873       | 13.000      | 41.115      | 0.689       | 20.269      |
| AdaBoost| 27.986      | 0.864       | 15.314      | 39.868      | 0.710       | 22.919      |
| Proposed model | 23.219    | 0.900       | 10.896      | 35.800      | 0.750       | 17.681      |

In addition, the impact of feature type on the performance of prediction models are presented in Figure 2. The result showed that by combining sliding window approach and statistical features (StatFeature) provided lower RMSE as compared to prediction models with sliding window-based attributes, except for SVR and DT for PH 30 minutes (Figure 2a), and SVR and AdaBoost for PH 60 minutes (Figure 2b). By integrating sliding window approach and statistical features as attributes for prediction models, the average of RMSE was reduced as much as 0.668 and 1.688 for PH 30 and 60 minutes respectively as compared to prediction models with sliding window-based attributes only.

Figure 2. Impact of feature type on the model performance for PH (a) 30 minutes and (b) 60 minutes.

4. Conclusion and Future Works

In this study, the proposed XGBoost model was utilized to predict future blood glucose of 30 and 60 minutes ahead of time. The proposed model has been tested to the dataset consist of 5 (five) children who develop T1D. The proposed XGBoost model showed highest performance by generating lowest RMSE and MAPE as compared to other prediction models. The RMSE of proposed XGBoost are 23.219 mg/dL and 35.800 mg/dL for prediction horizon (PH) 30 and 60 minutes respectively, while the MAPE of proposed model are 10.896 and 17.681 for PH 30 and 60 minutes respectively. Furthermore, the proposed model also generated highest r (correlation coefficient) compared to other models, they are 0.900 and 0.750 for PH 30 and 60 minutes respectively. In addition, by integrating
sliding window approach and statistical features as attributes for prediction models, the average of RMSE was reduced as much as 0.668 and 1.688 for PH 30 and 60 minutes respectively as compared to prediction models with sliding window-based attributes only.

The dataset in this study was limited to small set of type 1 diabetes for children, so it is difficult to generalize the robust prediction model to be applied for different purposes. The other type of datasets such as that collected from other CGM sensor or glucose meter need to be utilized in the near future. Furthermore, extending the comparison with other prediction models could be presented in the near future.

Acknowledgments
This research was financially supported by the Ministry of Trade, Industry and Energy (MOTIE) and the Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D program (Grant number N0002301).

References
[1] World Health Organization 1999*Definition, Diagnosis, and Classification of Diabetes Mellitus and Its Complications. Part I: Diagnosis and Classification of Diabetes Mellitus*
https://apps.who.int/iris/handle/10665/66040 retrieved at September 1st 2019
[2] American Diabetes Association 2018*Introduction:Standards of medical care in diabetes—2018 Diabetes Care* 41 pp S1–S2 DOI:10.2337/dc18-S1
[3] Bode B W and Battelino T 2010 Continuous glucose monitoring*Int. J. Clin. Pract. Suppl.* 64 pp 11–5 DOI: 10.1111/j.1742-1241.2009.02272.x
[4] Torres I, Baena M G, Cayon M, Ortego-Rojo J and Aguilar-Diosdado M 2010 Use of sensors in the treatment and follow-up of patients with diabetes mellitus*Sensors* 10 pp 7404–20 DOI:10.3390/s100807404
[5] Pérez-Gandía C, Facchinetti A, Sparacino G, Cobelli C, Gómez E J, Rigla M, de Leiva A and Hernando M E 2010 Artificial Neural Network Algorithm for Online Glucose Prediction from Continuous Glucose Monitoring*Diabetes Technol. Ther.* 12(1) pp 81-8 DOI: 10.1089/dia.2009.0076
[6] Ben Ali J, Hamdi T, Di Costanzo V, Fnaiech F and Ginoux J 2018 Continuous blood glucose level prediction of Type 1 Diabetes based on Artificial Neural Network*Biocybernetics and Biomedical Engineering* 38(4) pp 828-40 DOI:10.1016/j.bbe.2018.06.005
[7] Hamdi T, Ben Ali J, Di Costanzo V, Fnaiech F, Moreau E and Ginoux J 2018 Accurate prediction of continuous blood glucose based on support vector regression and differential evolution algorithm*Biocybernetics and Biomedical Engineering* 38(2) pp 362-72 DOI:10.1016/j.bbe.2018.02.005
[8] Park D, Kim S, An Y and Jung J-Y 2018 LiReD: A Light-Weight Real-Time Fault Detection System for Edge Computing Using LSTM Recurrent Neural Networks*Sensors* 18(7) p 2110 DOI:10.3390/s18072110
[9] Alfian G, Syafrudin M, Yoon B and Rhee J 2019 False Positive RFID Detection Using Classification Models*Applied Sciences* 9(6) p 1154 DOI:10.3390/app9061154
[10] Ben Taieb S, Bontempi G, Atiya A F and Sorjamaa A 2012 A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition*Expert Systems with Applications* 39(8) pp 7067–83 DOI:10.1016/j.eswa.2012.01.039
[11] Evaluation of Counter-regulatory Hormone Responses during Hypoglycemia and the Accuracy of Continuous Glucose Monitors in Children with TIDM
https://public.jaeb.org/directnet/study/167 retrieved at September 1st 2019
[12] Chen T and Guerstrin C 2016 XGBoost: A Scalable Tree Boosting System*Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining - KDD ’16* pp 785–94
[13] Pedregosa F et al 2011 Scikit-learn: Machine Learning in Python The J. of Machine Learning Research 12 pp 2825-30