Evolution in Electrical and Electronic Engineering Vol. 4 No. 2 (2023) 457-463
© Universiti Tun Hussein Onn Malaysia Publisher's Office

EEEE

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/eeee
e-ISSN : 2756-8458

Blood Glucose Prediction Based on ARIMA Time-Series Machine Learning Model

Rishalini Saravanan[1], **Farhanahani Mahmud**[1,2*]

[1]Faculty of Electrical and Electronic Engineering,
Universiti Tun Hussein Onn Malaysia, Batu Pahat, 86400, MALAYSIA

[2]Microelectronics and Nanotechnology Shamsuddin Research Centre (MiNTSRC),
Institute of Integrated Engineering (I2E),
Universiti Tun Hussein Onn Malaysia, Batu Pahat, 86400, MALAYSIA

*Corresponding Author Designation

DOI: https://doi.org/10.30880/eeee.2023.04.02.056
Received 02 July 2023; Accepted 05 September 2023; Available online 30 October 2023

Abstract: Glucose levels prediction is a difficult task commonly faced by people with diabetes, a chronic health condition that affects how a human body synthesizes food. The glucose levels in the human body depend on a variety of factors, so the patient always assumes the risk of making incorrect calculations. Nowadays, using new technologies such as Artificial Intelligence (AI) or Machine Learning (ML), these calculations can be supported and eased by the application of prediction systems. Time series modelling involves developing models used to describe the observed time series and understand the "why" behind its dataset. In recent years, there has been a growing trend of applying machine learning algorithms to time-series predictions. Machine learning approaches have been applied to the prediction of blood glucose levels in several studies. However, it is hard to compare the performance of different prediction approaches used in these references, either classical regression or machine learning based models, since different datasets were used by different studies. In this work, through the simulation of an open-loop insulin delivery system based on Cobelli type 1 diabetic model, time-series blood glucose datasets were generated and were used to train auto regressive integrated moving average (ARIMA) ML model for the blood glucose prediction. This work is believed could give more insight in improving the diabetes management and treatment. The study's end goal is to examine the outcomes and performance of the constructed machine learning-based system. The performance of the machine learning model was evaluated through the Mean Squared Error (MSE), Mean Average Error (MAE) and Root Mean Squared Error (RMSE). It was found that the mean errors of MSE = 643, MAE = 19.83 and RMSE = 25.04 for 70:30 of train and test data splitting were lower than the other two ratios of 80:20 and 90:10 after the seasonality removal was conducted.

Keywords: ARIMA, Time-series Blood Glucose Prediction, Cobelli Model, Type 1 Diabetes Melitus

*Corresponding author: farhanah@uthm.edu.my
2023 UTHM Publisher. All rights reserved.

1. Introduction

High blood glucose levels are a chronic feature of diabetes mellitus. Patients with diabetes are either incapable of properly utilizing insulin or unable to produce enough insulin to maintain a healthy blood glucose level. The pancreas produces the enzyme insulin, which helps cells absorb glucose from the blood. The subtypes of diabetes include Type 1, Type 2, and gestational diabetes. Type 1 diabetes (T1DM) is characterized by insufficient or absent insulin production by the pancreas [1]. In contrast, type 2 diabetes (T2DM) occurs when the pancreas stops producing insulin and the body develops insulin resistance [2]. Gestational diabetes mellitus (GDM) is a transient condition that occurs only during pregnancy. It affects between 3% and 20% of pregnant women and increases the risk of developing chronic diabetes in both mother and child [3]. Frequent urination, increased thirst, and increased appetite are all indicators of high blood sugar. If diabetes is not treated, a number of complications may develop. Acute complications include hyperosmolar hyperglycemia, diabetic ketoacidosis, and even fatality. Serious long-term consequences include cardiovascular disease, stroke, chronic kidney disease, foot ulcers, and vision loss. In the T1DM research community, predicting blood glucose levels accurately has been a persistent obstacle.

It is significant to conduct this blood glucose level testing for diabetes management in order to identify blood sugar levels are high or low. In glucose-regulating systems like the artificial pancreas, accurate blood glucose predictions serve as the basis for the control algorithms. In order to produce forecasts for blood glucose levels with particularities in the input signals and underlying models used, numerous research investigations have already been carried out. In this case, further study on a suitable mathematical model is required to improve the efficiency of development of artificial pancreas as an alternative for diabetes treatment.

A time series is a set of observations made through time, whether it be daily, weekly, monthly, or annually. Time series modelling involves developing models used to describe the observed time series and understand the "why" behind its dataset [4]. In recent years, there has been a growing trend of applying machine learning algorithms to time-series predictions [5]. Thus, a study on a blood glucose prediction using a machine learning model in time series will be conducted through this work. This work is believed could give more insight in improving the diabetes management and treatment.

2. Materials and Methods

Figure 1 shows the general flowchart of the overall work development procedure. In consistent with the flowchart in Figure 1, the development of the work was started with a review of relevant research in order to establish the background and context of this study and to obtain a better understanding of the current state of knowledge and ideas in the field of machine learning.

In this work, the Cobelli diabetic model was chosen and used as an essential tool to generate ten time-series blood glucose datasets for T1DM using MATLAB simulation for this machine learning tasks. The Cobelli model is a compartmental model of glucose kinetics and insulin action that represents the input–output relationship between insulin infusion (as the input) and glucose concentration (as the output) [6]. Then, the machine learning algorithm will be conducted, that is by using the Autoregressive Integrated Moving Average (ARIMA). Then, the machine learning model performance evaluation was evaluated through the Mean Squared Error (MSE), Mean Average Error (MAE) and Root Mean Squared Error (RMSE).

2.1 Autoregressive Integrated Moving Average (ARIMA) Modelling

ARIMA, which is known as the 'Autoregressive Integrated Moving Average' is a statistical model that is widely used for time series prediction. ARIMA combines an auto regression (AR) component, which models the relationship between the current time step and previous time steps, and a moving average (MA) component, which models the leftover mistakes from a previous model, to model the time series data. It is a class of models that 'explains' a given time series based on its own previous values, that is, its own lags and the lagged forecast errors, so that the resulting equation can be used to forecast future values.

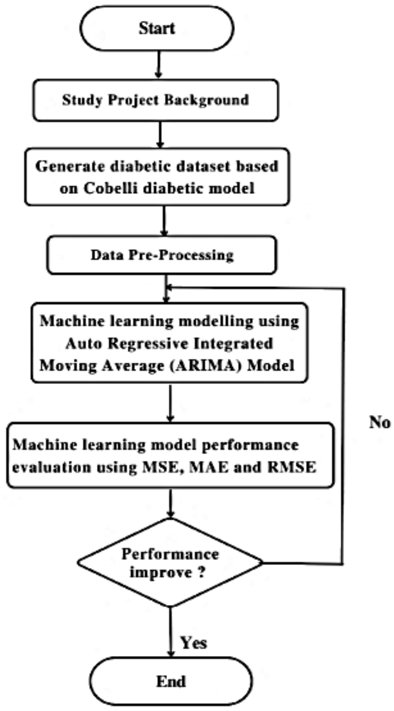


Figure 1: Flowchart of the overall methodology

2.1.1 ARIMA Parameters

ARIMA models can be built in an array of software tools, including Python with specific parameters [7]. The ARIMA model has 3 parameters and characterized by the terms: p, d, q :

i) p — Auto regressive feature of the model
A model with AR of order p regresses on its own p past values. In other words, it is possible to describe the current value of the series Vt as a linear combination of the p previous values plus a random error. We suppose that the p past values and the upcoming values are connected. The term "order" refers to the quantity of previous inputs utilized to forecast the following value. It is commonly referred to as "p". These earlier values are referred to as "lagged variables" and are utilized to calculate the upcoming value.

ii) d — Differencing order
Usage of raw observation differencing (e.g., subtraction of an observation from an observation at the previous time step) to make a time series stationary.

iii) q — Moving average feature of the model
The dependency between an observation and a residual error from a moving average applied to lag observations is used in the Moving Average model. In other words, MA is a combination of the past error terms up to the q step into a linear model to represent the value to be predicted. The formula for the error or residual is res = predicted value – true value.

2.1.2 ARIMA Seasonality and Stationarity

There are many different forms of seasonality, including time of day, daily, and weekly, monthly and yearly. The simplest way to determine if there is a seasonal component is to plot and examine the data, possibly at various scales and with trend lines added.

Stationarity can be described with precise mathematical terms, but for our purposes, it refers to a series that appears flat, has no trend, a constant variance over time, a constant autocorrelation structure over time, and no periodic fluctuations or seasonality. There are stationary and non-stationary ARIMA models that can be used for forecasting. However, it is really crucial to see if time series is stationary or non-stationary. In general, non-stationary data are unpredictable and cannot be modelled or predicted. The results obtained by using non-stationary time series may be inaccurate in a way that they may suggest a relationship between two variables where none exists. The non-stationary data must be converted into stationary data in order to obtain reliable and consistent results. In stationary time-series data, a sample observation's properties or value are independent of the timestamp at which it was observed. For instance, given a hypothetical dataset of the year-by-year population of a region, if one observes that the population doubles or increases by a constant number each year, then these data are non-stationary.

2.1.3 Converting Non-Stationary to Stationary Time Series Data

Predicting stationary series is relatively straightforward, and the ensuing forecasts are more exact. Consequently, converting non-stationary to stationary data resolves this issue because it eliminates persistent autocorrelation, thereby leaving the predictors (series lags) in the prediction models virtually independent. The ADF (Augmented Dickey Fuller) test is the most popular and common statistical test conducted to check the stationarity. It can be used to find the series' unit root and, in turn, determine whether or not the series is stationary. The null hypothesis is therefore rejected if the P-Value in the test is less than the significance level (0.05), thus the data is said to be stationary meanwhile if the P-Value in the test is more than the significance level (0.05) then it is a non-stationary data.

2.1.4 Performing Seasonal Decomposition

Using differencing, which determines the difference between the present value and its value in the previous season, we may eliminate seasonality from the data [6]. To make the time series stationary and maintain its statistical features throughout time, this is done. When we are in a certain season, seasonality causes the mean of the time series to be different. Therefore, its statistical properties are not constant. The mathematically equation of seasonal differencing is :

$$d(t) = y(t) - y(t-m) \qquad (3.1)$$

Where d(t) represents the differenced data point at time t, y(t) represents the value of the series at time t, y(t-m) represents the value of the data point from the previous season, and m represents the extent of one season. In this case, m = 24 because the seasonality is 24 hours. If the null hypothesis of no stable seasonality is rejected at a significance level of 0.1%, p value < 0.001, then the series is considered seasonal and the test result is displayed.

2.1.5 Choosing ARIMA Model parameters

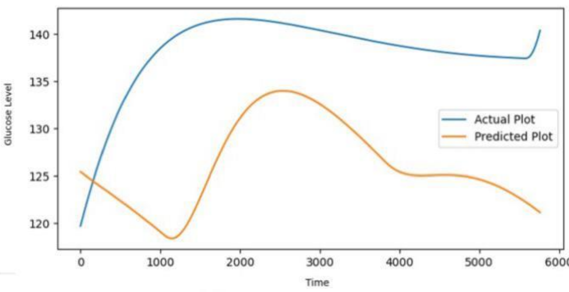
The ARIMA-model was implemented in Python, and the statistical programming language's auto.arima function was used to determine model-specific parameters. It uses the Bayes information criteria to ascertain p and q and the Philips-Perron unit root test to find parameter d. When using this Auto ARIMA, the model will produce the ideal p, d, and q values that are acceptable for the data set and would result in better predictions. Auto ARIMA simplifies this task as it gives the suitable parameters based on the generated dataset. The Auto ARIMA model function is more effective at determining the ideal p, d, and q values than the typical ARIMA implementation, which requires differencing and plotting ACF and PACF plots. In this work, the auto ARIMA (p=2, q=1, d=1) was implemented. It has two autoregressive (AR) terms, one moving average (MA) term, and one differencing term. The ARIMA (2, 1, 1) model is a flexible model that can be used to model a wide range of different sorts of data. It is a relatively simple model to understand and implement, and it can be used to produce precise predictions.

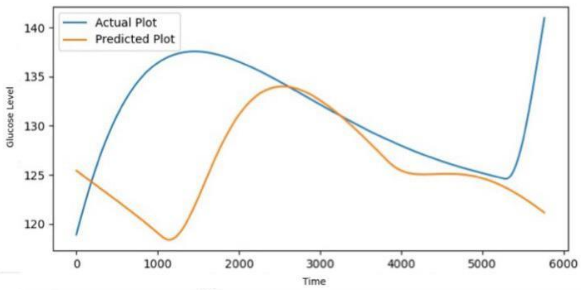
3. Results and Discussion

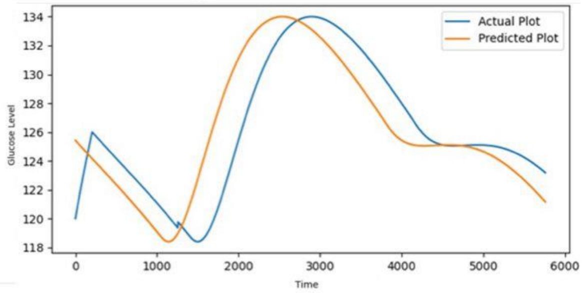
In this discussion the data obtained is for 70:30 split of training and testing and after the seasonality removal as it was found that all the ten datasets used are seasonal. Table 1 shows a part of the obtained results of actual dataset from the testing data and the graph of the prediction of blood glucose value.

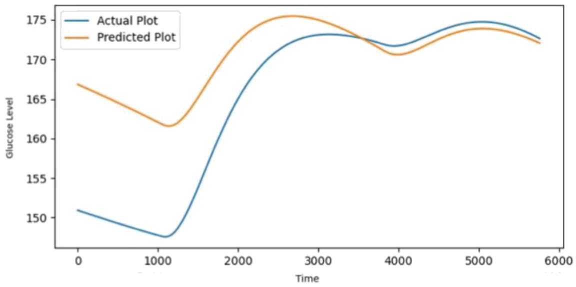
Based on the Table 1, the graphs consisting of Data 1, Data 2 and Data 5 are the prediction results of glucose values for normal blood glucose levels. Normal blood glucose levels are between 70 mg/dL and 100 mg/dL however the prediction seems to appear slightly higher. Whereas Data 6 and Data 9 from Table 1 are the results of glucose values which results in hyperglycemic blood glucose levels. Generally, hyperglycemic levels are when the blood glucose values are greater than 125mg/dl.

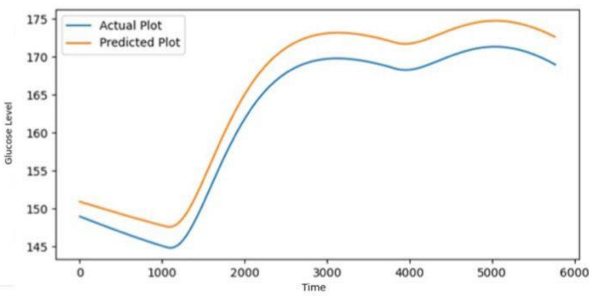
Table 1 : Comparison of the blood glucose values between the actual dataset from the testing data and its prediction data from selected datasets of Data 1, 2, 5 6 and 9

<table>
<tr><th>Dataset</th><th>Results</th></tr>
<tr><td>Data 1</td><td>


</td></tr>
<tr><td>Data 2</td><td>


</td></tr>
<tr><td>Data 5</td><td>


</td></tr>
<tr><td>Data 6</td><td>


</td></tr>
<tr><td>Data 9</td><td>


</td></tr>
</table>

Although the graphs appear to be quantitatively comparable yet, the changes in the prediction graphs are noticeable. Although the accuracy of the prediction graphs is not 100%, it can be concluded that the 70:30 split of data is more precise compared to 80:20 and 90:10 of data splitting.

3.1.1 ARIMA Performance Evaluation

Evaluation of the model's performance is a significant part of any machine learning model. Data splitting is frequently applied in machine learning to prevent overfitting [8]. In this ARIMA model, the Mean Squared Error (MSE), Mean absolute error (MAE) and Root Mean Squared Error (RMSE) are used to evaluate the performance of the model. The MSE, MAE and RMSE results of 10 datasets with 90:10, 80:20 and 70:30 ratio of training and testing data before and after the seasonality test was conducted for comparison purposes.

Idealistically, if there are an infinite amount of training data, the ability to optimize the model's performance is extremely high. The model would be best quality ever created. And if there were an infinite amount of testing data, there would be complete confidence in the model's reliability too. The 80:20 division is widely used including the 90:10 ratio which is at random. The current research yielded insufficient data, so the 70:30 split is the most appropriate for this model in which it works by training on 70% of the data and testing on 30% of the data [7].

When compared to MAE, MSE and RMSE penalize big prediction errors. The RMSE is often recommended over MAE and MSE when evaluating model efficacy [7]. This is due to the fact that developers frequently seek to reduce the occurrence of large outliers in their predictions, and MAE is considered insufficient for assessing overall performance of models. A lower MAE, MSE, and RMSE value indicates that the model used is more accurate and precise [7].

The ARIMA performance evaluation was conducted before and after the seasonality removal. This was done to observe if any differences occur in the dataset values. Based on the analysis of the above Table 2, it can be presumed that the results of MSE, MAE and RMSE with ratio of 70:30 after conducting seasonality removal has shown a higher performance compared to 80:20 and 90:10 ratio of training and testing. The mean error for all the calculation reduced by 20% after the seasonality removal was conducted where for MSE was 991.74 and reduced to 643.40, MAE was 26.55 and reduced to 19.83 and lastly RMSE was 30.93 and reduced to 25.04 which is lower than the other two ratios of 80:20 and 90:10. Here, smaller the value of MSE, MAE and RMSE indicates higher performance of the model. Therefore, it is found that other than stationarity, the seasonality also plays a very crucial role in this ARIMA model.

Table 2: Summary of Mean Error for MSE, MAE and RMSE

Training to testing ratio, before/ after seasonality removal	MSE	MAE	RMSE
90:10 before seasonality removal	5568.28	84.31	84.35
90:10 after seasonality removal	2121.46	33.26	48.47
80:20 before seasonality removal	3900.68	58.27	60.83
80:20 after seasonality removal	1293.08	27.21	34.84
70:30 before seasonality removal	991.74	26.55	30.93
70:30 after seasonality removal	643.40	19.83	25.04

4. Conclusion

Effective blood glucose prediction is necessary for the development of an artificial pancreas as a substitute for diabetes treatment. The main idea of this study was to extend current research on predicting blood glucose values mainly for type 1 diabetic patients. This approach is designed to ease diabetic patients for better understand and keep track of glucose levels in order to improve the treatment of type 1 diabetes. Therefore, this study proposed an effective method for predicting blood glucose levels. Ten datasets were used to evaluate the effectiveness of the proposed machine learning algorithm. The first objective of this study was to generate a blood glucose dataset from Cobelli diabetic model. Thus, by using the MATLAB software the dataset needed were successfully simulated and obtained. Next, the second objective of this work was to design a supervised machine learning model for the prediction of blood glucose based on time series modeling. This has been achieved as the machine learning model using ARIMA has been designed in the Google Colaboratory and the programming language that has been used in Python programming language. Besides that, the last objective of this work was to evaluate the performance of machine learning for the prediction of blood glucose. In order to conduct the evaluation, metrics for regression which involve calculating an error score to summarize the predictive skill of a model has to done. This has been achieved since the performance of the machine learning model has been evaluated through the MSE, MAE and RMSE. After calculating the mean error of MSE, MAE and RMSE it can be concluded that the good prediction performance has been achieved for 50%.

For the recommendation, this work could be continued in a variety of ways. A relatively simple upgrade should be done to test the ARIMA models using a larger dataset, to be clearer a collection of data for a few months of observing the patient. This is mainly because it is more on personalized prediction, as a different patient has various trends of data. Therefore, it could help to enhance the accuracy of prediction. By doing so, the performance of the ARIMA models would be further established and it might even lead to new concepts for obtaining better results. The quantity and number of data features employed in this study were both limited. Thus, building new data models produce even greater performance may be assisted by having a data collection with a greater variety of variables. The accuracy of the prediction could be enhanced by taking into consideration measures like activity and sleeping patterns, weight, or even pulse.

Acknowledgement

The authors would like to thank the Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia for its support through TIER 1 Grant No. Q359.

References

[1] Romesh Khardori. "Type 1 Diabetes Mellitus: Practice Essentials, Background, Pathophysiology," 3, July 2023. [Online]. Available: Medscape https://emedicine.medscape.com/article/117739-overview. [Accessed Jul. 6, 2023]

[2] L.Mei Qing. “Assessment of Cobelli Type 1 Diabetic Model for Meal Simulation with An Open-Loop Insulin Administration,” Undergradute Thesis, Dept. Elect. Eng., Universiti Tun Hussein Onn Malaysia, 2019

[3] Yang, J., Li, L., Shi, Y., & Xie, X., "An ARIMA Model with Adaptive Orders for Predicting Blood Glucose Concentrations and Hypoglycemia." IEEE Journal of Biomedical and Health Informatics, 2019 pp.1251–1260. https://doi.org/10.1109/jbhi.2018.2840690.

[4] Cobelli. C, Dalla Man. C, Sparacino. G, Magni. L, De Nicolao. G, and Kovatchev. B. P, (2009) “Diabetes: Models, signals, and control,” IEEE Rev. Biomed. Eng., vol. 2, pp. 5496, 2009, doi: 10.1109/RBME.2009.2036073.

[5] Ekram Mohammad Faudzi1. “Optimization of a PID Controller System for Cobelli Diabetic Model Based Glucose-Insulin Control,” Undergradute Thesis, Dept. Elect. Eng., Universiti Tun Hussein Onn Malaysia, 2022

[6] Etienne, B, "Time Series in Python: Dealing with seasonal data, " April 2023. [Online]. Available: Medium https://medium.com/towards-data-science/time-seriesin-python-part-2-dealing-with-seasonal-data-397a65b74051. [Accessed June 16, 2023]

[7] C. Dalla Man, R. A. Rizza, and C. Cobelli, “Meal simulation model of the glucose-insulin system,” IEEE Trans. Biomed. Eng., vol. 54, no. 10, pp. 1740–1749, Oct. 2007, doi: 10.1109/TBME.2007.893506.

[8] Finan, D. A., Zisser, H., Jovanovic, L., Bevier, W. C., & Seborg, D. E. “Identification of Linear Dynamic Models for Type 1 Diabetes: A Simulation Study. 2006, pp. 503–508, doi: https://doi.org/10.3182/20060402-4-br-2902.00503.