Supervised Machine Learning Model for Diabetic Students' Glucose Levels Classification System

Mona Alotaibi(✉) and Mike Joy

Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK
{mona.alotaibi,m.s.joy}@warwick.ac.uk

Abstract. Accurate and timely blood glucose prediction is essential for students who have been diagnosed with diabetes to prevent hypoglycemia and hyperglycemia episodes throughout the school days. Continuous Glucose Monitoring (CGM) is utilized in clinical situations to manage blood sugar levels effectively which is a form of sensor that can be incorporated in the Internet of Things (IoT). In addition, machine learning (ML) models can precisely be used for predicting glucose levels using a CGM system. We aim to develop a decision support system based on ML algorithms and the IoT, for which we have applied Random Forest, Logistic Regression, AdaBoost (Ada), Multi-layer perceptron (MLP), and linear SVM algorithms, followed by a majority voting algorithm. Evaluation metrics such as the Confusion Matrix, Accuracy, Precision, F1 score and Recall, were implemented to evaluate the performance of each algorithm. The experimental results demonstrate that the accuracy of the majority voting model after five-fold cross-validations is 99.61%, which ideally achieves higher performance than any single model.

Keywords: Internet of things · Machine learning · Diabetic students · Majority voting algorithms · Predictive model · Continuous glucose monitoring

1 Introduction

The chronic disease diabetes is described by abnormal blood glucose levels (BGLs), which can be either higher (hyperglycemia) or lower (hypoglycemia) than the normal range (80 mg/dl to 140 mg/dl), according to the American Association of Clinical Endocrinologists (AACE) [1]. Keeping BGLs within the normal range is crucial for diabetes treatment [2], and BGL predictive models might help to achieve this aim, by predicting severe glycaemic episodes and allowing patients to take preventative measures in advance.

Type 1 diabetes is a case of a childhood health condition that significantly impacts academic performance and educational achievement [3]. Previous research has demonstrated that students with chronic diseases, such as diabetes, have more difficulties with scientific performance and learning than non-diabetic students. We performed an investigative study [4], the purpose of this study was to investigate the academic performance

K. Arai (Ed.): IntelliSys 2023, LNNS 825, pp. 516–524, 2024.
https://doi.org/10.1007/978-3-031-47718-8_35

of diabetic students in Saudi Arabian high schools and to highlight the academic problems that diabetic students, teachers, parents of children with diabetes, and school administrators, may encounter. The findings of the investigative study indicate that diabetes significantly impacts the academic performance of students, and diabetic students face a variety of academic difficulties, including attendance problems, concentration, and school settings such as diabetes care plans, and measures to monitor blood glucose.

Traditionally, the educational levels of students are assessed by exams, homework, and contributions. However, students with chronic diseases are assessed by these methods regardless of performance issues which those diseases may contribute to. Due to the lack of previous research addressing these issues, the aspect of this research focuses on Saudi Arabian education, with the aim of enhancing student safety and assisting teachers in evaluating diabetic students.

Fig. 1. Proposed system design.

To monitor students' health problems and improve assessments of diabetic students, we proposed a system based on the machine learning (ML) model and the Internet of Things (IoT). Figure 1 shows the proposed system design; the targeted users of the system include teachers, administrators, and diabetic students. This system comprises wearable IoT sensors for diabetic students that can be worn inside or outside a classroom, such as Continuous Glucose Monitoring (CGM). These sensors track the glucose levels of the students. So, the students will be monitored during the school day. A decision support system is then used to analyse the glucose readings after they are delivered to a cloud server. A smart device installed with a mobile application is used to provide real-time notifications to a teacher and the administrative team. The decision support system sends notifications and alerts if any abnormal readings are detected with guided instructions to avoid the risk of hypoglycaemia and hyperglycaemia. Moreover, to assist teachers throughout the student assessment period, the system provides a visualization of the attendance, homework, exam, and contribution scores of the students, and generates

charts. The first phase of the system development was to create the ML model. Different ML algorithms have been implemented such as Random Forest, Logistic Regression, AdaBoost, Multi-layer Perceptron (MLP), and linear Support Vector Machine. Then, a majority voting algorithm was applied.

This paper is structured as follows. In Sect. 2, a brief overview of the related work is provided. Section 3 describes the adopted methodology which includes details about the dataset and pre-processing, explaining the used ML algorithms and experiment setup. In Sect. 4, an analysis of the model's performance is discussed. The final section presents the conclusion and future work.

2 Related Work

There are a variety of previous studies which have utilized machine learning algorithms (ML) and the internet of things (IoT) to predict ypoglycaemia and hyperglycemia events. The author in [5] proposed a smart healthcare device for diabetic patients which used artificial intelligence (AI) based algorithms. A microprocessor, an insulin pump, and a continuous glucose monitoring (CGM) system comprise the system. Long short-term memory (LSTM) architecture was implemented as the method for machine learning. This system demonstrated efficiency in terms of the reliability and speed of the communication link between the CGM, the pump, and the patient. Moreover, the system displayed the user's glucose levels every 5 min. In addition, it alerted the user in the event of a serious situation and allowed the user to share a comprehensive report with the healthcare center. Normalized mean square error (MSE) and root mean square error (RMSE) were calculated to compare the performances. The MSE was 0.000915 while the RMSE was 0.03025.

Another study was conducted by Bhargav et al. [6] to evaluate several ensemble machine learning (ML) models for generalized BGL prediction and their innovative integration with decision tree (DCT) models. On the data, the decision tree, random forest, extra trees, gradient boost, Adaboost, and bagging models were evaluated. A novel two-stage model including decision tree (DCT) and Adaboost (DCT-Aboost) was designed, with the DCT model's predictions providing input data to the Aboost model for the final BGL prediction. The results demonstrated that the DCT-Aboost model outperformed and it achieved a test RMSE of 2,207.

3 Materials and Methods

This section describes the methods, the dataset, data pre-processing, supervised ML algorithms, and the experimental setup.

3.1 Dataset

Dataset Description

The dataset was downloaded from the Dexcom website [7]. This company develops, manufactures, and sells continuous glucose monitoring devices for the treatment of

diabetes. In addition, it has an API (Application Programming Interface) which provides access to a dataset.

The Dexcom API supports different languages such as Python and provides multiple libraries used in ML. One of the developers' environments in this API is a sandbox environment which contains simulated CGM data and is available for all registered developers to test their applications.

As shown in Table 1, the original dataset contains raw data of the system time, display time, value, realtime value, smoothed value, status, trend, and trend rate. The dataset was downloaded as CSV files and the total number of CSV files was 84 from different years. Then, they were combined in one CSV file. Finally, the total number of rows in this file was 647437. Furthermore, one important attribute was added to the dataset which is the "_class" column to classify the glucose reading depending on the value and trend columns. The class values were coded from critical low (1), low (2), normal (3), high (4), to critical high (5).

Dataset Pre-Processing

The main aim of data pre-processing is to enhance model accuracy and data quality, minimize noise and reduce the chance of overfitting [8]. To perform the pre-processing steps, a Python application was developed to clean the data. The first pre-processing task was taking care of unnecessary data such as system time and display time. So, these columns were removed from the dataset. The next pre-processing step was to remove the duplicated columns, where the column value and real-time value were the same. As a result, the column "realtime" was deleted. In addition, the pre-processing included removing 10% of null columns, thus, smoothed value and status were deleted because they were empty. After that, features with low variance were deleted, and the trend rate column was removed because the value of this column is 0. Furthermore, the column "trend" consisted of categorical data, so the column was converted into numeric data by using the Pandas get_dummies function for the hot encoding [9]. As an illustration, if the trend is Flat, hot encoding transforms Flat into [0,0,1,0,0,0,0]. Table 2 shows the final dataset after the pre-processing process.

3.2 ML Algorithms

Supervised ML

Several studies have examined the application of ML models for diabetes prediction and diagnosis [10]. The novelty of this study is the classification of glucose readings based on the value and trend parameters. In addition, the performance of six different supervised ML models in terms of prediction is presented. To classify the values and trend of blood glucose, Random Forest, Logistic Regression, AdaBoost (Ada), Multi-layer perceptron (MLP), linear SVM, and majority voting techniques were used.

Algorithms for supervised ML require external support. The input data is divided into separate training and testing datasets. The training dataset comprises the output variables necessary for classification or prediction. All algorithms learn the same pattern from the training data set in order to predict or classify the test data set [11].

Table 1. Dataset

System Time	Displaytime	Value	Realtime Value	Smoothed value	Status	Trend	Trendrate	_Class
2015–06-30T23:59:10	2015–06-30T17:01:42	156	156			Flat	0	4
2015–06-30T23:54:10	2015–06-30T16:56:42	152	152			Flat	0	4
2015–06-30T23:49:10	2015–06-30T16:51:42	151	151			Flat	0	4
2015–06-30T23:19:11	2015–06-30T16:21:43	148	148			fortyFiveUp	0	5
2015–06-30T23:04:10	2015–06-30T16:06:43	133	133			Flat	0	3
2015–06-30T22:59:10	2015–06-30T16:01:43	129	129			Flat	0	3
2018–01-25T20:09:00	2018–01-25T13:11:32	158	158			singleUp	0	5

Table 2. Final dataset.

Value	Double down	Doubleup	Flat	Forty five down	Forty five up	Single down	Single up	Class
156	0	0	1	0	0	0	0	4
148	0	0	0	0	1	0	0	5
146	0	0	0	0	1	0	0	5
140	0	0	0	0	1	0	0	3
133	0	0	1	0	0	0	0	3
142	0	0	0	0	0	1	0	4
63	0	0	0	1	0	0	0	1
73	0	0	1	0	0	0	0	2

Random Forest: Random Forest is a simple ensemble algorithm with a high level of accuracy that applies to both regression and classification issues utilizing many decision trees and the "bagging" approach. It is a method for ensemble learning that outperforms a single decision tree. This tree category has the benefit of averaging the results to prevent overfitting. Specifically, the Random Forest algorithm picks data at random from a given dataset. Subsequently, a decision tree is constructed for each sample, which is then used to make predictions. A majority voting procedure is then used to determine the final prediction [12, 13].

Logistic Regression: The Logistic Regression ML algorithm is an adaptive regression method that employs a Boolean combination of binary variables to create predictions [14]. The algorithm is used for a classification problem to determine a single Boolean expression that predicts a binary outcome.

Using Logistic Regression models, the impact of predictor variables on categorical outcomes is used. In most situations, the result is binary, such as the existence or absence of a disease, in which case the model is referred to as a binary logistic model. When a Logistic Regression model has only one predictor variable, the model is referred to as a Simple Logistic Regression. Numerous or Multivariable Logistic Regression refers to a model that combines categorical and continuous variables as predictors when there are multiple predictors [15].

AdaBoost (Ada): AdaBoost generates a collection of component classifiers by keeping weights over the training set and iteratively making adjustments to these weights after each boosting iterative process. The weights of the training set that are incorrectly classified by the latest component classifier are increased, whereas the weights of the training set that are correctly classified are decreased [16].

Multi-layer Perceptron (MLP): Artificial Neural Networks (ANNs) can execute model function prediction and manage linear/nonlinear functions by learning from data associations and generalising to unknown cases. MLP is a common ANN, and is a robust modelling tool that employs a supervised training approach utilising data examples with known outcomes. This technique develops a model of a nonlinear function that permits the prediction of output data based on input data [17].

Linear Support Vector Machines (SVM): SVMs are used for both classification and regression. In the SVM model, data points are displayed on the area and divided into groups, and points with similar characteristics are grouped. In a linear SVM, a provided data set is represented as a p-dimensional vector that may be split by a maximum of p-1 hyperplanes. These hyperplanes divide the data space or define the data group boundaries for classification or regression issues [18].

Majority Voting Ensemble: The final classification is based on the majority vote of the five models mentioned in this section, which are merged into an ensemble model (hard voting). Each model predicts each case, and the prediction that obtains more than half of the votes is the final prediction [19].

3.3 Experiments Setup

In our experiment, we implemented the above five ML algorithms and applied a majority voting algorithm. Accuracy, Precision, F1 score, and Recall were used as the evaluation metrics for each algorithm. In addition, we measure the accuracy of the majority voting model after five cross-validations. The dataset was divided into 70% for the training and 30% for the test set. We used scikit-learn, and Pandas in Python software for pre-processing, data cleaning, and training and testing the models.

3.4 Results and Discussion

Table 3 shows all models' performances for each ML algorithm. In terms of accuracy indicators, the MLP model has the highest accuracy, Precision, F1 score and Recall more than other models which were 99.23, 99.24, 99.23 and 99.23, respectively. The accuracy of Random Forest, linear SVM, AdaBoost and Logistic Regression were 98.69, 95.94, 95.11, and 94.86, respectively. Next, the accuracy of the majority voting algorithm after five-fold cross-validations was 99.61.

Table 3. Supervised ML model performance.

Algorithm	Accuracy	Precision	F1 score	Recall
Random forest	98.69	98.84	98.56	98.69
Logistic regression	94.86	92.22	93.51	94.86
AdaBoost	95.11	90.88	92.84	95.11
MLP	99.23	99.24	99.23	99.23
Linear SVM	95.94	96.17	95.75	95.94

In addition, A confusion matrix was used to evaluate the performance. A confusion matrix of n x n dimension related to a classifier illustrates the predicted and actual classification, where n is the number of different classes [20].

As shown in Table 4 in terms of the confusion matrix of the majority voting model, 200 cases of critical high class were misidentified as a normal class. The results revealed that the majority voting model outperformed the five single supervised ML models.

Table 4. Confusion matrix of the majority voting model.

Classes	Critical low	Low	Normal	High	Critical high
Critical low	884	0	0	0	0
Low	0	4495	0	0	0
Normal	0	0	122039	0	0
High	0	0	0	54939	0
Critical high	0	0	200	0	8251

4 Conclusion and Future Work

Students with diabetes have faced issues in academic achievement and school settings due to the implications of repeated hypoglycemia and hyperglycaemia cases. Separating diabetic students from this risk through a system that predicts the degree of glucose readings based on ML models and the IoT is a school priority. In the first phase of the system development, we developed the ML model to integrate it with web and mobile applications. We applied Random Forest, Logistic Regression, Ada, MLP, and linear SVM algorithms. Then, a majority voting algorithm was applied to achieve the best performance of the model. The result of the experiment reveals the accuracy of the majority voting model as 99.61%, which was better than applying a single model. This study was limited by the absence of using different types of Continuous Glucose Monitoring (CGM). Our future work is to develop web and mobile applications and integrate the ML model with them.

References

1. American Diabetes Association: Standards of medical care in Diabetes—2007. Diabetes Care **30**(Suppl 1), S4–S41 (2007)
2. Ajjan, R., Slattery, D., Wright, E.: Continuous glucose monitoring: a brief review for primary care practitioners. Adv. Ther. **36**, 579–596 (2019)
3. Persson, E., Persson, S., Gerdtham, U.G., Steen Carlsson, K.: Swedish childhood diabetes study group: Effect of type 1 diabetes on school performance in a dynamic world: new analysis exploring Swedish register data. Appl. Econ., **51**(24), pp. 2606–2622 (2019)
4. Alotaibi, M., Joy, M.: Internet of things support system for diabetic students: an exploratory study. In: Arai, K. (eds). In: Proceedings of the Future Technologies Conference (FTC) 2022, Volume 2. FTC 2022 2022. Lecture Notes in Networks and Systems, **560**. Springer, Cham (2023)
5. Almulla, S.K., Zaatar, O., Ahmed, H.S., Jarndal, A.: Smart Artificial-Intelligence based Self-Care-Device for diabetic patients. In 2022 Advances in Science and Engineering Technology International Conferences (ASET) pp. 1–5. IEEE (2022)
6. Bhargav, S., Kaushik, S., Dutt, V.: A combination of decision trees with machine learning ensembles for blood glucose level predictions. In Proceedings of International Conference on Data Science and Applications: ICDSA 2021, **2** (pp. 533–548). Springer Singapore (2022)
7. Dexcom API Home, https://developer.dexcom.com/home, last Accessed 24 Feb. 2023

8. Maharana, K., Mondal, S. and Nemade, B.: A review: Data pre-processing and data augmentation techniques. In: Global Transitions Proceedings (2022)
9. Paper, D., Paper, D.: Predictive modeling through regression. Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python, pp. 105–136 (2020)
10. Alexiou, S., Dritsas, E., Kocsis, O., Moustakas, K., and Fakotakis, N.: An approach for personalized continuous glucose prediction with regression trees. In 2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) pp. 1–6. IEEE (2021)
11. Kotsiantis, S.B., Zaharakis, I., and Pintelas, P.: Supervised machine learning: A review of classification techniques. Emerg. Artif. Intell. Appl. Comput. Eng. **160**(1), pp. 3–24 (2007)
12. Ye, Y., Xiong, Y., Zhou, Q., Wu, J., Li, X., and Xiao, X.: Comparison of machine learning methods and conventional logistic regressions for predicting gestational diabetes using routine clinical data: a retrospective cohort study. J. Diabetes Res. (2020)
13. Denil, M., Matheson, D., and De Freitas, N.: Narrowing the gap: Random forests in theory and in practice. In International Conference on Machine Learning (pp. 665–673). PMLR (2014)
14. Kooperberg, C., Ruczinski, I., LeBlanc, M.L., Hsu, L.: Sequence analysis using logic regression. Genetic epidemiology, **21**(S1), pp. S626–S631 (2001)
15. Nick, T.G., Campbell, K.M.: Logistic regression. Topics in Biostatistics, pp. 273–301 (2007)
16. Kuncheva, L.I., Whitaker, C.J.: Using diversity with three variants of boosting: Aggressive, conservative, and inverse. In Multiple Classifier Systems: Third International Workshop, MCS 2002 Cagliari, Italy, June 24–26, 2002 Proceedings **3** pp. 81–90. Springer Berlin Heidelberg (2002)
17. Taud, H., Mas, J.: Multilayer Perceptron (MLP). In: Camacho Olmedo, M., Paegelow, M., Mas, JF., Escobar, F. (eds). Geomatic approaches for modeling land change scenarios. Lect. Notes Geoinformation Cartogr. Springer, Cham (2018)
18. Kaur, H., and Kumari, V.: Predictive modelling and analytics for diabetes using a machine learning approach. Appl. Comput. Inform. **18**(1/2), pp. 90–100 (2020)
19. RahmaAtallah, A.A.M.: Heart disease detection using machine learning majority voting ensemble method, pp. 1–6 (2019)
20. Visa, S., Ramsay, B., Ralescu, A.L., Van Der Knaap, E.: Confusion matrix-based feature selection. Maics **710**(1), 120–127 (2011)