A Machine Learning Model Based Web App to Predict Diabetic Blood Glucose

Masuda Begum Sampa (sampa.stat@gmail.com)
University of Science and Technology Chittagong
Topu Biswas
University of Science and Technology Chittagong
M Rakibul Hoque
University of Dhaka
M Nazmul Hossain
University of Dhaka
Ashir Ahmed
Kyushu University

Research Article

Keywords: blood glucose, NCDs, machine learning, non-invasive, Boosted Decision Tree Regression model

Posted Date: January 19th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2488325/v1

Abstract

Aim of this study is to use machine learning approaches for predicting blood glucose based on basic non-invasive health checkup test results, dietary information, and socio-demographic characteristics and to develop a web application to predict blood glucose easily. We evaluated the performance of five widely used machine learning models. Data have been collected from 271 employees of Grameen Bank complex, in Dhaka, Bangladesh. This study used continuous blood glucose data to train the model and predicted new blood glucose values using the trained data. Finally, we developed a blood glucose prediction web application. The Boosted Decision Tree Regression model showed the best performance among other models based on the Root Mean Squared Error (RMSE) 2.30, this RMSE is better than any reported in the literature. This study developed a blood glucose prediction model and web application which is easier, more convenient, and more efficient for people. People can also easily check their blood glucose values using our app, especially in remote areas of developing countries that lack adequate skilled doctors and nurses. By predicting blood glucose, this study can help to save medical costs and time and to reduce health management costs. Our system can be helpful in achieving SDGs, Universal Health Coverage and thus reducing overall morbidity and mortality.

1 Introduction

Diabetes mellitus (DM) which is characterized by high blood glucose, is one of the rapidly increasing diseases throughout the world [1]. Diabetes is a Non-Communicable Disease (NCDs). It is the root or mother of all diseases worldwide. Almost all organs of the body are affected by uncontrolled diabetes. The excess glucose circulating through the body in the bloodstream over time leads to severe long-term-mortality-related complications such as cardiovascular diseases, kidney diseases, eye problems, diabetic neuropathy, and diabetic retinopathy [2]. The occurrence of diabetes has a strong relation with diseases like Wheeze, Edema, and Oral disease [3]. Early detection of blood glucose level value can help to develop awareness among pople to protect them from diabetes and can help to save a lot of public health resources every year. Early detection of blood glucose value also plays an important role in the effective management of diabetes.

Optimal and sustainable control of blood glucose levels (BGL) is the aim of type-1 diabetes management. The automated prediction of BGL using machine learning algorithms is considered as a promising tool that can support this aim. Machine learning (ML) is a computational method for automatic learning from experience and improves the performance to make more accurate predictions [4, 5]. A future glucose value depends not only on past observations of glucose values, but also on insulin dosage, carbohydrate intake and other life data. Many factors affect blood glucose levels. A large number of factors that are known to be important in development and progression of DM such as obesity, biomarkers (body fluids: urine, blood, saliva) [5]. In this context, this work presents an approach through ML techniques to predict blood glucose values utilizing non-invasive basic health measurements. Data from 271 urban corporate people were used to train the machine learning models to predict future glucose values. We did random blood glucose test among 271 corporate employees.

People who are working in urban areas, especially in the private sector have significant workloads and remain seated for a long time to complete their tasks, they are more likely to develop NCDs. In addition, little chance exists to engage in physical activities among the urban people in Bangladesh because of a lack of playgrounds, parks, walkable footpaths, and safe roads for cycling [6]. The prevalence of risk factors for developing NCDs is also higher among urban than rural people in Bangladesh [6]. Therefore, it is important to control and prevent the seriousness of NCDs by getting regular health checkups. However, most people are not interested in spending money and time on preventive healthcare services. Corporate people in Bangladesh lack health insurance and high health awareness, do not get routine mandatory health checkups and are not habituated to using ICT-based healthcare services. Moreover, to get a health checkup, they need to visit a hospital in traffic-congested area and wait in a long, laborious queue [7].

The health status of an individual depends on blood glucose and it is considered to be a risk factor for the development of NCDs [8, 9]. Therefore, blood glucose should be measured routinely at basic health check-ups. Because its prediction could be very helpful in preventing various NCDs. As the reduction of NCDs management cost is the main goal of health policy [10], studies are needed to determine the blood glucose regularly in a cost-effective way. An accurate predictive model can help to identify a risky population [11]. By using the prediction model designed by the machine learning approaches to test individual blood glucose measurement rapidly will save the cost and time of doctors and patients as well.

First, to our knowledge application of machine learning approaches in blood glucose prediction in developing countries are very rare. That is why blood glucose is chosen for this research. Second, different algorithms will work differently on different types of data in various diseases, such as different types of cancers, and diabetes. Therefore, a different investigation is needed for different types of data and different types of diseases in order to identify the most accurate algorithms [12].

The machine learning methods for predicting blood glucose have not yet been practically established for clinical data from developing countries such as Bangladesh. There is a lack of research on predicting blood glucose based on basic non-invasive clinical measurements, dietary information, and socio-demographic characteristics by using machine learning approaches in Bangladesh. Specifically, yet, no study focused on the urban corporate people in Bangladesh.

Therefore, this study aims to use machine learning approaches to predict blood glucose value based on basic non-invasive health checkup test results, dietary information, and socio-demographic characteristics. This study used several state of the art machine learning approaches to evaluate the predictive power of these techniques and to predict personalized blood glucose value. The models are evaluated regression-wise. The performance of the model was compared with other popular state-of-the-art models for predicting health parameters. The results show the performance of the developed model over the other models.

We evaluate our model using the standard Root-Mean-Squared Error (RMSE) metric. The goal of predicting health checkup test measurements is very helpful to reduce health management costs.

We have developed a personalized machine learning model that predicts the blood glucose value. The model has been integrated into an app, BloodGlucosePrediction Calculator(link is external and opens in a new window), that will allow individuals with non-invasive basic health measurements to predict their blood glucose value—the key to preventing or controlling the major complications of the diabetes disease.

2 Methods

2.1 Existing related studies

There are a number of studies on blood glucose level prediction that have used different features, a different number of subjects, and a different number of features [1]. This paper aims to deliver high accuracy solution with basic non-invasive health check up items considering 271 urban corporate employees. The previous works done in this field is summarized below.

In [13], a recurrent ANNs; Elman recurrent ANNs have been used for making blood glucose level (BGL) predictions using the previous blood glucose values. They used a virtual data set (Case 002) compiled from AIDA simulator. Root Mean Square Error (RMSE) was used for performance calculation. The obtained averages of RMSE are 6.43 mg/dL, 7.45 mg/dL, 8.13 mg/dL and 9.03 mg/dL for Prediction Horizon (PH) respectively 15 min, 30 min, 45 min. Further work, more case studies were suggested.

In [14], focused on type 1 diabetes patients; neural network models have been trained on glucose signals of a large and heterogeneous cohort of patients and then applied to infer future glucose-level values on a completely new patient.

In [3], the machine learning model was trained by patients of diabetes 2 data set. The Dataset comprises 23 features which is collected by the Egyptian National Research Center from diabetic patients. The results of the developed system show that the prediction accuracy is 84%.

In [15], type 2 diabetes melitus was predicted based on the Pima Indian Diabetes using the improved K-means algorithm and the logistic regression algorithm. The dataset comprises 8 features. The accuracy of the proposed model was 95.42%. Further work was suggested to develop an application based on the proposed model.

In [4], they have utilized machine learning technique in Pima Indian diabetes dataset to develop trends and detect patterns with risk factors using R data manipulation tool. To classify the patients into diabetic and non-diabetic they have developed and analyzed five different predictive models.

In [16], the OhioT1DM dataset, contains data from six patients with type 1 diabetes who participated in an IRB-approved study for eight weeks each, between March2016 and April 2017 have been used. They aplied LSTM and Neural Attention Models for Blood Glucose Prediction. The LSTM showed the best performance.

In [5], papers on diabetes prediction were reviewed. This review study summarized that taking into account several characteristics of the dataset, such as dimensionality, low number of instances compared to number of features or even the type of the dataset itself (genetic or clinical), can affect significantly the performance of the algorithm. Hence, an algorithm with the best performance in one dataset could easily have lower prediction accuracy compared to other algorithms in different datasets.

There are many other studies that are aimed to predict BGL by using different sets of features, applying mathematical and machine learning models. But none of the models has delivered 100% accuracy. Therefore, in this paper, we aim to achieve better accuracy or prediction performance through continousnous blood glucose monitoring data by using non-invasive basic health check up tests, dietary information, and socio-demographic characteristics.

Most of the previous machine learning-based researches in healthcare was conducted in developed countries [17]. However, the application of supervised machine learning in medical data to predict diseases, survivability of diseases, different types of health checkup test results by using sample data from Bangladesh is very little.

This study used machine learning (ML) approaches because clinical input data are not completely independent and complex interactions exist between them. Conventional statistical models have limitations to consider these complex interactions but ML can consider all possible interactions between input data. Machine learning prediction models can incorporate all the input variables with marginal effect and variables with unknown associations with the targeted outcome variable. Through machine learning prediction models, we incorporated both well-known risk factors of high blood glucose such as age, BMI, calorie intake, etc. and factors without clear associations to it as well [9]. Machine learning algorithms are used to identify patterns in datasets and to iteratively improve in performing this identification with additional data [18]. Machine learning algorithms have been extensively used in various domains such as advertisement, agriculture, banking, online shopping, insurance, finance, social media, travel, tourism, marketing, consumer behavior, and fraud detection. It is also used to analyze current and historical facts in order to make predictions about future events. In the healthcare field, machine learning is used in prevention, diagnosis and treatment phases [19]. Our model is used to predict future blood glucose value in fast, early, and easy way, helping users to increase awareness about diabetic.

In this paper we present an efficient web-based blood glucose prediction app. A web-based app was integrated with a machine learning prediction model based on a trained, ‘Boosted Decision Tree regression model’ to early diagnose the blood glucose level. Visual Studio was used to develop the web app. The web-based application was implemented based on the machine learning predictive model API and POST url. Automated web-besed blood glucose diagnosis sytem in this study aimed at aiding health communities especially in highly remote areas to early diagnose blood glucose in a fast, and easy way.

2.2 Sample

Data have been collected from the employees who work in the Grameen bank complex, Dhaka, Bangladesh. The Grameen Bank Complex holds 18 different institutions, such as Grameen Bank, Grameen Communications, other non-government organizations, and private companies, with more than 500 workers. The researchers have collected data from 271 employees (n = 271) to predict the blood glucose. For machine learning approaches we normally expect a big sample size. However, some studies used a small sample size, e. g 300 [20] and 118 [21]. It is to be mentioned here that sometimes small sample size is associated with higher classification accuracy [22].

Grameen Communications, Bangladesh, and Kyushu University, Japan, have jointly developed a human-assisted Portable Health Clinic (PHC) system [23]. A PHC is an eHealth system that aims to provide affordable primary healthcare services to prevent severity or to control non-communicable diseases (NCDs). A PHC system has four modules: (a) a set of medical devices, (b) a software system to collect and archive medical records, (c) healthcare workers to make the clinical measurements and explain ePrescriptions, and (d) ICT-trained call center doctors. Consumers come to the service point, and a health checkup is conducted by pre-trained healthcare workers. If needed, the consumer is connected to the call center doctors for a consultancy. The clinical measurements addressed by a PHC are as follows: (1) blood pressure, (2) pulse rate, (3) body temperature, (4) oxygenation of blood (SpO2), (5) arrhythmia, (6) body mass index (BMI), (7) waist, hip, and W/H ratio, (8) blood glucose, (9) blood cholesterol, (10) blood hemoglobin, (11) blood uric acid, (12) blood grouping, (13) urinary sugar, and (14) urinary protein.

The test items included (except arrhythmia, blood cholesterol, blood hemoglobin, blood grouping, urinary sugar, and urinary protein because there were many missing cases in these measurements) in this PHC system were used as input factors in this study except for the blood glucose measurement which is set as an output factor.

2.3 Measurements

Clinical measurements are obtained through direct diagnosis using PHC instruments operated by well-trained nurses or healthcare professionals. Data on dietary information and socio-demographic characteristics were collected during interviews by using a standard questionnaire.

2.4 Regression predictive modeling

As the targeted output variable of this study is a continuous variable, the regression predictive model will be applied and our objective is to predict the value of the blood glucose of an individual. Because a regression predictive model predicts a quantity, the skill of the model must be reported as an error in those predictions. There are many evaluation criteria to estimate the performance of a regression predictive model, but the most common is to calculate the Root Mean Squared Error (RMSE).

This study evaluated the most commonly used machine learning models, specifically, Boosted Decision Tree Regression, Decision Forest Regression, Bayesian Linear Regression, and Linear Regression. These models were chosen for comparison in this study due to their popularity in the medical data prediction.

This study has chosen these algorithms to see if the prediction accuracy can be further improved. Details of each models are described below:

2.4.1 Linear regression

Linear Regression is a very simple machine learning method in which each data point consists of a pair of vectors: the input vector and the output vector. It is the simplest, oldest, and most commonly used correlational method. This method fits a straight line to a set of data points using a series of coefficients multiplied to each input, like a weighting function, and an intercept. The weights are decided within the linear regression function in a way to minimize the mean error. These weight coefficients multiplied by the respective inputs, plus an intercept, give a general function for the outcome, uric acid measurement. Thus, linear regression is easy to understand and quick to implement, even on larger datasets. The disadvantage of this method is that it is inherently linear and does not always fit real-world data [24]. The LRM has the fol- lowing form

$$U_{pred} = \beta^{t} . x_{in}$$

where β represents the vector of coefficients, which are cal- culated by applying the least-squares method [25].

2.4.2 Boosted Decision Tree Regression

Boosting is a popular machine learning ensemble method [26]. Boosting means that each tree is dependent on prior trees. The algorithm learns by fitting the residual of the trees that preceded it. Thus, boosting in a decision tree ensemble tends to improve accuracy with some small risk of less coverage. In Azure Machine Learning, boosted decision trees use an efficient implementation of the MART gradient boosting algorithm. Gradient boosting is a machine learning technique for regression problems. It builds each regression tree in a step-wise fashion, using a predefined loss function to measure the error in each step and correct for it in the next. Thus the prediction model is actually an ensemble of weaker prediction models. In regression problems, boosting builds a series of trees in a step-wise fashion, and then selects the optimal tree using an arbitrary differentiable loss function [27]. Like Random Forest it uses many smaller, weaker models and brings them together into a final summed prediction. However, the idea of boosting is to add new models to the ensemble in a sequence for a number of sequences. In each iteration, a new weak model is trained with respect to the whole ensemble learned up to that new model. These new models, iteratively produced, are built to be maximally correlated with the negative gradient of the loss function that is also associated with the ensemble as a whole. In this approach, a performance function is placed on the GBM in order to find the point at which adding more iterations becomes negligible in benefit, i.e. adding more simple models, in this case, Decision Trees no longer reduces the error by a significant margin. It is at this point that the ensemble sums all of the predictions into a final overall prediction [24]. The Boosted Trees Model is a type of additive model that makes predictions by combining decisions from a sequence of base models. More formally we can write this class of models as:

g(x) = f0(x) + f1(x) + f2(x)+.....

where the final classifier gg is the sum of simple base classifiers fifi. For boosted trees model, each base classifier is a simple decision tree. This broad technique of using multiple models to obtain better predictive performance is called model ensembling.

2.4.3 Neural Network

It is a widely used machine learning algorithm. The Neural Network is a network of connected neurons. The neurons cannot operate without other neurons, with whom they are connected. Usually, they are grouped in layers and process data in each layer and pass forward to the next layers. The last layer of neurons is making decisions. The basic neural network, which is also known as multi-layer perceptron (MLP), is used for comparison with 1 hidden layer of 500 neurons, which is a reasonable number in neural network-based approaches [28].

2.4.4 Decision Forest Regression

This regression model consists of an ensemble of decision trees. A collection of trees constitutes a forest. Each tree in a regression decision forest outputs a Gaussian distribution as a prediction. Aggregation is performed over the ensemble of trees to find a Gaussian distribution closest to the combined distribution for all trees in the model [29]. This technique generates a number of decision trees during training which is allowed to split randomly from a seed point. This results in a “forest” of randomly generated decision trees whose outcomes are ensembled by the Random Forest Algorithm to predict more accurately than a single tree does alone. One problem with a single decision tree is overfitting, making the predictions seem very good on the training data, but unreliable in future predictions [24]. By using decision forest regression, we can train a model with a relatively small number of samples and get good results.

2.4.5 Bayesian Linear Regression

In recent years, Bayesian learning has been widely adopted and even proven to be more powerful than other machine learning techniques. Bayesian linear regression allows a fairly natural mechanism to survive insufficient data or poorly distributed data. It allows putting a prior on the coefficients and on the noise so that in the absence of data, the priors can take over. Bayesian linear regression provides us information about which parts of it fit confidently to the data, and which parts are very uncertain. The result of Bayesian linear regression is a distribution of possible model parameters based on the data and the prior. This allows us to quantify uncertainty about the model, if we have fewer data points, the posterior distribution will be more spread out.

Microsoft Azure Machine Learning Studio was used for implementation of these models. For evaluating the performance of the models, Root Mean Squared Error (RMSE) values from each model were used. The RMSE of a model is the average difference between the model’s prediction and the actual outcome [24].

Each model was trained on a 70% training sample to ensure each model were trained uniformly. We split data according to training set ratio = 0.7, test set ratio = 0.3. We did not use Cross-Validation method because K-fold Cross-Validation (CV) produces strongly biased performance estimates with small sample sizes [22].

The IPO (input-process-output) model for predicting blood glucose based on socio-demographic characteristics, dietary information and some basic health checkup test results is shown in Fig. 1.

2.5 Ethical approval

The authors obtained ethical approval from the National Research Ethics Committee (NREC) of the Bangladesh Medical Research Council with approval no. 18325022019.

3 Results

3.1 Description of the study population

Data from a total of 271 employees of Grameen bank complex were collected during the health checkup provided by PHC service. Descriptive statistics were used to describe the baseline characteristics of the participants. The descriptive statistics of participants are shown in Table1.

Table 1. Summary statistics of the selected continuous predictors used in machine learning.

Number	Variables	n	Minimum	Maximum	Mean	Std. Deviation
1	Age	271	34	77	49.61	7.39
2	Height (cm)	271	140.00	184.00	163.05	7.45
3	Weight (kg)	271	44.20	114.40	67.52	10.06
4	BMI (kg/m2)	271	18.39	40.53	25.37	3.20
5	Waist (cm)	271	63.60	118.00	90.24	7.80
6	Hip (cm)	271	80.00	127.00	94.54	6.29
7	Waist/Hip Ratio	271	0.64	1.11	0.96	0.06
8	Body Temperature (° F)	271	92.12	99.64	96.07	1.15
9	SpO2	271	93	99	97.67	1.17
10	Systolic BP (mmHg)	271	92	180	126.68	14.88
11	Diastolic BP (mmHg)	271	59	108	81.71	8.43
12	Pulse Rate (bpm)	271	51	123	80.27	11.66
13	Blood uric acid	271	3.10	11.00	6.63	1.54
14	Blood Glucose (mg/dl)	271	66.60	392.40	128.02	56.92

Table 2: Summary statistics of the selected categorical predictors used in machine learning.

Number	Categorical variables	Description	Categories/Levels	Frequency	%
1	*Gender*	Gender of the participant	Male = 1;	225	83.0
			female = 0	46	17.0
2	*Education*	Education completed by the participant	1 = No education (no school entered);	10	3.7
			2 = Primary school completed;	30	11.1
			3 = Secondary school completed;	11	4.1
			4 = High school completed;	23	8.5
			5 = Vocation school completed;	1	0.4
			6 = College/University completed;	63	23.2
			7 = Higher (Master or Doctor) completed	133	49.1
3	*Drinks*	Drinking sugar contained drinks (Coke, Fanta, Soda, Fruit Juice, other Sweet/Sugar contained drinks) three or more times a week	Yes=2;	26	9.6
			No=1	245	90.4
4	*Eating fast foods*	Eating fast foods such as Pizza, Hamburger, Deep Fried Foods (e.g. Singara, Samosa, Moglai Parata, etc.) three or more time a week	Yes=2;	49	18.1
			No=1	222	81.9

The mean age of participants is 49.61, most of the participants are of aged 50 years. According to BMI most of the respondents have BMI =25.37 i.e., most of them are overweight. Because of the range of BMI defined by WHO from 25 – 29.9 as overweight. The blood glucose of most of the people is on the borderline (128.02 mg/dl whereas the normal range 140 mg mg/dl). Therefore, they need to check blood glucose regularly.

Table 2. Summary statistics of the selected categorical predictors used in machine learning.

83% of respondents are male and most of them have completed College/ University degree. Among 271 respondents 9.6 % reported that they drink sugar contained drinks (Coke, Fanta, Soda, Fruit Juice) three or more time a week and 18.1% reported that they eat fast foods (Pizza, Hamburger, and Deep Fried Foods) three or more time a week.

3.2 Prediction performance assessment

To examine the prediction performance of the regression predictive technique by using Ml, the main evaluation criteria used, the root mean squared error (RMSE). The results are shown in Table 3. The Boosted Decision Tree Regression model showed the best performance among other models.

Table 3. Comparison of modeling techniques ranked from best to worst based on RMSE.

Model name	Root Mean Squared Error (RMSE)	Mean absolute error (MAE)	Coefficient of determination (R^2)
Neural Network	45.54	36.24	0.18
Decision Forest regression	22.99	17.46	0.79
Linear Regression	44.66	35.59	0.21
Boosted Decision Tree Regression	2.30	1.30	0.99
Bayesian Linear Regression	44.49	34.50	0.22

Note: The Mean Absolute Error (or MAE) is the sum of the absolute differences between predictions and actual values. On the other hand, Root Mean Squared Error (or RMSE) measures the average magnitude of the error by taking the square root of the average of squared differences between prediction and actual observation. That means, it indicates how close the predicted value is to the actual vale. There isn't a cutoff or bench mark in RMSE value. The smaller the value, better the prediction.

The Boosted Decision Tree Regression is the best predictive model in terms of RMSE in comparison to other models. The score model obtained by using The Boosted Decision Tree Regression model is shown in figure 2.

4 Discussion

Machine learning algorithms can identify the pattern in a dataset that may not be apparent directly. Thus, machine learning can provide useful information and support to the medical staff by identifying patterns that may not be readily apparent [30]. There are several advantages of choosing Machine learning algorithms over conventional statistical methods for designing a prediction model. 1) ML algorithm can handle noisy information 2) they can model complex, nonlinear, relationships between variables without prior knowledge of a model [31]. Therefore, it is possible to include all information from the dataset during the analysis [9]. Finally, ML can consider all potential interactions between input variables whereas conventional statistical analysis assumed that the input variables are independent [32], though, in real-world many input variables are inter-related in complex ways, whether these ways are known or not. Machine learning algorithms can be used to identify high-risk individual cases and can help medical staffs for clinical assessment [32].

Machine learning uses techniques that enable machines to use the experience to improve at tasks. Through machine learning, data feed into an algorithm or model, use this data to train and test a model. Then the model is deployed to do an automated rapid predictive task or to receive the predictions returned by the model. In many clinical studies**,** Gradient boosting machine learning algorithm has been successfully used to predict cardiovascular diseases [11]. The gradient boosting decision tree (GBDT) method by Friedman [33] predicted BMI with accuracy 0.91 [34]. In the current study boosted decision tree regression is found as the best predictive model followed by decision forest regression. Both of these are popular ensemble learning methods.

In this study, a prediction model was designed for improving blood glucose prediction by including not only well-known relevant factors of high blood glucose, such as age, BMI, but also factors with unknown associations with it. The noninvasive test items used in PHC service were used as input factors except for the blood glucose which is set as an output factor and except for blood uric acid. This study developed a mechanism to predict blood glucose with RMSE, 2.30, this RMSE is better than any reported in the literature. Results can provide useful insights for understanding the observed trend in population health and to inform future strategic decision-making for improved health outcomes. By using the prediction model designed by the machine learning approach to test individual blood glucose will save the cost and time of doctors and patients as well. Our model should be a powerfull tool to predict blood glucoes with limited medical resources.

The comparison of the results found from this study to those in previous related works is very important. Most of the previous studies reported performance measurements as a function of classification accuracy, which may not be directly compared to this study with a regression approach to build a predictive model for the continuous variable (blood glucose measure).

The current study developed a blood glucose prediction model by using machine learning approaches and including personal characteristics, dietary information and basic non-invasive clinical measurements. The data that we used in this paper was collected by using portable and cheap devices. Health records of 271 employees, age 34-77 years, male (83%), female (17%) have been collected. We

found that blood glucose can be predicted with 2.30 RMSE value. Among the five machine learning algorithms, Boosted Decision Tree Regression was found as the most effective one.

This is the first study of predicting laboratory test results of health measurements or health checkup items in Bangladesh. To our knowledge, to collect primary data from corporate employee through field survey and by using PHC system for health checkup measurements is the first time in Bangladesh. The current study developed a blood glucose prediction model based on personal characteristics, dietary information and some non-invasive basic health checkup test results. Machine learning Blood glucose prediction using only 17 noninvasive health measurements, dietary information, and social-demographic information is one of our contribution. Low-input dimension provides the advantage of keeping the necessary time to train the models relatively small. Ours is the first study to build model that predicts blood glucose using health checkup data from a developing country. This study empirically compared five state-of-art machine learning algorithms: Boosted Decision Tree Regression, Decision Forest Regression, Bayesian Linear Regression, and Linear Regression to predict blood glucose and identified the best performance method which showed the minimum RMSE value. This study predicts blood glucose based on basic health checkup data by semi experts with affordable, cheap, portable indicative devices and we found RMSE value was very low, 2.30, which is better than any reported in the literature and was smaller than the previous study [18]. This study also provided the web deployment to measure the blood glucose.

If we can determine blood glucose by using the developed ML prediction model, healthcare workers of PHC service do not need to carry the blood glucose measuring instruments. The findings can be helpful in achieving SDGs, Universal Health Coverage and thus reducing overall morbidity and mortality. By using the prediction model designed by the machine learning approaches to measure individual blood glucose will save the cost and time of doctors and patients as well. This prediction model can also be applied in other institutions.

This model could be used to predict blood glucose in situations with limited medical resources.

This study developed the blood glucose value prediction website (Figure 3). The link is as follows:

https://bloodglucoseprediction.azurewebsites.net/webform1

5 Conclusions

This study provides a measure in reducing NCDs and hence can be a good component in the concerned national or global plan. We developed a blood glucose prediction model based on personal characteristics, dietary information and some basic clinical measurements related to NCDs. Such a blood glucose prediction model is useful for improving awareness among high-risk subjects. The blood glucose prediction model can help to provide health services on early detection and cost-effective management of non-communicable diseases. Predicted glucose values can be used for early hypoglycemic or hyperglycemic alarms.

In future, we aim to apply our model and app to predict blood glucose of remote people in Bangladesh and other developing countries.

There are several limitations to this study. First, the number of samples we studied needs to be enlarged for training the prediction model in the future. Second, this study is limited to a particular area among a group of employees who work in a corporate office. Our prediction model is not confirmed on the data from other institutes. Although the framework achieves high performance on Grameen bank complex data, we believe this data-based strategy is also fit for predicting blood glucose on other institutional people. A future study could also include additional features (e.g. work stress, everyday physical activity, eating red meat). We conclude that this study served as a successful case to open discussions on further applications of this combined approach to wider regions and various health checkup measurements.

Abbreviations

NCDs: Non-communicable diseases

PHC: Portable Health Clinic

RMSE: Root Mean Squared Error

Declarations

Acknowledgements

This research work has been supported by multiple organizations. JSPS KAKENHI, Grant Number 18K11529, and the Future Earth Research Fund, Grant Number 18-161009264 jointly financed the core research. The Institute of Decision Science for a Sustainable Society (IDS3), Kyushu University, Japan, provided travel expenses for data collection, and Grameen Communications, Bangladesh, provided technical assistance.

Author contributions

M.B.S. conducted the experiments, concluded results and fndings, and wrote the whole manuscript. A.A. and M.R.H. assisted in data collection and provided guidance in data privacy. T.B. and M.N.H. assisted for concluding the results and commented on the findings. All the authors have reviewed the manuscript.

Data availability

The data and materials are available on request from M.B.S.

Conflicts of Interest

The authors declare that there is no conflict of interest.

References

1. Asad M, Qamar U, Zeb B, Khan A, Khan Y. Blood Glucose Level Prediction with Minimal Inputs Using Feedforward Neural Network for Diabetic Type 1 Patients. 2019 [cited 2020 May 27]; Available from: https://doi.org/10.1145/3318299.3318354
2. Zarkogianni K, Mitsis K, Litsa E, Arredondo MT, Fico G, Fioravanti A, et al. Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring. Med Biol Eng Comput [Internet]. 2015 Dec 1 [cited 2020 Jul 3];53(12):1333–43. Available from: https://link.springer.com/article/10.1007/s11517-015-1320-9
3. B SG, Sarkar S, Mitra P, Ghosh S. Early Predictive System for Diabetes Mellitus Disease. In: Perner P, editor. Springer International Publishing Switzerland; 2016. p. 420–7.
4. Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. Appl Comput Informatics [Internet]. 2019;(xxxx):0–5. Available from: https://doi.org/10.1016/j.aci.2018.12.004
5. Kavakiotis I, Tsave; O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J. 2017;15:104–16.
6. Zaman MM, Rahman MM, Rahman MR, Bhuiyan MR, Karim MN, Chowdhury MAJ. Prevalence of risk factors for non-communicable diseases in Bangladesh: Results from STEPS survey 2010. Indian J Public Health. 2016;60(1):17–25.
7. Sampa MB, Hossain MN, Hoque MR, Islam R, Yokota F, Nishikitani M, et al. Influence of Factors on the Adoption and Use of ICT-Based eHealth Technology by Urban Corporate People. J Serv Sci Manag. 2020;13(1):1–19.
8. Kim S, Chang Y, Yun KE, Jung HS, Lee SJ, Shin H, et al. Development of Nephrolithiasis in Asymptomatic Hyperuricemia: A Cohort Study. Am J Kidney Dis. 2017;70(2):173–81.
9. Lee S, Choe E, Park B. Exploration of Machine Learning for Hyperuricemia Prediction Models Based on Basic Health Checkup Tests. J Clin Med. 2019;8(172).
10. Hunter DJ., Reddy KS. Non-communicable diseases. N Engl J Med. 2013;369:1336–43.
11. Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevska O. Predictive analytics with gradient boosting in clinical medicine. Ann Transl Med. 2019;7(7).
12. Noohi NA, Ahmadzadeh M, Fardaei M. Medical Data Mining and Predictive Model for Colon Cancer Survivability. Int J Innov Res Eng Sci. 2013;2(2).
13. Ben Ali J, Hamdi T, Fnaiech N, Di Costanzo V, Fnaiech F, Ginoux JM. Continuous blood glucose level prediction of Type 1 Diabetes based on Artificial Neural Network. Biocybern Biomed Eng. 2018;38(4):828–40.
14. A Multi-Patient Data-Driven Approach to Blood Glucose Prediction. IEEE Access [Internet]. 2019 [cited 2020 May 27];7. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8723121

15. Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. Informatics Med Unlocked [Internet]. 2018;10(August 2017):100–7. Available from: https://doi.org/10.1016/j.imu.2017.12.006
16. IEEE Xplore Full-Text PDF: IEEE Xplore [Internet]. 2019 [cited 2020 May 27]; Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8856940
17. Gu D, Li J, Li X, Liang C. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. Int J Med Inform. 2017;98:22–32.
18. Lynch CM, Abdollahi B, Fuqua JD, de Carlo AR, Bartholomai JA, Balgemann RN, et al. Prediction of lung cancer patient survival via supervised machine learning classification techniques. Int J Med Inform. 2017;108:1–8.
19. Misawa D, Fukuyoshi J, Sengoku S. Cancer Prevention Using Machine Learning, Nudge Theory and Social Impact Bond. Int J Environ Res Public Health. 2020;17(3).
20. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform. 2017;97:120–7.
21. Zelič I, Kononenko I, Lavrač N, Vuga V. Induction of decision trees and Bayesian classification applied to diagnosis of sport injuries. J Med Syst. 1997;21(6):429–44.
22. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One. 2019;14(11).
23. Sampa MB, Hossain N, Hoque R, Islam R, Hossain MN, Hoque R, et al. A Framework of Longitudinal Study to Understand Determinants of Actual Use of the Portable Health Clinic System. In: Streitz N. KS, editor. HCI International [Internet]. Springer, Cham; 2019 [cited 2019 Nov 25]. p. 323–32. Available from: https://link.springer.com/chapter/10.1007/978-3-030-21935-2_24#citeas
24. Lynch CM, Abdollahi B, Fuqua JD, de Carlo AR, Bartholomai JA, Balgemann RN, et al. Prediction of lung cancer patient survival via supervised machine learning classification techniques. Int J Med Inform. 2017;108(August):1–8.
25. Zarkogianni K, Mitsis · K, Litsa · E, Arredondo M-T, Fico · G, Fioravanti · A, et al. Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring. Med Biol Eng Comput. 2015;53:1333–43.
26. Wu J, Roy J, Stewart WF. Prediction Modeling Using EHR Data. Med Care. 2010;48(6):S106–13.
27. Manna S, Biswas S, Barman S. A statistical approach to predict flight delay using gradient boosted decision tree. In: International Conference on Computational Intelligence in Data Science(ICCIDS). 2017.
28. Li X, Ding Q, Sun JQ. Remaining useful life estimation in prognostics using deep convolution neural networks. Reliab Eng Syst Saf. 2018;172(June 2017):1–11.
29. Criminisi A, Shotton J, Konukoglu E. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Found Trends Comput Graph Vis. 2012;7(2–3):81–227.

30. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: A comparison of three data mining methods. Artif Intell Med. 2005;34(2):113–27.
31. Perai AH, Moghaddam HN, Asadpour S, Bahrampour J, Mansoori G. A comparison of artificial neural networks with other statistical approaches for the prediction of true metabolizable energy of meat and bone meal. Poult Sci. 2010;89(7):1562–8.
32. Singal AG, Mukherjee A, Elmunzer J, DR Higgins P. Machine Learning Algorithms Outperform Conventional Regression Models in Predicting Development of Hepatocellular Carcinoma. Am J Gastroenterol. 2013;108(11):1723–30.
33. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29(5):1189–232.
34. Hu M, Nohara Y, Wakata Y, Ahmed A, Nakashima N, Nakamura M. Machine Learning Based Prediction of Non-communicable Diseases to Improving Intervention Program in Bangladesh. Eur J Biomed Informatics. 2018;14(4).

Figures

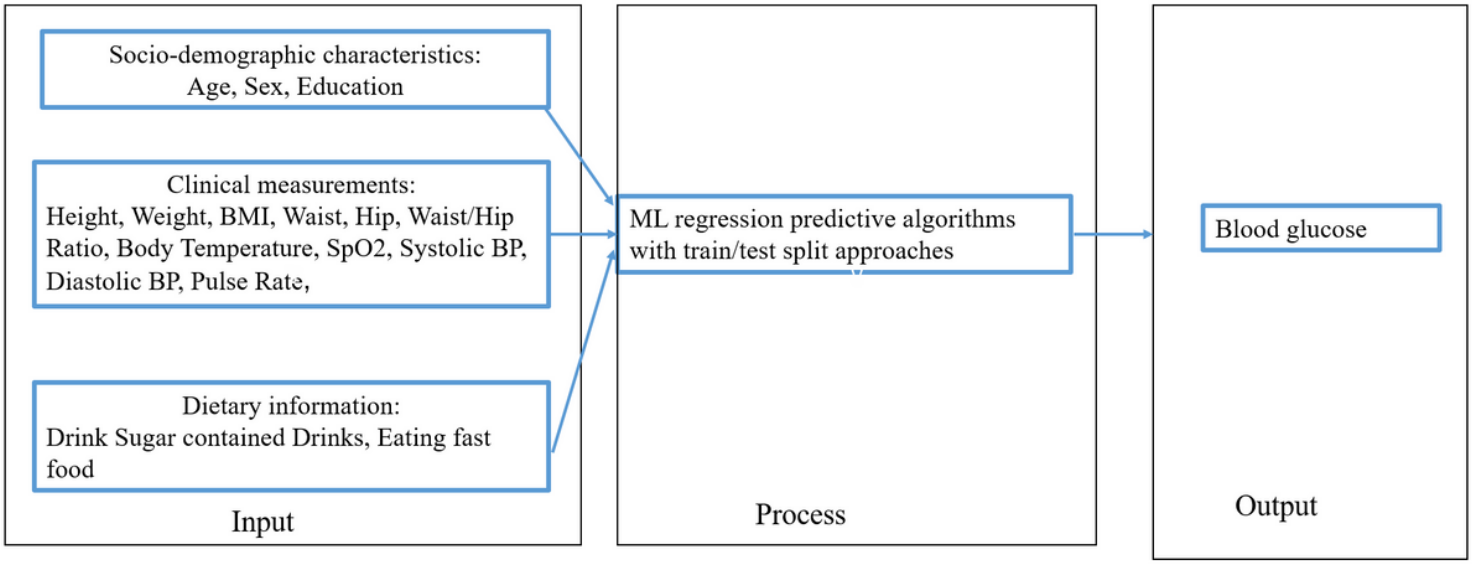

Figure 1

The input-process-output model used for predicting blood glucose after processing 17 input variables by machine-learning (ML) algorithms.

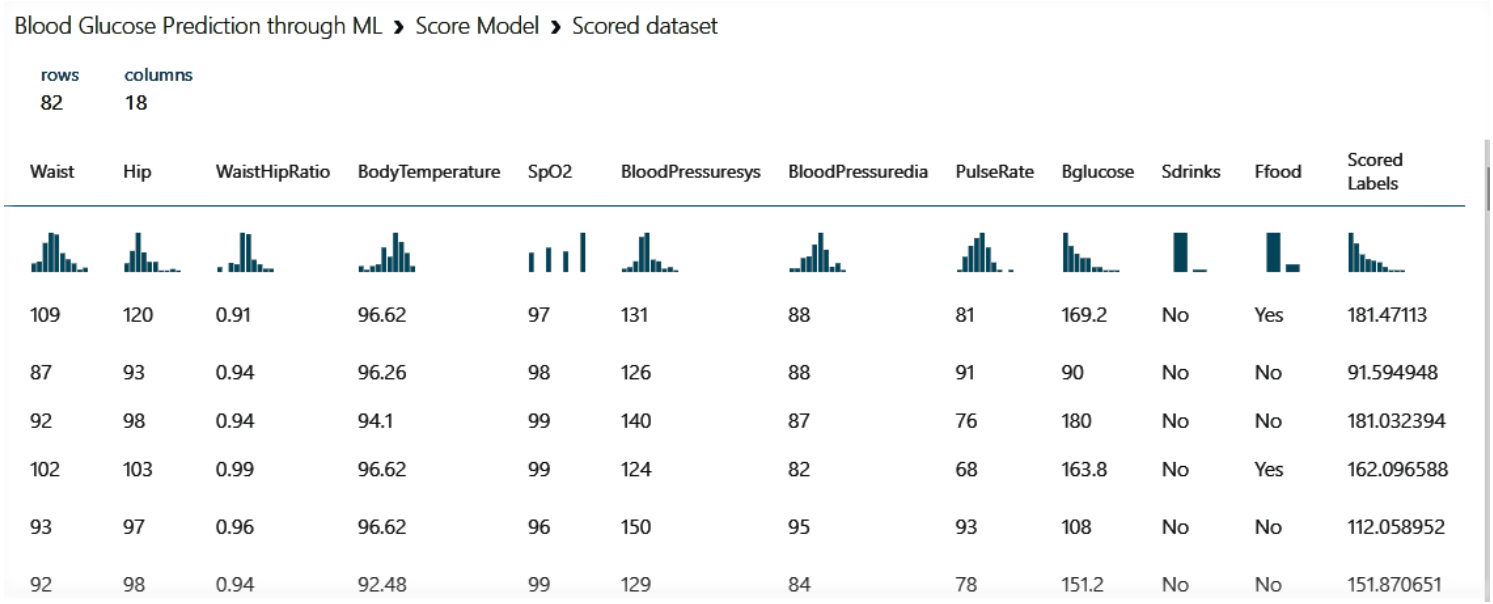

Figure 2

Partial view of the score model (prediction blood glucose) obtained by the Boosted Decision Tree Regression.

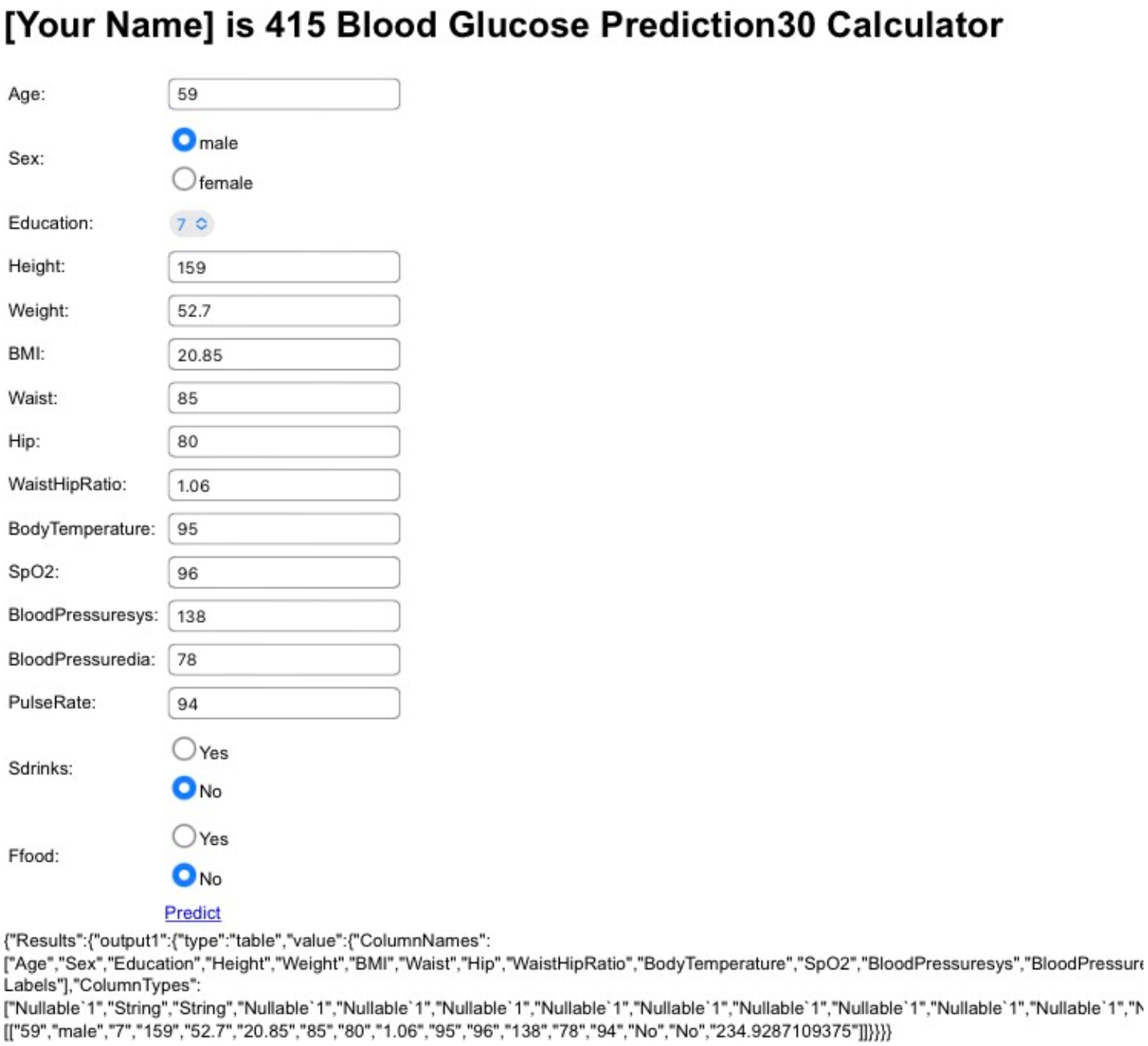

Figure 3

Screenshot view of the website (to predict blood glucose value) obtained by integrating the best performed Boosted Decision Tree Regression model in the application.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.pdf