Analysis on Diabetes Mellitus- Blood Glucose Prediction Using Machine Learning Algorithms

Radhika Gupta[1] and Dr. Maria Anu V [2,*]
[1]School of Computer Science And Engineering, VIT Chennai
[2] School of Computer Science And Engineering, VIT Chennai
E-mail address: radhika.gupta2022@vitstudent.ac.in
E-mail address: mariaanu.v@vit.ac

Abstract

Diabetes is a hazardous and serious condition that is caused due to abnormality of glucose level in blood. Nowadays, diabetes prediction at early stage is beneficial for the health care of mankind, so that it can save many lives on this earth. This crucial disease has become dangerous for human' mortality therefore it should be cured soon. Imbalanced diet, irregular exercise, obesity, genetic inheritance, blood pressure etc. are some causes that results in diabetes. Excess level sugar and carbohydrate in our diet also causes the risk of diabetes that may result too many other diseases such as heart diseases, eye disease, nerve disorder etc. So for the detection of the symptoms of this disease many new technologies, machine learning algorithms and data-sets are present in our health-care system, which are helpful for prediction of this sickness. This literature survey will help to evaluate the different machine learning algorithms and techniques such as Logistic Regression , Random Forest ,Support Vector Machine , Decision Tree that are used by different researchers. This paper will proceed the PIMA dataset which is fetched originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The comparison of the different machine learning models are further processed into boosting classifier algorithms like Ada-Boost, Gradient Boosting and Stacking Classifiers that results in the prediction of diabetes. Ada-Boost Classifier achieves higher accuracy of 82.47% in the prediction of diabetes.

Keywords: *Continuous glucose prediction, neural network, insulin, boosting, algorithms, machine learning, symptoms.*

Introduction

Diabetes is fatal condition which is caused due to increase level of glucose in our body, as nowadays it is becoming widespread metabolic sickness in the world .Glucose gives us energy to build our body and through energy body is able to perform all the regular functions .But if glucose level is disorder in our body then it harms our body in many aspects and can also cause to death. It is becoming dangerous situation for modern health care system, as it is mostly affecting the adult generation by more than 9%. Recent researches have predicted that this will soon increase by 56% in the coming decades and that will increase the fatality rate. This highlights the importance for the treatment of diabetes by new upcoming modern technologies.

Type1 and Type2 diabetes are determined by the amount of hormone called insulin produced by pancreas inside the body.

Type1: This happens when body is unable to produce hormone called insulin and that increases the level of glucose in body. Insulin is very essential for body as it permits glucose in blood to

enter our cells and regulate our body. Type1 spreads very rapidly in body as the cause of this diabetes is not known yet. And the prevention of this diabetes is only insulin injection as the proper medicine is not defined yet by doctors and scientist. If the proper dose of insulin injection is not received by body then that organ fails to operate and death is the result.
Type2: When pancreas is unable to make sufficient amount of insulin or the insulin produced didn’t work perfectly causes regularly rise in blood glucose. Type2 diabetes is complicated as it can cause damage to organs such as eyes heart. But by taking the proper medical cure at right time can reduce the risk of loss of organs .This can be caused by genetic age factor, heavy weight etc. Proper diet regular exercise are some factor that can reduce the risk of developing diabetes.
The cases of both type of diabetes are rising rapidly but Type2 is more complex as 90 percent of people are affected by it, whereas only 10 percent of population is suffering from Type1. Patients under the age 30 are suffering fromType1 diabetes and Type2 affects mostly people belonging to the age of middle years. BMI and age are key factors that are helpful in prediction of diabetes. In today’s world one third of the people are unaware of their diabetes and sometimes undetected diabetes becomes the reason of their fatality. Therefore diabetic patients should regularly examine their blood glucose level at-least twice in a year. Hence, advanced detection at early stage is beneficial for healthy and secure lifestyle of human beings.

Motivation

The purpose of choosing this domain is very important by seeing all the current scenario of diseases which are increasing vertically every day. People are facing the problem of diabetes and are not able to get proper guide and cure to rectify this diseases .So by selecting it and performing researches diabetic patients could get best result and boon by their health. The pain of diabetic’s people and their struggle towards food (glucose) is very sensitive that can attract everyone. Hence if researches are performed in positive manner then it can save many lives of many peoples and this is the main motivation for everyone to save life on this earth.

Literature Survey

There are many methodologies and techniques used by different researchers in the case of blood glucose level prediction. [1] Nikos Fazakis and his researchers worked on KDD (Knowledge discovery in database) process to evaluate and analysis the early danger of diabetes. They performed it by using SVM model in which they collected and analyzed all the data-sets features and all verification. They studied on Finnish Diabetes Risk Score for conducting their effective learning. After performing this analyses they resulted ensemble weighted voting model achieved 0.883 area under the curve .The voting method which is processed by researchers has worked significantly high in predicting the risk of disease. The limitation of this research is that it can perfectly evaluate the less patients, but it is inaccurate for large patients in future generation. [2] Egidio Gomes Filho and Placido Rogerio Pinheiro studied on outlined and hybrid strategy which enables diabetes detection on GDM. In this method disease is measured by Bayesian networks and the weight of disease is tabled by multi-attribute utility theory (MAUT). This study resulted many rules as the system’s accuracy is taken as domain. But the limitation is that it is not applicable for medical diagnosis and treatments. Therefore in future this model should be formed into single unit, so that it can be applicable for accepting remote areas and improve all the health-care networks. [3] Maryamsadat Shokrekhodaei in this survey paper promotes F1-Score Confusion Matrix that are accessed for the prediction of diabetic patients by different models .This method utilizes different wavelengths to increase reactivity of the glucose solution .They used F1 model and Clarke error model for obtaining their results .The result of this categorized by regression methods in which F1 obtained 98% while Clarke error accurate the result by 99.69% and both these results are involved into range category. [4] Namho Kim , Da Young Lee and researchers used real data-sets for their study by creating a clustering tailored model that can detect HbA1c diabetic monitoring information . Their proposed model gave 0. 88% and 0. 86% while calculating estimated models and CGM based HBA1c achieved the target. [5] Jivan Parab and Marlon Sequeira proposed their research paper on non-invasive monitor on blood glucose and blood urea in chronic kidney disease. In this paper researchers use the internal system to cure blood glucose patients and this technique provides sufficient results in detection of glucose level in body. ANN and Partial least Square model are being used in this research to overcome the disadvantage of accuracy. BP-ANN can provide result even if less data information is accessible Bland-Altman was estimated by PLSR and BP-ANN model by urea estimation method [6] Yusra Obeidat studied on artificial neural network that detected best accurate result of 5.68 in insulin pattern. In this research accuracy was developed by Tensor Flow data and even insulin pattern performed best but KNN is applicable for MSE with 6.39 .In future this research will proceed to inhalation device that can help in balancing the blood glucose level detection. [7]V.K.Daliya studied in research about the age, gender, BMI and other different blood serum measurements of the diabetic patient .It utilizes Optimized Multivariate Linear Regression method, by using the data-set with Root Mean Square Error. This provides accuracy result which is sharply associated with actual result of 15 units. This model doesn’t account any change in lifestyle and blood sample measurements as these samples are taking at regular time. [8] Nahla H. Barakat, Andrew P. Bradley, they proposed SVMs to advocate diabetic with extended Black box model for accurate detection. After detecting they believed that their research is valid and simple as their rules are beneficial in resulting. [9] Alessandro Aliberti investigated in his research paper about the prediction model, which is applied on different types of patients and then their glucose level values and records are kept separately to predict completely new patient in future. The paper works on LSTM and auto regression prediction model of neural network. NAR provides satisfactory performance for short-term prediction, while LSTM gave best result for every type of term prediction .Multi-patient method can be beneficial in future researches with real-time data for future generation. [10] Liyan Jia and Zhiping Wang suggested PE_DIM model that is applied on missing inequities and then it is extended to accuracy results in prediction of blood glucose .This uses Local Median-Based Gaussian Bayes which utilizes maximum use of data for recognizing missed values, but it produces many duplication of data PE_DIM produces satisfactory results for average rank data. This model is not suitable for large dimension of data in future.[11] John Daniels studied on different methodologies to examine the different prediction horizons levels of blood glucose .This paper researches on how different components affect the blood glucose prediction .It uses Support Vector Regression for prediction that gives higher accuracy results for hyperglycemia, when data is taken from last two weeks on an average .It provides accuracy results for small data-sets ,but it’s performance might be affected by large dimension of data-sets. [12] Hoda Nemat Heydar Khadem proposes three different deep learning and ensemble learning methods for prediction of blood glucose level between ranges of 30-60minutes .It takes use of Baseline Model Ensemble and Non-Ensemble Model in that Baseline provides satisfactory accuracy for both prediction horizons .And Ensemble model performs better recovery results then Non-Ensemble model. These models are developed by CGM data to enhance accuracy in the prediction of diabetes .[13] Jelena Tasic Gyorgy Eigner and Levente Kovacs studied on various new methods that uses different methodologies to upgrade blood glucose control by lowering chances of hyperglycemia or hypoglycemia in Type1 diabetic patients .This research uses Model prediction Control Approach and Bayesian Optimization Approach, in which Bayesian Optimization Approach successfully reduced the hypoglycemia range and also less amount of time is required in compare to MPC. This proposed technique can accurately interchange the data-sets to achieve enough glucose level without chance of hypoglycemia within small data-sets.[14]Arthur Bertachi and his other research partners proposed the points of blood glucose regularity in patients of Type1 diabetes . When they are unaware of unwanted exercise performed by them in their daily schedule. This uses Insulin feedback loop algorithm that alerts the patient to eat carbohydrate so that body can be prevented by the risk of hypoglycemia .This algorithm satisfies the target by detecting patients to consume correct amount of carbohydrates. And it also reduces the risk of hypoglycemia even if mixed food and unwanted exercise is performed by body. [15] Usama Ahmed and Ghassan F Issa researched on early prediction of diabetic patients by using fusion model and Machine learning algorithms. By dividing data-sets in two modules of intervals between 70 as training data-sets and 30 to testing data-sets. It utilizes Support Vector Machine that assess, where the diagnosis report is accurate or inaccurate while performing the prediction on patients .Some different classifiers have been used like Random Forest, Logistic Regression .The recommended fusion model preforms highest satisfactory result of 94.87, which is more effective than previously used models. These suggested model and data-sets can be stored on cloud storage for the benefit of upcoming future diabetic patients.

Methodology

(A) Diabetes Prediction

The primary goal of our model is to create a model which gives accurate result with high accuracy, which will be accomplished through use of different algorithms. A huge amount of data is generated in the medical industry, and it is very useful to use those data in early disease prediction.

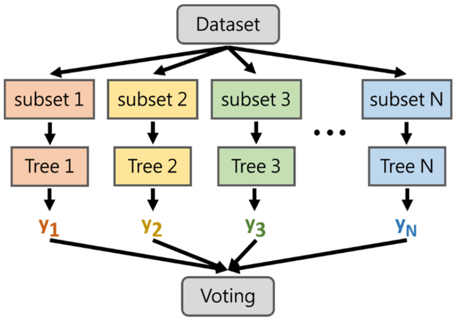
(B) Techniques for treating Diabetes

The attributes of data-set such as age, sex, cholesterol level, etc. are classified using Logistic Regression, K-Nearest Neighbors SVM, Naive Bayes and RF approach. The whole data-set is divided into both two section: testing and training in the amount of 70% and 30% of data respectively. Then performance and accuracy of trained model is evaluated using testing data-set.

(i) Random Forest (RF):

A Random Forest is a combination of many random forest algorithms which is termed as forest and later the forest is trained through bagging. More number of trees gives the highest accuracy of the algorithm. Its accuracy and performance is better than decision tree algorithm.

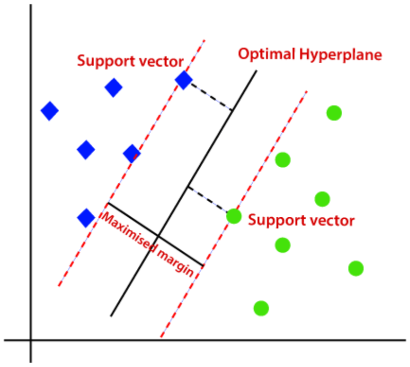
Fig 1: Random Forest



(ii) Support Vector Machine (SVM):

It is applied for solving both regression as well as classification problems. It creates hyper plane which is best decision boundary, which separates the data into classes. Since we are building our model in the medical data field, the data-set can be non-linear. So, Support Vector Machine can be a good option.

Fig 2: SVM



(i) Decision Tree :

Decision Tree is simple tree structure algorithm that is used to solve both regression as well classification but majorly uses for classification. In Decision Tree branches are decision process, leaf gives final output and tree structure‵s nodes describes features of the dataset. In this repeated method is used to form tree into subdivision of attributes until sample condition is satisfied.

(iv) Logistic Regression: It assumes a relationship between input variable and single output variable. It is use to find value based on other value. It is one of the powerful decision making algorithms that predicts the probability of the instant whether it is located in this class-interval or not.

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}},$$

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^i \log\, h_\theta\left(x^i\right) + \left(1 - y^i\right)\log\left(1 - h_\theta\left(x^i\right)\right)\right]$$

A. Boosting Techniques: Gradient Boosting Classifier: It combines all the weak model and then produce a decision tree of higher predicting model. It also takes use of statistical method to erase all the errors from the previously made model and make a new errorless prediction model. This is called so that it always remains close to gradient error and forecasting model.

Ada-Boost Classifier: It is one of the powerful ensemble methods of machine learning which is also called adaptive boosting. If boosting is applied then bias and variation are decreased. It takes ideas of old classifiers to give new results of classifiers and also loaded weight on the each data values. In this algorithms we use decision tree as a primary estimator to perform its output.

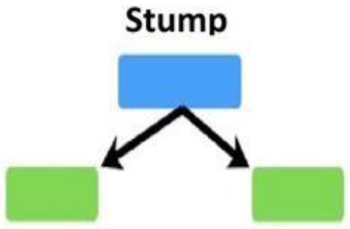
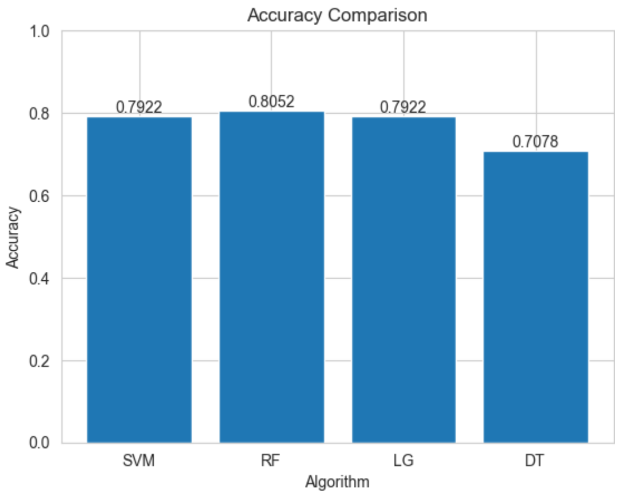


Fig 3: Ada-Boost Classifier

Stacking Classifier: It is an ensemble algorithms that increases the accuracy of performance by combing many nodes to form a new model. In this all the model that are used is equally weighted to form a predicting model. This model is also called stacked generalization as it has two or more models as base models.

Result Analysis



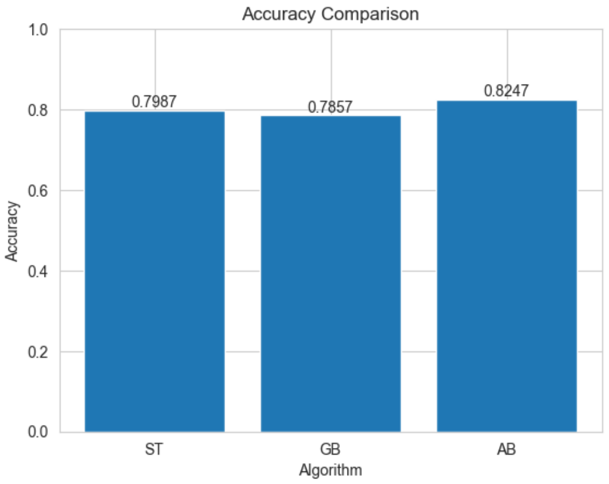


Fig 4: Comparison accuracy graph of algorithms

Fig 5: Comparison accuracy graph of boosting techniques

Conclusion

In this study, we highlighted the problem and challenges for identifying diabetes like how difficult and complicated for a doctor to identify diabetes and also as the prospective of patient, the test for identifying a diabetes is expensive. In all the above it is observed that the pre-processing of data is very important. It is also clearly visible that the most trusted data set for all the researchers are

PIMA Indian diabetes data-set. In this research, we proposed an automated system for predicting diabetes using machine learning algorithms. These attributes have been used to train and classify using ML algorithms like Logistic Regression, Support Vector Machine and Random Forest along with the boosting techniques that increases the final accuracy as 82.47%. The main disadvantage is that the dataset is not properly pre-processed and limited use of data availability. In future, we can develop a web application that overcomes the problem of longitudinal data, where people can identify if they have any diabetic patient or not by giving input to the model manually.

References:

[1] John Daniels; Pau Herrero; Pantelis Georgiou IEEE Journal of Biomedical and Health Informatics ear: 2022 | Volume: 26, Issue: 1 Cited by: Papers (1)

[2] Hoda Nemat; Heydar Khadem; Mohammad R. Eissa; Jackie Elliott; Mohammed Benaissa IEEE Journal of Biomedical and Health Informatics Year: 2022 | Volume: 26, Issue: 6 |

[3] Henock M. Deberneh and Intaek Kim . Prediction of Type 2 Diabetes Based on Machine Learning Algorithm Int. J. Environ. Res. Public Health 2021, 18, 3317. https://doi.org/10.3390/ijerph18063317

[4] (jsparab@unigoa.ac.in) Jivan Parab; Marlon Sequeira; Madhusudan Lanjewar; Caje Pinto; Gourish Naik IEEE Journal of Translational Engineering in Health and Medicine Year: 2021 | Volume: 9 Cited by: Papers (1)

[5] Usama Ahmed;Ghassan F. Issa; Muhammad Adnan Khan; Shabib Aftab;Muhammad Farhan Khan; Raed A. T. Said;Taher M. Ghazal Munir Ahmad IEEE Access Year: 2022 | Volume: 10 Cited by: Papers (2)

[6] Maryamsadat Shokrekhodaei; David P. Cistola; Robert C. Roberts; Stella Quinones IEEE Access Year: 2021 Cited by: Papers (6)

[7] Egidio Gomes Filho; Plácido Rogério Pinheiro; Mirian Caliope Dantas Pinheiro; Luciano Comin Nunes; Luiza Barcelos Gualberto Gomes IEEE Access Cited by: Papers (9)

[8] G. Geetha, K.Mohana Prasad, Prediction of Diabetics using Machine Learning International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-5, January 2020

[9] Daniels, John, Pau Herrero, and Pantelis Georgiou. "A Multitask Learning Approach to Personalized Blood Glucose Prediction." IEEE Journal of Biomedical and Health Informatics 26, no.1(2021): 436-445.

[10] Nemat, Hoda, Heydar Khadem, Mohammad R. Eissa, Jackie Elliott, and Mohammed Benaissa. "Blood Glucose Level Prediction: Advanced Deep-Ensemble Learning Approach." IEEE Journal of Biomedical and Health Informatics (2022).

[11] Tašić, Jelena, György Eigner, and Levente Kovács. "Review of Algorithms for Improving Control of Blood Glucose Levels." In 2020 IEEE 18th International Symposium on Intelligent Systems and Informatics (SISY), pp. 179-184. IEEE, 2020

[12] Beneyto, Aleix, Arthur Bertachi, Jorge Bondia, and Josep Vehi. "A new blood glucose control scheme for unannounced exercise in type 1 diabetic subjects." IEEE Transactions on Control Systems Technology 28, no. 2 (2018): 593-600.

[13] Ahmed, Usama, Ghassan F. Issa, Muhammad Adnan Khan, Shabib Aftab, Muhammad Farhan Khan, Raed AT Said, Taher M. Ghazal, and Munir Ahmad. "Prediction of diabetes empowered with fused machine learning." IEEE Access 10 (2022): 8529-8538.

[14] Aliberti, Alessandro, Irene Pupillo, Stefano Terna, Enrico Macii, Santa Di Cataldo, Edoardo Patti, and Andrea Acquaviva. "A multi-patient data-driven approach to blood glucose prediction." IEEE Access 7 (2019): 69311-69325.

[15] Jia, Liyan, Zhiping Wang, Siqi Lv, and Zhaohui Xu. "PE_DIM: An Efficient Probabilistic Ensemble Classification Algorithm for Diabetes Handling Class Imbalance Missing Values." IEEE Access 10 (2022): 107459-107476.