IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.8, Issue.3, Mar 2020 E-ISSN: 2347-2693

A Novel Prediction of Diabetes by Automatic Insulin Therapy Using Machine Learning Algorithm

B. Vinothkumar[1*], M. Ramaswami [2]

[1,2] Department of Computer Applications,Madurai Kamaraj University, Madurai,Tamilnadu

DOI: https://doi.org/10.26438/ijcse/v8i3.1823 | Available online at: www.ijcseonline.org

Received: 19/Feb/2020, Accepted: 07/Mar/2020, Published: 30/Mar/2020

***Abstract*—** Diabetes mellitus is one of the world's fast-growing diseases. Differentiation is among the most important decision-making approaches in many real-world problems. In this work, the main objective is to classify the diabetic patient's data into various levels based upon the values. This will help to assist the required dose which should be provided to the patients through an automatic insulin pump. The efficiency of the different classifiers is measured to assess the reliability of the classification. In this analysis, four common algorithms for machine learning, namely Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression, Random forest, and decision tree, for the estimation of diabetic mellitus on data from the adult population. Based on the comparison of performance parameters like precision, recall, F1-score, and accuracy the algorithms are ranked and selected the best among all. The accuracy value of Logistic Regression is the highest among the other algorithm, therefore Logistic Regression performs best with the patient data in forecasting diabetes mellitus.

***Keywords*—** Diabetes mellitus, Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression, Random forest, decision tree

I. INTRODUCTION

Diabetes is one of the world's common diseases which is growing rapidly. It is a significant health problem in most nations. The estimated number of diabetes patients was found at 171 million in 2000. Estimates will double the number by 2030[1]. The National Health and Nutrition Examination Survey organization had taken the survey from 1999–2000 and estimated that 18–20% of a sample population over the age of 65 had diabetes mellitus (DM) and 40% had either DM or early disease [2].

If a person's glycemic state was predicted in time slot earlier, and if he were alerted for any impending hypo/hyperglycemia levels, he could take measures for the prevention to avert additional difficulties. Therefore, the forecasting of glucose levels is extremely essential. The accuracy of prediction is affected by the presence of various noise components in the CGM sensor data and in the lack of adaptive, personalized methods of real-time tuning in the algorithms for prediction. Despite the wide range of work that has been done in predictive monitoring by various research groups such as Pappada et al (2008) [3], Cobelli et al (2009) [4], and PerezGandia et al (2010) [5], many challenges are still there in denoising of errors in CGM signal, uncertainties resulted in inter-patient and intra-patient variation impacting the accuracy of plasma glucose prediction.

Machine learning approaches produce effective results by creating predictive models from observational medical databases obtained from diabetic patients to obtain information. Extracting information from such data might be helpful in forecasting patients with diabetes. Different machine learning methods are capable of predicting mellitus diabetes. Nonetheless, choosing the best strategy to forecast based on those characteristics is very difficult. We use four common machine learning algorithms for the study, namely Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression, Random Forest and Decision Tree, to predict diabetic mellitus in adult's population results.

II. Continuous Glucose Monitoring (CGM)

Continuous Glucose Monitoring (CGM) systems continue to increase worldwide penetration, efficiency, and usability characteristics, and the correlation between real-time use of CGM and improved results keeps growing. The method is not yet generally known but there is an abundance of evidence supporting its use. The data available by CGM can allow for far more fine-tuned changes in insulin dosing and other treatments than spot-checking from self-monitoring of blood glucose (SMBG) can also provide. CGM systems for automated data collection have sparked interest in non-invasive glucose monitoring as an additional tool for collecting glucose level information.

Table 1 lists out the currently available continuous glucose monitoring devices with alarm systems, which have been released from the Food and Drug Administration (FDA) department, USA. with their approval.

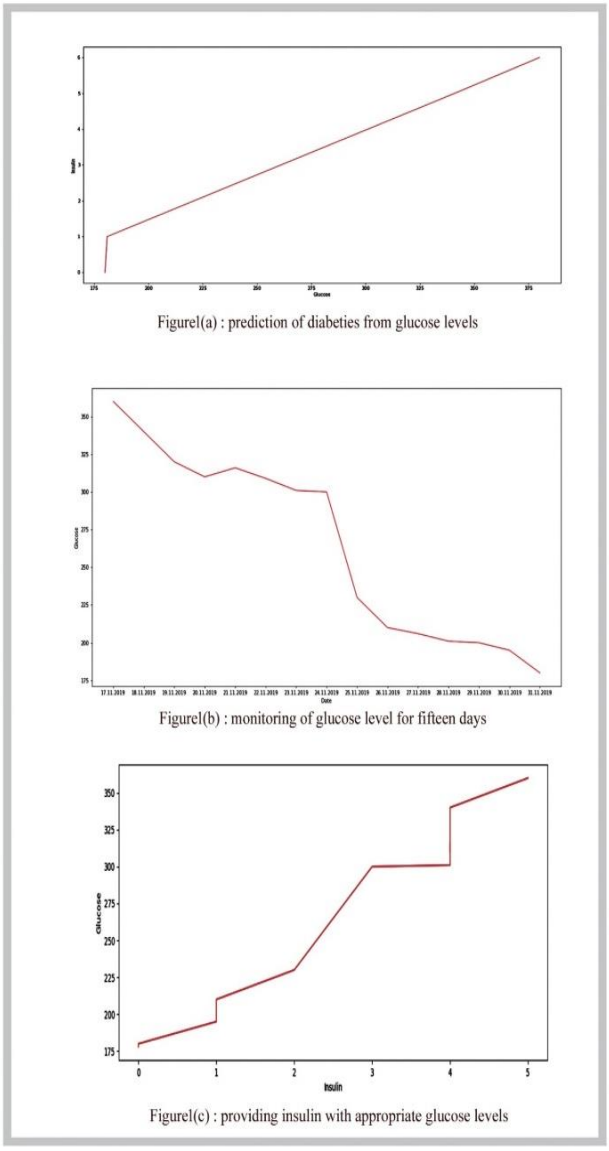
Table 1: Commercial CGM system

Names of the CGMS	Alert Generation
Medtronic CGMS Gold	***45Minutes Predictive alert (50% Accuracy)***

Guardian Real Time	***Threshold alert***
Gluco Watch Bio grapher	***30Minutes Predictive alert (24% Accuracy)***
Dexcom Seven	***Threshold alert***
Free Style Navigator	***Threshold alert***

III. TIME SERIES ANALYSIS

The study of the time series involves techniques for analyzing data from time series to derive meaningful statistics and other data characteristics. Thus, glucose prediction based on CGM tests enables the patient to make clinical choices based on expected potential glucose levels rather than the current levels, thereby reducing the risk of hypo-and hyper-glycemic incidents. CGM data was first studied as a version of the time series using repeated test periods of ten minutes.



Figurel(a) : prediction of diabeties from glucose levels

Figurel(b) : monitoring of glucose level for fifteen days

Figurel(c) : providing insulin with appropriate glucose levels

Figure 1 : Time Series Analysis for Diabeties

The details were collected by tracking fifty patients in free-living conditions for 15 days. For every five minutes, the average glucose value was recorded by the CGM system. The observing time was split into four-time intervals of landmarks. Glucose and insulin levels are directly proportional to one another.

That is, Glucose level $\propto$ insulin value

Because insulin is directly proportional to glucose, we should use different insulin amounts dependent on glucose levels. By the process of consuming insulin regularly depending upon the glucose value, diabetes can be controlled. The above time series analysis is prepared with the real-world dataset for diabetes.

IV. MACHINE LEARNING TECHMIQUES

Naïve Bayesian

Sorting of the algorithm [6, 7], a probabilistic classifier relying on Bayes' theorem with the predictors assuming equality. Naïve Bayesian process takes the database as input, conducts class mark analysis and forecasts using the Bayer's theorem. It tests the probability of the class of input data and helps to forecast the uncertain type of sample data. It is an efficient classification method, ideally suited for large datasets. The version of the Bayes Theorem measures the latter likelihood for each class using the formula below.

$$P(c/x)=\frac{P(x/c)P(c)}{P(x)}$$

$$PP(c/x) = P(x1/c) \text{ x } P(x2/c)\text{x } \ldots \text{x} P(xn/c)\text{xP(c)}$$

P(c|x) is the posterior probability of class (target) given predictor (attribute).

- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

Decision tree

It is a supervised method of learning which is used to solve problems of classification. The decision tree [8, 9] is a technique that splits the given dataset iteratively into two or more sample data. The method's purpose is to forecast the target variable's class value. The decision tree can help separate the data set and construct the model of judgment to determine the uncertain class labels A decision tree with binary as well as continuous variables can be constructed. The decision tree determines the root node optimally, depending on the maximum entropy value. It gives an advantage to the decision tree in selecting the most appropriate theory among the training data set. Issues faced during the development of a decision model are the specification of the separating factor, divides, stop parameters, pruning, testing study, quality and quantity, split order, etc.

The architecture of the decision is a tree system, where the set of nodes is part of a structure. It involves nodes for judgment (split node with the condition) and nodes for the leaf choosing the correct attributes-root node to start the split is a challenging task among the various attributes in the package. The Node of Judgment may have 2 or more divisions. The first node, called the root node, begins. The model forecasts the greatest characteristic obtainable as the root node, or the greatest forecaster node from the set of nodes obtainable. Depending on the degree of child node impurity there are many options to choose the right characteristic to be as the root node. Quality Measures [10] are entropy, giniindex, and assignment error. These measures are taken and an estimate is made to select the best spilt for all properties.

Logistic Regression

In statistics, Logistic regression is a type of regression where the reliant variable is categorical, called binary reliant variable-that is, in which only two variables, "0" and "1" can be taken, Which are like pass / fail, win / lose, alive / dead or healthy / sick results. Logistic regression has been used in different fields, including machine learning, most fields of psychology, and social sciences. For eg, the Trauma and Injury Severity Score (TRISS), (TRISS), originally developed using logistic regression, which is commonly used to estimate the mortality of injured patients. Numerous other medical scales used to assess a patient's condition were developed using logistic regression. The methodology can also be used in engineering, particularly to predict the probability that a given operation, device or product may fail. It is also used in marketing techniques such as projecting the inclination of a consumer to buy a product or preventing a subscription. It can be used in economics to forecast the probability of a person choosing to be in the labor force, and a company transaction is about to estimate the possibility of a defaulting borrower on a mortgage. Conditional random fields, which apply logistic regression to sequential results, are used in the analysis of natural languages.

Support Vector Machine (SVM)

This is one of the most common methods suggested by J. Platt et al., for classification. A Support Vector Machine (SVM) is an excluded classifier, characterizing the data formally by splitting a hyper plane. SVM isolates entities in specified classes. It can also recognize and recognize cases that aren't evidence assisted. SVM does not take care to disperse the acquiring data of each class The one extension of this algorithm is to perform a regression analysis to create a linear function, and another extension is learning to rank elements to produce a classification for each element.

SVM is a group of linked, managed to learn methods used for sorting and regression in medical diagnosis [11, 12]. At the same time, SVM minimizes the error in analytical grouping and maximizes the geometric margin. The Maximum Margin Classifiers are called SVM. SVM is a general algorithm based on assured probability constraints of the philosophy of predictive learning, i.e. the so-called concept of formal risk minimization. SVMs can do non-linear sorting efficiently utilizing what is named the kernel trick, converting their input data into large-sized function spaces implicitly. The kernel trick allows the classifier to be constructed without the space of the function being clearly known.

Random Forest (RF)

RF algorithms use a method identified as bagging, which is used to resample data examples many times to make diverse training subsets from the training data. Decision trees are then formed from each training subset before trees ensembles are established. That tree then casts a unit vote for the results of an incoming class mark for the data case. RF is scalable and needs only minimal computing resources [13].

RF makes plenty of decision trees, which is quite different since the algorithm for decision tree. When the RF forecasts a new object based on certain characteristics, each tree in RF gives its classification outcome and ‘vote’ and then the forest's potential output will be the leading quantity of taxonomies. In the regression problem, the average outcome value for all decision trees is the RF output [14].

V. METHODOLOGY

CGM sensor real-time data can be gathered by processors and transmitted via MQTT protocol. MQTT is targeted for the exchange of messages between the networks. The collected data are preserved in MQTT's cloud server.

The patient history gathered could be retrieved on a Mobile / PC API. The glucose level of the patient can be periodically monitored through the dashboard to boost the patient's health status. The doctor or caretaker considers the patient's glucose level and the insulin is automatically administered to the patients with the aid of an insulin pump.

Through utilizing machine learning methods, the patients' different glucose levels are estimated and the medication is immediately administered to the patients when insulin treatment during an emergency and the alert is sent to their families and their physicians in emergency cases.

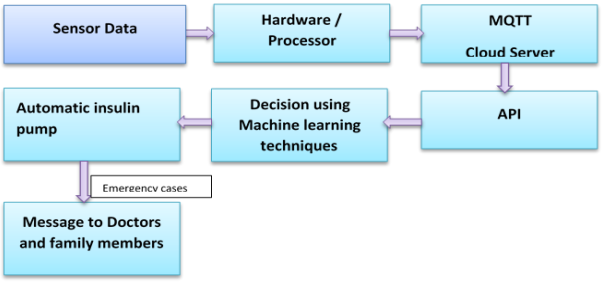


Figure 2. Frame work of Diabetes Prediction using various machine algorithms

VI. DATASET ATTRIBUTES

The attributes given have some impact on diabetes growth, so all of them were gathered from the dataset and used in the following steps for further cleaning.

Table 1: Data Set Attributes

Attribute	Description	Type
Pregnant	A record of the number of times the patient pregnant	Numeric
Plasma Glucose	Plasma glucose concentration measured using a two-hour oral glucose tolerance test (mm Hg)	Numeric
Diastolic BP	Diastolic blood pressure	Numeric
Triceps SFT	Triceps skin fold thickness (mm)	Numeric
Serum–Insulin	Two-hour serum insulin (muU/ml)	Numeric
BMI	Body mass index(weight Kg/height in $(mm)^2$	Numeric
DPF	Diabetes pedigree function	Numeric
Age	Age of the patient (years)	Numeric
Class	Diabetes on set within five years	Nominal

VII. PERFORMANCE MEASUREMENT

The standards of the produced findings were assessed in terms of the classification report by specific machine learning algorithms. The derived functions were categorized using the numerous machine learning algorithms to make diabetes prediction, including Naïve Bayesian, Decision tree, Logistic Regression, Help Vector Machine and Random Forest Algorithm. Table 2 demonstrates how the different algorithms used in this analysis are working.

Table 2: Evaluation parameters of Machine Learning Algorithm

Parameters	Naïve Bayesian	Decision Tree	Logistic Regression	Support Vector Machine	Random Forest
Precision	0.78	0.8	0.75	0.61	0.87
Recall	0.83	0.82	0.86	1	0.89
F1-core	0.79	0.81	0.86	0.77	0.91
Accuracy	0.75	0.79	0.81	0.64	0.8

The derived features were listed utilizing the numerous machine learning algorithms to forecast diabetes, including Naïve Bayesian, Decision Tree, Logistic Regression, SVM and Random Forest. To choose the high-performance algorithm, the values of all the test parameters of the different machine learning algorithms are compared.

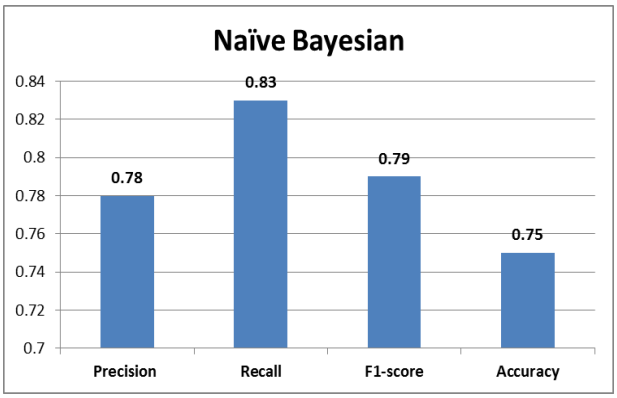


Figure 3: Performance measures of Naive Bayesian algorithm

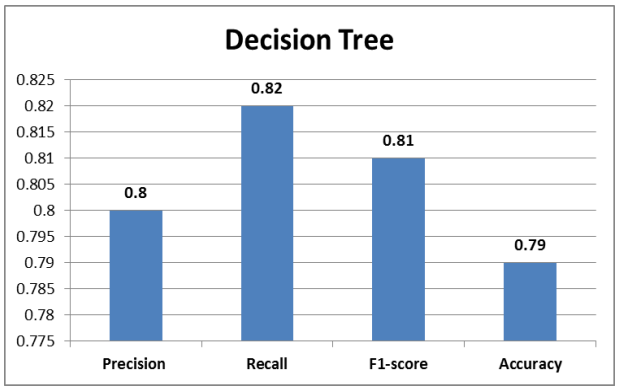


Figure 4: Performance measures of Decision Tree algorithm

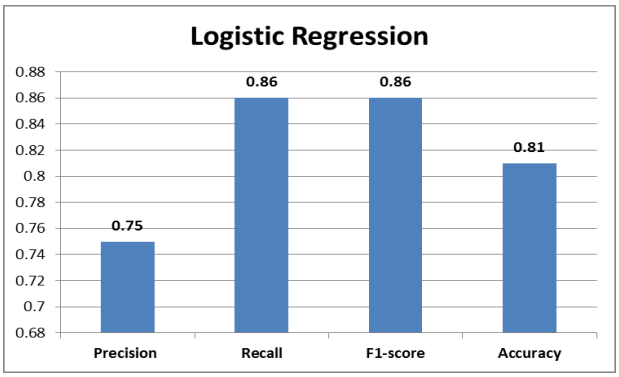


Figure 5: Performance measures of Logistic Regression algorithm

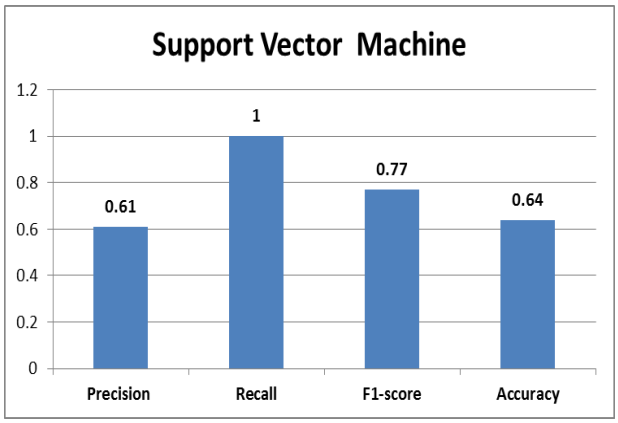


Figure 6: Performance measures of Support Vector Machine (SVM) algorithm

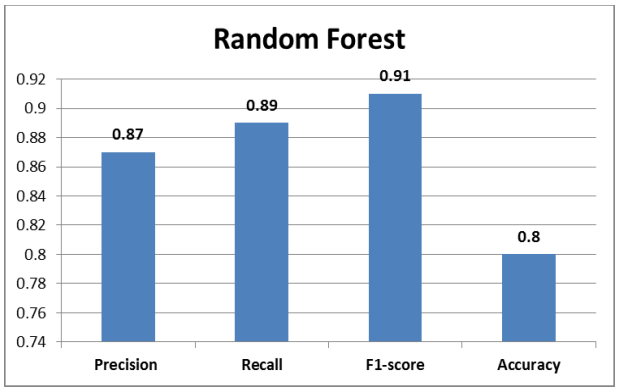


Figure 7: Performance measures of Random Forest algorithm

Figures 3, 4, 5, 6 and 7 display comparison of the machine learning algorithms based on classification consistency, accuracy, recall, and F1-score meaning and accuracy.

Automatic Insulin Pump

The pump configuration requires a minimum glucose level as well. In the casualty location, numerous patients by T1D aim for constricted glucose controller through an aim set at 80–100 mg/dL (4.4–5.5 mmol/L), that might be too small for the hospital situation [15] and [16]. No huge randomized measured trials have studied greatest glucose goal levels for hospitalized patients with T1D; though, an organized analysis of 19 studies (9 randomized and 10 observational) described that in surgical non-critically hostile hospitalized patients, the complete amount of toxicities be able to expressively cheap in protection of glucose attentions among 100 and 180 mg/dL (5.5–10 mmol/L) [17]. The ADA [18] and Endocrine Culture [19] approaches for the management of hyperglycemia in non-critically ill hospitalized patients have suggested that patients with T1D or T2D tend to target for fasting; BG < 140 mg / dL (7.8 mmol / L) and spontaneous glucose < 180 mg / dL (10 mmol / L) premeal. Since 2017, the ADA's Standards of Medical Care in Diabetes altered inpatients' aim glucose, endorsing stages among 140 and 180 mg/dL (7.8–10 mmol/L) for most ICU as well as non-ICU medical surgery patients with diabetes. Stricter thresholds below 140 mg / dL (7.8 mmol / L) may be perfect for definite circumstances, such as those with cardiac surgery, than severe ischemic cardiac actions or neurological actions, if the goals could be met without severe hypoglycemia [20].

Table 2: Insulin dose daily based on glucose values and nutritional intake

BG before meals	**Dose**
< 180 mg/dL (< 10 mmol/L)	No correction
181–220 mg/dL (10.1–12 mmol/L)	1 unit
221–260 mg/dL (12.1–14 mmol/L)	2 units
261–300 mg/dL (14.1–16 mmol/L)	3 units
301–340 mg/dL (16.1–18 mmol/L)	4 units
341–380 mg/dL (18.1–20 mmol/L)	5 units
>380 mg/dL (.20.1 mmol/L)	6 units, notify physician

Table 3: Insulin dose hourly based on glucose values and nutritional intake

Check BG Every hour	
BG ,100 mg/dL (5.5 mmol/L)	Hold basal infusion rate, check BG every 30 min
BG 101–140 mg/dL (5.6–7.7 mmol/L)	Decrease basal rate by 25%.
BG 141–180 mg/dL (7.8–10 mmol/L)	Maintain basal rate
BG 181–220 mg/dL (10.1–12.2 mmol/L)	Increase basal rate by 25%
BG .220 mg/dL (.12.2 mmol/L)	Increase basal rate by 25–50% and give 2–4 units as bolus insulin.

CGM systems can be aggressive (intravascular—venous and arterial), marginally invasive (intravenous), and non-invasive (transdermal). Glucose is restrained in interstitial fluid exhausting the glucose oxidase technique over fluorescence or else measured intravenously complete electrochemistry, fluorescence, mid-infrared spectroscopy, or electrochemical impedance spectroscopy [21]. All measured glucose levels are registered and stored in the data file, the measured value (MV) using the control algorithm is aligned with the setpoint (SP). The outcome is made based on the conditions, to determine the dosage necessary. The table is given the disparity in dose by respect to the dissimilarity in BG values.

VIII. RESULT AND DISCUSSION

Figures 2,3,4,5 and 6 display the accuracy recall, F1-score and precision results achieved in every algorithm for machine learning. Figure 7 displays the algorithm's global comparison.

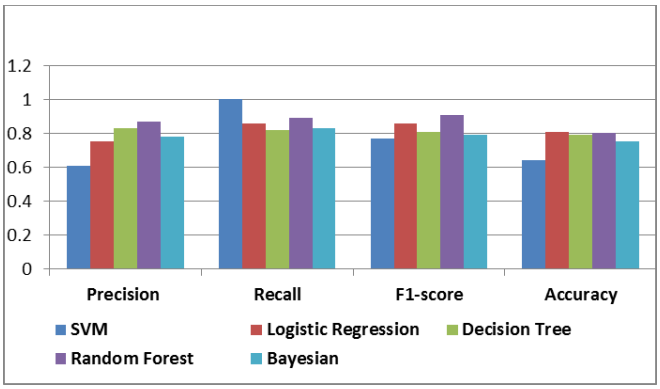


Figure 7: Performance measures of machine learning algorithms

By Fig. 7, It is apparent that Random Forest algorithm accuracy calculations are higher than other algorithms. Random forest algorithms yield high accuracy values compared with all other algorithms. The SVM yields low precision value. With regard to the recall, factor SVM provides greater value than other algorithms. For recall test SVM achieves higher value than any other algorithm. Random forest algorithm generates high value when it comes to F1-score ranking. Random forest algorithm results in a high F-measure value compared to the logistic regression algorithm.

The Logistic Regression Model's accuracy was found to be around 81 percent, likewise, the accuracy of Random Forest was found to be around 80 percent, the accuracy of the decision tree model was found to be 79 percent, the accuracy of Naïve Bayesian was found to be 75 percent and the accuracy was found to be 64 percent in the SVM model.

XI. CONCLUSION

Data mining aims at retrieving details from data preserved in the database and creating clear and understandable pattern explanations. Big Data Analysis in deployment is a comprehensive way to achieve better outcomes for all communities, such as the quality and affordability of health care services. Non-communicable diseases like diabetes, are one of India's major health threats. This work, therefore, helps with the discovery of the right machine learning algorithm for diabetes prediction. Use CGM, along with a good understanding of meal quality and form of a bolus, can also promote optimum use with CSII faster asparts. Currently, there is limited evidence about the therapeutic use of faster aspart with CSII, and further studies are needed to optimize its potential benefits in pump therapy. Ultimately, by evaluating the overall accuracy of the various machine learning algorithms, the Logistic Regression algorithm achieves greater accuracy than the Random Forest algorithm which is higher than the Decision Tree algorithm. Relative to all other machine learning algorithms for diabetic prediction, it is very clear from the result obtained that Logistic Regression performs much better in terms of all the different performance steps.

REFERENCES

[1] W. Rathmann, G. Giani, Global prevalence of diabetes: estimates for the year 2000 and projections for 2030, Diabetes Care 27 (10) (2004) 2568–2569

[2] The National Health and Nutrition Examination Survey (NHANES), ⟨http:// www.cdc.gov/nchs/nhanes.htm⟩, Last accessed December 2012.

[3] Pappada, Scott M., Brent D. Cameron, and Paul M. Rosman. "Development of a neural network for prediction of glucose concentration in type 1 diabetes patients." Journal of diabetes science and technology 2.5 (2008): 792-801.

[4] Cobelli, C., Dalla Man, C., Sparacino, G., Magni, L., De Nicolao, G., & Kovatchev, B. P. (2009). Diabetes: models, signals, and control. IEEE reviews in biomedical engineering, 2, 54-96.

[5] Pérez-Gandía, C., Facchinetti, A., Sparacino, G., Cobelli, C., Gómez, E. J., Rigla, M., ... & Hernando, M. E. (2010). Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. Diabetes technology & therapeutics, 12(1), 81-88.

[6] Ash C, Farrow JAE, Wallbanks S, Collins MD. Phylogenetic heterogeneity of the genus bacillus revealed by comparativeanalysis of small subunit ribosomal RNA sequences. Lett Appl Microbiol. 1991;13:202–6.

[7] Audic S, Claverie JM. The significance of digital gene expression profiles. Genome Res. 1997;7:986–95.

[8] Quinlan JR, Rivest RL. Inferring decision trees using the minimum description lenght principle. Inform Comput.1989;80(3):227–48.

[9] Agrawal R, Ghosh S, Imielinski T, Iyer B, Swami A. An interval classifier for database mining applications. 1992. pp.560–73.

[10] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Belmont: Wadsworth International Group; 1984.

[11] Cortes, C., Vapnik, V., "Support-vector networks", Machine Learning, 20(2),pp. 273-297, 1995.

[12] V. Vapnik, "The Nature of Statistical Learning Theory." NY: Springer- Verlag. 1995.

[13] Rodríguez-Rodríguez, I., Rodríguez, J. V., Chatzigiannakis, I., & Zamora Izquierdo, M. Á. (2019). On the Possibility of Predicting Glycaemia 'On the Fly'with Constrained IoT Devices in Type 1 Diabetes Mellitus Patients. Sensors, 19(20), 4538.

[14] Zou, Q., Qu, K., Ju, Y., Tang, H., Luo, Y., & Yin, D. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 9, 515.

[15] Mendez CE, Umpierrez GE. Management of type 1 diabetes in the hospital setting. Curr Diab Rep 2017;17:98

[16] Yogi-Morren D, Lansang MC. Management of patients with type 1 diabetes in the hospital. Curr Diab Rep 2014;14:458

[17] Murad MH, Coburn JA, Coto-Yglesias F, et al. Glycemic control in non-critically ill hospitalized patients:a systematic reviewandmeta-analysis. J Clin Endocrinol Metab 2012; 97:49–58.

[18] Moghissi ES, Korytkowski MT, DiNardo M, et al.; American Association of Clinical Endocrinologists; American Diabetes Association. American Association of Clinical Endocrinologists and American Diabetes Association consensus statement on inpatient glycemic control. Diabetes Care 2009; 32:1119–1131

[19] Umpierrez GE, Hellman R, Korytkowski MT, et al.; Endocrine Society. Management of hyperglycemia in hospitalized patients in non-critical care setting: an Endocrine Society clinical practice guideline. J Clin Endocrinol Metab 2012; 97: 16–38.

[20] Noschese ML, DiNardo MM, Donihi AC, et al. Patient outcomes after implementation of a protocol for inpatient insulin pump therapy. Endocr Pract 2009;15:415–424

[21] Adamson TL, Eusebio FA, Cook CB, LaBelle JT. The promise of electrochemical impedance spectroscopy as novel technology for the management of patients with diabetes mellitus. Analyst (Lond) 2012;137:4179–4187

