ORIGINAL PAPER

Machine learning models for non-invasive glucose measurement: towards diabetes management in smart healthcare

Harshita Agrawal[1] · Prateek Jain[2] · Amit M. Joshi[1] [ORCID]

Abstract

The patients of diabetes require to observe and control their glycemic profile through continuous glucose level monitoring. The blood glucose measurement is possible through invasive, minimally invasive and non-invasive methods. Invasive method is traditional method for instant glucose measurement where glucose is measured by taking blood samples from the body. However, the repeated finger pricking increases the risk of blood-related infections and trauma. Hence, the development of non-invasive real time device is essential for smart healthcare to manage glucose-insulin balance. The paper presents machine learning models for non-invasive glucose measurement. So, various machine learning algorithms including Logistic Regression, KNN, Gaussian Naive Bayes, Linear Regression, Multi-polynomial Regression, Neural Network, XGBoost, Decision Tree, Random Forest and Support Vector Machine are applied on two dataset which are PIDD (UCI repository) and iGLU dataset (iGLU device). The comparative analysis is carried out where accuracy, training time, recall, precision, f-1 score and AUC curve is measured for classification algorithms. For regression algorithms, measures like accuracy, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are used for comparison purpose. Random forest with 84% accuracy and 68% recall, 76% precision and 72% f1-score for PIDD and Decision tree with 70% accuracy, 8% mean absolute error (MAE) and 8.5% root mean square error (RMSE) for iGLU dataset gives best results. Clark grid analysis has also been done where all the values fall under zone A which gives 100% accuracy and the device is useful for medication purpose. The proposed work has been also compared with similar methods and the proposed work has excellent results in terms of MAD, mARD, RMSE and AvgE. The device would be ideal as non-invasive solution for continuous glucose monitoring.

Keywords Continuous glucose measurement · Classification · Regression · Smart healthcare

1 Introduction

The pervasiveness of diabetic patients has become double from 2010 globally [1]. Diabetes prevalence in the world was noted to be 9.3 % (463 million) in 2019 and it is expected to grow to 10.2 % (578 million) by the year 2030 and to 10.9 % (700 million) by the year 2045 [2]. Diabetes is a chronic disease which is caused when the pancreas stops producing insulin or the body is not able to utilize the insulin produced by the pancreas [3]. Insulin is the hormone which helps the body cells to absorb glucose from the blood. In diabetes, the glucose level in the blood increases. Diabetes can be divided in three parts mainly which includes type-1, type-2 and gestational diabetes [4]. Negligence in the treatment of diabetes can cause some serious health issues. It may cause strokes, nerve damages, heart diseases, kidney diseases and blindness. Diabetes could be administered by doing physical exercises, maintaining a proper diet and using proper dosage of insulin.

Intake of oral medications can also be helpful for controlling during early stages of diabetes. In most cases of adults, 5 % Type 1 diabetic patients have been considered almost in all diagnosed case while 90-95 % Type 2 diabetic patients have been considered for treatment. Hence, it is necessary to develop the device to measure blood glucose for quick and continual diagnosis of diabetes [5]. Diabetes requires continuous glucose level monitoring. If the existing invasive methods are used for this purpose, repeated finger pricking increases the risk of blood-related infections and

✉ Amit M. Joshi
amjoshi.ece@mnit.ac.in

[1] Malaviya National Institute of Technology, Jaipur, Rajasthan, India

[2] School of Electronics Engineering, VIT AP University, Amaravati, Andhra Pradesh, India

trauma [6]. Hence, development of non-invasive real time devices becomes very essential. There are very few such devices available in the commercial market right now and those which are available are very expensive [7, 8].

Coronavirus Disease 2019 (COVID-19) has been proclaimed as a pandemic by the World Health Organization (WHO) on January 30, 2020. At this time, 23 million people are infected by corona virus globally and the global mortality rate is 3.4 % [9]. Past viral pandemics have also seen the connection of diabetes to higher morbidness and death rate. During 2002-2003 upsurge of Severe Acute Respiratory Syndrome (SARS-CoV-1), diabetes was considered as an independent risk factor and demise. Likewise, in 2009 during outbreak of Influenza A (H1N1), risk of hospitalization was tripled and risk of emergency admission was quadrupled because of diabetes [10]. Diabetes was pervasive in about 50 % of population at the time of upsurge of Middle East Respiratory Syndrome Coronavirus (MERS-CoV) and in comparison to entire population, diabetic patients chances proportion of having basic or serious MERS-CoV ranged from 7.2 to 15.7. Death rate in patients, who had MERS with diabetes was 35 % [11]. However, higher risk of developing COVID-19 or a more severe form of this disease belongs to several risk groups specially having hypertension, chronic respiratory diseases, cardiovascular diseases (CVD), Diabetes Mellitus (DM) or Metabolic Syndrome (MS) seem to play vital role. Hence, continuous blood glucose monitoring becomes very essential for measurement of diabetes for COVID-19 patients.

As the population is increasing day by day and resources are limited, the requirement of implementation of smart cities is also increasing. Smart healthcare is a core component of a smart city [12]. Many information technologies for example, internet of things (IoT), cloud computing, big data, artificial intelligence are utilized by smart healthcare to increase the efficiency of healthcare and make it more user-friendly [13]. We need smart healthcare solution to detection the diabetes in non-invasive manner for rural areas and remote locations as instant medical facilities are not much available there [14]. Smart healthcare for diabetes is represented with Fig. 1.

The significant difficulties associated with the advancement of genuinely non-intrusive glucose observing innovations are identified with accuracy, ease of use and applicability for home use among an assortment of individuals. Only such a device that overcomes the mix of these difficulties may give a noteworthy achievement in this field and improve the personal satisfaction of a huge number of individuals with diabetes around the world [15]. It is very challenging as it obviates frequent visit to patient's bedside, particularly if the patient is serious and receiving intravenous insulin which is not safe due to corona. But if the patient is not critically ill, then we can give a non-invasive device which is easy to operate so that he/she able to self monitor the blood glucose for diabetes management [16]. The mHealth application would also allow the glucose reading can be sent to remotely located the doctor. Therefore the idea is to develop the continuous glucose monitoring device which is convenient, affordable, harmless and provides real time measurement. This can be done by using a non-invasive device and applying intelligent post processing algorithms on the data collected by this device and integrating it with a microcontroller programmed with best suited machine learning algorithm to give high accuracy and fast results.

In the past, many works have already been reported for the glucose measurement. They can be invasive, noninvasive, or minimally invasive. There has been several attempted for continuous glucose monitoring based on

Fig. 1 Smart healthcare for Diabetes
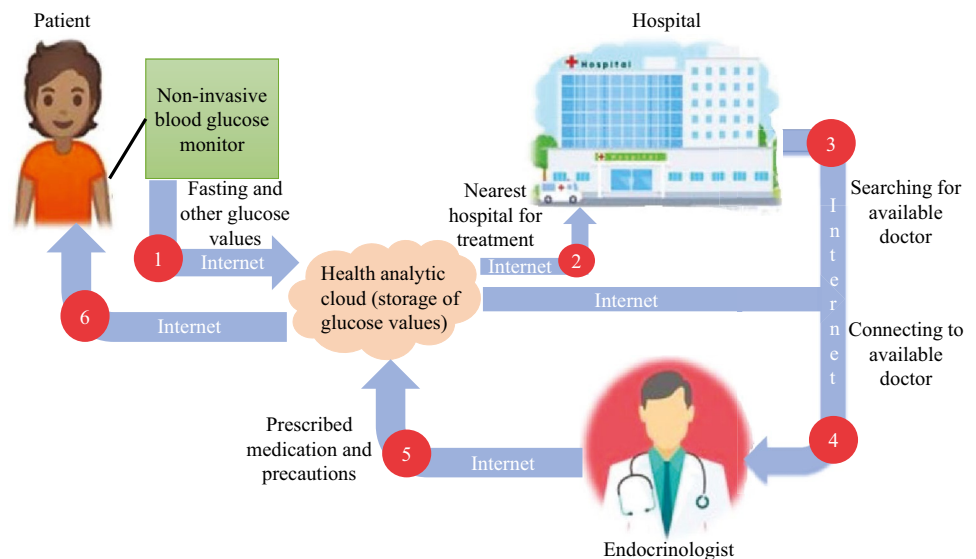
the non-invasive technique. They are technically based on optical and non-optical methods. Some of the optical techniques used methods based on Raman Spectroscopy, NIR spectroscopy, PPG method etc. Fig. 2 gives an overview of measurement techniques for blood glucose measurement systems. After the acquisition of data from sensors, many researchers concentrated to develop the optimized computing model to predict the glucose level precisely. In this way, Sejdinović et al. demonstrated the development of artificial neural network for classification of prediabetic and type 2 diadetic patients [17]. The testing of developed model has been performed using certain ratio of samples. Alić et al. presented the developed expert system for the classification of metabolic syndrome (MetS) [18]. A feed-forward artificial neural network (ANN) is presented for MetS classification. A different artificial neural network is also implemented for lactose intolerance prediction [19]. The physiological behaviors model is elaborated for glucose-insulin regulatory mechanism [20]. Some neural networks of different computing model are also introduced for other medical perspectives such as segregation of cancer and normal patients [21]. A lot

of simpler and sophisticated models have been introduced for computing apart. Still, it is required to have optimized model for fast and precise computations.

The report is organized in the following manner. Section 2 covers the literature review. The novel contribution of this work is represented in Sect. 3. Section 4 presents the proposed methodology which comprises of the process of data acquisition, data preprocessing, brief theory of all the machine learning algorithms that have been applied as classifiers or regressors and the performance evaluation is carried out. The proposed method has been compared with state of art work in the literature. Simulation results and comparative analysis of all the applied machine learning algorithms through tables, charts and graphs have been presented in Sect. 5.

PPG is one of the non-invasive measurement technique. The sensor is used which is similar as working principle of pulse oximeter and helps to record PPG signal. Paul et al. [12] developed PPG based blood glucose monitoring with help of pulse oximeter. The light was sent through transmitter and the prediction of glucose was made as per light
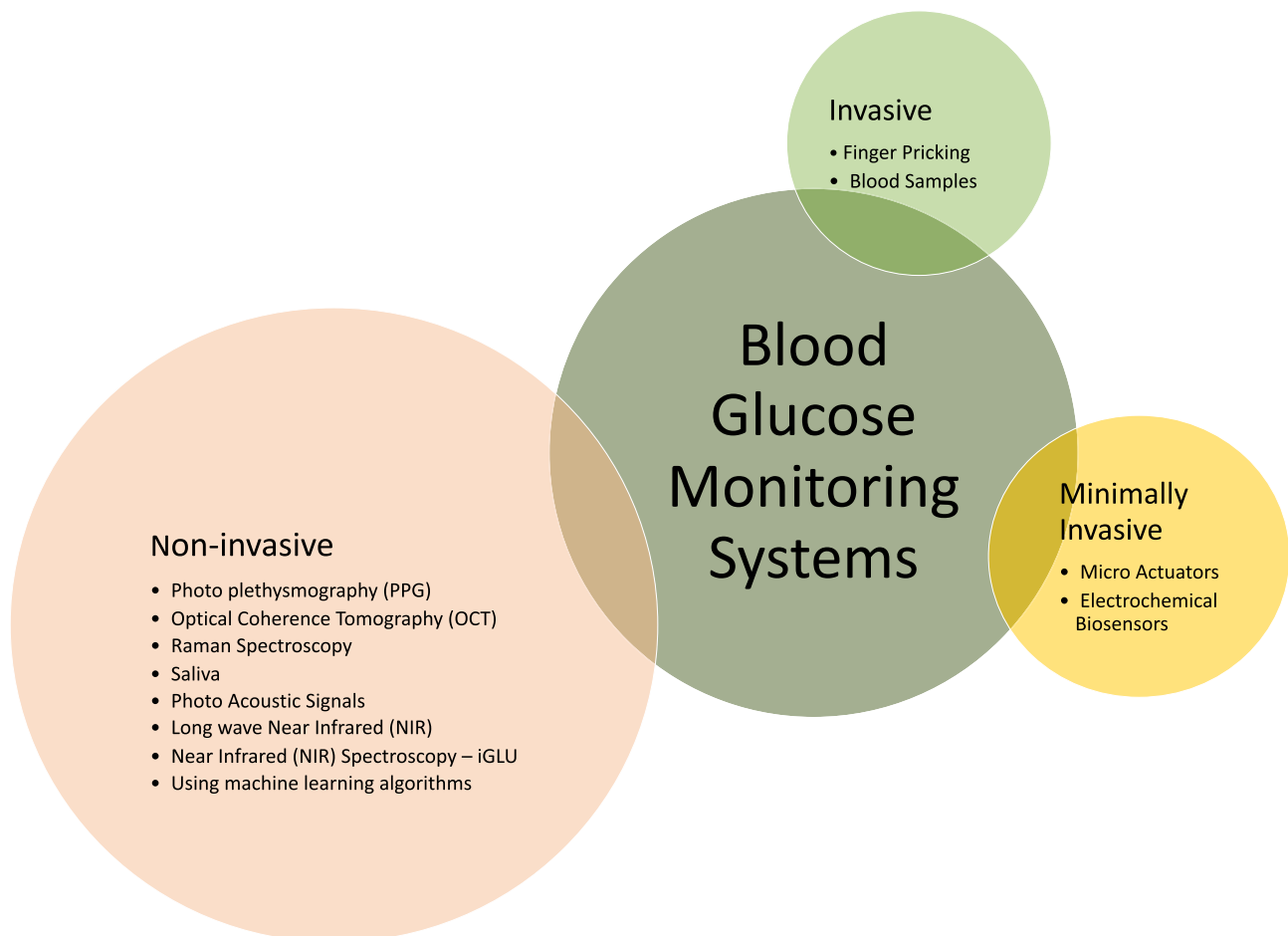

Fig. 2 An overview of Blood Glucose measurement techniques

intensity at the receiver. The change in voltage values was observed as per glucose concentration. In similar manner, Monte-Moreno [22] designed PPG based sensor to extract the information from PPG and the blood glucose was estimated using machine learning models. The continuous glucose monitoring was attempted by using wearable microsystem based minimally invasive approach [23]. It was first wearable glucose measurement device which was used to extract the glucose from the human skin. Optical Coherence Tomography (OCT) is non-invasive approach which helped to have glucose estimation as per OCT slope [24]. The non-invasive approach with Raman spectroscopy was explored which is based on chemical process and the interaction with molecules [25]. The glucose estimation using saliva was attempted using non-invasive method [26, 27]. Ramashyamam et al. [28] suggested NIR spectroscopy based glucose estimation using PPG. The specific wavelength of 935 nm,950 nm and 1070 nm was used for blood glucose prediction with FPGA using Artificial Neural Network. The blood glucose measurement was investigated with painless approach on micro-controller [29]. The diabetes management system with insulin pump was explored for better healthcare to maintain the glycemic profile. The pulsed laser diodes were introduced to collect the photo acoustics signals which are subsequently used for glucose estimation [30].

The intelligent Glucometer iGLU has been developed with optical method using machine learning models. The device was integrated with Internet of Medical Things (IoMT) framework to store the data for the remote monitoring purpose [14]. The device was based on NIR spectroscopy with three channel data was collected which subsequently processed through regression models for the glucose measurement. There are several solution have been developed till date but most of them suffer from accuracy. The PPG approach was used to measure the light intensity variation as per blood volume. In PPG, the light intensity may vary according to blood volume change in the body hence it may not provide accurate glucose value. The wearable microstrip solution is larger in size so it may not convenient for continuous glucose measurement. The OCT based technique would take considerable amount of time for glucose concentration estimation and also results in low specificity and low sensitivity. The Raman spectroscopy based solution requires larger space and portability would be always the issue. The saliva based glucose detection may not be

much advisable because saliva sample always vary among the people. LASER based solution may not applicable for frequent glucose measurement. Therefore, the short NIR spectroscopy is considered as the best approach for continuous glucose measurement and it also mitigate all the previous drawbacks. The general flow of the model is represented in Fig. 3.

2 Research challenges and novel contribution

The noninvasive measurement of blood glucose helps to have the continuous glucose measurement in smart healthcare system. The accurate value of glucose prediction is really the challenging task in non-invasive measurement. Presently, the available solutions are costly and requires complex mathematical models in order to process for the instant diagnosis. The following research challenges are being addressed in the present manuscript: (1) The best non-invasive solution is defined for the precise glucose measurement. (2) The efficient machine learning models are used for the measurement of blood glucose. (3) The smart healthcare solution using IoMT framework to have continuous glucose monitoring. (4) The cost effective solution is required to be developed for all types of patients (diabetic, pre-diabetic and healthy).

The main Contribution of the paper is as follows:

1. The non-invasive measurement of blood glucose is carried out on two diabetes data-sets which are PIDD (Pima Indian Diabetes Data-set) and data-set collected from a intelligent glucometer device iGLU.
2. The machine learning based regression approaches are implemented to get the accurate predicted blood glucose measurement value.
3. Clarke Error Grid analysis has also been conducted on iGLU data-set to validate the feasibility of proposed solution with the outcome of 100% values lie in the A-B zone which are clinically accepted.
4. The proposed machine learning model is also compared with other related work on the basis of various performance evaluation criteria such as MAD, mRAD, RMSE, AvgE to find the most suitable and accurate algorithm for diabetes detection and prediction of glucose level.

Fig. 3 General Flow of Model

3 Proposed methodology

The paper presents a machine learning model for detection of diabetes and prediction of glucose level. The optimised regression models are used for calibration to have accurate glucose prediction from the non-invasive iGLU device [31]. The proposed model is designed with the purpose to create a non-invasive, wearable, painless, precise, and low-cost device with high accuracy [32, 33]. Figure 4 presents the process flow of proposed model.

Various machine learning algorithms have been used on the data acquired from UCI repository (PIDD) [34] and the iGLU device (iGLU dataset) [35]. We have applied various regression and classification algorithms on these data sets individually. Further, the results of these models based on various evaluation criteria such as MAE, RMSE and Accuracy Score are compared. For classification, Recall, F1-Measure and ROC curves have also been used for comparison. Figure 5 represents process of conceptual framework for proposed methodology for glucose measurement.

The following steps have been followed to get the desired results.

1. **Data Acquisition:** This step involves collection of data. We have used two datasets out of which one is an open source data set and the other has been collected by iGLU device.
2. **Data Prepossessing:** This involves processing the data in such a way that it becomes ready for machine learning algorithm implementation [36]. This include arranging data into proper attributes, making sure that the data is complete and removing the unwanted data.

Fig. 5 Process of Conceptual Framework

3. **Applying Machine learning algorithms:** This step involves the application of various machine learning algorithms on the datasets and training the model.
4. **Performance Evaluation:** Here the results of the applied algorithms are evaluated and the performance of each algorithm in terms of accuracy is checked.
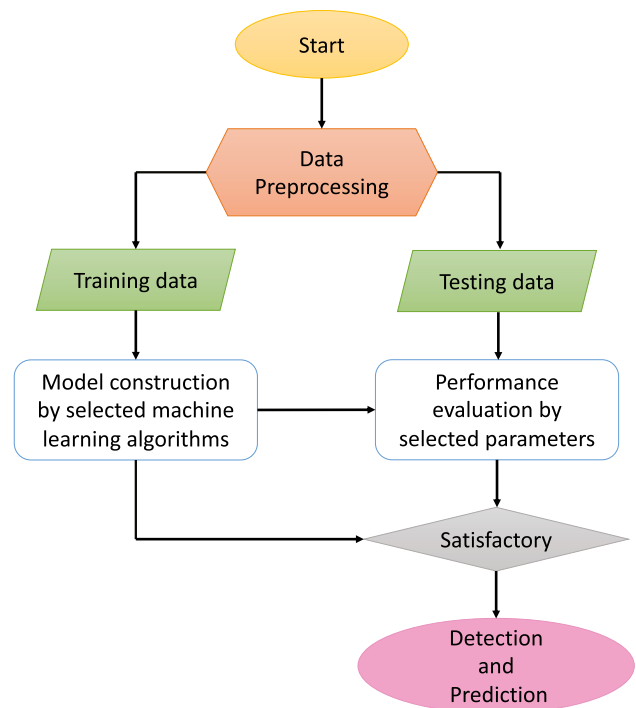5. **Comparative Analysis:** Here the comparison of all the applied machine learning algorithms is done to under-

Fig. 4 Process Flow of Proposed Model
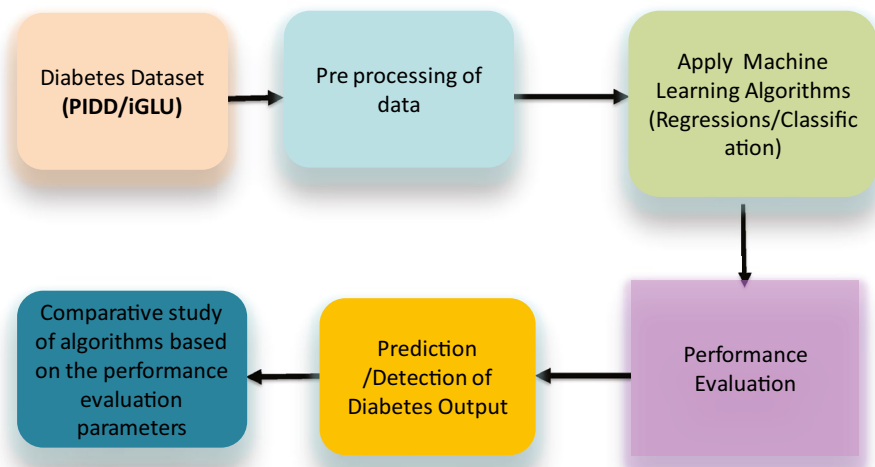
Table 1 Dataset description

Database	No. of Attributes	No. of Instances
PIDD	8	768
iGLU	3	99

stand which one of them is the best for diabetes detection and prediction of glucose level.

3.1 Data acquisition

In this work, two datasets have been taken to work on which are described below:

Pima Indian Diabetes Dataset (PIDD) The proposed machine learning model is evaluated on an open source data namely pima indian diabetes dataset which is acquired from the UCI repository. The dataset has medical details of 768 female patients. Table 1 represents the dataset description and the Table 2 represents the attributes description for PIDD.

iGLU dataset The workflow of proposed model with iGLU device is presented in Fig. 6. The data is accumulated by three fingers by placing them between the emitter and detector lined with the pads. The design of the pads is such that the emitters and detectors are placed underneath the surface of the pads. Detectors with daylight blocking filters are packaged in a such a way that they are not affected by sweat. Hence, the probability of a flawed measurement is reduced to minimum. Table 3 represents the attributes description for iGLU dataset.

3.2 Data preprocessing

This involves two steps. First, we filled the missing values (data cleaning) [37]. Then we selected features and labels and then we modified the data by using scaling. For PIDD, we used auto scaling whereas for iGLU dataset we made our own scaling function to scale the data so that it can fit in the model in best way.

Table 2 Attributes and Abbreviaton Used for PIDD

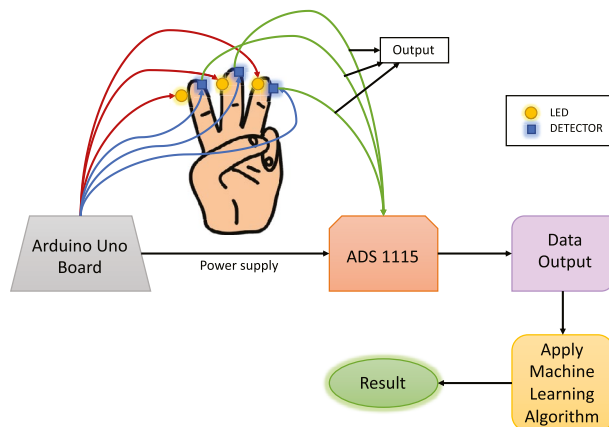| S No.	Attributes	Abbreviation of Attributes
1	Number of Times Pregnant	pr
2	Plasma glucose concentration	pl
3	Distolic blood Pressure (mm Hg)	pr
4	Skin Fold Thickness(mm)	sk
5	2-Hour serum insulin(U/ml)	in
6	BMI (weighted $gl(height/mm)^2$	ma
7	Diabetes pedigree function	pe
8	Age in Years	ag
9	Class '0' or '1'	cl

Fig. 6 Workflow of proposed model with iGLU device

3.3 Applying machine learning algorithms

The machine learning algorithms have been implemented for the blood glucose measurement. The logistic regression is applied to classify the whether the patient has diabetes or not. We have used Multiple Polynomial Regression (degree = 2 to degree = 7) for 3 independent variables and its corresponding output variable (dependent variable). We have chosen the most suitable degree which gives best results. Further, SVM is used as classifier (support vector classification algorithm) and regressor (support vector regression algorithm) both to detect the disease and prediction of glucose level in our work. Subsequently, KNN (K Nearest Neighbors), Decision Tree, Random Forest, Gradient Boost, Gaussian naive Bayes and Deep Neural Network have been used for the glucose measurement. The block diagram of various machine learning models for both data-set is shown in Fig. 7.

4 Performance evaluation and results discussion

The performance has been evluated for both datasets for diabetes by applying regression and classification type of machine learning algorithm. Logistic Regression, KNN, Gaussian Naive Bayes are applied for classification purpose and Linear Regression, Multi-polynomial Regression are applied for regression purpose and Neural

Table 3 Attributes and Abbreviation Used for iGLU dataset

S no.	Attributes	Abbreviation of Attributes
1	Voltage 1	ch 1
2	Voltage 2	ch 2
3	Voltage 3	ch 3
4	Predicted Value	val

Fig. 7 Applied machine learning algorithms in accordance to dataset
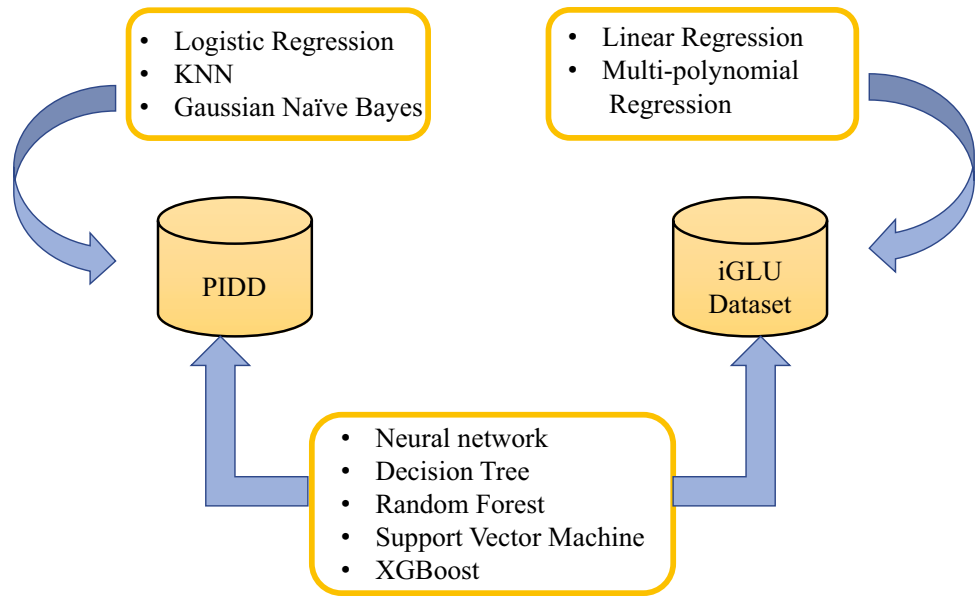
Network, XGBoost, Decision Tree, Random Forest and Support Vector Machine are applied as both classifiers and regressors.

4.1 Performance evaluation

The effectiveness of the model is measured using various quantified parameters. The performance of classification is measured with Accuracy, Confusion Matrix, Precision, Recall and ROC-AUC curve, whereas the performance of Regression is evaluated using Accuracy, MAE, RMSE. All these parameters used for performance evaluation are explained further in brief.

(a) **Confusion Matrix:**
Confusion Matrix is summary of predictions made by the classifier [38]. It keeps the count of correct and wrong predicted values. For two class problem, there are four parameters:

- True Positive (TP) : Actual positive value is predicted positive.
- False Negative (FN) : Actual positive value is predicted negative.
- False Positive (FP) : Actual negative value is predicted positive.
- True Negative (TN) : Actual negative value is predicted positive.

(b) **Precision:**
Precision is measurement of how many positive predictions are made for actual positive values. The formula is,

$$P = \frac{TP}{(TP + FP)}$$

(c) **Recall/true positive rate/sensitivity:**
Recall is measurement of correct positive predictions from all positive predictions made and is defined as,

$$R = \frac{TP}{TP + FN}$$

(d) **Accuracy:**
The total number of correct predictions made out of all. It shows overall effectiveness of classifier. The formula is,

$$AC = \frac{(TP + TN)}{TP + FN + FP + TN}$$

(e) **F-1 Score:**
F-1 Measure is a combination of precision and recall and is defined as follows:

$$F-1Score = \frac{(2 * Precision * Recall)}{(Precision + Recall)}$$

(f) **ROC-AUC curve:**
AUC - ROC is a curve that measures the performance of classification problem at various thresholds settings. ROC is a likelihood curve and AUC represents a degree/measure of separability. A receiver operating characteristics (ROC) curve tells us how good a classifier can differentiate between two classes (i.e. whether the patient has diabetes or not). Better classifier can differentiate accurately. As greater the value of AUC under the curve defines the better the performance of

Table 4 Comparison of Algorithm in terms of Accuracy and Time

Classifier	Train Accuracy	Test Accuracy	Training Time(sec)
Logistic Regression	76%	82%	0.05784
KNN	74%	77%	0.13156
Linear SVM	76%	82%	2.83625
XGBoost	72%	82%	0.10871
Decision Tree	71%	75%	0.00200
Neural Network	73%	71%	2.16293
Random Forest	75%	84%	0.16556
Gaussian Naive Bayes	75%	79%	0.00100

classifier. ROC curve is plotted as TPR (true positive rate or Sensitivity) on y-axis against with FPR (false positive rate or 1-Specificity) on x-axis.

(g) ***Mean absolute error (MAE):***

Mean absolute error (MAE) is average over the test data of the absolute differences between actual values and predicted values. It measures the mean amount of the errors in the data of predictions, without taking directions into consideration.

(h) ***Root mean square error (RMSE):***

As name says RMSE is the square root of mean of squared error. Error is defined as differences between actual value and predicted value. It is a quadratic scoring rule, that is, this score represents/reflects the closeness of an anticipated probability distribution to the detected/realized output. It measures the mean amount of the error.

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n}(y_j - y'_j)^2}{n}}$$
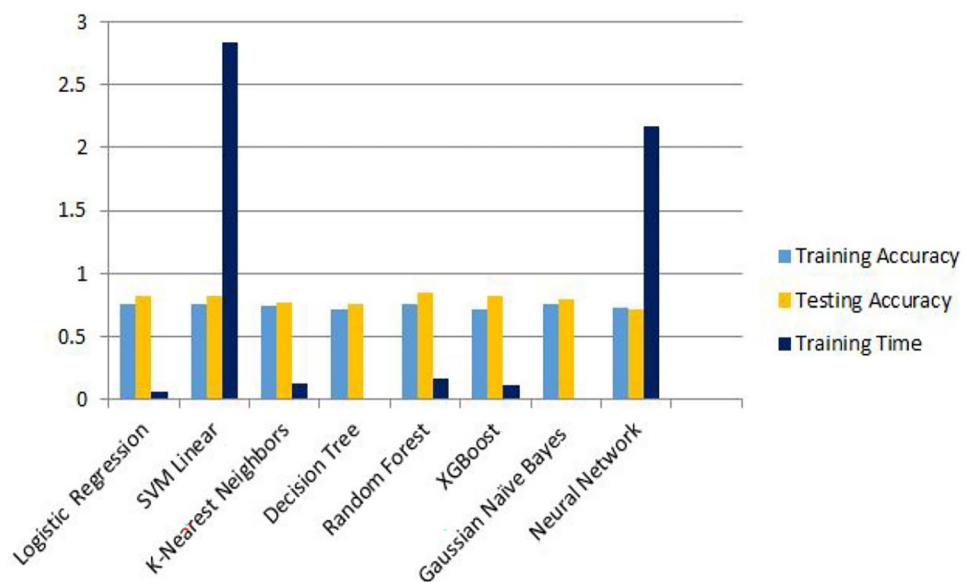
4.2 Comparative analysis

The results of application of classification and regression on the two datasets have been compared by forming tables. Their evaluation based on the above mentioned parameters has been shown in the form of charts, tables and confusion matrix. This gives us a clear indication for selecting the best algorithm to be used for detection of diabetes and prediction of glucose values.

4.3 Analysis of Pima Indian diabetes dataset

According to experimental studies, the datasets have been distributed between 80-20 % (614-154) for training and testing purpose. We have applied various classification algorithms like Logistic Regression, Gaussian Naive Bayes, Support Vector Machine (Linear), Gradient Boost, Neural Network, KNN, Random Forest, Decision Tree and obtained the parameters like Accuracy, ROC-AUC, Confusion Metrics. From confusion metrics, we have calculated Precision, Recall and F1-Score. ROC curve is plotted for each algorithm and the classifier which have covered more area is better. Table 4 shows training accuracy, testing accuracy and training time for each algorithm. Among all the applied algorithms, Random Forest Algorithm has the highest accuracy 84% by taking 0.16 secs for training. Being the simplest classifier, Logistic regression has performed effectively with 82 % accuracy by taking less time than Random Forest which is 0.05 secs. However, these results can be enhanced by applying larger updated data.

Figure 8 shows the comparison of applied algorithms in terms of training-testing accuracy and training time. These bars show that SVM Linear and Neural Network take more

Fig. 8 Comparison of Algorithms in terms of Accuracy and Time

PREDICTED DIABETES		
	NO	YES
NO	348	45
YES	100	121

(ACTUAL DIABETES)

Fig. 9 Training cm

PREDICTED DIABETES		
	NO	YES
NO	344	49
YES	95	126

(ACTUAL DIABETES)

Fig. 12 Training cm

PREDICTED DIABETES		
	NO	YES
NO	98	9
YES	19	28

(ACTUAL DIABETES)

Fig. 10 Testing cm

PREDICTED DIABETES		
	NO	YES
NO	97	10
YES	18	29

(ACTUAL DIABETES)

Fig. 13 Testing cm

than 2 secs to train the model. All algorithms providing more than 70 % accuracy. Also 4 out of 8 algorithms gives more than 80 % accuracy which is quite well. Accuracy alone is not enough for choosing the better classifier. So, we have calculated confusion metrics and ROC curve for different algorithms.

(a) **Logistic Regression Algorithm**

Confusion matrix obtained from LR is represented in Figs. 9 and 10. We train the preprocessed data through Logistic Regression Algorithm and get the following results. AUC-ROC curve obtained for Logistic Regression algorithm with resultant AUC = 0.87 (Fig. 11).

(b) **SVM Linear Algorithm**

Confusion matrix obtained from SVM is represented in Figs. 12 and 13 We train the preprocessed data through Support Vector Machine (Linear) Algorithm and get the following results. AUC-ROC curve obtained for SVM Linear algorithm with resultant AUC = 0.86 (Fig. 14).
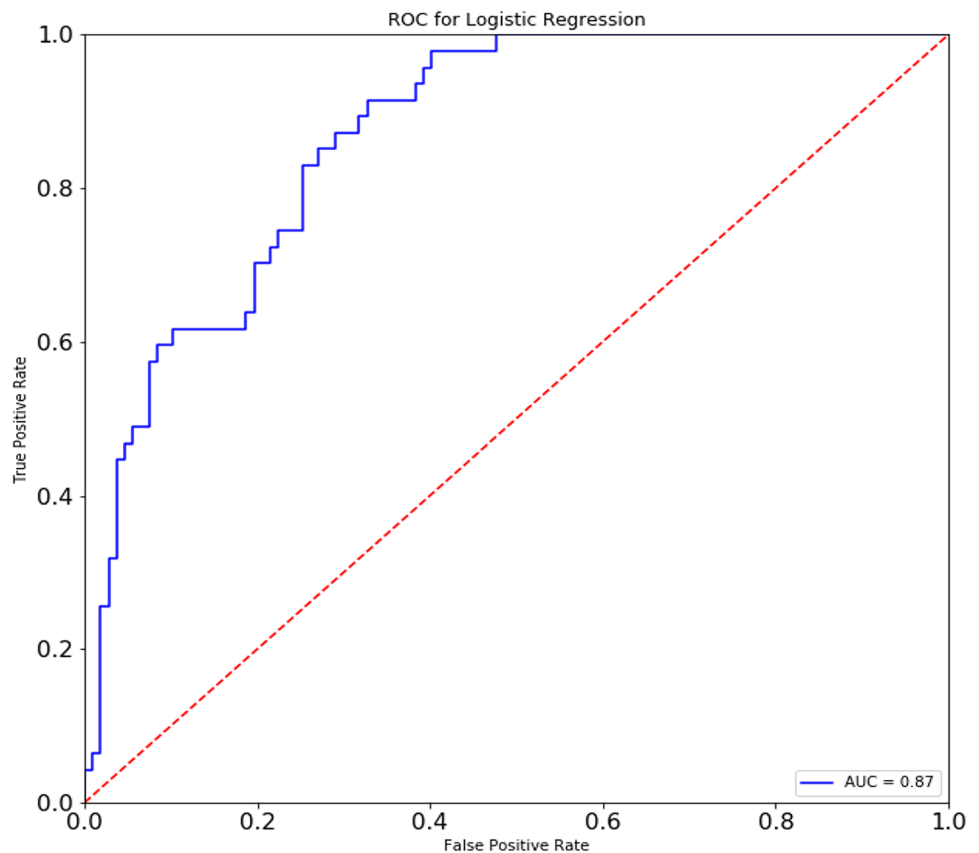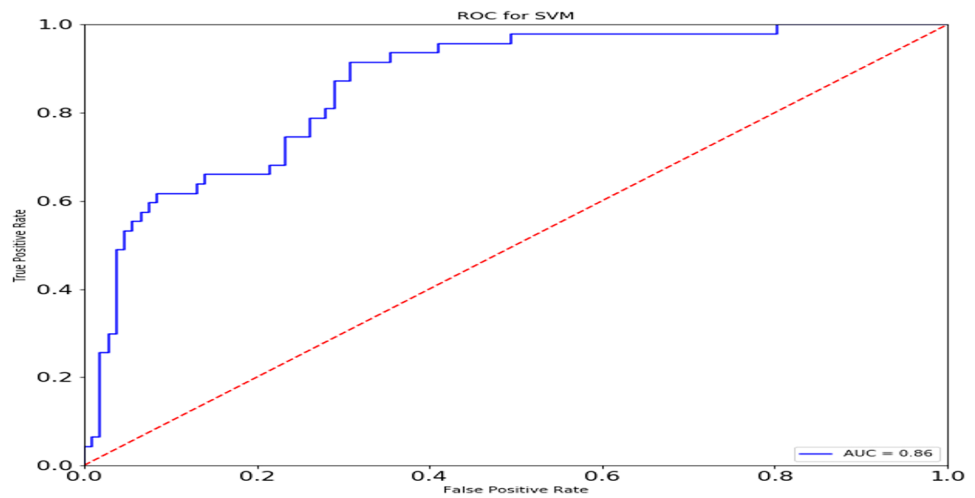
Fig. 11 ROC for Logistic Regression

Fig. 14 ROC for SVM Linear

Fig. 15 Training cm

| ACTUAL DIABETES | PREDICTED DIABETES | |
	NO	YES
NO	392	1
YES	3	218

Fig. 16 Testing cm

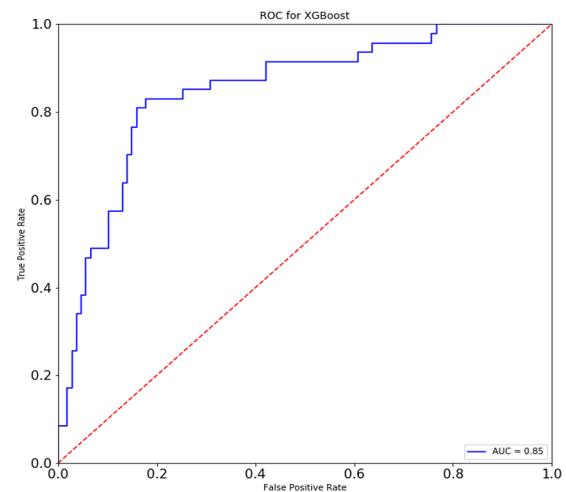| ACTUAL DIABETES | PREDICTED DIABETES | |
	NO	YES
NO	91	16
YES	12	35

(c) **XGBoost Algorithm**

Confusion matrix obtained from XGB is represented in Figs. 15 and 16. We train the preprocessed data through XGBoost Algorithm and get the following results. AUC-ROC curve obtained for XGBoost algorithm with resultant AUC = 0.85 (Fig. 17).

(d) **Gaussian Naive Bayes Algorithm**

Confusion matrix obtained from GNB is represented in Figs. 18 and 19. We train the preprocessed data through Gaussian Naive Bayes Algorithm and get the following results. AUC-ROC curve obtained for Gaussian Naive Bayes algorithm with resultant AUC = 0.84 (Fig. 20).

(e) **KNN Algorithm**

Confusion matrix obtained from KNN is represented in Figs. 21 and 22. We train the preprocessed data through KNN Algorithm and get the following results. AUC-ROC curve obtained for KNN algorithm with resultant AUC = 0.83 (Fig. 23).

Fig. 17 ROC for XGBoost

Fig. 18 Training cm

| ACTUAL DIABETES | PREDICTED DIABETES | |
	NO	YES
NO	329	64
YES	85	136

Fig. 19 Testing cm

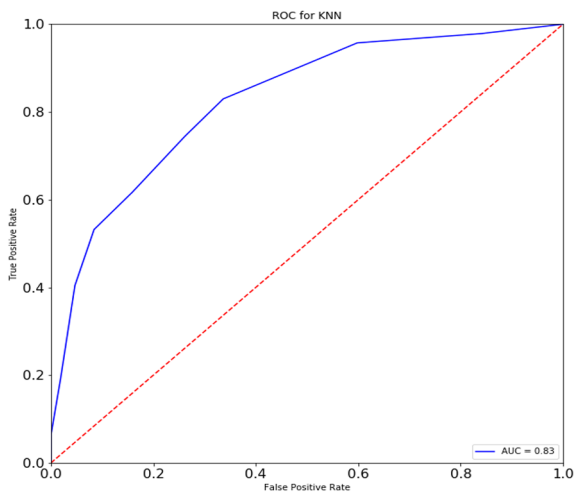| ACTUAL DIABETES | PREDICTED DIABETES | |
	NO	YES
NO	93	14
YES	18	29

Fig. 20 ROC for GNB

(f) **Decision Tree Algorithm**

Confusion matrix obtained from DT is represented in Figs. 24 and 25. We train the preprocessed data through Decision Tree Algorithm and get the following results. AUC-ROC curve obtained for Decision Tree algorithm with resultant AUC = 0.81 (Fig. 26).

(g) **Random Forest Algorithm**

Confusion matrix obtained from RF is represented in Figs. 27 and 28. We train the preprocessed data

		PREDICTED DIABETES	
		NO	YES
ACTUAL DIABETES	NO	352	41
	YES	71	150

Fig. 24 Training cm

		PREDICTED DIABETES	
		NO	YES
ACTUAL DIABETES	NO	346	47
	YES	86	135

Fig. 21 Training cm

		PREDICTED DIABETES	
		NO	YES
ACTUAL DIABETES	NO	91	16
	YES	22	25

Fig. 25 Testing cm

		PREDICTED DIABETES	
		NO	YES
ACTUAL DIABETES	NO	90	17
	YES	18	29
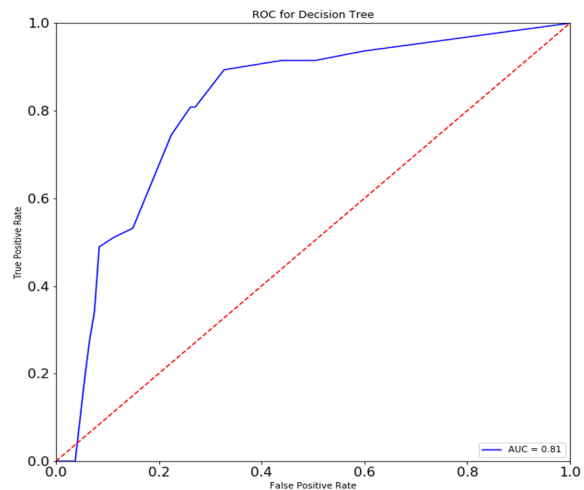
Fig. 22 Testing cm

Fig. 26 ROC for Decision Tree

Fig. 23 ROC for KNN

		PREDICTED DIABETES	
		NO	YES
ACTUAL DIABETES	NO	365	28
	YES	61	160

Fig. 27 Training cm

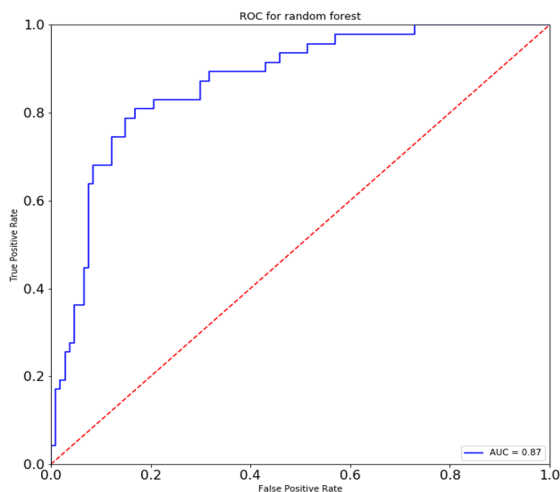| PREDICTED DIABETES		
	NO	YES
ACTUAL DIABETES NO	97	10
ACTUAL DIABETES YES	15	32

Fig. 28 Testing cm

Fig. 29 ROC for Random Forest

through Random Forest Algorithm and get the following results. AUC-ROC curve obtained for Random Forest algorithm with resultant AUC = 0.87 (Fig. 29).

(h) **Neural Network Algorithm**

Confusion matrix obtained from NN is represented in Figs. 30 and 31. We train the preprocessed data through Neural Network Algorithm and get the following results. AUC-ROC curve obtained for Neural Net-

PREDICTED DIABETES		
	NO	YES
ACTUAL DIABETES NO	334	59
ACTUAL DIABETES YES	107	114

Fig. 30 Training cm

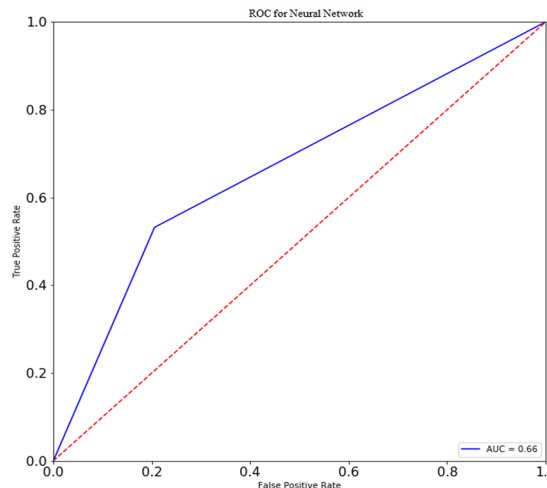| PREDICTED DIABETES		
	NO	YES
ACTUAL DIABETES NO	85	22
ACTUAL DIABETES YES	22	25

Fig. 31 Testing cm

Fig. 32 ROC for Neural Network

work algorithm with resultant AUC = 0.66 (Fig. 32). From confusion matrix, precision, recall and f-1 score is also calculated for training and testing both (Table 5). Figure 33 represents comparison of algorithms in terms of AUC. Random Forest and Logistic Regression both algorithms have same AUC =0.87. These two are better classifiers in comparison to others. These results are better in comparison to a previous work done on the same dataset [38].

4.4 Analysis of iGLU dataset

The dataset has been split into 80:20 for training and testing purpose for best results. In other distributions, it becomes highly prone to over-fitting. Among the applied all the algorithms, best results are obtained by Decision Tree with accuracy of 70.64 % having mean absolute error of 7.89 % and root mean square error of 8.56 % which is lowest among all algorithms. Gradient boost regression with certain parameters gives the lowest relative absolute error 6.82 % which is quite well. The Spyder tool has been used to take these measurements. Table 6 shows the summary of prediction for different algorithm. However, these results can be improved by applying a updated, large sized dataset.

Figure 34 shows comparison of regression algorithms in terms of MAE, RMSE and Accuracy. Logistic Regression is best with low value of least root mean square error. Linear Regression and Neural Network are not providing good result in our work. Table 7 shows the comparison of applied algorithms based on training time and training-testing score. Though Neural network regressor takes longest time to run but Decision Tree gives best testing score of 70.64 % by taking much lesser time.

Table 5 Summary of prediction for different algorithms

Classifier AUC	Train			Test		
	Recall	Precision	F1-Score	Recall	Precision	F1-Score
XGBoost **0.85**	0.9864	0.9954	0.9908	0.7446	0.6862	0.7142
Decision Tree **0.81**	0.6787	0.7853	0.7281	0.5319	0.6097	0.5681
Random Forest **0.87**	0.7239	0.8510	0.7823	0.6808	0.7619	0.7190
G. Naive Bayes **0.84**	0.6153	0.6800	0.6460	0.6170	0.6744	0.6444
SVM Linear **0.86**	0.5701	0.7200	0.6363	0.6170	0.7435	0.6743
Logistic Regression **0.87**	0.5475	0.7289	0.6253	0.5957	0.7567	0.6666
K-Nearest N. **0.83**	0.6108	0.7417	0.6699	0.6170	0.6304	0.6236
Neural Network **0.66**	0.5158	0.6589	0.5768	0.5319	0.5319	0.5319

Fig. 33 Comparison of Algorithms in terms of AUC
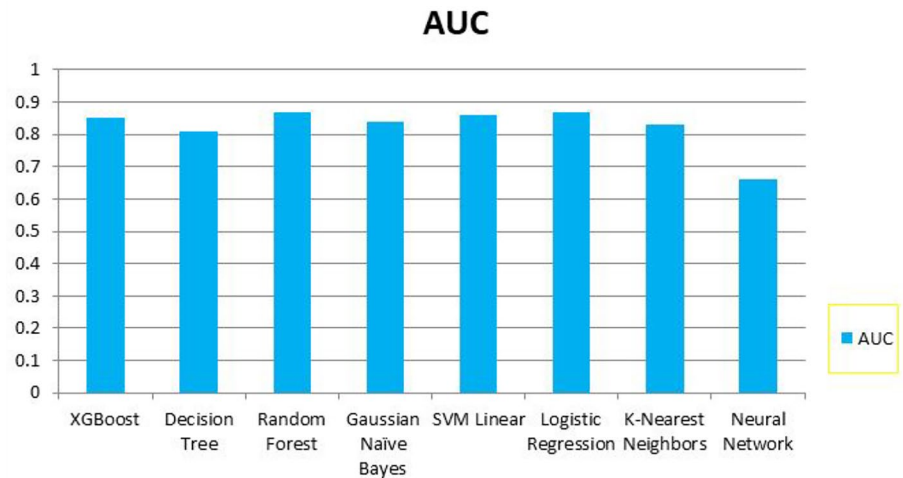
Table 6 Summary of Prediction for Different Algorithms

Regression Algorithm	MAE	RMSE	Accuracy
XGB Regression	0.0682	0.0914	0.6655
Linear Regression	0.1316	0.1614	0.0234
Polynomial Regression	0.1525	0.1238	0.0689
Support Vector Regression	0.1082	0.1342	0.2788
Random Forest	0.0735	0.0885	0.6866
Neural Network	0.1304	0.1637	0.0000
Decision Tree	0.0789	0.0856	0.7064

Fig. 34 Comparison of Algorithms in terms of Accuracy, MAE, RMSE

Figure 35 represents comparison of algorithms in terms of training-testing score and training time. As we can see, all algorithms take less than 1 sec to train the model. XGBoost, Random Forest and Decision Tree gives good accuracy while other algorithms (Linear Regression, Polynomial Regression, Support Vector Regression, Neural Network) does not provide more than 30 % accuracy for training and testing both as our dataset is too small and chances of over-fitting increase in case of these algorithms.
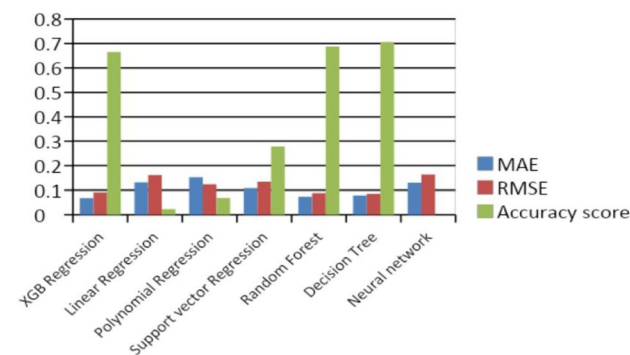
Table 7 Comparison of Algorithm in terms of Accuracy and Time

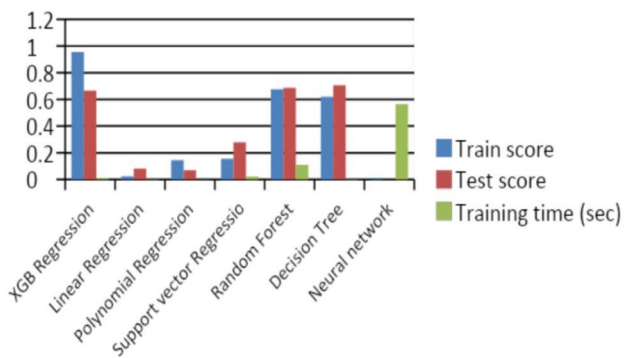| Regression Algorithm	Train Accuracy	Test Accuracy	Training Time (sec)
XGB Regression	0.9546	0.6655	0.0120
Linear Regression	0.0234	0.0817	0.0022
Polynomial Regression	0.1446	0.0689	0.0010
Support Vector Regression	0.1550	0.2788	0.0219
Random Forest	0.6759	0.6866	0.1096
Decision Tree	0.6205	0.7064	0.0027
Neural Network	0.0111	0.0000	0.5649

Fig. 35 Comparison of Algorithms in terms of Score and Training Time

4.5 Clarke error grid analysis

Clarke Error Grid is used to quantify the clinical accuracy of the values predicted by the device or method under the test in comparison of reference glucose values (values obtained from clinically approved method) [39]. Predicted values are plotted on y- axis whereas reference values are plotted on x-axis. There are five zones A, B, C, D and E in the grid. Zone 'A' (accepted) represents the predicted value which deviate 20 % from the reference value or in hypoglycemic range (<70 mg/dl). Zone 'B' (benign errors) represents the values which are deviated from reference values more than 20 % but are clinically uncritical. Zone 'B' lies below and above the zone 'A' in the grid. Zone 'A' and 'B' are clinically accepted. Zone 'C', 'D' and 'E' values will lead to wrong decisions and are potentially dangerous.

Figure 36 represents Clarke Error Grid Analysis of iGLU datset used in our work. 86 % values are in zone 'A' and 14 % values lies in zone 'B' that are clinically accepted. In zone 'C','D' and 'E' there are no values.

4.6 Comparison with previous work

We have also computed Root Mean Square Error (RMSE) and Mean Absolute Error or Mean Absolute Deviation (MAD) which we have mentioned before. Now, we have also computed Average error (AvgE) and mean absolute relative difference (mARD). The average error 9.03 % and mean absolute deviation 6.82 % represents the accuracy of our model. Table 8 represents comparison of our work with previous work.

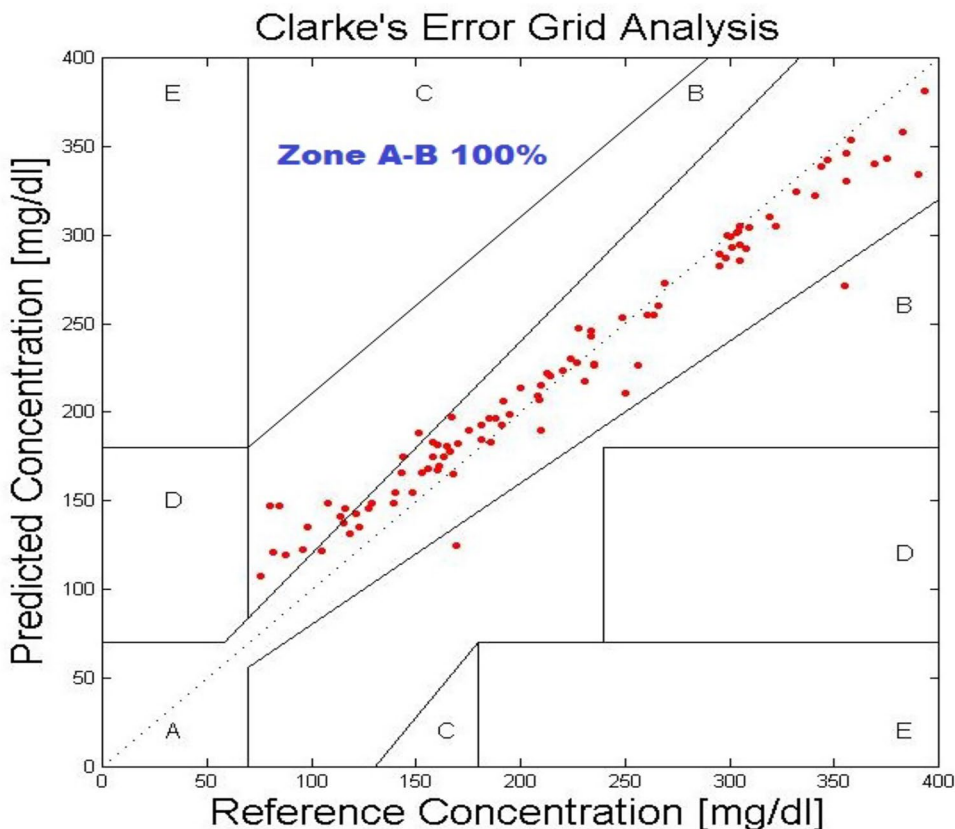**Fig. 36** Clarke Error Grid of iGLU dataset

Table 8 Comparison with previous work

Works	Category	MAD (mg/dl)	mARD (%)	RMSE (mg/dl)	AvgE (%)	CEG (A&B)
Habbu et al. [1]	Non-invasive	-	-	-	-	100%
Joshi et al. [14]	Non-invasive	9.42	4.86	13.57	4.88	100 %
Enejder et al. [25]	Non-invasive	7.8 ± 1.8	-	-	7.7	-
Wang et al. [23]	Minimally invasive	-	-	-	-	82.6%
Haider et al. [32]	Non-invasive	-	8.0	-	-	-
Song et al. [33]	Non-invasive	-	8.3	19	-	100 %
Jain et al. [3]	Non-invasive	9.89	7.32	11.56	7.30	100 %
Proposed Work	**Non-invasive**	**6.82**	**10.64**	**9.14**	**9.033**	**100%**

5 Conclusion and future scope

The paper presents a machine learning models for blood glucose measurement using non-invasive technique on two different data sets. The comparative analysis of machine learning algorithms mainly as Logistic Regression, Decision tree, K Nearest Neighbors (KNN), Random Forest, SVC (linear), Gradient Boost, Gaussian Naïve bayes and Neural Network have been used to detect the diabetic samples from the PIMA Indian diabetes data (PIDD) and XGB Regression, Linear Regression, Multi-Polynomial Regression, SVR, Random Forest, Decision Tree and Neural Network to predict the glucose level using the data set collected by the iGLU device. The performance of these algorithms is compared on the basis of MAE, RMSE, Accuracy, ROC, Precision, F-1 Measure and Recall parameters obtained from the dataset. Random Forest and Logistic Regression has AUC value of 0.87 which suggest our model is good at diabetes detection. Also Decision Tree with 70 % accuracy and RMSE 8.56 % can be used for prediction of glucose level for most accurate results. However, these results can be improved by increasing the size of dataset. Clarke error grid analysis is also done where all values lies in zone 'A' and 'B' so the model has able to predict diabetes correctly. The further work is proposed to address the security and privacy issues for continuous glucose measurement. The efforts would also put forward to integrate robust mechanism of insulin drug delivery for type -1 diabetes patients.

Author contributions All the authors contributed to the various stages of the study and the manuscript preparation.

Data availability The data that was used to prepare the manuscript is available for further consideration upon request.

Code availability Not applicable.

Declarations

Ethical approval NA

Conflicts of interest The authors report no potential conflict of interest relevant to this article.

References

1. Habbu S, Dale M, Ghongade R. Estimation of blood glucose by non-invasive method using photoplethysmography. Sadhan a. 2019;44(6):135.
2. Saeedi P, Petersohn I. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. 9th ed. 2019. (vol 157). https://doi.org/10.1016/j.diabres.2019.107843.
3. Jain P, Joshi AM, Mohanty SP. iGLU: An intelligent device for accurate non-invasive blood glucose- level monitoring in Smart Healthcare. IEEE Consumer Electronics Magazine. 2020;9(1):35–42.
4. Joshi AM, Jain P, Mohanty SP. Everything you wanted to know about continuous glucose monitoring. IEEE Consumer Electronics Magazine. 2021;10(6):61–6.
5. Ahmadi MM, Jullien GA. A wireless-implantable microsystem for continuous blood glucose monitoring. IEEE Transactions on Biomedical Circuits and Systems. 2009;3(3):169–80.
6. Jain P, Joshi AM, Mohanty SP. iGLU 1.0: an accurate non-invasive near-infrared dual short wavelengths spectroscopy based glucometer for smart healthcare. arXiv:1911.04471 [Preprint]. 2019. Available from: https://arxiv.org/abs/1911.04471.
7. Sarkar K, Ahmad D, Singha SK, Ahmad M. Design and implementation of a noninvasive blood glucose monitoring device. In: 21st International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh vol. 2018. 2018. p. 1–5.
8. Buda A, Addi MM. A portable non-invasive blood glucose monitoring device. In: 2014 IEEE Conference on Biomedical Engineering and Sciences (IECBES), Kuala Lumpur. 2014. p. 964–69. https://doi.org/10.1109/IECBES.2014.7047655.
9. World Health Organization. Coronavirus disease 2019 (COVID-19) situation report e 123. 2020.
10. Joshi AM, Shukla UP, Mohanty SP. Smart healthcare for diabetes during COVID-19. IEEE Consumer Electronics Magazine. 2020;10(1):66–71.
11. Singh AK, Gupta R, Ghosh A, Misra A. Diabetes in COVID-19: Prevalence, pathophysiology, prognosis and practical considerations [published online ahead of print, 2020 Apr 9]. Diabetes Metab Syndr. 2020;14(4):303-310.
12. Paul B, Manuel MP, Alex ZC. Design and development of non invasive glucose measurement system. In: Proccedings on 1st

International Symposium on Physics and Technology of Sensors. 2012. p. 43–6.

13. Sundaravadivel P, Kougianos E, Mohanty SP, Ganapathiraju MK. Everything you wanted to know about smart health care: evaluating the different technologies and components of the Internet of Things for better health. IEEE Consumer Electronics Magazine 2017;7(1):18–28.

14. Joshi AM, Jain P, Mohanty SP, Agrawal N. iGLU 2.0: a new wearable for accurate non-invasive continuous serum glucose measurement in IoMT framework. In: IEEE Transactions on Consumer Electronics, vol 66, no 4. Nov. 2020. p. 327–35. https://doi.org/10.1109/TCE.2020.3011966.

15. Lin T. Non-Invasive glucose monitoring: a review of challenges and recent advances. Current Trends in Biomedical Engineering & Biosciences. 2017;6. https://doi.org/10.19080/CTBEB.2017.06.555696.

16. Joshi AM, Jain P, Mohanty SP. iGLU 3.0: a secure noninvasive glucometer and automatic insulin delivery system in IoMT. In: IEEE Transactions on Consumer Electronics. https://doi.org/10.1109/TCE.2022.3145055.

17. Sejdinović D, et al. Classification of prediabetes and type 2 diabetes using artificial neural network. In: Badnjevic A, editor. CMBEBIH 2017. IFMBE Proceedings, vol 62. Springer, Singapore; 2017.

18. Alić B, et al. Classification of metabolic syndrome patients using implemented expert system. In: Badnjevic A, editor. CMBEBIH 2017. IFMBE Proceedings, vol 62. Springer, Singapore; 2017.

19. Spahić, et al. Lactose intolerance prediction using artificial neural networks. In: Badnjevic A, Škrbić R, Gurbeta Pokvić L, editors. CMBEBIH 2019. IFMBE Proceedings, vol 73. Springer, Cham; 2019.

20. Imamović E, et al. Modelling and simulation of blood glucose dynamics. 2020 9th Mediterranean Conference on Embedded Computing (MECO). 2020 p. 1–4.

21. Spahić, L., Ćordić, S. Prostate tissue classification based on prostate-specific antigen levels and mitochondrial DNA copy number using artificial neural network. In: Badnjevic, A., Škrbić, R., Gurbeta Pokvić, L. (eds) CMBEBIH 2019. CMBEBIH 2019. IFMBE Proceedings, vol 73. Springer, Cham.

22. Monte-Moreno E. Non-invasive estimate of blood glucose and blood pressure from a photoplethysmograph by means of machine learning techniques. Artif Intell Med. 2011;53(2):127–38.

23. Wang G, Poscente M, Park S, Andrews C, Yadid-Pecht O, Mintchev M. Wearable microsystem for minimally invasive, pseudo-continuous blood glucose monitoring: the e-Mosquito. IEEE Transactions on Biomedical Circuits and Systems. 2017. p. 1–9. https://doi.org/10.1109/TBCAS.2017.2669440.

24. Amrane S, Azami N, Elboulqe Y. Optimized algorithm of dermis detection for glucose blood monitoring based on optical coherence tomography. In: Proccedings on 10th International Conference on Intelligent Systems: Theories and Applications 2015. 2015. p. 1–5.

25. Enejder A, Scecina T, Jeankun O, Martin H, Wei-Chuan S, Slobodan S, Horowitz G, Feld M. Raman Spectroscopy for noninvasive glucose measurements. J Biomed Opt 2005;10: 031114. https://doi.org/10.1117/1.1920212.

26. Agrawal RP, Sharma N, Rathore MS, Gupta VB, Jain S, et al. Noninvasive method for glucose level estimation by saliva. J Diabetes Metab. 2013;4:266. https://doi.org/10.4172/2155-6156.1000266.

27. Demitri N, Zoubir AM. Measuring blood glucose concentrations in photometric glucometers requiring very small sample volumes. IEEE Trans Biomed Eng. 2017;64(1):28–39. https://doi.org/10.1109/TBME.2016.2530021.

28. Ramasahayam S, Haindavi K, Chowdhury S. Noninvasive estimation of Blood glucose concentration using near infrared optodes. Smart Sensors Meas Instrum. 2015;12:67–82. Springer.

29. Heller A. Integrated medical feedback systems for drug delivery. AlChE J. 2005;51(4):1054–66.

30. Pai PP, Sanki PK, De A, Banerjee S. NIR photoacoustic spectroscopy for non-invasive glucose measurement. In: Proccedings on 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2015. p. 7978–981.

31. Jain P, Maddila R, Joshi AM. A precise non-invasive blood glucose measurement system using NIR spectroscopy and Hubers' regression model. Opt Quant Electron 2019;51(2):51. US: Springer.

32. Ali H, Bensaali F, Jaber F. Novel approach to non-invasive blood glucose monitoring based on transmittance and refraction of visible laser light. IEEE access. 2017;5:9163–74.

33. Song K, Ha U, Park S, Bae J, Yoo HJ. An impedance and multiwavelength near-infrared spectroscopy IC for non-invasive blood glucose estimation. IEEE J Solid-State Circuits. April 2015;50(4):1025–37.

34. https://archive-beta.ics.uci.edu/.

35. Jain P, Joshi AM, Agrawal N, Mohanty S. iGLU 2.0: a new non-invasive, accurate serum glucometer for smart healthcare. arXiv:2001.09182 [Preprint]. 2020. Available from: http://arxiv.org/abs/2001.09182.

36. Pancholi S, Joshi AM. Novel time domain based upper-limb prosthesis control using incremental learning approach. arXiv:2109.04194 [Preprint]. 2021. Available from: http://arxiv.org/abs/2109.04194.

37. Kokate P, Sidharth P, Joshi AM. Classification of upper arm movements from EEG signals using machine learning with ICA analysis. arXiv:2107.08514 [Preprint]. 2021. Available from: https://arxiv.org/abs/2107.08514.

38. Mir A, Dhage SN. Diabetes disease prediction using machine learning on big data of healthcare. Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). IEEE; 2018. p. 1–6.

39. Jain P, Pancholi S, Joshi AM. An IoMT based non-invasive precise blood glucose measurement system. 2019 IEEE International Symposium on Smart Electronic Systems (iSES)(Formerly iNiS). 2019. p. 111–16.