Precision Insulin Delivery: Predictive Modelling for Bolus Insulin Injection in Real-Time

V.K.R. Rajeswari Satuluri, Vijayakumar Ponnusamy*
Department of ECE, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

***Abstract*—Insulin is recommended for patients with Diabetes Mellitus (DM). It is challenging for doctors to prescribe accurate bolus insulin before every meal due to real-time factors such as the size of the meal, skipping a previous meal, and physical activity, which can risk the patient towards hyperglycemia or hypoglycemia. Previous studies executed insulin predictions where the methods did not consider the cases of controlled glucose levels, type of insulin prescribed, time of insulin-induced, and data detersion that can alter the predictions. To address these problems, our work has proposed an insulin predictive model from the integration of Internet of Things (IoT) devices, i.e., Continuous glucose monitoring (CGM) sensor and insulin pumps with rapid-acting insulin type where the insulin dosage with corresponding Current Blood Glucose levels (CBG) and improved Next Blood Glucose levels (NBG) are chosen. The dataset is subjected to data detersion where pre-processing, Exploratory Data Analysis (EDA), and Feature Selection is performed. Machine Learning (ML) models are applied on curated dataset where Decision Tree (DT)-Bagging algorithm, performed the best with a Mean Absolute Error (MAE) of 1.54 and a Mean Square Error (MSE) of 4.15. Performance metrics of the current study imply its suitability in medical applications for accurate prediction of real-time insulin dosage.**

Keywords*—Continuous glucose monitoring; bolus insulin prediction; data curation; data detersion; diabetes mellitus; exploratory data analysis; feature selection; machine learning; pre-processing

I. Introduction

DM is an abnormality where irregular blood glucose levels arise due to the inadequacy of insulin secretion from the pancreas, insulin action in the body, or both [1]. Prediction of insulin is important for making informed decisions to maintain blood glucose levels [2].

A. Background

Previous methods of predicting insulin dosage based on invasive blood glucose collection methods have not considered the type of insulin which varies for every person that can alter the readings [3-4]. Other challenges are fluctuating glucose levels with respect to lifestyle factors such as skipping the previous meal, meal size, uncontrolled food habits, or physical activity. Prescribed insulin dosage may lead to overdosage and underdosage in these cases. Therefore, there is a need of a prediction model for insulin dosage in real-time which can be achieved from IoT devices by acquiring real-time blood glucose from CGM sensor and an insulin pump data to deliver accurate insulin dosages.

A study implemented a Gradient-boosting classifier to predict diabetes and linear regression for predicting insulin dosage from the Pima diabetes dataset and University of California (UCI) insulin dataset. An accuracy of 100% with the Gradient-boosting classifier and 78% with linear regression is achieved [5]. Deep reinforcement learning is implemented for bolus insulin advisors. It is observed that Time in range (TIR) increased for volunteers with bolus insulin advisor from TIR=74.1%±8.4% to 80.9% ± 6.9%, 54.9% ±12.4% to 61.6 ±14.1 [6]. A study discussed predicting insulin levels from 36 months of patient data by implementing Recurrent Neural Network (RNN)-Long Short-Term Memory (LSTM) and Artificial Neural Network (ANN) with an accuracy of 90% [7]. In a similar study of obtaining high MAE on predicting insulin dosage based on predicted glucose levels, Support Vector Regression (SVR) provides MAE of 28mg/dL on CBG, 21mg/dL on average daily glucose levels, and 3.8mg/dL on insulin required in next 24hrs. The study concludes that predicting accuracy is hard because glucose and insulin are highly erratic [8]. A study attempted to predict the initial inpatient Total Daily Dose (TDD). Ensemble learning model, i.e., Ridge regularization, Lasso regression, Random Forest, Gradient boosted DT is implemented where an Area Under Receiver Operating Curve (AUROC) of 0.85 and Area Under Precision-Recall Curve (AUPRC) of 0.66 is achieved [9]. Another study proposed glucose prediction with an accuracy of 98.7% and insulin dosage delivery prediction by employing an ANN. MSE calculated for ANN is 5.79. Feature Selection is carried out to identify the best features for insulin prediction. Data is set to zero when the patient takes no insulin during data processing. This may result in input data variation due to incorrect data patterns [10]. A dataset containing full CBG and insulin-prescribed information is vital for predicting insulin dosage. Other essential parameters, such as carb ratio, Body Mass Index (BMI), and correction factor, must be considered for predicting insulin dosage. In a similar study of initial insulin estimation during hospital admission, an ensemble algorithm with regression, Random Forest (RF), and gradient boosting is applied to classify patients who require more than six units of insulin and TDD. Receiver operating characteristic curve (ROCC) of 0.84 with 95% confidence interval (CI), Area under the curve (AUC) of 0.65 with a 95% CI, and MAE with 12 units of insulin is achieved [11]. MAE obtained is too high for insulin dosage prediction. In a study of gestational diabetes for predicting insulin levels, the Oral Glucose Tolerance Test (OGTT) was considered an independent predictor. Area Under the Curve (AUC) for the prediction of insulin treatment was found to be 0.77[12]. The algorithm can predict insulin and glucose levels by considering other parameters such as BMI, PBG, NBG, and CBG. Weight, fasting blood glucose, and gender are fed into an ANN algorithm for predicting insulin dosage, where an average accuracy of 96.5% and an average prediction error of 4% are achieved [13]. A neural network (NN) based bolus insulin prediction is attempted in a study from CGM. The NN is trained to learn Standard Formula (SF) parameters by examining the Blood Glucose Risk Index (BGRI). The parameters chosen are the Optimal bolus insulin calculator (SF-OPT), found to be 0.40, and the Neural Network Correction factor (SF-NNC), 0.37. Optimal-Neural network corrector (OPT-NNC), i.e.,0.30, Scheiner -Neural network corrector (SC-NNC), i.e., 0.23, Pettus and Edelman (PE-NNC) which is found to be 0.20[14].

The research gaps identified from the above literature are as follows:

1) The existing methods have predicted insulin by considering blood glucose values and their prescribed insulin dosage. The methods haven’t focused on evaluating the cases of improved blood glucose levels w.r.t the prescribed insulin dosage. Therefore, the prediction may not be accurate in real-time.

2) Existing literature hasn't focused on the data detersion process [5-14]. Data detersion is vital for fixing ambiguities, errors, and any irrelevant data that may contribute to weakening the model. It is required for generating reliable visualizations and accurate models.

3) Various types of insulin are suggested for patients such as short-acting, ultra short-acting, intermediatory and long-acting insulin with different onset or peak times. The existing methods haven’t focused on the type of insulin for predicting insulin dosage. As every type of insulin varies, the predictions are inaccurate and prone to hyperglycemia or hypoglycemia in real-time.

4) Meal intake is a potential discrepancy that influences the prescribed insulin dosage. Existing methods haven’t focused on considering meal intake before the prescribed insulin regimen for accurate prediction of insulin dosage.

There is a need to create a multidisciplinary approach for predicting bolus insulin dosage by considering the parameters, i.e., insulin type, meal influence, CBG, improved NBG, and the corresponding insulin dosage. This attempt trains the model accurately by considering the suitable insulin dosages w.r.t the CBG. Data detersion is required to ensure that the reliability and accuracy is achieved by removing the outliers, inconsistencies to avoid skewed results by improving the quality of data for insightful information. This is our rationale to implement an advanced method of insulin prediction based on CBG and NBG [15] and data detersion methods. To our knowledge, this work is the first attempt from the existing literature to apply various methods of data detersion and prediction of bolus insulin from CBG and NBG levels, i.e., blood glucose recorded after half an hour of inducing bolus insulin. The outcome of the prediction is applicable in making informed clinical decisions, treatment titrations, changes in lifestyle habits, evidence-based dosage recommendations based on the patient's historical data, treatment outcomes, and the patient's response to the drug and clinical trials of insulin drug dosage. The novel contribution of the work is as follows:

1) The novelty of the proposed work is to create a prediction model from CBG and improved NBG for predicting accurate bolus insulin dosage.

2) Among all ML models, a striking improvement with 39.7% (from 3.12 to 1.88) in MAE and 72.7% (from 17.52 to 4.78) in MSE with ANN is achieved after applying Feature Selection.

3) After applying data detersion, the datasets improved the performance with 47.4% (from 3.12 to 1.64) in MAE and 76.2% (from 17.52 to 4.16) in MSE with the ANN model.

4) Bagging and boosting enhanced the performance of the dataset when compared with non-bagging and non-boosting models. An improvement of 35.5% in MAE (from 2.39 to 1.54) and 78.1% MSE (from 18.96 to 4.15) with DT-Bagging is achieved. Similarly,10% in MAE (from 2.39 to 2.13) and 56.6% MSE (from 18.96 to 8.22) with DT-Boosting is achieved.

The proposed work is organized as follows: Section II presents Material and Methods where data collection and cohort, data preparation, and data detersion are carried out. Data pre-processing, EDA, and feature selection are employed in data detersion method. Section III presents the Results and Discussion section, where bolus insulin prediction, predictive analysis, and validation of the DT-Bagging model are executed. The paper ends with Section IV, an exposition on the conclusion.

II. MATERIALS AND METHODS

This section presents the experimental workflow, starting with data collection, as illustrated in Fig. 1. In data preparation, essential features and a dataset are extracted from the CGM sensor. The procedures and processes for data detersion i.e., data pre-processing, EDA, feature selection, are executed.ML algorithms are applied on the data detersion applied datasets for validation.

A. Data Collection and Cohort

Ethical clearance was obtained from the SRM Medical College Hospital and Research Centre, Kattankulathur-603203, Tamil Nadu, India (Ethical clearance number:-8274/IEC/2022). A publicly available ‘closed-loop control to range system’ public dataset was obtained from JCHR-JAEB center for health research which was coordinating center. The study was carried out in seven clinical centers (Sansum Diabetes Research Institute; USA, Montpellier University Hospital; France, Shafer Institute for Endocrinology and Diabetes; National Centre for Childhood Diabetes; Schneider Children’s Medical Centre of Israel; Sackler Faculty of Medicine; Israel, Barbara Davis Centre for Childhood Diabetes; Colorado) where ethical clearance was approved by respective review boards. The written informed consent was obtained from each patient or parent, with assent obtained as required. The full protocol is available online (www.clinicaltrials.gov/ct2/show/NCT01271023).The study is designed and conducted according to ethical principles that comply with in the Declaration of Helenski. In this work, patient data anonymization was strictly performed by omitting the patient’s name, address, and other personal details. The dataset for the proposed work was created considering the datetime and glucose values. Following are the study protocols followed by JDRF and the proposed work:

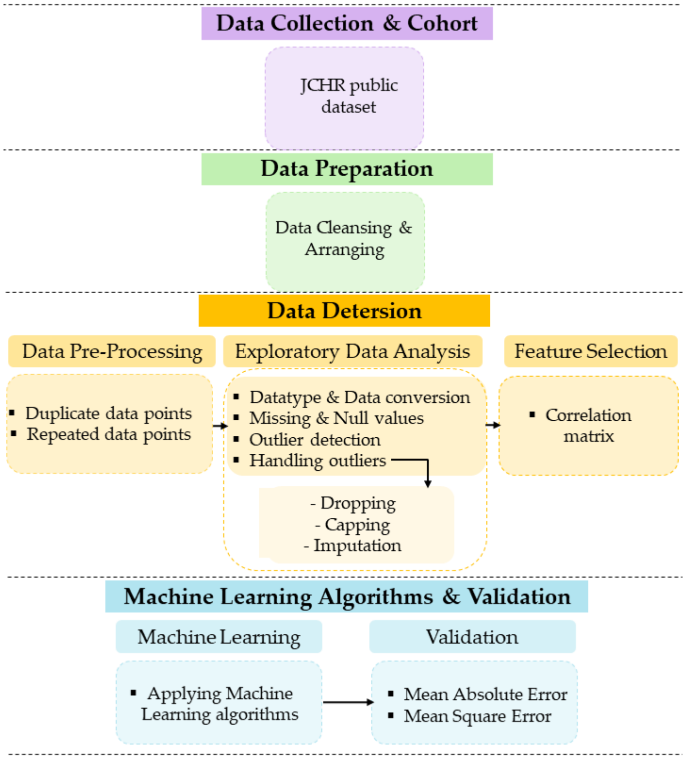


Fig. 1. Flow chart of data analyzation.

1) Eligibility: Clinically diagnosed T1DM patients for at least one year and using insulin pumps for at least six months are chosen for this study because the majority of TIDM patients are required to be on insulin pumps on a daily basis whereas T2DM and Gestational DM patients consume oral medications for regulating blood glucose levels. Patients with proper mental health and cognition for the study are chosen.

2) Sample size: In this work, sample size is chosen based on the patients whose blood glucose levels were improved. Many studies have chosen sample sizes of 13,20,25,56 [6,10,16-18] to predict insulin dosage. Therefore, glucose values at the time of bolus infusion, meal time, and amount of insulin dosage given are focused on a total sample size of 60 patients.

3) Inclusion and exclusion criteria: The inclusion criteria for this study are male and female groups aged 12 to 63 years who were on insulin pumps and CGM sensors without any break. Other age groups discontinued the treatment, and few were from the exclusion criteria. Pregnant and lactating women are excluded. Patients with diabetic ketoacidosis in the last six months, patients with Hypoglycemic episodes with unconsciousness, seizure disorder, and patients who have Coronary artery disease, active infection, muscular condition, and Cystic fibrosis are excluded due to the possibility of potential bias. The patient's name, address, and other personal particulars are entirely omitted.

B. Data Preparation

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations. Data from rapid-acting insulin type is considered in this work. A similar meal size is given to all the patients during lunch, breakfast, and dinner time. Bolus insulin dosage is given before 15 minutes of meal intake. Blood glucose levels are noted. In data preparation, features required for this work are selected from Dataset1. An instance of Dataset1 is presented in Table I. The dataset consists of P.ID, Age, Sex, BMI, CBG, bolus insulin given, bolus date time (Bdt), NBG, CR, CF, and Basal Infusion (BI) as features. NBG is the blood glucose value collected after 30 min and 60 min of bolus infusion. NBG and CBG are noted for every mealtime, i.e., breakfast, lunch, and dinner time. The shaded portion in the table depicts a record of a patient whose condition improved after bolus insulin treatment. The records where the blood glucose levels are improved (all shaded portion of the dataset) after 30 min of bolus infusion are chosen and created in a separate dataset, i.e., Dataset 2.A total of eight features, i.e., Age, Sex, BMI, CBG, bolus infused, NBG, CR, and CF, are selected from Dataset1 and created into Dataset2. The features are selected based on the previous works [5, 10-11,13-14,17-18] that align with predicting bolus insulin dosage for further processing.

C. Data pre-processing, Exploratory Data Analysis, and Feature Selection

Data pre-processing, EDA, and feature selection are executed on Dataset2 by implementing Python software.

1) Data pre-processing: Data pre-processing is the next step after data collection [19,20-21]. At this step, duplicate/repeated data points are removed. In this work, Dataset 2 is checked for duplicate and repeated values at each row. No duplicates or repeated values are found in Dataset2. Therefore, EDA is applied to the dataset.

2) Exploratory data analysis: ML models perform best after applying EDA on the dataset [19,22-23]. Therefore, in this work, EDA is considered the next step after data collection and pre-processing. The primary objective of EDA is to test the data for the nature of data distribution, outliers, anomalies, and complexity. It is a tool to visualize the data for manipulation. It helps in developing parsimonious models and implements clinically relevant variables [19,22]. EDA is applied on Dataset2, where the following steps are implemented:

TABLE I. AN INSTANCE OF DATASET1 CONSTRUCTED

P.ID	Age	Sex	BMI	Bolus Insulin at Breakfast					Bolus Insulin at Lunch					Bolus Insulin at Dinner					Carbohydrate Ratio	Correction Factor
				CBG	Bolus	Bdt	NBG (30min)	NBG (60min)	CBG	Bolus	Bdt	NBG (30min)	NBG (60min)	CBG	Bolus	Bdt	NBG (30min)	NBG (60min)		
4	39	M	33.7564	154	4.25	07-05-2011 9:47:35	110	92	200	2.35	07-05-2011 11:10:42	262	270	145	3.55	07-05-2011 19:07:41	134	129	7	90
12	*40*	*F*	*27.8896*	*134*	*11.25*	*14-06-2012 8:59:46*	*174*	*217*	*379*	*0.35*	*14-06-2012 13:07:10*	*298*	*200*	*110*	*11*	*14-06-2012 19:10:23*	*119*	*110*	*13*	*27*
25	*32*	*M*	*21.9612*	*128*	*4.4*	*01-04-2012 9:00:19*	*115*	*166*	*185*	*4.3*	*01-04-2012 13:01:06*	*182*	*195*	*130*	*4.25*	*01-04-2012 19:01:16*	*119*	*110*	*12*	*20*

P.ID=patient ID;BMI=body mass index;CBG=current blood glucose;Bdt=bolus datetime;NBG=next blood glucose;

a) Datatype and data conversion: The 'Sex' attribute is converted into an integer for ease of analysis. The data type and description of the dataset are thoroughly visualized. The dataset consists of float64 and int64 datatypes suitable for further processing.

b) Identifying missing / null values: The dataset was created by eliminating inactive sensor readings; therefore, null values are not found in the dataset. If a dataset consists of null values, they must be filled by calculating the mean for numerical data and mode for categorical data.

c) Detecting outliers: Outliers are data entry errors, measurement errors, experimental errors, and sampling errors. Outliers are detected using the Interquartile range (IQR) visualization method [24-25]. The Dataset2 is checked for outliers for each feature. Outliers are found at NBG and CBG values. Boxplots for CBG and NBG are depicted in Fig. 2. Outliers detected are detailed in Table IV. It can be observed from Fig. 2(a) that CBG consists of five outliers with a max outlier value of CBG at 383 for patients aged 52 and a min outlier value of CBG at 280 for patients aged 45 as detailed in Table II. It can be observed from Fig. 2(b) that NBG consists of four outliers with a max outlier value of NBG at 370 for a patient aged 52 and a min outlier value of NBG at 268 for a patient aged 12 as detailed in Table II.

IQR is calculated as,

$$IQR = Q3 - Q1 \qquad (1)$$

where,

IQR=Interquartile range, Q3=third quartile representing 75th percentile, Q1=first quartile representing 25th percentile

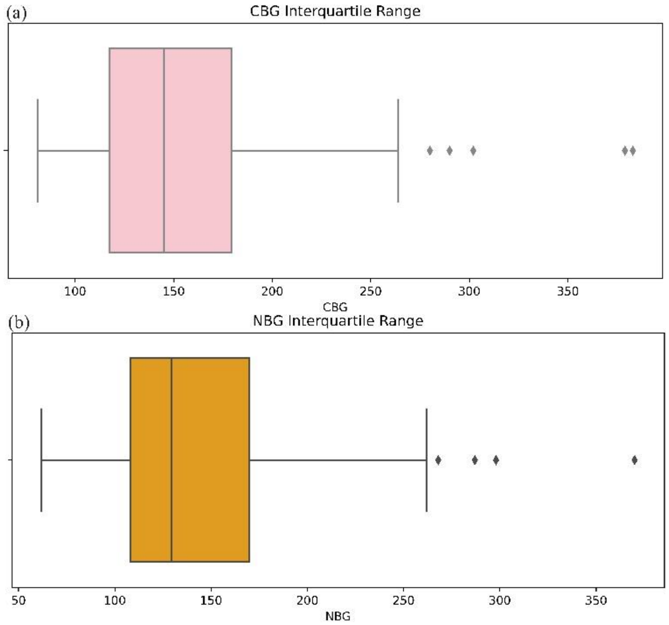


Fig. 2. Outlier detection using the IQR method. (a) Outlier detection in the CBG; (b) Outlier detection in the NBG.

The formula for the outer boundary limit is calculated as,

$$UBL = Q3 + (1.5 \times IQR) \quad (2)$$

$$LBL = Q1 - (1.5 \times IQR) \quad (3)$$

where,

UBL= Upper Boundary Limit, LBL= Lower Boundary Limit,1.5 is the decision range closer to the Gaussian distribution of outlier detection [30].

TABLE II. OUTLIER DETECTED FROM THE DATASET2

Attribute	No.of outliers	Max outlier value	Min outlier value
CBG	5	383 at age 52	280 at age 45
NBG	4	370 at age 52	268 at age 12

CBG=current blood glucose; NBG=next blood glucose

- Handling the outliers: The outliers are handled by considering the lower limit and upper limit boundaries from Eq. (6) and Eq. (7). The outliers are handled by three methods, i.e., dropping the outliers, capping the outliers, and imputing the outliers [26-27]. All approaches are applied on the Dataset2.
- Dropping the outliers: Outliers are dropped in this method. It is done by replacing outliers with a null value to differentiate from other data, and the null values are dropped. In this work, the outliers in the Dataset2 are first transformed into null values. The null values are then dropped from the dataset. A separate dataset ' droppedDataset2' file is created.
- Capping the outliers: The outliers are capped by setting a limit in the dataset. Capped values replace outliers identified above the upper limit and below the lower limit. The upper limit and lower limit is calculated from Eq. (4) and Eq. (5) as,

$$CUL = M + (3 \times SD) \quad (4)$$

$$CLL = M - (3 \times SD) \quad (5)$$

where,

CUL= Capping Upper Limit, CLL=Capping Lower Limit, M=Mean

Capped values on Dataset2 are detailed in Table III. It can be observed from Table III that on Dataset2, the lower limit for capped value is -41.69 for CBG and -46.69 for NBG. Any value falling below the lower limit will be capped at -41.64 for CBG, and -46.69 for NBG. Similarly, any value falling above the upper limit is capped at 362.67 for CBG 341.14 for NBG. The outliers from Dataset2 are capped, and a separate dataset is created as ‘cappedDataset2’.

TABLE III. CAPPING OUTLIERS ON DATASET2

Attribute	Capped lower limit value	Capped upper limit value
CBG	-41.64	362.67
NBG	-46.69	341.14

CBG=current blood glucose; NBG=next blood glucose

- Imputing the outliers: The imputation of outliers is carried out by identifying the upper and lower limits in the dataset. The mean value of the feature in the dataset replaces outliers found above the upper limit and below the lower limit. The upper limit and lower limit is calculated from Eq. (6) and Eq. (7) as,

$$UL = Dataset2 > [Q3 + (1.5 \times IQR).max()] \quad (6)$$

$$LL = Dataset2 > [Q3 - (1.5 \times IQR).min()] \quad (7)$$

where,

UL=Upper Limit, LL=Lower limit

TABLE IV. IMPUTING OUTLIERS ON DATASET2

Attribute	Mean value
CBG	140.35
NBG	126.86

CBG=current blood glucose; NBG=next blood glucose

Imputed values from Dataset2 can be observed in Table IV. It can be inferred from Table IV that any value falling above the upper limit and below the lower limit is imputed by mean value, i.e., 140.35 for CBG and 126.86 for NBG. A separate dataset 'imputedDataset2' file is created.

Further analysis is carried out on the three separately created datasets, i.e., droppedDataset2, cappedDataset2, and imputedDataset2.

Feature Selection: After applying EDA on droppedDataset2, cappedDataset2, and imputedDataset2, important features are chosen to increase the performance of a model. At this step, features are selected by implementing a heatmap correlation matrix. This work implements a correlation matrix for finding the related features and patterns in a dataset. The features are highly correlated if the heatmap value is close to 1 [28]. It can be visualized from Fig. 3(a), 3(b), and 3(c) that CBG and NBG are highly correlated, whereas bolus is the target variable. Hence, CBG, NBG, and bolus are selected features from the correlation matrix and CR from [10]. ANN is applied to the dataset to test the correlation matrix's performance. It can be observed from Table V that before using feature selection, MAE of 3.12 and MSE of 17.52 were obtained with ANN. Similarly, after applying feature selection, an MAE of 1.88 and MSE of 4.78 are achieved. An improvement of 39.7% on MAE and 36.9% is observed on the dataset after applying Feature selection. The droppedDataset2, cappedDataset2, and imputedDataset2 are refined by dropping uncorrelated features, i.e., Age, Sex, BMI, CF, and by selecting CBG, NBG, Bolus, and CR.

3) Choosing the Machine Learning Model: In a few studies, LR and logistic regression are implemented [29,12], whereas other notable models such as SVR, RF, RR, LAR, and gradient boosting are explored [5,9,11,29]. Some literature has explored ensemble methods, i.e., bagging, boosting, and DT [10] and ANN [10,12,13,14]. It was inferred from the studies that the performance of the ML algorithm depends on the type of the dataset and methodologies [5-14]. Therefore, in this proposed work, k-NN, k-NN bagging, k-NN boosting, DT, DT bagging, DT boosting, and ANN are compared for validation.

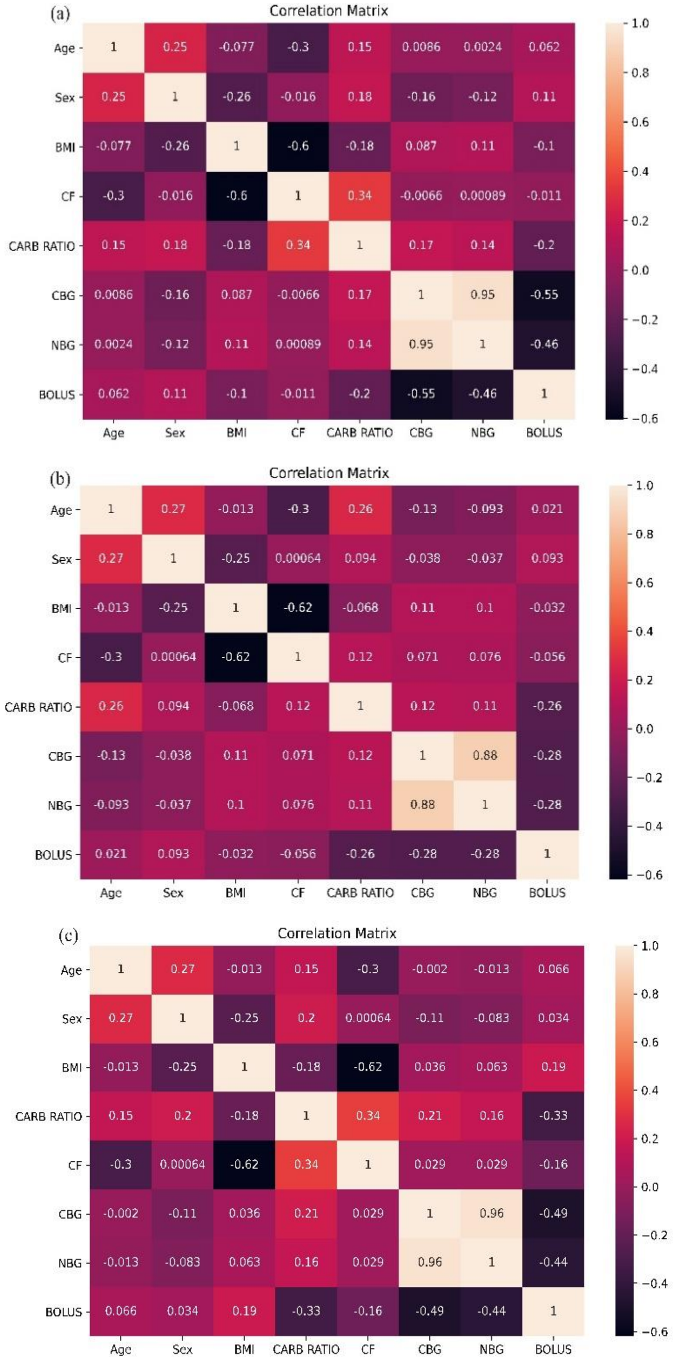


Fig. 3. Correlation matrix. heatmap of (a) Droppeddataset2; (b) Cappeddataset2; (c) Imputeddataset2.

TABLE V. PERFORMANCE METRICS OF ARTIFICIAL NEURAL NETWORKS BEFORE AND AFTER FEATURE SELECTION ON DATASET2

Dataset	MAE	MSE
Before Feature Selection	3.12	17.52
After Feature Selection	1.88	4.78

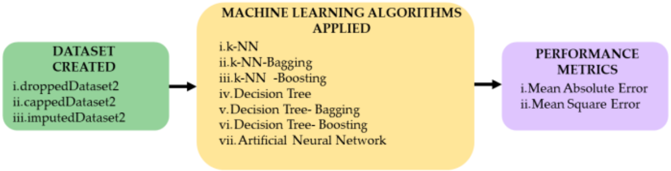
MAE=mean absolute error; MSE=mean squared error

III. RESULTS AND DISCUSSION

This section presents the results for predicting bolus insulin by applying ML algorithms on curated datasets, i.e., droppedDataset2, cappedDataset2, and imputedDataset2.

A. *Bolus Insulin Prediction Based on Current Blood Glucose and Improved Next Blood Glucose Levels*

The process flow of the work is depicted in Fig. 4. All three datasets are subjected to different ML algorithms for predicting bolus insulin. The models are validated by evaluating MAE and MSE. The performance of each dataset is compared with a recent work carried out to predict insulin [10].



DATASET CREATED
i.droppedDataset2
ii.cappedDataset2
iii.imputedDataset2

MACHINE LEARNING ALGORITHMS APPLIED
i.k-NN
ii.k-NN-Bagging
iii.k-NN -Boosting
iv.Decision Tree
v.Decision Tree- Bagging
vi.Decision Tree- Boosting
vii.Artificial Neural Network

PERFORMANCE METRICS
i.Mean Absolute Error
ii.Mean Square Error

Fig. 4. Illustration of pipelines from dataset, machine learning algorithms and performance metrics.

1) Performance Metrics on Data Detersion Applied Models: Metrics implemented to measure the curated models are MAE and MSE. Absolute Error (AE) is the difference between the target and the predicted value, as mentioned in Eq. (8). MAE is an average of AE, as mentioned in Eq. (9). Squared error (SE) is the difference between the square of the target and the predicted value. MSE is the average mean of SE as mentioned in Eq. (10). The performance metrics are given as,

$$AE = \left|BI_{pred} - BI_{tgt}\right| \tag{8}$$

$$MAE = \frac{1}{M}\sum_{I=1}^{M}\left|BI_{pred} - BI_{tgt}\right| \tag{9}$$

$$MSE = \frac{1}{M}\sum_{i=1}^{M}\left|BI_{tgt} - BI_{pred}\right|^{2} \tag{10}$$

where,

BI =Bolus insulin, M = number of observations, BI_{pred}= predicted bolus insulin level, BI_{tgt}= target bolus insulin level.

2) Predictive Analysis: The dataset consists of 60 samples where CR, CBG, NBG, and Bolus Insulin are considered as input features. The ML models are trained by splitting the dataset into 80% for training and 20% for testing. ML algorithms are applied to the dataset before and after implementing feature selection. It can be inferred from Table VI that the performance metrics are high before feature selection, and the dataset performed the best with an MAE of 1.88 and MSE of 4.78 after feature selection.

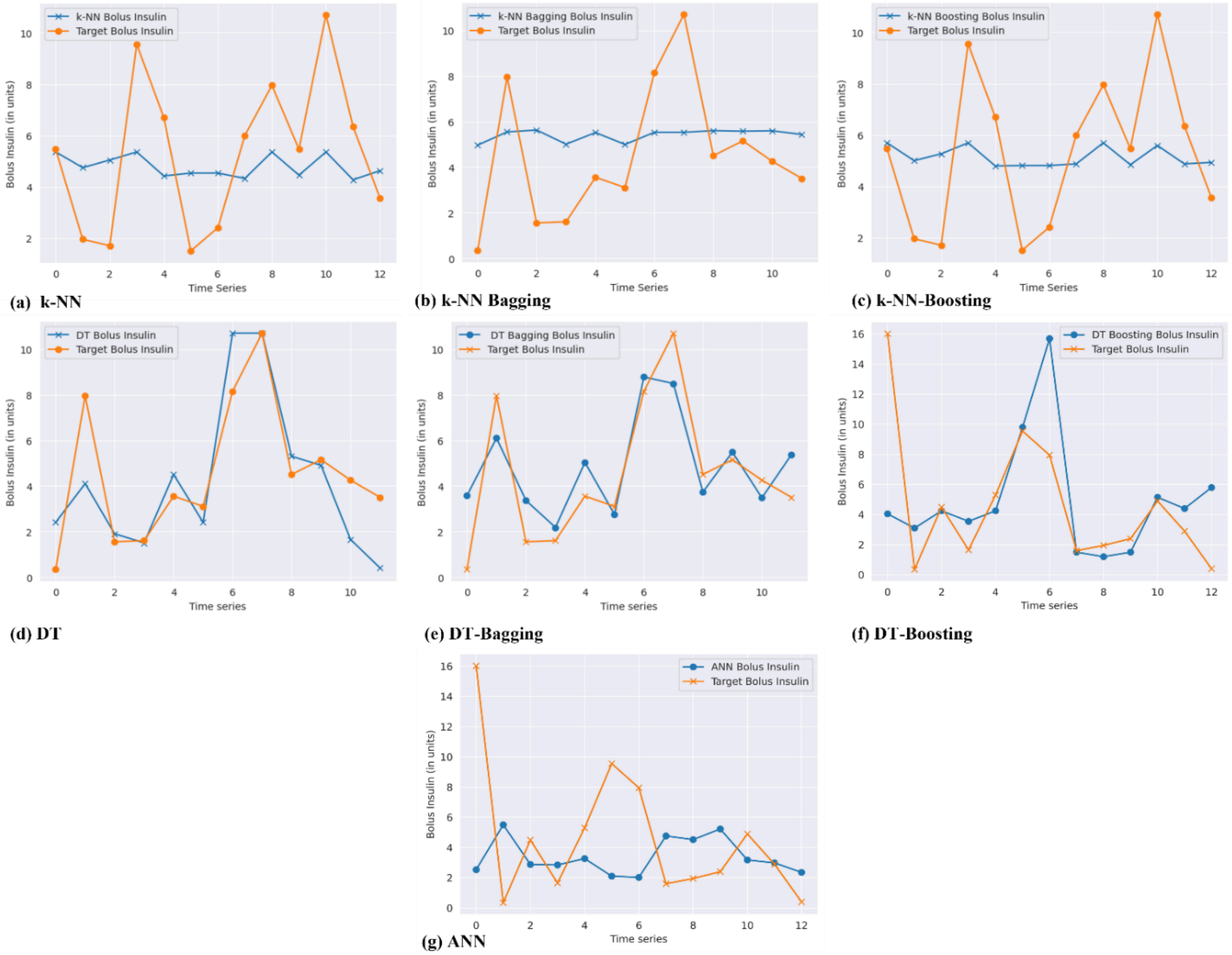
K-NN is a regression algorithm where the predicted dependent variable is the average of k-nearest neighbors [30]. k-NN is applied to the curated datasets. Total neighbors, i.e., n_neighbors=21, are considered with ‘uniform weights, ‘brute’ algorithm, and ‘Minkowski’ tree metric with power ‘p=2’. imputedDataset2 performed best with MAE as 2.43 and MSE as 7.40 when compared with droppedDataset2 where MAE as

2.40, MSE as 8.08, and cappedDataset2 with MAE as 2.61, MSE as 10.47 is achieved. A difference in the trend of target and predicted bolus insulin can be observed in Fig. 5(a). Therefore, the model cannot be recommended for prediction of bolus insulin.

TABLE VI. ANALYSIS OF CALIBRATION ON DIFFERENT MODELS

Reference		**Machine Learning Algorithms**							
		Metrics	***k-NN***	***k-NN -Bagging***	***k-NN -Boosting***	***DT***	***DT-Bagging***	***DT-Boosting***	***ANN***
Recent work [23]		MAE	2.33	2.36	2.40	2.57	2.41	2.45	**2.12**
		MSE	6.47	6.65	7.16	11.35	8.59	10.66	**5.79**
Proposed Work									
Before Feature Selection		MAE	3.06	2.95	3.85	2.42	2.03	**2.34**	3.12
		MSE	14.59	14.04	18.47	13.41	12.52	**12.14**	17.52
After Feature Selection		MAE	2.51	2.51	2.54	2.62	2.48	2.11	**1.88**
		MSE	8.39	8.20	8.45	13.41	7.28	11.51	**4.78**
Data Detersion Applied	droppedDataset2	MAE	2.40	2.34	2.33	2.06	2.10	2.19	**1.64**
		MSE	8.08	6.53	7.48	6.59	7.16	9.67	**4.16**
	cappedDataset2	MAE	2.61	2.63	2.80	2.39	**1.54**	2.13	3.65
		MSE	10.47	10.10	10.28	18.96	**4.15**	8.22	18.69
	imputedDataset2	MAE	2.43	2.63	**2.33**	**2.04**	2.12	2.22	2.75
		MSE	7.40	9.73	**7.25**	**5.13**	7.51	10.26	11.75

MAE=mean absolute error; MSE=mean squared error; DT=decision tree; ANN=artificial neural network



(a) k-NN

(b) k-NN Bagging

(c) k-NN-Boosting

(d) DT

(e) DT-Bagging

(f) DT-Boosting

(g) ANN

Fig. 5. Target and predicted bolus insulin from different models, I.E., (a) K-NN (b) K-NN Bagging (c) K-NN Boosting (d) DT (e) DT-Bagging (f) DT-Boosting (g) ANN.

K-NN with the Bagging-Ensemble algorithm combines two or more models [30]. A bagging regressor is applied on k-NN where the dataset is divided into many subsets, and the model is fitted on each subset independently. Predictions are made by aggregating individual predictions on the subsets [31-32]. K-NN is the base estimator with n_estimators=20. K-NN Bagging is applied on the curated datasets where droppedDataset2 performed best with MAE of 2.34, MSE of 6.53 when compared to the cappedDataset2 with MAE of 2.63, MSE of 10.10 and imputedDataset2 with MAE of 2.63, MSE of 9.73. droppedDataset2 performed best among curated datasets and even with comparison to recent work on predicting insulin [10]. The pattern of target and predicted bolus insulin on droppedDataset2 with k-NN Bagging can be observed in Fig. 5(b). It can be inferred that with an MAE of 2.34 and MSE of 6.53, the pattern of k-NN bagging is similar to k-NN with a decrease of 0.1 in MAE and a 0.87 increase in MSE. Differences can be observed between target and predicted bolus insulin. Therefore, this model cannot be recommended to predict bolus insulin.

K-NN with Boosting is an ensemble learning model learned from previous mistakes of weak classifiers sequentially [30]. The advantage of the model is to tune the weak into a robust model. It is an iterative method of increasing the efficiency of binary classifiers [31-32]. The base estimator is k-NN with n_estimators=100, a learning_rate of 0.3, and a ‘square’ loss. k-NN Boosting is applied on the curated datasets where imputedDataset2 performed best with MAE of 2.33 and MSE of 7.25 when compared to droppedDataset2 with MAE of 2.33, MSE of 7.48, and cappedDataset2 with MAE of 2.80, MSE of 10.28. The imputedDataset2 performed best in MAE with an increase of 0.9 in MSE when compared to [10]. The trend of target and predicted bolus insulin on imputedDataset2 can be observed in Fig. 5(c), where the pattern of target and predicted bolus insulin is similar to k-NN and k-NN Bagging. Therefore, this algorithm cannot be suggested for the prediction of bolus insulin.

DT model utilizes a set of binary rules to evaluate target value. Each tree has a simple model with branches, nodes, and leaves [33]. DT is applied on the curated datasets, i.e., where droppedDataset2 performed the best with MAE of 2.39, MSE of 18.96, and imputedDataset2 with MAE of 2.04, MSE of 9.13. droppedDataset2 performed best when compared to [10]. It can be inferred from Fig. 5(d) that the pattern of target and predicted bolus insulin performed better than other datasets. Therefore, it can be considered for the prediction of bolus insulin.

DT Model with Bagging is an ensemble model with DT as the base estimator where n_estimators=20.Bagging is applied on the curated datasets where the cappedDataset2 performed the best with MAE of 1.54 and MSE of 4.15 when compared to droppedDataset2 with MAE of 2.10 and MSE of 7.16 and imputedDataset2 with MAE of 2.12 and MSE of 7.51. Curated datasets performed the best compared to recent work on predicting insulin [10]. It can be inferred from Fig. 5(e), with MAE of 1.54 and MSE of 4.15, that the target and predicted bolus insulin follow a pattern. DT model with bagging can be implemented for predicting real-time insulin levels. This prediction is supportive of insulin pump therapy with minimum error.

DT Model with Boosting is an ensemble model with DT as the base estimator where n_estimators=20. Boosting is applied on DT to the curated datasets where capped Dataset2 performed best with MAE of 2.13 and MSE of 8.22 when compared to droppedDataset2 with MAE of 2.19 and MSE of 9.67, and imputedDataset2 with MAE of 2.22 and MSE of 10.26. The curated dataset performed better when compared to recent work [10]. It can be inferred from Fig. 5(f) that the model has a similar pattern to DT-Bagging, with an increase of 0.59 in MAE and 4.07 in MSE. As the former model, i.e., DT-Bagging, performs better than DT-Boosting, the former model can be considered for bolus insulin prediction.

ANN is applied where input dimensions of four, kernel initializer as ‘normal’ and ‘relu’ activation layer is considered. Hidden layers of 10 are considered with an epoch of 1000, batch size of 50, and verbose of 1. ANN is applied on the curated datasets where MAE and MSE obtained on droppedDataset2 are 1.64 and 4.16, performing the best compared to recent work [10]. MAE, MSE obtained on droppedDataset2 is 3.65, 18.69, and MAE, MSE obtained on imputedDataset2 is 2.75, 11.75. It can be observed from Fig. 5(g) that droppedDataset2 performed best when compared with cappedDataset2 and imputedDataset2. Due to the higher MAE and MSE of the ANN algorithm than DT with bagging, this model cannot be implemented for real-time prediction of bolus insulin.

DT with bagging performed the best with an MAE of 1.54 and MSE of 4.15. This model is recommended for predicting bolus insulin in real time. The findings of the proposed work are: (i) Feature selection plays a significant role in enhancing the performance of the dataset. An improvement of 39.7% (from 3.12 to 1.88) in MAE and 72.7% (from 17.52 to 4.78) in MSE with ANN is achieved after applying Feature Selection. (ii) The model's performance is enhanced with the data detersion process, where an improvement of 47.4% (from 3.12 to 1.64) in MAE and 76.2% (from 17.52 to 4.16) in MSE with the ANN model. (iii) Applying bagging and boosting enhanced the dataset's performance compared to non-bagging and boosting models. An improvement of 35.5% in MAE (from 2.39 to 1.54) and 78.1% MSE (from 18.96 to 4.15) with DT-Bagging is achieved. Similarly, 10% in MAE (from 2.39 to 2.13) and 56.6% MSE (from 18.96 to 8.22) with DT-Boosting is achieved. Therefore, it can be implied that the integration of AI and data science for the data detersion process boosts the performance of the models. The validation of the DT-Bagging model on cappedDataset2 is presented in the further section.

3) Validation of DT- Bagging Model: Bagging is an ensemble method combining several decision trees to optimize performance. DT with bagging architecture is presented in Fig. 6. The training dataset ‘T_r’is divided into several subsets of data, i.e., $T_{1......}$ T_n which can be chosen randomly with replacement. Multiple learning models are generated by training each learner in the ensemble structure with the subsets. The subset of data is implemented for training the decision tree. Prediction from each DT model, i.e.,

$DT\ Model_1\ DT\ Model_n$ are aggregated, and insulin dosage is finally predicted. DT-Bagging is derived in Eq. (11).

$$p(o) = \frac{1}{B_P}\sum_{n=1}^{B_p} p_n\,(o) \qquad (11)$$

where,

$p(o)$ =predicted output, B_P =bootstrapping sets,$p_n(o)$=weak learners

The plot against the target and predicted bolus insulin is depicted in Fig. 7. The X-axis represents the target bolus insulin, whereas the Y-axis represents the predicted bolus insulin. It can be inferred that the target and predicted bolus insulin data points are closer to the trendline, defining a high correlation. DT-Bagging is validated by performing an error analysis. Error analysis evaluates MAE between the target and predicted insulin levels, as illustrated in Table VII. The model is tested with a new dataset of 20 samples where 13 samples are tabulated in Table VII. It can be inferred that the maximum variance achieved is 2.49, and the minimum variance is 0.08, falling under the tolerance limit of ±5 from IEC60601-2-12 of insulin pump protocol [34]. As the performance of the model with MAE of 1.50 is in the clinically acceptable range, the developed model is suitable for deploying insulin pumps.

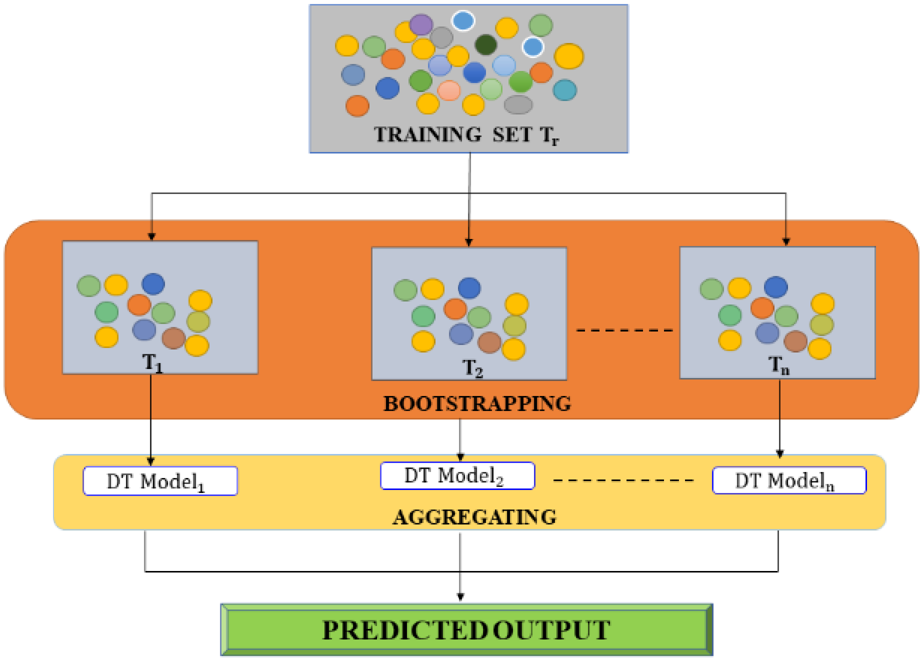


Fig. 6. Decision tree-bagging architecture.

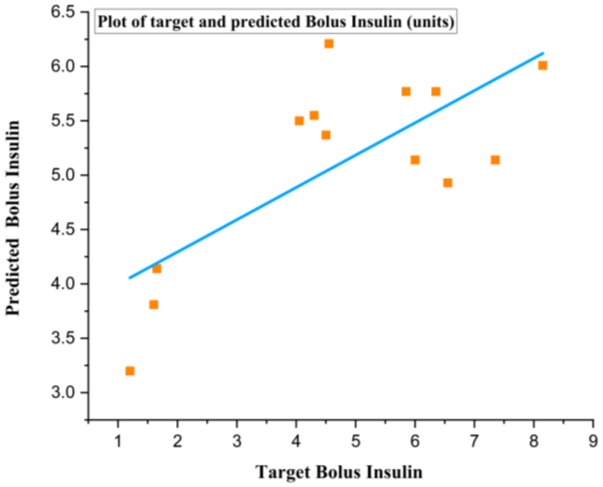


Fig. 7. The plot of target and predicted insulin levels on the proposed data detersion process on logcappeddataset2.

The proposed study is compared with previous approaches to insulin prediction in Table VIII. All the datasets curated from the data detersion process obtained the best performance with MAE and MSE compared to previous literature [9-11, 13].

The data detersion process proposed in the current work obtained the best performance with an MAE of 1.50 and MSE of 4.15.

The implications of the study outcomes are to make informed clinical decisions, treatment titrations, changes in lifestyle habits, and evidence-based dosage recommendations. It can be applied at the development stage of insulin clinical trials and drug dosage.

The model can be deployed in an insulin pump and can be integrated with the CGM device. Insulin dosage can be predicted in real-time based on blood glucose levels. The model can be deployed with a customized regimen considering an individual's health conditions and physical activity. The likelihood of successful outcomes due to improvement in treatment efficacy can be expected from the model's performance in bolus insulin prediction. Therefore, with the proposed work, adverse side effects such as hyperglycemia and hypoglycemia can be controlled, and balanced blood glucose levels can be achieved with better diabetes management.

TABLE VII. VALIDATION OF PROPOSED DATA DETERSION ON CAPPEDDATASET2

Target Insulin ($\boldsymbol{BI_{tgt}}$) units	Predicted Insulin ($\boldsymbol{BI_{pred}}$) units	Absolute Error$\lvert \boldsymbol{BI_{pred}} - \boldsymbol{BI_{tgt}} \rvert$ units
6	5.14	0.86
5.85	5.77	0.08
6.35	5.77	0.58
1.65	4.14	2.49
8.15	6.01	2.14
4.05	5.5	1.45
1.6	3.81	2.21
7.35	5.14	2.21
6.55	4.93	1.62
4.3	5.55	1.25
4.55	6.21	1.66
4.5	5.37	0.87
1.2	3.2	2
Mean Absolute Error (MAE) $\frac{1}{N}\sum_{I=1}^{N}\lvert BG_{pred} - BG_{tgt}\rvert$		1.50

BG_{pred}=predicted blood glucose; BG_{tgt}=target blood glucose

TABLE VIII. COMPARISON OF NON-INVASIVE APPROACHES IN NIR-SPECTROSCOPY WITH THE CURRENT STUDY

Reference	**Data Detersion Applied**	**Methodology**	**Performance Metrics**
Liu et al.[21]	No	Random Forest	MAE=4.1
Y.Obeidat, et al.[23]	k-NN Imputation	ANN	MAE=5.79
Nguyen et al.[24]	No	Ensemble Machine Learning algorithm	MAE=12
Zahran et al.[26]	No	ANN	Prediction error=4%
Proposed Work			

droppedDataset 2	Dropping	ANN	MAE=1.64
			MSE=4.16

logcappedDataset2	Capping	DT-Bagging	MAE=1.54
			MSE=4.15

imputedDataset 2	Imputation	k-NN-Bagging	MAE=2.12
			MSE=7.51

ANN=artificial neural network

IV. CONCLUSION

The strength of the proposed work is in (i) Bolus insulin prediction from CBG and improved NBG from previous literature are implemented in the current study [23]. (ii) Feature Selection is done to select correlating features between independent and dependent variables. An improvement of 37.9% is observed before and after applying Feature Selection on MAE and MSE from the DT-Bagging algorithm. (iii) Implementing Bagging on DT has improved the performance by 15% in both MAE and MSE, thus enhancing the model's performance. To understand the performance of the original dataset, ML algorithms are applied after which feature engineering is implemented. This attempt was to analyze if feature engineering could make any improvement in the prediction. To improve the performance after feature engineering, the original dataset was subjected to three ways of data detersion process to cure the data on which ML algorithms are applied. The limitation of the proposed work is the size of the dataset created. As CBG and improved NBG are considered from the dataset of 24,170 rows of bolus infusion, only 60 data showed improvement in NBG levels. Therefore, the model is built on a small dataset of size 60. To deploy the algorithm in a real-time scenario in an insulin pump, uncertainties and artifacts such as integration with CGM device and other health complications. The study is conducted only on T1DM with insulin pumps of at least six months and excluded patients with Diabetic and Coronary disease complications, making the proposed study less generalizable to a large population. Future work is to create a model on the massive dataset by considering CBG and improved NBG levels from different public datasets and predict bolus insulin dosage.

ACKNOWLEDGMENT

The authors are grateful and want to thank the JDRF Artificial Pancreas Project Consortium-Public dataset archive, for allowing them to use the datasets and giving the consent to publish research findings from the study.

Funding: This research received no external funding

Data Availability Statement: The data that support the findings of this study are openly available as 'Inpatient Evaluation of an Automated Closed-Loop Control-to-Range System (CTR)' in [JCHR] at [https://public.jaeb.org/datasets/diabetes][35].

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

[1] American Diabetes Association. "Diagnosis and Classification of Diabetes Mellitus." Diabetes Care, vol. 33, no. Supplement_1, 30 Dec. 2010, pp. S62–S69, www.ncbi.nlm.nih.gov/pmc/articles/PMC2797383/, https://doi.org/10.2337/dc10-s062.

[2] Freeman, Andrew M, and Nicholas Pennings. "Insulin Resistance." Nih.gov, StatPearls Publishing, 2019, www.ncbi.nlm.nih.gov/books/NBK507839/.

[3] Benyó, Balázs, et al. "Classification-Based Deep Neural Network vs Mixture Density Network Models for Insulin Sensitivity Prediction Problem." Computer Methods and Programs in Biomedicine, vol. 240, 1 Oct. 2023, p. 107633, www.sciencedirect.com/science/article/pii/S0169260723002985, https://doi.org/10.1016/j.cmpb.2023.107633. Accessed 18 Oct. 2023.

[4] Aiello, Eleonora M., et al. "A Novel Model-Based Estimator for Real-Time Prediction of Insulin-On-Board." Chemical Engineering Science, vol. 267, 5 Mar. 2023, p. 118321, www.sciencedirect.com/science/article/pii/S000925092200906X, https://doi.org/10.1016/j.ces.2022.118321.

[5] "An Expertise System for Insulin Dosage Prediction Using Machine Learning Techniques." IJIREEICE, ijireeice.com/papers/an-expertise-system-for-insulin-dosage-prediction-using-machine-learning-techniques/.

[6] Zhu, Taiyu, et al. “An Insulin Bolus Advisor for Type 1 Diabetes Using Deep Reinforcement Learning.” *Sensors*, vol. 20, no. 18, 6 Sept. 2020, p. 5058, https://doi.org/10.3390/s20185058.

[7] Gupta, Ketan, and Nasmin Jiwani. “Prediction of Insulin Level of Diabetes Patient Using Machine Learning Approaches.” *Papers.ssrn.com*, 18 Jan. 2022, papers.ssrn.com/sol3/papers.cfm?abstract_id=4205251.

[8] Nguyen, Minh, et al. “Machine Learning for Initial Insulin Estimation in Hospitalized Patients.” *Journal of the American Medical Informatics Association*, vol. 28, no. 10, 19 July 2021, pp. 2212–2219, https://doi.org/10.1093/jamia/ocab099.

[9] Reddy & Shashil,. “Machine Learning for Initial Insulin Dosage Prediction in Hospitalized Patients.” *Journal of Engineering Sciences*. 2022,vol.13,no.3, ISSN num-ber:0377-9254.

[10] “A System for Blood Glucose Monitoring and Smart Insulin Prediction | IEEE Journals & Magazine | IEEE Xplore.” *Ieeexplore.ieee.org*, ieeexplore.ieee.org/document/9393950/. Accessed 17 Dec. 2023.

[11] Pesl, Peter, et al. “An Advanced Bolus Calculator for Type 1 Diabetes: System Architecture and Usability Results.” IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 1, Jan. 2016, pp. 11–17, https://doi.org/10.1109/jbhi.2015.2464088. Accessed 7 June 2022.

[12] Eleftheriades, Makarios, et al. “Prediction of Insulin Treatment in Women with Gestational Diabetes Mellitus.” Nutrition & Diabetes, vol. 11, no. 1, June 2021, https://doi.org/10.1038/s41387-021-00173-0. Accessed 9 Nov. 2021.

[13] Zahran, Bilal. “A Neural Network Model for Predicting Insulin Dosage for Diabetic Patients”. *The International Journal of Computer Science and Information Security* (IJCSIS).2016, 14.

[14] Cappon, Giacomo, et al. “A Neural-Network-Based Approach to Personalize Insulin Bolus Calculation Using Continuous Glucose Monitoring.” Journal of Diabetes Science and Technology, vol. 12, no. 2, Mar. 2018, pp. 265–272, https://doi.org/10.1177/1932296818759558.

[15] Battelino, Tadej, et al. “Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations from the International Consensus on Time in Range.” Diabetes Care, vol. 42, no. 8, 8 June 2019, pp. 1593–1603, care.diabetesjournals.org/content/42/8/1593, https://doi.org/10.2337/dci19-0028.

[16] De Farias, João Lucas Correia Barbosa, and Wallace Moreira Bessa. “Intelligent Control with Artificial Neural Networks for Automated Insulin Delivery Systems.” Bioengineering, vol. 9, no. 11, 8 Nov. 2022, p. 664, https://doi.org/10.3390/bioengineering9110664. Accessed 3 May 2023.

[17] Guzman Gómez, Guillermo Edinson, et al. “Application of Artificial Intelligence Techniques for the Estimation of Basal Insulin in Patients with Type I Diabetes.” International Journal of Endocrinology, vol. 2020, 2020, p. 7326073, pubmed.ncbi.nlm.nih.gov/33204261/, https://doi.org/10.1155/2020/7326073. Accessed 17 Dec. 2023.

[18] “A Fuzzy Logic Based Approach for the Adjustment of Insulin Dosage for Type 1 Diabetes Patients.” Www.bracu.ac.bd, 5 Feb. 2018, www.bracu.ac.bd/fuzzy-logic-based-approach-adjustment-insulin-dosage-type-1-diabetes-patients. Accessed 17 Dec. 2023.

[19] Komorowski, Matthieu, et al. “Exploratory Data Analysis.” PubMed, Springer, 2016, pubmed.ncbi.nlm.nih.gov/31314267/. Accessed 17 Dec. 2023.

[20] Abhishekmamidi. “Exploratory Data Analysis and Data Pre-processing Steps”. www.abhishekmamidi.com/2019/08/exploratory-data-analysis-and-data-preprocessing-steps.html.

[21] Nguyen, Leah. “EDA, Data Preprocessing, Feature Engineering: We Are Different!” Medium, 1 Apr. 2022, medium.com/@ndleah/eda-data-preprocessing-feature-engineering-we-are-different-d2a5fa09f527.

[22] Chatfield, Chris. “Exploratory Data Analysis.” European Journal of Operational Research, vol. 23, no. 1, Jan. 1986, pp. 5–13, https://doi.org/10.1016/0377-2217(86)90209-2. Accessed 26 Mar. 2019.

[23] Pramanik, Jitendra el.at, “Exploratory Data Analysis using Python”.*International Journal of Innovative Technology and Exploring Engineering*. pp. 4727–4735.2019.

[24] Payne, Walker. “How to Analyze Blood Glucose Data with Python Data Science Packages.” Medium, 1 Dec. 2021, towardsdatascience.com/how-to-analyze-blood-glucose-data-with-python-data-science-packages-4f160f9564be.

[25] Bergenstal, Richard M. “Understanding Continuous Glucose Monitoring Data.” PubMed, American Diabetes Association, 2018, www.ncbi.nlm.nih.gov/books/NBK538967/.

[26] Rawlings, Renata A., et al. “Translating Glucose Variability Metrics into the Clinic viacOntinuousGLucoseMOnitoring: AGRaphicalUSerINterface forDIabetesEValuation (CGM-GUIDE©).” Diabetes Technology & Therapeutics, vol. 13, no. 12, Dec. 2011, pp. 1241–1248, https://doi.org/10.1089/dia.2011.0099.

[27] Czerwoniuk, Dorota, et al. “GlyCulator: A Glycemic Variability Calculation Tool for Continuous Glucose Monitoring Data.” Journal of Diabetes Science and Technology, vol. 5, no. 2, Mar. 2011, pp. 447–451, https://doi.org/10.1177/193229681100500236.

[28] Da Poian Victoria, Theiling Bethany et.al, “Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry”. Frontiers in Astronomy and Space Sciences.(10). 2023.https://doi.org/10.3389/fspas.2023.1134141.

[29] Cappon, Giacomo, et al. “A Neural-Network-Based Approach to Personalize Insulin Bolus Calculation Using Continuous Glucose Monitoring.” Journal of Diabetes Science and Technology, vol. 12, no. 2, Mar. 2018, pp. 265–272, https://doi.org/10.1177/1932296818759558.

[30] Mailagaha Kumbure, Mahinda, and Pasi Luukka. “A Generalized Fuzzy K-Nearest Neighbor Regression Model Based on Minkowski Distance.” Granular Computing, 25 Sept. 2021, https://doi.org/10.1007/s41066-021-00288-w.

[31] Bühlmann, Peter. “Bagging, Boosting and Ensemble Methods.” *Handbook of Computational Statistics*, 21 Dec. 2011, pp. 985–1022, https://doi.org/10.1007/978-3-642-21551-3_33.

[32] Baskin, Igor I., et al. “Bagging and Boosting of Regression Models.” Tutorials in Chemoinformatics, 23 June 2017, pp. 249–255, https://doi.org/10.1002/9781119161110.ch16.

[33] Celal Bayar, Manisa, et al. “Makale Hakkında Regression Analyses or Decision Trees?” *Journal of Social Sciences*, vol. 18, no. 4, 2020, dergipark.org.tr/en/download/article-file/1295121.

[34] “IEC 60601-2-24:2012 | IEC Webstore.” Webstore.iec.ch, webstore.iec.ch/publication/2635.

[35] “PUBLIC STUDY WEBSITES.” *Public.jaeb.org*, public.jaeb.org/datasets/diabetes.