Indian Journal of Engineering & Materials Sciences
Vol. 28, June 2021, pp. 240-249

Implementation of machine learning model-based to predict T2DM risk using heart rate variability features

Shashikant Rajaram Rathod[a*], Uttam Chaskar[a], Leena Phadke[b] & Chetan Kumar Patil[a]

[a]Department of Instrumentation and Control Engineering, College of Engineering, Pune 411 005, India

[b]Department of Physiology, Smt. Kashibai Navale Medical College and General Hospital, Pune 411 041, India

Received: 23 February 2021; Accepted: 05 May 2021

Non-invasive early diabetes prediction has been gaining much premarkable over the last decade. Heart rate variability (HRV) is the only non-invasive technique that can predict the future occurrence of the disease. Early prediction of diabetes can help doctors start an early intervention. To this end, the authors have developed a computational machine learning model to predict type 2 diabetes mellitus (T2DM) risk using heart rate variability features and have evaluated its robustness against the HRV of 50 patients data. The electrocardiogram (ECG) signal of the control population (n=40) and T2DM population (n=120) have been recorded in the supine position for 5 minutes, and HRV signals have been obtained. The time domain, frequency domain, and non-linear features have been extracted from the HRV signal. A decision support system has been developed based on a machine learning algorithm. Finally, the decision support system has been validated using the HRV features of 50 patients (Control n=10 and T2DM n=40). HRV features are selected for the prediction of T2DM. The decision support system has been designed using three machine learning models: Gradient boosting decision tree (GBDT), Extreme Gradient boosting (XGBoost), Categorical boosting (CatBoost), and their performance have been evaluated based on the Accuracy (ACC), Sensitivity (SEN), Specificity (SPC), Positive predicted value (PPV), Negative predicted value (NPV), False-positive rate (FPR), False-negative rate (FNR), F1 score, and Area under the receiver operating characteristic curve (AUC) metrics. The CatBoost model offers the best performance outcomes, and its results have been validated on 50 patients. Thus the CatBoost model can be use as a decision support system in hospitals to predict the risk of T2DM.

Keywords: Heart rate variability, Type 2 diabetes mellitus, Gradient boosting decision tree, Extreme gradient boosting, Categorical boosting

1 Introduction

Diabetes mellitus is usually known as diabetes. The primary reason behind this disease is that the body is unable to metabolize glucose properly. Some researchers in diabetes have reported that from 1980 to 2014, diabetes disease has risen from 4.7% to 8.5%. The number will further increase to 25%, 51% in 2030 and 2045, respectively[1]. There are three types of diabetes: • Type 1 diabetes which is found in children; in this type, the pancreas cannot produce insulin[2,3] • Type 2 diabetes which is a common type of diabetes found in adults. Around 85% of the world population has this type of diabetes. It occurs if the body cannot convert glucose into energy due to a lack of insulin production[4]. • Gestational diabetes which is found in pregnant women due to inadequate insulin secretion[5]. Although there is no proper cure for diabetes, it can be prevented and controlled if early indications of diabetes is possible. An early indication of the disease is possible using the HRV signal. HRV shows time variation between RR intervals of ECG signals. The pacemaker of heart is the sinoatrial node (SA) which generates cardiac impulses influenced by the autonomous nervous system (ANS). ANS consists of sympathetic nervous system (SNS) and parasympathetic nervous system (PNS), which control the heart rate. SNS and PNS balance the normal heart rate which is strongly influenced by different body factors. Therefore, the status of ANS can be analyzed using the HRV signal. If any disease developed in the body, it might affect ANS, so HRV gets affected, and an early indication of that disease is possible. HRV is a simple and non-invasive measurement technique and indicates stages of the diseases. In this study, HRV signal has been analyzed mainly using machine learning algorithm. The main objective of the study is to design a machine learning-based decision support system for physicians using HRV features that can be used as an initial screening test tool to predict T2DM risk. The contributions of the present study are

*Corresponding author (E-mail: rshashikant.bme@yahoo.com)

comprehensive analysis between the three types of HRV methods, namely time domain, frequency domain, and non-linear method. We have been concern about the breathing rate (BR) effect on HRV for diabetes prediction. The time domain, frequency domain, and non-linear features of HRV, along with the breathing rate have been used to predict the risk of T2DM. We have proposed a machine learning model as a decision support system for the prediction of T2DM risk. The proposed model is based on real and authentic HRV data. We have proposed a cost-effective screening tool to detect the risk of T2DM patients. Contribution to the validation of the machine learning model on patients (n=50) in the hospital.

2 Materials and Methods

2.1 Participants and data collection

The present study was performed by following the protocol given by the Taskforce of the European Society for Cardiology and the North American Society of Pacing and Electrophysiology[6]. The study was conducted at Smt. Kashibai Navale Medical College and General Hospital (SKNMCGH) Pune, Maharashtra, India, in collaboration with College of Engineering Pune (COEP), Pune India. The institutional ethical committee of SKNMCGH has approved the study. The participants were selected from the OPD of SKNMCGH by following American Diabetic Association guidelines[7]. Patients with a history of any acute or chronic diseases were safely excluded from the study. Before the electrocardiogram (ECG) recording, the procedure and objective of the study were informed to the participant, and informed consent was received. The ECG of selected subjects was recorded in the supine position for 15 minutes, and the last 5-minute segment was used for HRV analysis as shown in Fig. 1. The ECG was recorded using the data acquisition tool Chronovisor HRV DX system at sampling frequency 1000Hz, and HRV was analyzed using Chronovisor HRV software suite 1.1.487.

2.2 Feature extraction

In this section, the time domain, frequency domain, and non-linear features used for T2DM risk prediction have been discussed. The HRV features were derived using the RR time series interval of ECG signals.

2.2.1 Time domain features

In the time-domain analysis, simple statistical features were derived. The mean HR and mean RR features were obtained from the RR interval. The RR interval variability was represented using a standard deviation of normal to normal interval (SDNN) and root mean square standard deviation (RMSSD) features as presented in Table 1. Another important feature is the breathing rate (BR), which shows the effect of respiration on HRV and is represented in beats per minute. It is very important to consider BR while analyzing the HRV. The statistical difference between control and T2DM subjects was calculated using the Mann-Whitney U test. Statistical software tool Epi. Info. 7 was used for data analysis. In the present study, a p-value <0.05 was considered statistically significant, and the data were presented in the form of mean ± standard deviation[8,9].

2.2.2 Frequency domain feature

In the frequency domain method, a power spectral density estimator (PSD) calculates the frequency

Fig. 1 — ECG recording in a supine position.

Table 1 — Time-domain features

Features	Control (n=40)	T2DM (n=120)	p-Value
mean HR	71.73 ± 9.95	79.88 ± 12.09	0.0001#
mean RR	851.28 ± 111.26	767.84 ± 114.41	0.0001#
SDNN	71.82 ± 33.70	31.26 ± 15.08	<0.0001#
RMSSD	77.21 ± 54.69	70.05 ± 37.89	0.9529
BR	12.92 ± 2.71	15.15 ± 3.31	0.0001#

#-Significant difference, n= Number of samples, Mean HR- Average heart rate, Mean RR- Average RR interval, SDNN- Standard deviation of NN interval, RMSSD- Root mean square standard deviation, BR- Breathing rate

component of the RR interval series. The PSD estimation can be carried out using the Fast Fourier Transform method (FFT) or the Autoregressive modeling method (AR model). However, considering the complexity of the AR model, the results of the FFT method were used. The PSD estimator decomposes the RR interval into a frequency component using the FFT method[9]. The power in the frequency range of 0.04 to 0.15Hz is defined as a low-frequency power band. The power in the frequency range of 0.15 to 0.4Hz is defined as a high-frequency power band. The power in the frequency band was calculated in absolute (ms^2) and normalized unit (nu). The LF and HF power reflect the sympathetic and parasympathetic activity. In this study, total power (TP), LF power, HF power, LF nu, HF nu, and LF to HF power ratio were analyzed. The statistical difference between frequency-domain features of control and T2DM has been mentioned in Table 2.

2.2.3 Non-linear features

The nature of the biological signal is non-linear. Thus, the study of non-linear dynamics is important for analysis. The non-linear feature used in this study is as follows: Poincare plot, Detrended fluctuation analysis (DFA), Approximate entropy (AppEN), and Sample entropy (SampEN). The significant difference between non-linear features of control and T2DM subjects are presented in Table 3.

2.2.3.1 Poincare plot

The Poincare plot represents the present RR interval and the next RR interval, which shows the non-linear behaviors of RR interval variability[9]. The Poincare plot can be interpreted using standard deviation 1 (SD1), representing the short-term variability in RR interval, and standard deviation 2 (SD2), representing the long-term variability in RR interval[10].

2.2.3.2 Detrended fluctuation analysis (DFA)

DFA is used to assess the self-similarity properties of the RR interval. It also measures the correlation between different time series signals[11]. The fluctuation in the time series signal is represented by parameter α. Alpha (α) is called as the scaling exponent. The time-series signal is integrated and divided into segments of length n and $x_n(k)$ a least-square line is applied to each segment. Next, the integrated time series $x(k)$ is detrended from the next least square line of each segment. The detrended time series of RR interval is calculated by:

$$F(n) = \sqrt{\frac{1}{N}\sum_{K=1}^{N}(x(k) - x_n(k))^2} \quad \ldots (1)$$

DFA is plotted on a double log graph, and a linear relationship indicates the presence of fractal scaling. The value of α is closer to 1 for control subjects, and it may vary with disorders.

2.2.3.3 Approximate entropy (AppEN)

AppEN measures the regularity of the time-series signal. The value of AppEN is larger in the case of a control subject compared to diabetes subjects[12,13]. It is calculated as:

Table 2 — Frequency domain features

Features	Control (n=40)	T2DM (n=120)	*p*-Value
TP	3289.23 ± 2817.26	706.99 ±758.30	<0.0001#
LF power	807.54 ± 688.18	154.02 ± 194.49	<0.0001#
HF power	933.58 ± 997.94	219.53 ± 372.32	0.0001#
LF nu	51.08 ± 17.22	45.45 ± 15.78	0.039#
HF nu	48.19 ± 17.22	59.55 ± 62.89	0.0516
LF/HF	1.39 ± 0.99	1.59 ± 5.47	0.0477#

#-Significant difference, n= Number of samples, TP- Total power, LF- Low frequency, HF- High frequency, nu- Normalized unit

Table 3 — Non-linear features

Features	Control (n=40)	T2DM (n=120)	*p*-Value
SD1	42.76 ± 26.69	21.70 ±12.78	0.0001#
SD2	90.63 ± 42.22	38.06 ± 19.13	0.0001#
DFA α1	0.92 ± 0.24	0.89 ± 0.17	0.4067
DFA α2	1.13 ± 0.16	1.13 ± 0.20	0.5596
AppEN	1.45 ± 0.31	1.35 ± 0.33	0.1233
SampEN	42.76 ±26.69	21.70 ± 12.78	0.0001#

#-Significant difference, n= Number of samples, SD- Standard deviation, DFA- Detrended fluctuation analysis, AppEN- Approximate entropy, SampEN- Sample entropy

$$AppEN(m,r,N) = \frac{1}{N-m+1}\sum_{i=1}^{N-m+1} logC_i^m(r) - 1N-mi=1N-mlogCim+1(r) \quad \ldots (2)$$

where, C_m^i is the correlation integral

2.2.3.4 Sample entropy (SampEN)

SampEN measures the complexity of the time-series signal. It is actually like approximate entropy but a more refined version[12,13]. Higher values of SampEN represent the more irregularity in a time-series signal. Irregularity in a time series signal indicates the status of a patient. In the control subject, it was noted that irregularity is more as compared to diabetes subject. The sample entropy values can be calculated by:

$$SampEN(k,r,N) = -ln\left[\frac{A(k)}{B(k-1)}\right] \quad \ldots (3)$$

The variable $A(k)$ and $B(k)$ for all lengths k up to m and keeps track of template matches. In this study, $k = 0,1,2,\ldots\ldots m-1$ with $B(0) = N$, the length of the input time-series signal.

2.3 Dataset introduction

The present study uses two in-house HRV datasets. The dataset D1 comprises a control subject (n=40) aged 27.37 ± 6.73 and the diabetes subject (n=120) aged 53.59 ± 11.01. The dataset D2 comprised a control subject (n=10) aged 26.4 ± 6.29, and the diabetes subject (n=40) aged 54.60 ± 8.94. The dataset D1 is used to train the machine learning model, whereas dataset D2 is used to evaluate the performance of the machine learning model trained using dataset D1.

2.4 Data pre-processing

The preprocessing step includes outlier rejection and normalization of the dataset, which are described as follows:

The outliers are the observation that is deviated from their normal range[14]. The first criteria to remove outlier from the dataset is - 1). The data that falls outside the $\bar{x} \pm 3\sigma$ and 2). The data that falls outside of 1.5 times of an interquartile range, above 3rd quartile and below 1st quartile, are considered an outlier. The mathematical formulation of the outlier rejection system is written as-

$$N(x) = \begin{cases} x, & If \ > \bar{x} \pm 3\sigma \\ x, & If\ Q_1 - 1.5 \times IQR \times \leq x \leq Q_3 + 1.5 \times IQR \\ Reject, & Otherwise \end{cases} \quad \ldots (4)$$

where, x represents the attribute of the feature vector in the $n-$dimensional space, $x \in S^n$. $\bar{x}, \sigma, Q_1, Q_3, IQR$ is the average, standard deviation, first quartile, third quartile, and interquartile range of $x \in S^n$.

The min-max normalization technique was used for normalization, *i.e.*, rescaling the attribute value between zero mean and unit variance to achieve the normal distribution[15]. The mathematical representation of Min-max normalization is as follows-

$$Z(x) = \frac{x - x_{min}}{x_{max} - x_{min}}(p-q) + q \quad \ldots (5)$$

where, x is the n - dimensional attribute of the feature vector, $x \in S^n$, p is the new maximum value, and q is the new minimum value

2.5 Machine learning models

The dataset consists of a control subject and a diabetes subject. Thus, this problem can be considered as a binary classification assignment. The python data manipulation tool was used for implementing a machine learning model. We used three ensemble boosting algorithms in our study. The details about the algorithm are as follows:

2.5.1 Gradient boosting decision tree (GBDT)

Friedman extended the adaptive boosting concept by introducing the Gradient boosting decision tree (GBDT)[16]. The goal of GBDT is to improve the performance of the model by updating the weight of the samples during the training process. GBDT involves three elements: 1) optimization of the loss function, 2). Use of a weak learner to improve the performance, 3) develops an additive model to minimize the loss function.

The GBDT model initialized with the constant value, which minimizes the loss function. In each iteration of the training process, a negative gradient of the loss function is assessed as the residual value of the present model, and a new tree is formed to fit the residual. This new tree is trained to fit the present residual and added to the previous model. After this process, the residual is updated, and the process continues iteratively until the user-set maximum number of iteration conditions is reached. The GBDT algorithm is as follows:

Algorithm 1 — Gradient boosting decision tree
Input: Training set $\{(x_1,y_1),(x_2,y_2)\ldots\ldots(x_N,y_N)\}, x \in S^n$, y $\in S^n$, Loss function $L(y,f(x))$,
Output: Updated new tree $f_1(x)$

(1) Initialize model with a constant value $f_0(x) = argmin\ \sum_{i=1}^{n} L(y, \alpha)$;

(2) For m= 1, 2, 3….M

(a) Calculate the residual $res = -\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}$, $i = 1,2,3 \ldots . N$

(b) Fit a tree C_{mf} with the residual res using the training set $\{x_i, res)\}$

(c) $\alpha_{mf} = argmin \sum_{i=1}^{n} L(y_i, f_{m-1}(x_i) + \alpha)$

(d) Update the current model by $f_m(x) = f_{m-1}(x) + C_{mf}\, \alpha_{mf}$

(3) $f_1(x) = f_m(x)$

2.5.2 Extreme gradient boosting (XGBoost)

The XGBoost is a more regularized form of Gradient boosting proposed by Tianqi Chen in 2016[17]. The mathematics behind XGBoost is the same as the GBDT, but it shows the improved speed of tree construction and tree searching. The strength of XGBoost is parallel computing, L1 and L2 regularization, and second-order derivative of the loss function. It uses the advanced regularization of L1 and L2 algorithms, which improves the generalization capabilities of the model. The optimized XGBoost model with the N decision tree is represented by Eq. (6).

$$\hat{y_i} = \sum_{n=1}^{N} f_n(x_i) \quad \ldots (6)$$

The loss function is given by Eq. (7)

$$L(f) = \sum l(\hat{y_i}, y_i) + \sum \beta(f) \quad \ldots (7)$$

The first term represents the loss function which measures the predicted output y_i and tree output y_i. The second term β represents the regularization used as a penalty to avoid overfitting of the model. The β can be written as-

$$\beta(f) = \alpha T + \frac{1}{2}\|w\|^2 \quad \ldots (8)$$

where, T is the number of leaf nodes of the trees, and w is the weight of the leaf nodes. The final loss function equation is represented as-

$$L(f) = \sum_{j=1}^{T}\left[\left(\sum_{i=l} m_i\right) w_j + \frac{1}{2}\left(\sum_{i=l} n_i + \mu\right) w_j^2\right] + \alpha T \quad \ldots (9)$$

where, m_i , n_i represent 1st order, the 2nd Gradient of the loss function, respectively. The parameters μ and α represent the degree of regularization, which provides gradient direction to minimize the loss function and avoid overfitting.

2.5.3 Categorical boosting (CatBoost)

Catboost is a new machine learning method based on the Gradient boosting decision tree (GBDT), and Y and ex first proposed it in 2018[18]. It supports numerical, categorical, and text features. Various boosting techniques can solve the problem associated with the heterogeneous features, but CatBoost can handle the categorical data. CatBoost has the following advantages over the GBDT algorithm:

(1) It deals with the categorical data and uses the whole dataset for training. The GBDT uses Greedy target based statistics (GTBS), which can replace the categorical features with the average label, leading to overfitting the model. CatBoost adds the prior weight to GTBS, which reduces the overfitting of the model. For example, we have a dataset D with the features $D = \{x_i, y_i\}$, $i = 1, 2, 3 \ldots n$, if a permutation is $\sigma = (\sigma_1, \sigma_2 \ldots \ldots \sigma_n)$ then $x_{\sigma_p, k}$ is substituted with

$$\frac{\sum_{j=1}^{p-1}\left[x_{\sigma_j, k} = x_{\sigma_p, k}\right] Y_{\sigma_j} + \gamma \cdot P}{\sum_{j=1}^{p-1}\left[x_{\sigma_j, k} = x_{\sigma_p, k}\right] Y_{\sigma_j} + \gamma} \quad \ldots (10)$$

where, P is the prior value, and γ is the weight of the prior value. This method reduces the overfitting of the model.

(2) CatBoost combines multiple categorical features. When categorical features are converted into numerical values, it may lose some information. Thus, combining features may give new powerful features[18].

(3) In GBDT, each weak learner is trained based on the Gradient of the previous learner. Therefore, the Gradient of a weak learner in each iteration is biased, leading to overfitting the model. CatBoost can overcome this gradient bias using ordered boosting. Ordered boosting helps to avoid the predicted shift caused by gradient bias[18]. The algorithm of order boosting is as follows:

Algorithm 2 — Ordered boosting

Input:$\{(X_k,\ Y_k)\}_k^n$, I ; $\sigma \leftarrow$ permutation of $[1,\ n]$

$M_i \leftarrow 0$ for $i =\ 1 \ldots . n$;

for $t \leftarrow 1$ to I do

for $i \leftarrow 1$ to n do
$r_i \leftarrow y_i - M_{\sigma(i)-1}(X_i)$;
for $i \leftarrow 1$ to n do
$\Delta M \leftarrow Learn\ Model\left[(X_j, r_j): \sigma(j) \leq i\right]$;
$M_i \leftarrow M_i + \Delta M$
return M_n

Final model uses M_i to obtain the unbiased Gradient boosting by separately training the model with and without sample X_i .

2.6 Model optimization

The dataset was divided into 80% for training, and the other 20% were used for testing. We applied a grid search approach and 5-fold inner cross-validation to optimize the hyperparameters of the machine learning model. The inner 5-fold cross-validation was performed only on the training dataset. The optimized parameters of GBDT were as follows: 'subsample', 'max_features', 'learning_rate', 'criterion', 'random_state', 'loss.' The optimized parameter was used for XGBoost were as follows: 'max_depth', 'colssample_bytree', 'min_child_weight', 'learning_rate', 'random_state', 'gamma'. The following are the CatBoost parameter used for optimization: 'n_jobs', 'n_estimators', 'max_depth', 'criterion', 'random_state', 'bootstrap'. All the possible combinations of these hyperparameters were used before training and tested on the model. The machine learning model, which shows the best performance with the hyperparameters, was considered as the best model. The framework of the study is shown in Fig. 2.

2.7 Performance evaluation

The performance of the machine learning model was evaluated based on the various performance evaluation metrics: Accuracy (ACC), sensitivity (SEN), specificity (SPC), positive predicted value (PPV), negative predicted value (NPV), false-positive rate (FPR), false negative rate (NFR), F1 score, and area under the receiver operating characteristic curve (AUC). These metrics were evaluated using true positive (TP), false positive (FP), true negative (TN), and false-negative (FN). The performance metrics are shown in Table 4.

3 Results and Discussion

This section presents the results of the selected HRV features to predict the risk of T2DM. In this study, we have taken two datasets. The dataset D1 consists of 40 normal subjects and 120 diabetes subjects. The dataset D2 consists of 10 normal subjects and 40 diabetes subjects. The HRV features from dataset D1 were used to train the machine

Table 4 — Performance metrics

Metrics	Description
ACC	(TP+TN)/(TP+FP+TN+FN)
SEN	(TP)/(TP+FN)
SPC	(TN)/(TN+FP)
PPV	(TP)/(TP+FP)
NPV	(TN)/(TN+FN)
FPR	(FP)/(FP+TN)
FNR	(FN)/(FN+TP)
F1 Score	2TP/(2TP+FP+FN)
AUC	The area under the ROC curve

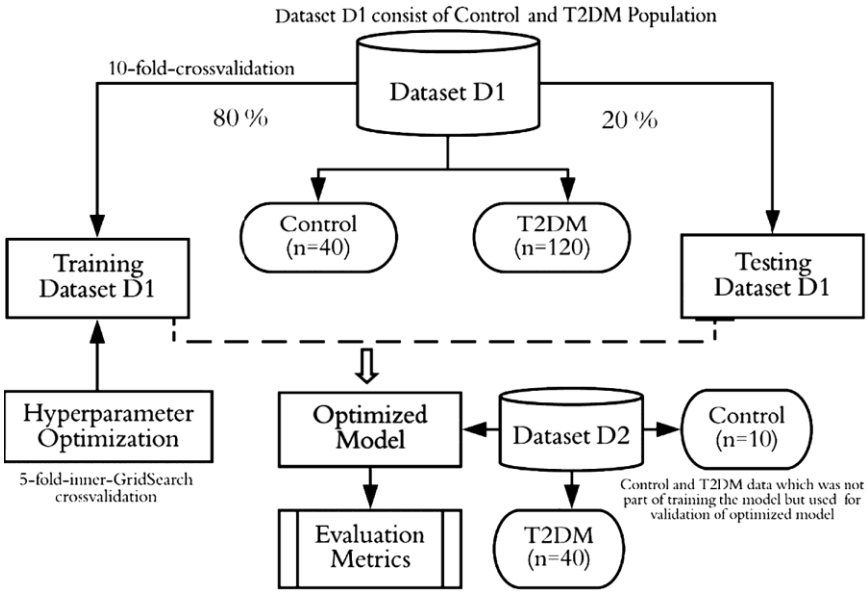

Fig. 2 — A complete framework of the study.

learning model, whereas dataset D2 was used as test data and used to evaluate machine learning models' performance. We have trained the machine learning model by the HRV features from dataset D1 and divided it into training data and testing data. The various traditional machine learning algorithms like Decision tree, Random forest, Naive Bayes, Support vector machine, and k-nearest neighbors were evaluated. The traditional machine learning algorithm has shown a good performance while training, but when dataset D2 was applied as test data, their predictive ability was lacking. Thus, the three best machine learning models GBDT, XGBoost, and CatBoost, which performed very well on train data and test data, were selected for the study. Before the training of the machine learning model was optimized using hyperparameters. The machine learning model results were evaluated based on performance measures. Performance of the model GBDT, XGBoost, and CatBoost using dataset D1 is shown in Table 5.

It can be observed that the optimized CatBoost model performs well as compared to other models and yields an accuracy of 91.6%, the sensitivity of 97.1%, the specificity of 76.9%, PPV of 91.8%, NPV of 90.9%, FPR of 0.23, FNR of 0.02, f1 score of 0.94 and AUC of 0.87. It was noted that the FPR and FNR of the CatBoost model were less, which indicates the ability of the model to predict correct classes.

The receiver operating characteristic (ROC) curve is another meaningful visualization way to compare the diagnostic performance of different models. It is a plot of 'sensitivity (TPR) versus '1-Specificity' (FPR)[19]. ROC indicates the performance of individual models, and the area under the ROC lies between 0.5 to 1, shows the classification ability. The AUC near to 1 represents the best machine learning model performance. ROC curves for GBDT, XGBoost, and CatBoost are given in Fig. 3.

3.1 Predictive ability evaluation and validation of the model

TheCatBoost model has shown better results on dataset D1. Now to know its predictive ability on unknown data, test data, or actual patient data, we have been using dataset D2. The dataset D2 was the test data for the CatBoost model. We hypothesized that the machine learning model trained to segregate two groups of control and diabetes based on the HRV features of dataset D1 would also differentiate the HRV features of dataset D2 into control and diabetes groups. Thus, HRV features of dataset D2 were used as input features for all the machine learning models which were trained using dataset D1.

The results given by the machine learning models were noted, and accordingly, true positive, false positive, true negative, and false negative were manually calculated. When the HRV features of dataset D2 were applied to the GBDT machine learning model trained using dataset D1, the values of true positive: 36, false-positive: 4, false-negative: 3, true negative: 7 were noted. When the HRV features of dataset D2 were applied to the XGBoost machine learning model trained using dataset D1, the values of true positive: 38, false-positive: 2, false-negative: 2, true negative: 8 were observed. When the HRV features of dataset D2 were applied to the CatBoost machine learning model trained using dataset D1, the values of true positive: 39, false-positive: 1, false-negative: 1, true negative: 9 were obtained. Thus, the CatBoost machine learning model has the highest predictive ability. The performance assessment of dataset D2 applied to the machine learning model developed using dataset D1 is shown in Table 6.

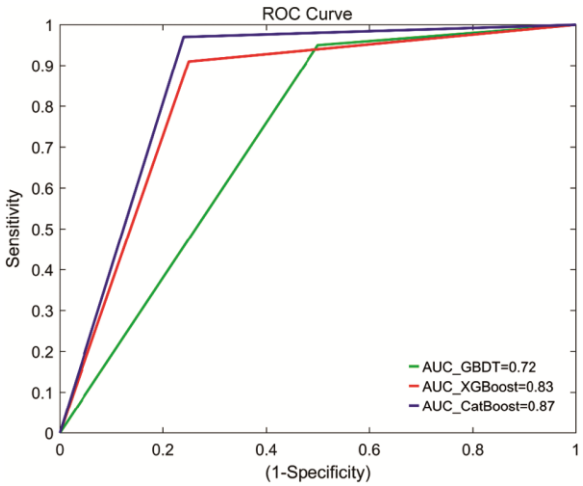

Fig. 3 — Area under ROC plot of the optimized model.

Table 5 — Performance evaluation of dataset D1

Model	ACC (%)	SEN (%)	SPC (%)	PPV (%)	NPV (%)	FPR	FNR	F1 Score	AUC
GBDT	84.3	95.8	50	85.1	80	0.50	0.04	0.90	0.72
XGBoost	87.5	91.6	75	91.6	75	0.25	0.08	0.91	0.83
CatBoost	91.6	97.1	76.9	91.8	90.9	0.23	0.02	0.94	0.87

Table 6 — Performance evaluation of dataset D2 on optimized model

Model	ACC (%)	SEN (%)	SPC (%)	PPV (%)	NPV (%)	FPR	FNR	F1 Score	AUC
GBDT	86	92.3	63.6	90	70	0.36	0.07	0.91	0.77
XGBoost	92	95	80	95	80	0.10	0.05	0.95	0.87
CatBoost	96	97.5	90	97.5	90	0.10	0.02	0.97	0.93

It can be observed that the test data from dataset D2 fed to the CatBoost machine learning model trained with dataset D1 has shown the highest ACC of 96%, SEN of 97.5%, SPC of 90%, PPV of 97.5%, NPV of 90%, FPR of 0.10, FNR of 0.02, F1 score of 0.97 and AUC of 0.93 as compared to other algorithms. The CatBoost model trained with dataset D1 has correctly predicted 39 diabetes subjects out of 40 diabetes subjects and 9 control subjects out of 10 control subjects. Thus, the FPR and FNR were reduced. The area under the ROC curve of all three models is shown in Fig. 4.

We have achieved the highest accuracy, sensitivity, and specificity with the CatBoost model. In some studies, the time domain and frequency domain characteristics of HRV have been used to predict diabetes. So far, to the best of our knowledge, no researcher has considered the time domain, frequency domain, and non-linear features along with the breathing rate factor in a single study. Various authors have achieved good prediction accuracy, but no study has implemented their machine learning model on actual patients. Thus, the implementation of the machine learning model on patients assures the validation of the model. We have found that most of the HRV features values are reduced in diabetic subjects compared to the control subjects, and their significant difference can help to predict T2DM at earlier stages.

The basic HRV features are the time-domain features. The author AL-Hazimi *et al.*[20] have found that the time domain features like SDNN, RMSSD, etc., were reduced in the diabetes group compared to the control group. Pfeifer *et al.*[21] have observed that parasympathetic activity was reduced in diabetes patients. Schroeder *et al.*[22] have found decreased time-domain parameters in the diabetes group. Kirvela *et al.*[23] performed time domain and frequency domain analysis of HRV and found a significant reduction in these parameters.

The author Seyd *et al.*[24] performed time domain and frequency domain HRV analysis. It was noted that the time domain parameters like mean HR, mean RR, SDNN, RMSSD, and the frequency domain parameter like TP, LF power, HF power were lower in diabetes patients. Chemla D *et al.*[25] used FFT and

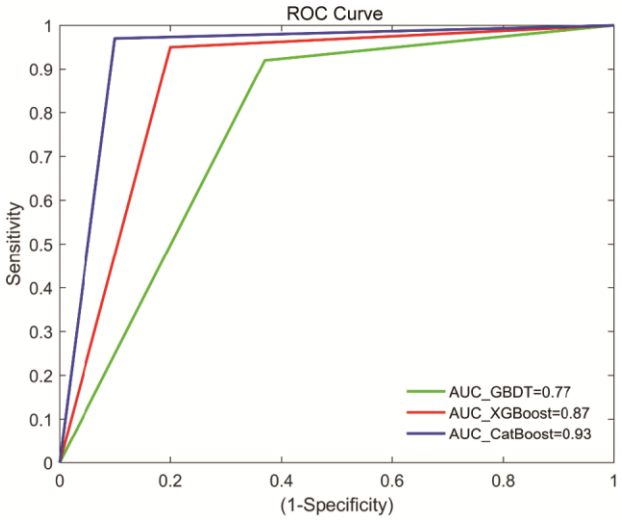

Fig. 4 — Area under ROC plot of results obtained from the optimized model on test data.

the autoregressive model to analyze the effect of HRV on diabetes patients. Javorka *et al.*[26] used linear and non-linear parameters used to predict type 1 diabetes mellitus. Faust *et al.*[27] analyzed the time domain, frequency domain, non-linear features and found that non-linear features provide prominent results in the diagnosis of diabetes. Acharya *et al.*[28,29] have proposed diabetes integrated index (DII) using non-linear HRV features, and an accuracy of 86% was observed with the AdaBoost model. Jian *et al.*[30] and Swapna *et al.*[31] have used higher-order spectrum (HOS) features to predict diabetes and obtained an accuracy of 79.9%, 90.5%, respectively. The summaries of diabetes prediction using HRV features are given in Table 7.

The best machine learning model was obtained using the CatBoost algorithm. Thus, this model can be considered as a decision support system for healthcare professionals. The graphical user interface (GUI) with the backend programming of the CatBoost algorithm was implemented using the python software as shown in Fig. 5. The user can manually extract the HRV features and enters the values of HRV features through the GUI. After clicking on submit button, users will be notified of the results as control or T2DM risk. Based on the results, patients will be advised to communicate with the doctors.

Table 7 — Comparison with the existing study of diabetes prediction using HRV

Features	Result	ML model validation on the number of patients
Non-linear[28]	Accuracy 86%	-
Non-linear[29]	Accuracy 90%	-
HOS[30]	Accuracy 79.9%	-
HOS[31]	Accuracy 90.5%	-
Time domain, Frequency domain, Non-linear features, and Breathing rate (Proposed Method)	Accuracy 91.6%	50 patients

ML: Machine learning, HOS-Higher order spectrum

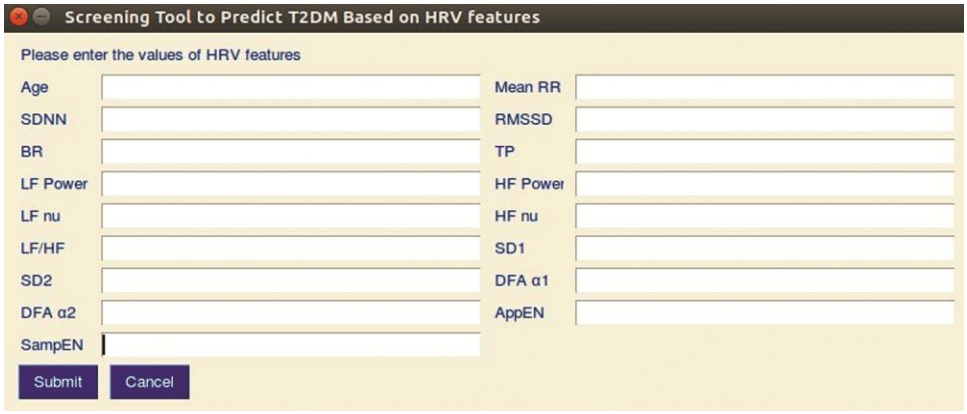

Fig. 5 — A graphical user interface to predict T2DM using HRV features for doctors in the hospital.

3.2 Consent regarding small sample size of control and diabetes dataset

The designed model can be considered as a preliminary model. We attempt to design the machine learning model with real and authentic data. However, it is challenging to find control and diabetes subjects unless they go through the pathological test. So it is difficult to increase the sample of control and diabetes subjects in less time constraint. We are still working to improve the sample size of control and diabetes subjects.

4 Conclusion

Type 2 diabetes mellitus is a long-term disease. Early prediction of diabetes can help doctors as well as patients take preventive measures. We have seen that HRV features have reduced in the diabetes group, and this reduction starts in early stages of diabetes. Thus, HRV features can be helpful to predict disease at an early stage. We can conclude that the CatBoost machine learning model is a better model to classify diabetes patients based on different performance metrics like ACC, SEN, SPC, PPV, NPV, FPR, FNR, F1 score, and AUC. The machine learning model was validated on 50 patients, and it has correctly predicted 48 patients out of 50 patients. Therefore, we can recommend this model as a decision support system to predict the risk of T2DM.

Acknowledgment

The authors would like to acknowledge their sincere gratitude to the Department of physiology, the central research lab, Smt. Kashibai Navale Medical College and General Hospital (SKNMCGH) Pune, India, where the study has been conducted. We would also like to thank the College of Engineering Pune (COEP) for providing fellowship during the study.

References

1 Sarwar N, Gao P, Seshasai S R, Gobin R, Kaptoge S, Di Angelantonio E, Ingelsson E, Lawlor D A, Selvin E, Stampfer M, & Stehouwer C D, *Lancet*, 375.9733 (2010) 2215.
2 Chiang Jane L, M Sue Kirkman, Lori M B Laffel, & Anne L Peters, *Diabetes Care*, 7 (2014) 2034.
3 Begum S A, Afroz R, Khanam Q, Khanom A, & Choudhury T S, *Journal of Paediatric Surgeons of Bangladesh*, 5.1 (2014) 30.
4 Crowshoe L, & Dannenbaum D, Green M, Henderson R, Hayward M N, Toth E, *Can J Diabetes*, 42 (2018) S296.
5 Centers for Disease Control and Prevention, *Atlanta, GA: US Dept of Health and Human Services CDC*, 201.1 (2011) 2568.
6 Electrophysiology, Task Force of the European Society of Cardiology the North American Society of Pacing, *Circ*, 93.5 (1996) 1043.
7 American Diabetes Association, *Diabetes Care*, 41.1 (2018) S13.
8 Acharya U R, Joseph K P, Kannathal N, Lim C M, & Suri J S, *Med Biol Eng Comput*, 44.12 (2006) 1031.
9 Faust O, Acharya U R, Molinari F, Chattopadhyay S, & Tamura T, *Biomed Signal Process Control*, 7.3 (2012) 295.

10 Brennan M, Palaniswami M, & Kamen P, *IEEE Trans Biomed Eng*, 48.11 (2001) 1342.

11 Peng C K, Havlin S, Hausdorff J M, Mietus J E, Stanley H E, & Goldberger A L, *J Electrocardiol*, 28 (1995) 59.

12 Fusheng Y, Bo H, & Qingyu T, *Nonlinear Biomed Signal Processing*, 2 (2001) 72.

13 Richman J S, & Moorman J R, *Am J Physiol Heart Circ Physiol*, 278 (2000) H2039.

14 Walfish S, *Pharm Technol*, 30.11 (2006) 82.

15 Patro S, & Sahu K K, *arXiv preprint arXiv: 1503.06462*, (2015) 1.

16 Natekin A, & Knoll A, *Front Neurorobot*, 7 (2013) 21.

17 Chen T, He T, Benesty M, Khotilovich V, Tang Y, & Cho H, *R package version 0.4-2*, 1.4 (2015).

18 Prokhorenkova L, Gusev G, Vorobev A, Dorogush A V, & Gulin A, *arXiv preprint arXiv: 1706.09516*, (2017) 1.

19 Hajian-Tilaki K, *Caspian J Intern Med*, 4.2 (2013) 627.

20 Al-Hazimi A, Al-Ama N, Syiamic A, Qosti R, & Abdel-Galil K, *Ann Saudi Med*, 22.5-6 (2002) 400.

21 Pfeifer M A, Cook D, Brodsky J, Tice D, Reenan A, Swedine S, Halter J B, & Porte D, *Diabetes*, 31.4 (1982) 339.

22 Schroeder E B, Chambless L E, Liao D, Prineas R J, Evans G W, Rosamond W D, & Heiss G, *Diabetes Care*, 28.3 (2005) 668.

23 Kirvela M, Salmela K, Toivonen L, Koivusalo A M, & Lindgren L, *Acta Anaesthesiol Scand*, 40.7 (1996) 804.

24 Seyd P A, Ahamed V T, Jacob J, & Joseph P, *Int J Biol Sci*, 4.1 (2008) 24.

25 Chemla D, Young J, Badilini F, Maison-Blanche P, Affres H, Lecarpentier Y, & Chanson P, *Int J Cardiol*, 104.3 (2005) 307.

26 Javorka M, Trunkvalterova Z, Tonhajzerova I, Javorkova J, Javorka K, & Baumert M, *Clin Neurophysiol*, 119.5 (2008) 1071.

27 Faust O, Acharya U R, Molinari F, Chattopadhyay S, & Tamura T, *Biomed Signal Process Control*, 7.3 (2012) 295.

28 Acharya U R, Faust O, Sree S V, Ghista D N, Dua S, Joseph P, Ahamed V T, Janarthanan N, & Tamura T, *Comput Methods Biomech Biomed Engin*, 16.2 (2013) 222.

29 Acharya U R, Faust O, Kadri N A, Suri J S, & Yu W, *Comput Biol Med*, 43.10 (2013) 1523.

30 Jian L W, & Lim T C, *J Med Imaging Health Inform*, 3.3 (2013) 440.

31 Swapna G, Rajendra Acharya U, Vinitha Sree S, & Suri J S, *Intell Data Anal*, 17.2 (2013) 309.