Integrated Science

Journal of Integrated
SCIENCE & TECHNOLOGY

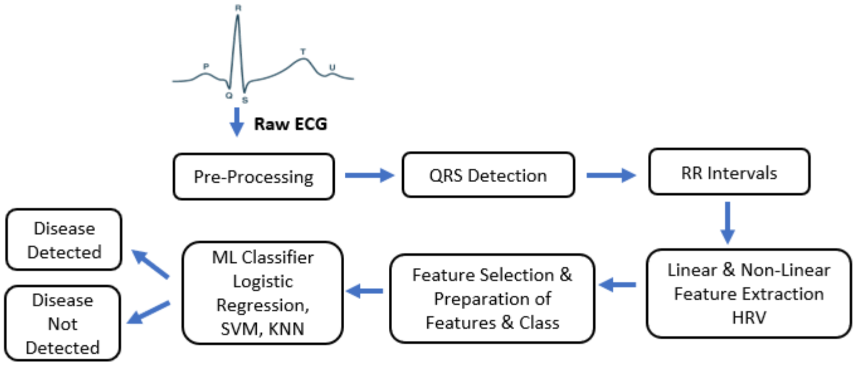
Detection of Cardiovascular Autonomic Neuropathy using Machine Learning Algorithms

Bhavin Mehta[1], Vijay Dave[2]

[1]Biomedical Engineering Department, L.D. College of Engineering, Ahmedabad-380015, Gujarat Technological University, Gujarat, India. [2]Biomedical Engineering Department, Government Engineering College, Gandhinagar-382028, Gujarat Technological University, Gujarat, India.

Received on: 02-Feb-2023, Accepted and Published on: 04-Apr-2023

ABSTRACT


R
P
T
U
Q
S
Raw ECG
Pre-Processing
QRS Detection
RR Intervals
Linear & Non-Linear Feature Extraction HRV
Feature Selection & Preparation of Features & Class
ML Classifier Logistic Regression, SVM, KNN
Disease Detected
Disease Not Detected

Cardiovascular autonomic neuropathy (CAN) is an asymptomatic and often overlooked complication of Diabetes mellitus (DM). It is found to be more prevalent in Type 2 diabetes (T2 DM) than in Type 1 diabetes (T1 DM). CAN can damage the sympathetic and parasympathetic cardiac and vascular nerve fibers, leading to reduced heart rate variability (HRV) and vascular dynamics. Depression in HRV is an indicator of CAN. Cardiovascular reflex tests (Ewing Test) are currently used as the diagnostic tool for CAN detection, but these tests cannot detect subclinical CAN and require patient cooperation. Therefore, we propose evaluating the feasibility of HRV features (linear and non-linear) over a 5-minute period, embedded within machine learning models to provide a comprehensive screening for patients with CAN. Our study used ECG datasets from PhysioNet, consisting of 100 patients (50 with T2 diabetes and 50 healthy individuals). We applied the Hamilton peak detection algorithm to the ECG signals to detect QRS peaks and extract heart rate signals. Linear and non-linear methods of HRV analysis were used to extract various features from every 5-minute segment of the HRV signal. These features were used to prepare a dataset, and k-fold cross-validation techniques were applied to separate training and testing datasets to make a classification model less biased. For our study, we used logistic regression, K-nearest neighbors (KNN), and support vector machine (SVM) as machine learning classification algorithms. To maximize performance, hyperparameter optimization techniques were used to select the right combination of hyperparameters for each classification model. The highest level of performance was achieved with accuracy measures of 96.67% (logistic regression), 93.33% (K-NN), and 93.33% (SVM), and these were found to be the most significant in the diagnosis of CAN in diabetic patients. The proposed machine learning-based classification models could predict the early onset of CAN and decrease the mortality rate due to myocardial infarction.

Keywords: *Diabetes, Heart rate variability (HRV), Linear Analysis, Hyperparameter optimization, Machine learning algorithms*

INTRODUCTION

Cardiovascular disorder is the most common cause of mortality and morbidity among patients with Diabetic Mellitus and microvascular complications in India. Around 415 million people are affected by Diabetes mellitus worldwide, and by 2040 it is expected to rise to approximately 640 million, as per the recent report of the International Diabetic Federation.[1] There are various non-communicable diseases like cardiovascular disorder (CVD), chronic respiratory illness, and diabetes estimated at 60% of total deaths.[2] India is the leading country with the highest burden of Cardiovascular Disorders worldwide. Clinical studies showed that risk factors of CVDs such as obesity, hypertension, and Diabetes are found more in the young Indian group as compared to the ethnic group. In urban communities, prevalence rates of CVDs risk have constantly been rising in India for the last 25 years. The study shows that one in ten people aged 18 years in India has increased blood glucose levels in India. India had more than 73 million cases

Corresponding Author: Bhavin Mehta, Research Scholar, Gujarat Technological University, Vijay Dave, Research Supervisor, Gujarat Technological University, Gujarat
Email: bhavin.bme@gmail.com, vpdave12@gmail.com

Cite as: *J. Integr. Sci. Technol., 2023, 11(4), 562. URN:NBN:sciencein.jist.2023v11.562*

ISSN: 2321-4635
http://pubs.thesciencein.org/jist

of Diabetes, the highest in any country across the globe. Diabetes has a high prevalence of 8.8 %, which is becoming a significant challenge in India. A high prevalence of Diabetes is due to rapid urbanization, globalization, inactive lifestyle, unhealthy diets, and obesity.[2]

Diabetes mellitus (DM) is a hyperglycemic state (high glucose level) that can lead to micro and macrovascular damage. Diabetic neuropathy affects 60 to 70 percent of people with diabetes and is further classified into autonomic, focal, peripheral, and proximal neuropathy. Our primary focus is diabetic neuropathy, which affects the nerve that controls heart functions. This type of neuropathy is known as cardiovascular autonomic neuropathy.[3] Cardiovascular autonomic neuropathy is an underdiagnosed complication of Diabetes mellitus that is asymptomatic, and it is more prevalent in Type 2 diabetes than in Type 1 Diabetes.[4] There is a high risk of cardiac arrhythmias and sudden death, and a condition similar to silent myocardial ischemia exists in around 20% of clinically diagnosed cases of diabetes. This condition leads to significant morbidity and mortality in patients with diabetes.[5] The subclinical stage of CAN may be detected within 1 to 2 years of diagnosis of diabetes which indicates a change in heart rate variability. The clinical symptoms of CAN have appeared after a long duration of diabetes. It remains undiagnosed until the CAN progresses to an advanced stage. So, the early detection and management of CAN are essential to prevent future complications.[6]

There is an involvement of the autonomic nerve in Diabetes. The first involvement is found in the vagus nerve, the longest autonomic nerve providing 75% of the body parasympathetic tone[7]. The preliminary sign of vagus dysfunction is resting tachycardia, exercise intolerance, left ventricular dysfunction at rest, and cardiac arrhythmias.[7] Later is affected by sympathetic nerve function in that orthostatic hypotension is the most evident manifestation. Two primary methods help to diagnose cardiac autonomic neuropathy. The first method is proposed by Ewing, which is called cardiovascular autonomic reflex tests (CARTs), the reference method for diagnosing CAN in Type 2 D.M. Patients. The Ewing Test consists of dynamic Heart rate and blood pressure, which indicate dysfunction of parasympathetic and sympathetic activities of the Autonomic nervous systems.[8] Another method is Heart rate variability, a non-invasive technique to diagnose diabetic CAN effectively.[9] Heart rate variability is the variation in the time intervals between adjacent heartbeats. It assesses the autonomic nervous system by measuring the changes in cardiac rhythm over time that reflect the changes in sympathetic and parasympathetic branches of ANS.[10, 11] A 5-minute ECG recording can serve as a marker for both sympathetic and parasympathetic nervous system activity, as measured by heart rate variability (HRV). The sympathetic branch is responsible for increasing heart rate and contractility, while the parasympathetic branch is responsible for decreasing heart rate and contractility.[12,13] A 5-minute ECG recording can provide enough RR intervals to accurately calculate HRV and assess changes in autonomic activity over time. Low values of HRV correlate with various events such as myocardial infarction, heart failure, coronary artery disease, sudden death, diabetes mellitus, acute and chronic stress, depression, metabolic syndrome, and emotional states.[14, 15, 16] Non-stationary/non–linear signal, heart rate variability is measured from the electrocardiography (ECG) signal. Heart rate is acquired by calculating the time interval between two ventricular contractions (or between two consecutive R waves (R.R. Interval)) from the ECG signal.[17] The ECG data acquired according to their duration is divided into the ultra-short term (less than 5 min), short term (5 min to 30 min), and long term (up to 24hr).

HRV analysis assesses the overall cardiac health in connection with heart rate regulation based on the status of ANS responsible for regulating cardiac activity.[18] HRV analysis can be performed using linear and non-linear methods. The linear method consists of time and frequency domain analysis of HRV signal. In contrast, a Non-linear method consists of Detrended fluctuation analysis (DFA), correlation dimension (CD), Approximate entropy (ApEn), Recurrence qualification analysis (RQA), Higher-order spectrum (HOS), Empirical mode decomposition analysis (EMD).[19, 20] A time-domain study only has one problem: statistical characteristics change with time. Frequency domain analysis will better evaluate an autonomic function, but the spectral power decreases with decreased power signal and SNR.[20] A Non-Linear is the most suitable method for exploring a non-linear and non-stationary ECG signal as it correctly predicts the onset of cardiovascular events such as Ventricular Tachycardia and congestive heart failure.[21]

Machine learning algorithms have decision-making abilities for computer-assisted automated illness detection. It is widely used to interpret data and transform their results into valuable information that helps to diagnose various disorders.[22] Machine learning algorithms find the mathematical function that produces a correct outcome (abnormality present or not) from the input training data(like ECG, HRV) and understands the hidden patterns in the input sequence. With the help of learned mathematical functions, the machine learning (ML) model can predict the output for new input data (which is anonymous to the model) with high accuracy. Moreover, by applying data analysis tools and machine or deep learning algorithms, numerous research identified the link between HRV and diabetes.[23] With the help of machine learning techniques, various heart-related disordered have been diagnosed very accurately. Many researchers have applied supervised machine learning algorithms, such as Support Vector Machine, Logistic Regression, Decision Tree, K-nearest neighbors (KNN), and many more classifiers for detecting cardiac disorders[24]. The only method to reduce mortality and morbidity associated with chronic diseases is to detect and treat them early. In traditional machine learning, feature extraction and classification are performed separately. Most CAN classifier models have low accuracy and have not extracted and included all Linear and Non-linear features of HRV for a CAN classification. Our study extracted all the essential features for HRV analysis and applied them for CAN classification.

The significant contribution of this study is:

- Detection of QRS Peak using Hamilton Algorithm and extract Heart rate signal.
- Applied Linear and Non-linear methods to extract various essential features of HRV analysis.

- Prepare a dataset of all the essential features and apply k-fold cross-validation techniques to separate training and testing datasets for a classification.
- Apply Hyperparameter optimization techniques to enhance the performance of the classifier.
- Various machine learning classifiers have been proposed to detect CAN.
- The performance of classifiers has been evaluated with various performance merits.

The current study is undertaken to detect cardiovascular autonomic neuropathy using various linear and non-linear methods of heart rate variability. The study is meant to develop and evaluate the performance of machine learning algorithms to predict early occurrences of cardiovascular autonomic neuropathy in type 2 diabetic patients.

METHODOLOGY

Data Set Description

In this study, we have used a dataset of Cerebral Vasoregulation in Diabetes from a PhysioNet [25] as a data source for ECG records. The dataset consists of electrocardiograms (ECG) of 50 patients (27 male and 23 female) with Type 2 diabetes, and 50 healthy volunteers (24 male and 26 female) were recorded, with the patients in a relaxed supine position for 5 min. Each ECG signal was recorded using Lead II with a sampling frequency of 1000 Hz. The subjects studied for diabetes were in the age group of 55 to 75 years, and the mean duration of diabetes for the patient groups was 12.5 years. The body mass index (BMI) and Glucose level of all 100 subjects were considered in the dataset.

Processing of ECG Signal

The Heart rate was calculated using the Hamilton peak detection algorithm from raw ECG signals [26]. The Hamilton peak detection algorithm derives the Heart rate variability by detecting the QRS Complex. **Figure 01** shows the generalized block diagram of the Hamilton Peak detection algorithm.

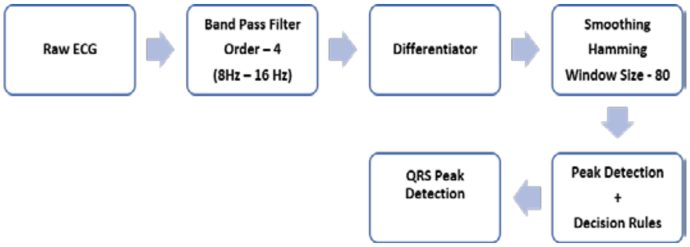

Raw ECG
Band Pass Filter Order – 4 (8Hz – 16 Hz)
Differentiator
Smoothing Hamming Window Size - 80
Peak Detection + Decision Rules
QRS Peak Detection

Figure 01: Block diagram of Hamilton peak detection algorithm

For QRS detection, preprocessing steps involve bandpass filter, rectification, averaging sliding window, followed by QRS Detection rules. The high pass, low pass, and derivative combined produce a bandpass filter with a passband from 8 Hz to 16 Hz. The bandwidth containing most of the QRS Complex energy and the response of the filtered signal is rectified, which gives better sensitivity for QRS detection. The averaging window kernel is Hamming, and the window size is 80ms wide for QRS Complex, which produces better results. QRS Peak detection rules are applied to find QRS complex based on threshold detection equation[26]. One buffer is created to store the eight most QRS peaks and noise peaks. The threshold detection equation is estimated QRS peak and noise peak heights, which are set between the noise peak and QRS peak buffer as shown in equation 1. Where T.H. is the threshold coefficient lies between 0.3125 and 0.475.

$$\text{Threshold Detection} = \text{Average Noise Peak} + \text{T.H.} * (\text{Average QRS Peak} - \text{Average Noise Peak}) \qquad (1)$$

In this study, we utilized various Python programming libraries to read a raw ECG signal, plot it, and extract the QRS complex using the Hamilton algorithm. We then calculated the heart rate from the R-R interval. An interval of two consecutive QRS Complexes is defined as an R.R. interval used to calculate heart rate by the following equation.

$$\text{Heart rate (BPM)} = 60/(\text{R.R.interval}) \qquad (2)$$

The parasympathetic and sympathetic nervous systems act accordingly to the situation that changes the heart rate. The sympathetic nervous system (SNS) increases heart rate, whereas the parasympathetic nervous system (PNS) works in opposition to the SNS. In Normal individuals, the variation in the heart rate is high and low in diseased conditions or aged individuals. The variation between consecutive heartbeat intervals is defined as Heart rate variability. HRV finds out activities of the Autonomic Nervous system (ANS) and Sinoatrial node (S.A. node) of the heart.[27]

Feature Extraction

In our study, feature extraction is the most decisive step for a Heart rate signal analysis and interpretation. We have performed Linear and Nonlinear Methods on heart rate signals to extract various features used later to classify Cardiac autonomic neuropathy using Machine learning algorithms. Linear Methods consist of Time-domain and Frequency-domain analysis to find the variation in heart rate signal.

Time Domain analysis

The time-domain analysis calculates heart rate for any given period by estimating the intervals between successive normal complexes (R-R or N-N time series). This method is ideal for the study of short-term recordings. This method is categorized into statistical and geometrical forms. The statistical operation calculates the mean and variance of R.R. interval of HRV data, where geometric methods are more complex and final results available in graphical form, which is challenging to interpret[11]. Our study extracted essential parameters from time-domain analysis, such as Mean H.R., STD HR, RMSSD, SDNN, NN50, pNN50, HRV Triangular index, TINN, and stress index, which are derived from R.R. intervals or differences between R.R. intervals. Time domain features such as SDNN, RMSSD, and pNN50 commonly assess HRV in patients with type 2 diabetes and CAN. SDNN measures the overall variability of the heart rate over a given period and has been shown to be reduced in patients with type 2 diabetes. RMSSD reflects the short-term variability of HRV, primarily influenced by parasympathetic activity. It is also reduced in patients with type 2 diabetes and is a sensitive marker of early ANS dysfunction. pNN50 measures parasympathetic activity and is decreased in patients with type 2 diabetes. The clinical significance of these time domain features in HRV for type 2 diabetes lies in their ability to provide information about the severity of autonomic dysfunction in diabetes patients. Reduced HRV has been associated with an increased risk of cardiovascular disease and mortality in diabetes patients[12,13,28]. Time-domain parameters and their relevance are shown in Table 1.[11,29,30]

Frequency Domain analysis

This method provides a better evaluation of autonomic function. It gives details of changes in the heart rate signals in terms of frequency and intensity.[31.] Power spectrum density (PSD) details the neurogenic heart rhythms representing how power (variance) is distributed as a function of frequency.[32.] Four spectral components can be calculated: 1) Ultra-low frequency (ULF) 2) Very-low frequency (VLF) 3) Low frequency (L.F.), 4) High frequency (H.F.) from frequency domain analysis.

Table 1: Time-domain (T.D.) features and their relevance

Sr. No.	List of T.D. Features	Description	Physiological Correlation	Formula
1	Mean H.R.	Mean of Heart rate	Indicate the role of the Autonomous nervous system	$\overline{HR} = \frac{1}{N}\sum_{n=0}^{N-1} HR(n)$
2	STD HR	Standard deviation of instantaneous heart rate values		SDHR = $\sqrt{\frac{\sum_{n=0}^{N-1}(\text{H.R.(n)}-\overline{HR})^2}{N}}$
3	RMSSD (ms)	The square root of the mean squared differences between successive R.R. interval	Estimate high-frequency variation in H.R. Estimate the Parasympathetic regulation of Heart	RMSSD = $\sqrt{\frac{1}{n-1}\sum_{j=1}^{n} \Delta RR_j^2}$ n: - Number of NNI ΔRRj: R.R. interval differences
4	SDNN (ms)	Standard Deviation of R.R. interval	Estimate of overall HRV Reflects all the cyclic Components are responsible for variability in the period of the recording. Indicate SNS activity.	$SDNN = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}\left(RR_j - \overline{RR}\right)^2}$ N: - Number of R.R. interval $\overline{RR}$: - Mean of R.R. interval series
5	NN50 (Count)	No. of successive R.R. interval pairs that differ more than 50ms	Sympathetic & Parasympathetic Nervous System	NN50 = $\frac{RR50}{n}$ pNN50: Ratio of NNI differences > 50 milliseconds RR50: - Number of NNI differences > 50 milliseconds n: Number of NNI differences
6	pNN50 (%)	RR50 divided by the total no. of R.R. interval expressed in %.		
7	HRV Triangular Index	Integral of R.R. interval histogram divided by the height of the histogram of all R.R. interval measured on a discrete scale with bins of 7.8125 ms.	Estimate of overall HRV	Tri = $\frac{n}{D(X)}$ n– no. of R.R. interval D(X) – Maximum of R.R. distribution
8	TINN (ms)	Baseline width of R.R. interval histogram		It is calculated by the interpolation method.
9	Stress Index	Geometric measure of HRV	Reflecting cardiovascular system stress	S.I. = $\frac{Amo*100\%}{2Mo*MxDMn}$ Amo – mode amplitude Mo – mode of most recurring R.R. interval MxDMn -The range of fluctuation reflects the HRV.

The characteristics of Ultra-low frequency are not fully understood and require long-term ECG recording. Our study has calculated various frequency domain parameters like Very Low frequency, Low-Frequency, High Frequency, and LF/HF ratio used to classify cardiac autonomic neuropathy using machine learning classification algorithms. The detailed description of various frequency domain parameters and their physiological correlation are shown in Table 2.[11,33]

Non-Linear Analysis

Non-linear analysis is most suitable for analyzing non-stationary and non-linear signals like ECG. In Linear analysis, there is a chance of loss of signal dynamics. So, researchers have developed a non-linear fractal theory based on a mathematical scheme that provides a better estimate of the cause of HRV fluctuations. The features extracted from this method can provide complete information about physiological status and support diagnosis, prognosis, and prevention of various cardiovascular diseases. Some essential features used for analyzing HRV are Poincare plot, Approximate entropy (ApEn), detrended fluctuation analysis (DFA). A poincare plot is a graphical method that plots every R-R interval against the previous interval and creates a scatter plot that analyzes a plot by fitting an ellipse. After fitting the ellipse, we may extract three non-linear metrics, S, SD1, and SD2.[34] The ellipse area representing total HRV correlates with Baroreflex sensitivity (BRS), H.F. and L.F. power, and RMSSD. DFA will be used to access fractal scaling features (alpha 1 and 2) of HRV. Parameter like alpha 1 and 2 indicate inherent fluctuation and vary according to cardiac disorders. ApEn better indicates heart rate signal disorder.[35,36] ApEn has a higher value for complex or irregular data and vice versa for cardiac impairment situations. A detailed description of various Non-Linear parameters and their physiological correlation are shown in Table 3.[36]

Table 2: Frequency-domain features and their relevance

Sr. No.	List of Features	Description	Physiological Correlation
1	VLF Peak	Peak Frequency range - 0.0033Hz – 0.04 Hz	Indicate hormonal and thermal regulation with vasomotor activity
2	VLF Power (ms2)	Absolute power of VLF band	
3	VLF Power (log)	Relative power of VLF band	
4	VLF Power (%)	Relative power of VLF band – (VLF Power/T.P.) * 100%	
5	L.F. Peak	Peak frequency range - 0.04Hz - 0.15 Hz	Indicating a presence of SNS and PNS It coordinates peripheral vasomotor and thermoregulation activity
6	L.F. Power (ms2)	Absolute power of L.F. Band	
7	L.F. Power (log)	Relative power of L.F. band	
8	L.F. Power (%)	Relative power of L.F. band – (L.F. Power/T.P.) * 100%	
9	H.F. Peak	Peak frequency range - 0.15Hz - 0.4 Hz	Indicate a response of Parasympathetic activities
10	H.F. Power (ms2)	Absolute power of H.F. band	
11	H.F. Power (log)	Relative power of H.F. band	
12	H.F. Power (%)	Relative power of H.F. band – (H.F. Power/T.P.) * 100%	
13	LF/HF (%)	It is a ratio of L.F. to H.F. Power	Indicate a sympathovagal balance An elevated LF/HF ratio indicates low vagal activity.

Table 3: Non - Linear features and their relevance

Sr. No.	List of Features	Description	Physiological Correlation
1	SD1 & SD2 (ms)	The standard deviation of perpendicular and along the line of identity	SD1 indicates parasympathetic activities and correlates with BRS. SD2 indicates an overall variability and correlates with BRS and L.F. Power.
2	SD2/SD1 (%)	The ratio of SD2 to SD1	Indicate a response of short and long-term variation in R.R. intervals. It measures autonomic balance and sympathetic activation.
3	ApEN	Approximate entropy	It measures the complexities and regularities of a time series sequence. High values represent high irregularity in the HRV.
4	SampEN	Sample entropy	It measures the complexities and regularities of a time series sequence. A Lower value represent increased regularity in the HRV.
5	DFA – Alpha 1, Alpha 2	Detrended fluctuation analysis	Indicates short and long-term fluctuations

Framework

Machine learning uses programmed algorithms that learn and optimize their operation by analyzing input data to make predictions within a certain range. It does not require explicit rule-based programming and human assistance.[37] The programmed algorithm finds a mathematical function that produces the correct response from the input training data and understands the hidden patterns. Based on that, it predicts the output for a new set of input data (test data) with high accuracy. Supervised machine learning algorithms have two types of problems 1) Classification Problems 2) Regression-based Problems.[3] In classification, the output variable is discrete such as “CAN detect” or “not detected(normal),” wherein the regression output variable gives a real value such as the risk of developing CAN disease for an individual. The classification algorithm works in a two-step process: the first step is a learning step in which the model is constructed from a training dataset, and the second step is the application of the constructed model to predict the class labels of the unseen testing dataset. In our study, we have used supervised machine learning algorithms for CAN prediction. Figure 02 represents the architecture of a proposed method for detecting CAN using HRV.Feature extraction, data preparation, hyperparameter optimization, and classification are the four major phases of the proposed model.

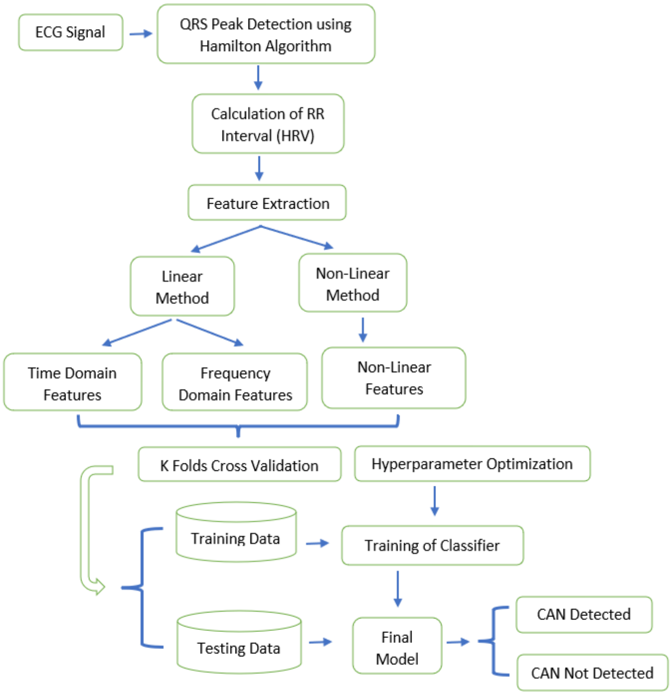


Figure 2 A framework of the designed method

A feature extraction phase extracted 30 essential features from Linear and Non-linear methods. We have added information on age, BMI, Years of Diabetic Mellitus, and Glucose value (mg/dl) with these essential features for further processing. We have applied different methods like handling missing values, categorical data encoding, and data standardization on all the extracted features and patient information in the data preparation phase. And then, the whole dataset is separated into training and testing datasets by applying the K fold cross-validation technique. The training dataset is applied to a classifier such as SVM, KNN, and Logistic Regression to prepare a model, and then the same model is evaluated by applying the testing dataset. After selecting a particular classifier, hyperparameter tuning increases the classifier's performance by choosing the best parameters for model creation. The below section provides complete detail on the cross-validation (CV) technique, hyperparameter optimization, and classifier.

Cross-validation

Cross-validation (CV) is a statistical resampling procedure that evaluates various machine learning models and tests their performance on limited data[38]. It is simple to comprehend and apply, and it has less bias than other approaches used to calculate a model's performance[39]. The CV had just one parameter, k, the number of groups that a given sample should be split into. This process is referred to as a K- fold CV. When an exact value for k is specified, it can be used to replace a k in the model's reference; for example, if we chose k = 5, it becomes 5-fold cross-validation. The k-Fold cross-validation technique's algorithm is as follows:

- Choose the number of folds – for example, a 5 or 10 usually less than a dataset length.
- Split a given dataset into k folds (equal parts)
- Choose k-1 folds for training; the remaining folds are used as a testing dataset.
- Train a model using a training dataset. Every iteration of CV trains a new model independently of the model on the previous iteration.
- Validate a model using a testing dataset and save the Score of each validation.
- Repeat a sequence for steps 3 to 5 for each k-fold time. Use a leftover fold as the test set and validate a model for each fold.
- To get the final results, average the results of each fold.

Hyperparameter Optimization

After selecting an appropriate machine learning model, this technique is applied to enhance the performance. Hyperparameters have various parameters that influence the learning process and have a significant impact on the performance of a model.[40] For example, hyperparameters in the SVM model include various kernels, maximum iteration number, gamma value, penalty parameters, and many more. All these parameters are tunable and directly affect the performance of a model during training. Hyperparameter optimization determines the best combination of hyperparameters that results in the maximum performance on a given dataset. Two methods are used for tuning hyperparameters: 1) Grid Search 2) Random Search. Grid Search is a widely used method to determine optimal values for a model. It works by experimenting with every conceivable combination of parameters in a model. A random search method takes a combination of hyperparameters' values to find the best solution for a model. But this method has one drawback, sometimes misses a significant value during hyperparameter tuning. Our study used the Grid search method to optimize a hyperparameter to enhance the machine learning algorithm's performance.

Logistic regression

Logistic regression is a simple and powerful analysis method for binary classification problems in supervised machine learning. A probability-based statistical model achieves very high performance for linearly separable events.[41] This method is considered the addition of ordinary regression and can model only a dichotomous variable representing an event that occurs or does not occur.[42] For binary classification in logistic regression, a threshold value is assigned to define the two classes. For example, if the probability value is higher than 0.50 for a given input, their value is classified into class 0; otherwise, 1. Figure 3 shows an illustration of the Logistic Regression.

Logistic regression is also used for multiple discrete outcomes called multimodal logistic regression. This model uses a logistic (sigmoid) function for binary classification[41] defined below in Equation 3, and their response range is bounded between 0 and 1. This model finds the relationship between dependent and independent categorical variables from a dataset. Here the dependent variable is a target class that we would be predicting through our model, and independent variables are the features used to predict the target class.

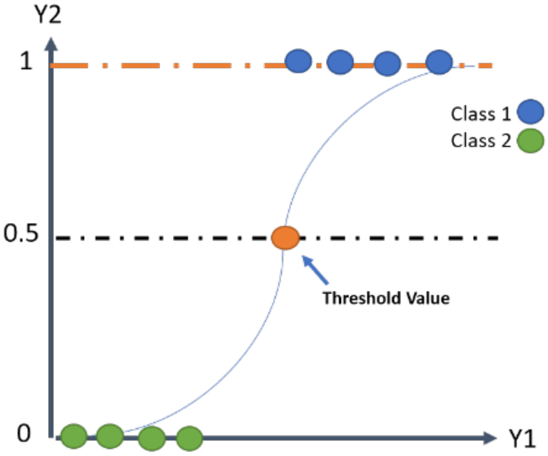


Figure 3 Logistic Regression

$$\text{Logistic Function} = \frac{1}{1+e^{-x}} \qquad (3)$$

Support Vector Machine (SVM)

SVM is the most robust and accurate method for binary classification problems in supervised machine learning. This model is used for both linear and non-linear data. It can handle multiple features with less risk of overfitting. In SVM, first, it maps each data into an n-dimensional feature space where n is denoted as the number of features available in the dataset. Based on that, it identifies the hyperplane that separates the data into two classes.[41] Figure 4 shows an illustration of the SVM Classifier. It maximizes marginal distance (a distance between a hyperplane and its nearest class instance) for both classes and minimizes classification errors. Each data point is plotted in an n-dimensional space, with each feature's value defined at a specific coordinate.

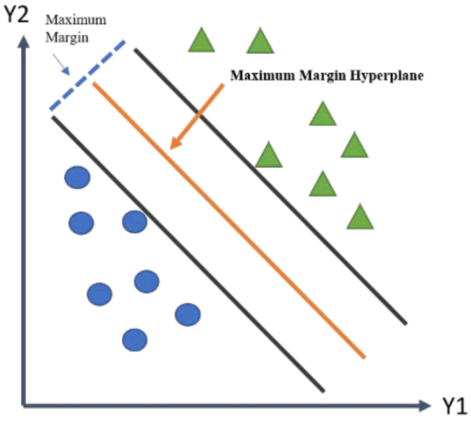


Figure 4 Support Vector Machine

We first need to find a hyperplane that differentiates two classes by their maximum margin, which helps future data classify more confidently in SVM Classification. Equations 4 and 5 represent a hyperplane; the maximum margin hyperplane separates two classes[41]. As a function, the kernel is used in SVM to solve complex non-linear problems. It helps to form a hyperplane and decision boundary for high-dimensional data without any complexity [37]. Through the kernel, the overfitting problem can solve by selecting the correct kernel. Many kernels are used in SVM, such as RBF (Universal Kernel for Small dataset), Polynomial, Gaussian RBF, Sigmoid, Linear, Hyperbolic Tangent, Graph, string, and tree Kernel function.

$$R(x) = w_o + w_1a_1 + w_2a_2 \qquad (4)$$

$$x = b + \Sigma_i \alpha_i y_i a(i) \times a \qquad (5)$$

In Equation 5 above, i - support vector, yi – training instance a(i) class value, b and α_i– numeric value defined by a learning algorithm.

K-nearest neighbors (KNN)

KNN is an instance-based and straightforward learning algorithm used for classification and regression-based problems in supervised machine learning.[41] In KNN, K is the number of nearest-neighbors considering taking a vote. Selecting a value of K can produce different classification results for the same data sample. Figure 5 shows an illustration of the KNN Classifier.

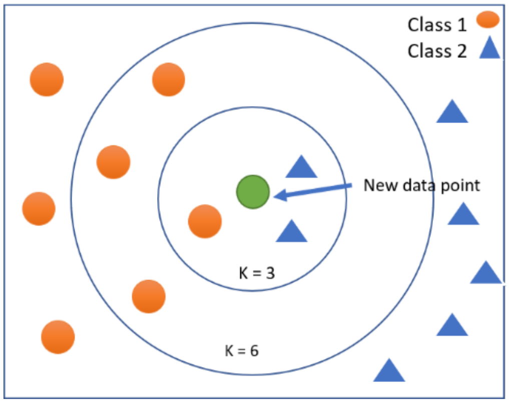


Figure 5 K-nearest neighbors (KNN)

KNN algorithm is always trying to predict the correct class for each testing data by calculating a distance between the testing and all training data, and based on that, the K value is selected, which is close to testing data.[42] KNN calculates the probability of testing data belonging to 'K' training data classes, and classes with the highest probability would be selected. Equations 6, 7, 8 represent three different methods used to calculate a distance for continuous and categorical data: 1) Euclidian, 2) Manhattan, and 3) Hamming distance.

$$\text{Euclidean Distance: -} \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \qquad (6)$$

$$\text{Manhattan Distance: -} \sum_{i=1}^{k}|x_i - y_i| \qquad (7)$$

Hamming Distance DH = $\sum_{i=1}^{k}|x_i - y_i|$ (8)

Where if x = y then D = 0 otherwise 1

Result and Discussion

Performance Evaluation Parameters

The proposed model accuracy is determined by setting the CAN Class value as a positive and notCAN (Normal) class value as a negative. The confusion matrix assesses different classification algorithms, also known as the error or contingency matrix[43]. The framework of a confusion matrix has four components: 1. True Positives (T.P.) – The classifier algorithm is correctly classified as CAN samples. 2. True Negatives (T.N.) – The classifier algorithm correctly classified as notCAN samples. 3. False Positive (F.P.) – notCAN samples were misclassified by a classifier. 4. False Negative (F.N.) – The classifier misclassified CAN samples. The confusion matrix considers the following measures to analyze classifiers' performance for various machine learning algorithms.[44]

Accuracy

It refers to a percentage of accurate estimates to total forecasts. It may be characterized as a capacity to predict an event's result accurately.

Accuracy = (TP+TN)/(TP+TN+FP+FN) (9)

Sensitivity/Recall/True Positive rate

The sensitivity calculates the percentage of adequately predicted positive observations by the class's total number of observations.

Sensitivity = TP/(TP+FN) (10)

Specificity

The specificity assesses the number of negative patterns in a class. The higher the specificity score, the more significant negative the classifier.

Specificity = TN/(TN+FP) (11)

Precision

It indicates the proportion of accurately forecasting positive patterns by a total predictive positive observation.

Precision = TP/(TP+FP) (12)

F1 Score

Precision and sensitivity are averaged in the F-measure. False positives and negatives are part of the process.

F1 Score = (2*T.P.)/(2*TP+FN+FP) (13)

False Positive rate/miss rate

It is referred to as a probability that the test would miss a true positive value.

False Positive rate = FP/(TN+FP) (14)

Classifier Discussion

Previous research has attempted to predict CAN using various methods. Linear and non-linear features were obtained from the HRV Signals, and It was found that non-linear approaches performed better in diagnosing CAN in T2 D.M. patients than linear methods.[45] Another study established a meta-ensemble model to examine the prediction of CAN using HRV. Jian & Lim et al. analyzed the HRV signal using principal component analysis(PCA).[46] Their values were fed into an SVM classifier, which yielded a diabetes detection accuracy of 79.93%. Swapna et al. applied Higher-order statistics (HOS) based features to diagnose CAN with a 90.5 percent accuracy.[24] Rajendra Acharya et al. obtained an accuracy of 92.64% by using a decision tree classifier.[24] Pachori et al. implemented logistic regression, and Evimp functions to detect co-occurrence of DM and CVD is 94.09%.[47] The previous works for automated detection of diabetes using HRV is given in Table 4.

Table 4. Summary of various methods using HRV Analysis

Authors	Method	Classifiers	Features	Accuracy(%)
U. Rajendra Acharya et al.[4]	HRV analysis using DWT	DT	8	92.64
		KNN	5	92.02
		NBC	13	62.58
		SVM poly 1	4	82.82
		SVM poly 2	6	85.28
		SVM Poly3	6	87.12
Acharya et al.[20]	Non-Linear	Perceptron - Adaboost	3	86
Swapna et al.[48]	Higher order spectral features	Bispectrum moments, entropies & weighted centres	8	90.5
Jian et al.[49]	HOS features	SVM	PCA	79.93
Acharya et al.[50]	HOS, Non-Linear	Least Squares – AdaBoost	4	90.0
Proposed Work	HRV Analysis (Linear & Non-Linear Method	Logistic Regression KNN SVM	27	96.67 93.33 93.33

In a current study, we have presented ML classification methods for assessing the risk of CAN in type 2 diabetic patients. The finding shows that ML models may detect the subclinical CAN early on and predict the chances of developing CAN in Type 2 diabetic patients. Selected classifiers are trained to discriminate between normal (notCAN) and CAN. During the learning of the model, extracted features of each class (i.e., normal and CAN class) from HRV signal are fed into a classifier as an input and the associated class label as output, and later on, each classifier is trained by 70% training dataset to understand the relationship among an input dataset (features) and both classes. During the testing phase, trained classifiers were tested by 30% of test data inputs to check whether a classifier generates a correct output class label. When unknown data input is provided to a trained classifier, the output label indicates which class (CAN or notCAN(normal)) data input belongs to.

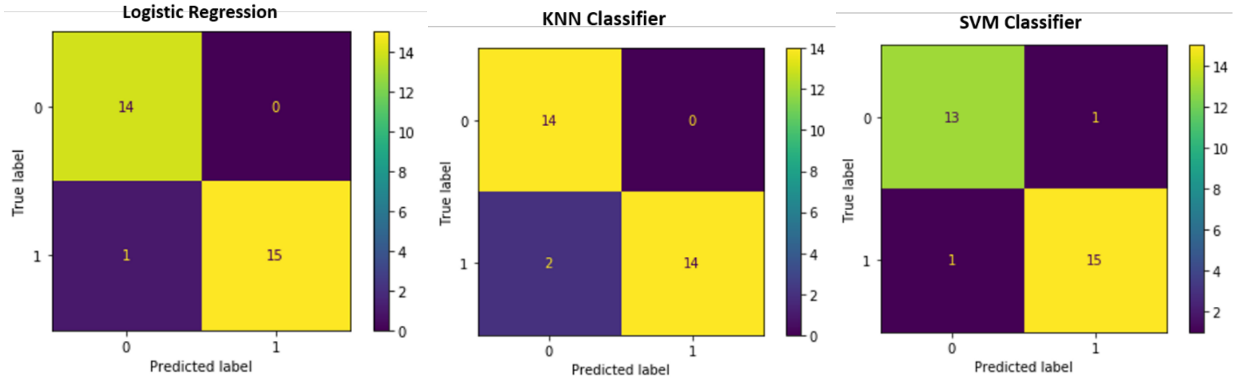


Figure 6 Confusion matrix of proposed classifiers

Table 4: Hyperparameter setting

Sr. No.	Classifier	Hyperparameters with their values
1.	Logistic Regression	C – 1, tolerance value – 0.01, solver - liblinear
2.	KNN	Leaf size – 5, neighbors – 3, weights – uniform, Distance - Euclidian
3.	SVM	C – 1, gamma value – 0.1, Kernel – RBF, max iteration - 10

Table 5: Analysis of Proposed classifier on CAN dataset

Classifier	Accuracy (%)	F1 Score (%)	Precision (%)	Sensitivity (%)	Specificity (%)	False Positive rate (%)
Logistic Regression	96.67	96.55	93.33	100	93.75	6.25
KNN	93.33	93.33	87.50	100	87.50	12.50
SVM	93.33	92.85	92.85	92.85	93.75	6.25

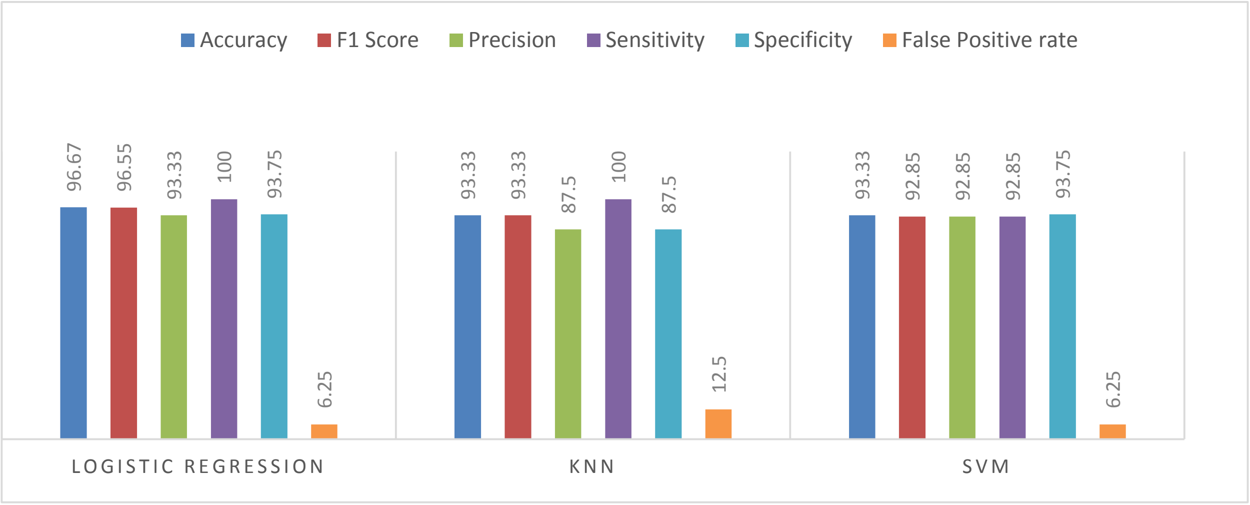


Figure 7 Graphical representation of Proposed Classifier

For validation of classifier, 5-fold CV method was applied. The whole dataset is divided into five equal parts. During a first fold, four parts of a dataset are used as classifier training, and the remaining one part will be used for testing, and based on that, evaluation parameters are calculated. The same procedure was followed five times using a different set of testing and training data in each fold, and from that, average performance measures from each fold were reported as a final performance measure. In our study, we have used 100 datasets out of those 70 datasets are used as training, and the remaining 30 datasets are used for testing. The greater the classifier's assessment parameters value, the more likely the classifier will correctly predict the class. The best combination of hyperparameters for classifiers was found by the Grid Search methods given below in Table 4 used for training and validating a model. The confusion matrices for the proposed methods are shown in Figure 6. It reveals that the provided model correctly recognized true positive and true negative occurrences. Table 5 gives a detailed summary of the evaluation parameters of all proposed classifiers. The CAN class provides accuracy, F1- Score, Precision, Sensitivity, Specificity, False positive rate. Figure 7 represents a graphical representation of the performance of the proposed model. The maximum accuracy of 96.67 % is achieved by Logistic regression while keeping the highest value for sensitivity and precision, and specificity. KNN classifier has an accuracy of 93.33%, similar to the SVM, and the highest sensitivity, 87.50 % precision and specificity, was measured. SVM classifier has high precision and specificity values compared to KNN.KNN classifier has a high false-positive rate of 12.50 % compared to Logistic regression and SVM.

Conclusion and future work

HRV is an early indicator of the onset of cardiac autonomic neuropathy and can identify Type 2 diabetes with high accuracy. In our study, we applied the Hamilton peak detection algorithm to detect heart rate signals, and we extracted important HRV features using both linear and nonlinear methods. To our knowledge, this is the first study to use all 30 essential features using linear and nonlinear methods, which can accurately distinguish between CAN and non-CAN (normal) classes. After that, we applied all the features to classifiers to detect CAN and normal classes. We achieved an accuracy of 96.67% in logistic regression, 93.33% in KNN, and 93.33% in SVM with the help of 5-fold cross-validation and hyperparameter optimization techniques to obtain the highest accuracy in detecting CAN using HRV. Our machine learning models have the potential to detect CAN at an early stage, which might help healthcare practitioners forecast CAN incidence in diabetic patients. However, the suggested methods were tested on small datasets, which is a limitation.

In future work, we will improve the model's performance by using a larger dataset and identifying unique parameters that indicate the presence of CAN. We can develop a mathematical equation for selecting these unique parameters based on essential features for classifying normal and CAN conditions. By fixing threshold values, we can separate the class between CAN detected and not detected. These same unique parameters can be applied to various deep learning algorithms for early and accurate prediction of CAN.

Acknowledgement

Authors are thankful to college for providing necessary facilities for carrying out this work.

Conflict of Interest

Authors do not have any conflict of interest for publication of this work.

References

1. K. Ogurtsova, J.D. da Rocha Fernandes, Y. Huang, et al. IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Res Clin Pract* **2017**, 128, 40–50.
2. A. Sreeniwas Kumar, N. Sinha. Cardiovascular disease in India: A 360 degree overview. *Med J Armed Forces India* **2020**, 76 (1), 1–3.
3. G. Swapna, K.P. Soman, R. Vinayakumar. Diabetes Detection Using ECG Signals: An Overview; *Deep Learning Techniques for Biomedical and Health Informatics,* **2020**, 299-327.
4. U. Rajendra Acharya, K.S. Vidya, D.N. Ghista, et al. Computer-aided diagnosis of diabetic subjects by heart rate variability signals using discrete wavelet transform method. *Knowl Based Syst* **2015**, 81 (2015), 56–64.
5. S. Banthia, D.W. Bergner, A.B. Chicos, et al. Detection of cardiovascular autonomic neuropathy using exercise testing in patients with type 2 diabetes mellitus. *J Diabetes Complications* **2013**, 27 (1), 64–69.
6. A.S. Abdalrada, J. Abawajy, T. Al-Quraishi, S.M.S. Islam. Prediction of cardiac autonomic neuropathy using a machine learning model in patients with diabetes. *Ther Adv Endocrinol Metab* **2022**, 13.
7. S. Sharma, V. Bhatia. Gut microbiota based treatment for Diabetes mallitus (T2DM): Challenges and Opportunities. *Chem. Biol. Lett.* **2021**, 8 (1), 31–39.
8. D.J. Ewing, C.N. Martyn, R.J. Young, B.F. Clarke. The value of cardiovascular autonomic function tests: 10 years experience in diabetes. *Diabetes Care* **1985**, 8 (5), 491–498.
9. H. Chuduc, K. Nguyenphan, D. Nguyenviet. A Review of Heart Rate Variability and its Applications. *APCBEE Procedia* **2013**, 7, 80–85.
10. E. Mejía-Mejía, K. Budidha, T.Y. Abay, J.M. May, P.A. Kyriacou. Heart Rate Variability (HRV) and Pulse Rate Variability (PRV) for the Assessment of Autonomic Responses. *Front Physiol* **2020**, 11.
11. F. Shaffer, J.P. Ginsberg. An Overview of Heart Rate Variability Metrics and Norms. *Front Public Health* **2017**, 5.
12. S. Borde, V. Ratnaparkhe. Optimization in channel selection for EEG signal analysis of Sleep Disorder subjects. *J. Integr. Sci. Technol.* **2023**, 11 (3), 527.
13. S. Ferdousi, P. Gyeltshen. Type 2 Diabetes Mellitus: Cardiovascular Autonomic Neuropathy and Heart Rate Variability. In *Type 2 Diabetes*; Stoian, A. P., Ed.; IntechOpen, Rijeka, **2021**.
14. R. Castaldo, P. Melillo, U. Bracale, et al. Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis. *Biomed Signal Process Control* **2015**, 18, 370–377.
15. A.K.F. Da Silva, M.P. Da Costa De Rezende Barbosa, F.M. Vanderlei, D.G.D. Christofaro, L.C.M. Vanderlei. Application of Heart Rate Variability in Diagnosis and Prognosis of Individuals with Diabetes Mellitus: Systematic Review. *Annals of Noninvasive Electrocardiology* **2016**, 21 (3), 223–235.
16. M.I. Stuckey, M.P. Tulppo, A.M. Kiviniemi, R.J. Petrella. Heart rate variability and the metabolic syndrome: a systematic review of the literature. *Diabetes Metab Res Rev* **2014**, 30 (8), 784–793.
17. I. Constant, D. Laude, I. Murat, J.-l. Elghozi. Pulse rate variability is not a surrogate for heart rate variability. *Clin Sci* **1999**, 97 (4), 391–397.
18. K. Nasim, H. Jahan Ara, A. Syed Sanowar. Heart rate variability - a review. *J. Basic Appl. Sci.* **2011**, 71–77..
19. B. Francesco, B. Maria Grazia, G. Emanuele, et al. Linear and nonlinear heart rate variability indexes in clinical practice. *Comput Math Methods Med* **2012**, 2012.
20. U.R. Acharya, O. Faust, S.V. Sree, et al. An integrated diabetic index using heart rate variability signal features for diagnosis of diabetes. *Comput Methods Biomech Biomed Engin* **2013**, 16 (2), 222–234.
21. B. Xhyheri, O. Manfrini, M. Mazzolini, C. Pizzi, R. Bugiardini. Heart Rate Variability Today. *Prog Cardiovasc Dis* **2012**, 55 (3), 321–331.

22. H.F. Jelinek, D.J. Cornforth, A. V Kelarev. Machine Learning Methods for Automated Detection of Severe Diabetic Neuropathy. *Journal of Diabetic Complications & Medicine* **2016**, 01 (02), 1–7.
23. M. Alkhodari, M. Rashid, M.A. Mukit, et al. Screening cardiovascular autonomic neuropathy in diabetic patients with microvascular complications using machine learning: a 24-hour heart rate variability study. *IEEE Access*, **2021**, 9, 119171-119187
24. G. Swapna, U. Rajendra Acharya, S. Vinithasree, J.S. Suri. Automated detection of diabetes using higher order spectral features extracted from heart rate signals. *Intelligent Data Analysis* **2013**, 17 (2), 309–326.
25. Cerebral Vasoregulation in Diabetes v1.0.0 https://physionet.org/content/cerebral-vasoreg-diabetes/1.0.0/ (accessed Oct 26, 2021).
26. P. Hamilton. Open source ECG analysis. *Comput Cardiol* **2002**, 29, 101–104.
27. S.R. Yadhuraj, B.G. Sudarshan, S.C. Prasanna Kumar, D. Mahesh Kumar. Analysis of linear and non-linear parameters of HRV for opting optimum parameters in wearable device. *Mater Today Proc* **2018**, 5 (4), 10644–10651.
28. T. Benichou, B. Pereira, M. Mermillod, et al. Heart rate variability in type 2 diabetes mellitus: A systematic review and meta-analysis. *PloS one*, 13(4), e0195166
29. A. Al-Hazimi, N. Al-Ama, A. Syiamic, R. Qosti, K. Al-Galil. Time-domain analysis of heart rate variability in diabetic patients with and without autonomic neuropathy. *Ann Saudi Med* **2002**, 22 (5–6), 400–403.
30. P. Kumar, A.K. Das, Prachita, S. Halder. Time-domain HRV Analysis of ECG Signal under Different Body Postures. *Procedia Comput Sci* **2020**, 167 (2019), 1705–1710.
31. O. Faust, U.R. Acharya, F. Molinari, S. Chattopadhyay, T. Tamura. Linear and non-linear analysis of cardiac health in diabetic subjects. *Biomed Signal Process Control* **2012**, 7 (3), 295–302.
32. M. Yilmaz, H. Kayancicek, Y. Cekici. Heart rate variability: Highlights from hidden signals. *J Integr Cardiol* **2018**, 4 (5), 1–8.
33. P. Seyd, V. Ahamed. Time and frequency domain analysis of heart rate variability and their correlations in diabetes mellitus. *Int. J. Biol. Life Sci.* **2008**, 4 (1), 24–27.
34. H.-H. A, E. Gospodinova, I. Domuschiev, N. Dey, A. Ashour. Nonlinear analysis of heart rate variability in Type 2 diabetic patients. *Fractal Geometry and Nonlinear Analysis in Medicine and Biology* **2016**, 1 (4), 134–139.
35. S.G. Caliskan, M. Polatli, M.D. Bilgin. Nonlinear analysis of heart rate variability of healthy subjects and patients with chronic obstructive pulmonary disease. *J Med Eng Technol* **2018**, 42 (4), 298–305.
36. G. Rojas-Vite, V. García-Muñoz, E.E. Rodríguez-Torres, E.L. Mateos-Salgado. Linear and nonlinear analysis of heart rate variability in essential hypertensive patients. *Eur. Physical J. Special Topics 2021* **2021**, 1–9.
37. R. Vicente, H.V. Ribeiro, R.R. Rosa, et al. A Fast Machine Learning Model for ECG-Based Heartbeat Classification and Arrhythmia Detection. *Frontiers in Physics* **2019**, 1, 103.
38. P. Refaeilzadeh, L. Tang, H. Liu. Cross-Validation. *Encyclopedia of Database Systems* **2009**, 532–538.
39. D. Berrar. Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* **2018**, 1–3, 542–545.
40. L. Yang, A. Shami. On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice. *Neurocomputing,* **2020**, 415, 295-316.
41. S. Uddin, A. Khan, M.E. Hossain, M.A. Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak* **2019**, 19 (1), 1-16.
42. I.H. Sarker, A.S.M. Kayes, P. Watters. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. *J Big Data* **2019**, 6 (1), 1-28.
43. P. Singh, N. Singh, K.K. Singh, A. Singh. Diagnosing of disease using machine learning. *Machine Learning and the Internet of Medical Things in Healthcare* **2021**, 89–111.
44. S.K. Arjaria, A.S. Rathore, J.S. Cherian. Kidney disease prediction using a machine learning approach: A comparative and comprehensive analysis. *Demystifying Big Data, Machine Learning, and Deep Learning for Healthcare Analytics* **2021**, 307–333.
45. O. Faust, U.R. Acharya, F. Molinari, S. Chattopadhyay, T. Tamura. Linear and non-linear analysis of cardiac health in diabetic subjects. *Biomedical Signal Processing and Control*; **2012**, 7, 295–302.
46. L.W. Jian, T.C. Lim. Automated detection of diabetes by means of higher order spectral features obtained from heart rate signals. *J Med Imaging Health Inform* **2013**, 3 (3), 440–447.
47. R.B. Pachori, M. Kumar, P. Avinash, K. Shashank, U.R. Acharya. An improved online paradigm for screening of diabetic patients using RR-interval signals. *J Mech Med Biol* **2016**, 16 (1), 1640003.
48. G. Swapna, U. Rajendra Acharya, S. Vinithasree, J.S. Suri. Automated detection of diabetes using higher order spectral features extracted from heart rate signals. *Intelligent Data Analysis* **2013**, 17 (2), 309–326.
49. L.W. Jian, T.C. Lim. Automated detection of diabetes by means of higher order spectral features obtained from heart rate signals. *J Med Imaging Health Inform* **2013**, 3 (3), 440–447.
50. U. Rajendra Acharya, O. Faust, N. Adib Kadri, J.S. Suri, W. Yu. Automated identification of normal and diabetes heart rate signals using nonlinear measures. *Comput Biol Med* **2013**, 43 (10), 1523–1529.