

Photoplethysmograph-based time-frequency and machine learning applications on biomedical signal analysis for medical diagnosis

Soumyadip Jana¹, Partha Sarathi Pal²

¹Department of Electrical Engineering, Techno Main Salt Lake, Kolkata, India

²Department of Electrical Engineering, Netaji Subhash Engineering College, Kolkata, India

Article Info

Article history:

Received Jul 19, 2024

Revised Oct 10, 2024

Accepted Oct 30, 2024

Keywords:

Biomedical signal analysis

Heart rate variability

Machine learning

PPG sensors

Time domain and frequency domain parameters

ABSTRACT

Machine learning (ML) integration in biomedical signal processing and medical diagnosis has the potential to revolutionize healthcare by improving diagnostic accuracy. This paper focuses on the applications of different ML algorithms for analyzing real-time physiological data collected from Photoplethysmography (PPG) sensors. Heart rate variability (HRV) analysis using electrocardiography (ECG) signals makes the process longer and bulky. Therefore, this paper demonstrates the real-time generation of HRV signals using a simple, low-cost, and non-invasive PPG sensor which is further processed using the Arduino ATMEGA328P microcontroller and then interfaced to a PC for display to investigate the usefulness of HRV feature analysis. HRV features have been computed using time domain analysis (TA), and frequency domain analysis (FA). At last, these TA and FA indices have been given to different ML models that could predict the gender, age group, and physiological conditions of a human being. Prediction of the physiological conditions using TA, FA, and ML models simultaneously makes the proposed approach more novel than the other existing methods. Comparative analysis of different ML approaches using ROC curves and confusion matrices has been shown to find the effectiveness and precision of different proposed models. It shows random forest ML approach has achieved 91% accuracy in identifying the physiological conditions. This simple yet accurate real-time PPG-based time-frequency ML system might be useful in medical assessment with faster response.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Soumyadip Jana

Department of Electrical Engineering, Techno Main Salt Lake

EM-4/1, Salt Lake City, Sector-V, Kolkata-700091, West Bengal, India

Email: soumyadipjana@gmail.com

1. INTRODUCTION

Biomedical signal processing and analysis play a crucial role in medical diagnosis and monitoring. In recent years, the advent of artificial intelligence (AI) and machine learning (ML) techniques has revolutionized the field, enabling more accurate and efficient medical diagnosis. Medical diagnosis using ML and physiological signals is an emerging field with great potential to improve healthcare outcomes. One such approach involves building models that diagnose medical conditions using heart rate variability (HRV) values from PPG sensors. PPG sensors typically consist of a light-emitting diode (LED) and photo detector placed on the skin, usually on the fingertip or earlobe, and measure various physiological parameters by illuminating the skin with light and detecting the resulting changes in blood volume. HRV, a physiological phenomenon, is a measure of the variation in time between successive heartbeats that reflects the balance

between the sympathetic (fight-or-flight) and the parasympathetic (rest-and-digest) branches of the autonomic nervous system, which regulate the heart rate.

These research studies span diverse applications of HRV analysis, showcasing the multidisciplinary nature of this field. Mojtahed *et al.* [1] have combined deep learning with wearable pressure and PPG data, enhancing accuracy in heart rate monitoring, which is vital for personalized health tracking. ML algorithms for detecting driver drowsiness based on HRV have been explored by CD Aswathi *et al.* [2] which contribute to the realm of driver safety technology. Bussas [3] have focused on employing ML to classify cognitive conditions using vital parameters, a crucial step in early diagnosis and intervention for cognitive disorders. Dolganov and Kublanov [4], [5] highlighted indicative factors of HRV for swift hypertension diagnosis, and evaluated the specificity of parameters in diagnosing arterial hypertension, providing valuable insights for hypertension management. Ishaque *et al.* [6] discussed the trends in HRV signal analysis, likely covering advancements and emerging methodologies. Wan-Hua *et al.* [7] investigated pulse wave forward peak detection and its cardiovascular applications, shedding light on novel diagnostic approaches. Matuz *et al.* [8] have employed ML models for HRV-based mental fatigue prediction, emphasizing the significance of HRV in mental health studies. Morales *et al.* [9] have shown sympathetic (fight-or-flight) and parasympathetic (rest-and-digest) branches of the autonomic nervous system, which regulate the heart rate. Wan-Hua *et al.* [7] and Sung *et al.* [10] have used different statistical features such as mean, root mean square values to provide useful information about the variability and regularity of the heart rate. Jepsen *et al.* [11] explained that HRV is constantly modulated through complex interactions between branches of the autonomic nervous system, the sympathetic nervous system, and the vagus nerve. Since the activity maintains nonlinear relationship therefore variation in the sympathetic activity or the vagal tone can change the response of the heart rate to the stimulation of any branch of the system. Birrenkott *et al.* [12] have shown that misshapen value between heartbeats is one of the first indicators of the existence of an anomaly in the patient's health which can reveal diverse conditions such as respiratory and cardiac arrest, systemic inflammatory response syndrome, renal insufficiency, cardiac insufficiency, systolic arterial pressure, among others. Papini *et al.* [13] have characterized Photoplethysmography (PPG) as one of the most popular technologies in the last decade for monitoring of the physiological conditions of a patient, and, because of its non-invasiveness, PPG has been largely applied to personal portable devices and pulse oximetry due to its convenience and capacity to perform continuous readings. In addition, the signal can provide information about both the cardiovascular and respiratory systems which creates vast viability of the utilization and easiness of the patient's physiological data acquisition. Albuquerque [14] efficiently incorporated PPG sensors into wristbands making these systems more accessible than the current electrocardiography (ECG) monitoring system which requires more number of electrodes to be attached to the patient's chest explained by researchers [15]-[17]. On the other hand, PPG signal does not have a complex hardware implementation and as well as the requirement of a reference signal. PPG is a non-invasive technique for measuring blood perfusion through tissues by the emission of light rays as explained by Peter *et al.* [18]. PPG signal extraction is considered simple; however, the components of this signal can provide valuable information about the cardiovascular system as described by Fan and Li [19]. He has explained the statistical time-domain indices obtained by the beat-to-beat determination. Four different nonlinear methods namely scaled amplified analysis (RSA), Higuchi fractal dimension (HFD), Displaced Flotation Analysis (DFA), and Exponential Generalized Hurst (GHE), have been applied by Karegar *et al.* [20] to extract resources for authentication of the ECG signal and study the nonlinear properties of this signal. The largest exponent of Lyapunov (LLE) has been used to extract useful characteristics of the PPG signal has been explained by Pham and Higa [21]. Quintero *et al.* [22] have rightly pointed out that the integration of ML algorithms with PPG data has the potential to enhance diagnostic accuracy, enable early detection of diseases, and facilitate personalized treatment strategies. The results of this study contribute to the growing body of research on ML-based approaches for biomedical signal analysis, paving the way for improved healthcare outcomes and patient care. Mejía-Mejía *et al.* [23] has shown time domain and frequency domain parameters can provide valuable information about the control of the cardiovascular system. PPG-derived inert-beat-intervals (IBI) have been proposed as a potential surrogate of HRV. Further, RR, an important marker extracted from HRV has been used to assess several cardiovascular, autonomic, and mental diseases using non-invasive indirect measurement. Priyadarshini *et al.* [24] have applied different machine-learning models to estimate PPG-based blood glucose, blood pressure, and activity detection. Mehrgardt *et al.* [25] have hypothesized that more accurate HR measurement results could be obtained by fusing data from complementing sensors and channels of varying modalities. A 3D-printed mechanical finger clip hosting multiple sensors, connected to an external electronics interface has been developed. Iqbal *et al.* [26] have used the deep learning method to use bilateral finger PPG sensors and CAD detection modules for the different classifications of coronary artery disease (CAD) patients. An uncertainty-based strategy for coronary artery disease screening from facial videos using

neural network decision-making has been proposed by Liu *et al.* [27]. Sadad *et al.* [28] have demonstrated cardiovascular disease detection based on PPG signal and using ML with cloud computing.

Hence a novel and suitable approach has been proposed in this paper combining the real-time PPG-based HRV with several time and frequency domain features and supervised ML models. Graphical analysis of time-frequency features in association with supervised learning models like logistic regression, K-nearest neighbors (KNN) algorithm, Naïve Bayes, support vector machines (SVM), decision trees, and random forests, have been used and known for their ability to handle complex datasets to provide accurate classification results. The ML models have been trained using the featured dataset of specific physiological conditions or states. By fitting the extracted parameters into these ML models, the aim was to establish a reliable and efficient system for medical diagnosis and to initially predict the physiological conditions of an individual. The models have been evaluated using confusion matrices and various performance metrics such as accuracy, sensitivity, and specificity to assess their effectiveness in distinguishing between different medical conditions and states.

2. METHOD

PPG signals have been collected from a group of more than 30 healthy Indian individuals both male and female within the age group of 18-35. A total of 300 HRV data are collected from these individuals during two months on different days of which 150 samples are for resting and 150 samples are for working/stressed conditions. The data were collected with a duration of 20 minutes including 10 minutes of adjustment, 5 minutes of rest (baseline), and 5 minutes of physical activity. As a type of physiological test, participants were needed to go upstairs and downstairs for 5 minutes to evaluate HRV under stress conditions. For the PPG signal, the PPG sensor connected to the Arduino board as shown in Figure 1 was placed two times in each subject both on the left and right index fingers consecutively. The subject was asked to sit down and make sure they were familiarized with the procedure. The PPG has been recorded after the device has been set up and the data has been imported into the MATLAB software for calculations as shown in Figure 2. Figures 2(a) and 2(b). In this study, time domain features have been calculated and analyzed from the HRV signal. Besides that, HRV features have been extracted which are the inter-beat interval (IBI) in a PPG signal. All the time-based HRV features have been calculated in the Python platform. Frequency domain parameters have been computed using the fast fourier transform (FFT) of the RR or inter beat (IB) interval data and obtained the power spectral density (PSD) by taking the absolute squared values of the FFT coefficients. The PSD represents the distribution of power (energy) across different frequency components in the RR interval signal. NumPy library has been used in Python code to perform the FFT and to calculate the PSD. Matplotlib library has also been used to create a frequency spectrum graph from the HRV data. From the different frequency domain approaches like FFT-based PSD, Poincare plots and nonlinear features like Shannon entropy analysis have also been accomplished for different physiological conditions. On the other hand, after computation correlation and dependency matrices of different time domain indices on mean HRV have been calculated using the Python platform. Most prominent distinctive features have been taken as input attributes for the different classifier models. In this work, out of 300 samples, 100 samples for each class i.e., a total of 200 samples have been considered to develop the training method for the different classifier algorithms, whereas separate 50 samples of each class, i.e., a total of 100 samples have been used to validate each of the proposed ML models. Binary logistic regression, KNN algorithm, random forest, SVM, and decision trees have been used for this purpose. All approaches have been done in the Python platform. Figure 3 represents the flowchart of the methodology adopted in this work.

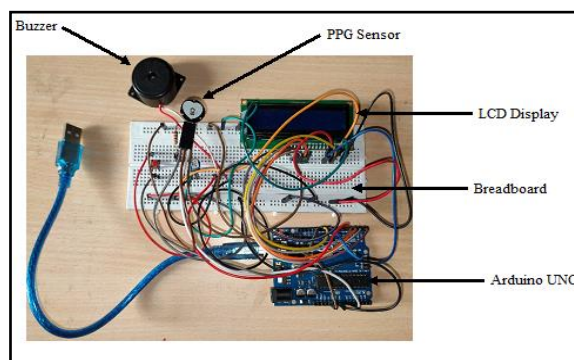


Figure 1. Hardware set up and components



Figure 2. The MATLAB software for calculations (a) final setup for the hardware implementation using the PPG and (b) data collection and monitoring using the serial monitor of Arduino

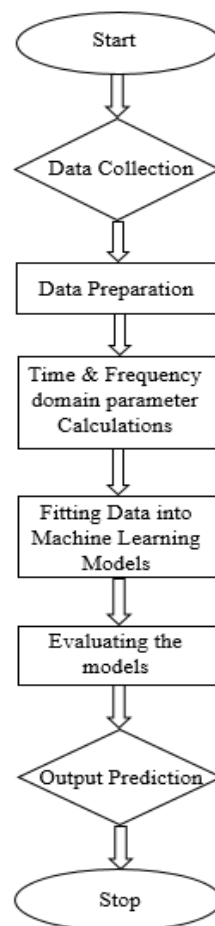


Figure 3. Flowchart of the proposed method

The distribution of the randomly selected HRV samples in milliseconds for resting and working/stress samples in box plots are shown in Figure 4 to represent the range of the HRV samples of different physiological conditions from different age groups and genders. Figures 4(a) and 4(b) depicts that resting persons have larger variations in their HRV values while working/stressed persons have less variations in their HRV ranges. This has been found as one of the key features for determining the different physiological conditions. The following HRV features as shown Table 1 have been computed.

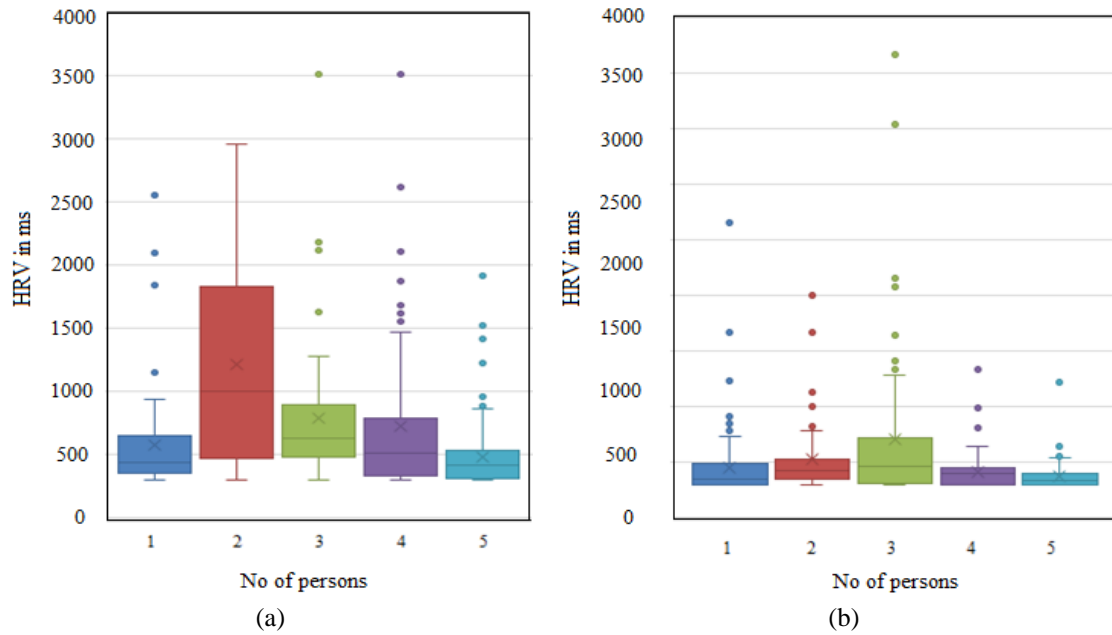


Figure 4. Variation of HRV for randomly chosen samples from resting (a) and working/stressed (b) persons of different age groups and genders

Table 1. Selected HRV features for extraction

Sl No	Processing method	HRV features	No of features
1	Time analysis	IBI, IBI _{difference} , HRV _{mean} , SD, SDNN, NN50, PNN50, NN20, PNN20, RMSSD, APV, SDDS, AVNN	13
2	Frequency analysis	SD ₁ , SD ₂ , SD ratio, LF power, HF power, VLF power, VHF power	7
3	Non-linear analysis	Shannon entropy	1
		Total	21

2.1.1. Time domain features

Figure 5 and Figure 6 show the variation of inter-beat interval (IBI) or peak-to-peak (PP) interval of randomly chosen samples from resting and working/stressed persons respectively of different age groups and genders. Figure 5 demonstrates that IBI values for resting people vary more in longer intervals while Figure 6 depicts that for working/stressed subjects IBI values vary with some lower intervals.

The difference between consecutive peaks, i.e., the variation in the time interval between successive heartbeats (IBI_{difference}) have also been shown in Figure 7 and Figure 8. From Figure 7 it has been found clearly that IBI_{difference} variation occurs less for resting people and Figure 8 shows larger variation in IBI_{difference} for working/stressed people.

All the time-based HRV features like standard deviation (SD), standard deviation of the normal-to-normal intervals (SDNN), normal-to-normal (NN) intervals that differ by more than 50 milliseconds (NN50), percentage of adjacent normal-to-normal intervals differing by more than 50 milliseconds (PNN50), normal-to-normal (NN) intervals that differ by more than 20 milliseconds (NN20), percentage of normal-to-normal (NN) intervals that differ by more than 20 milliseconds (PNN20), root mean square successive difference (RMSSD), average peak variability (APV), standard deviation of successive differences (SDDS), average NN interval (AVNN) which is the mean of the normal to the normal interval have been calculated in the Python platform and using the following equations that have been shown in Table 2.

Tables 3 and 4 show calculated time domain indices of randomly chosen samples from resting and working/stressed persons at different age groups and genders. It has been observed from the correlation and dependency matrices that the SD, SDNN, APV, RMSSD, and SDDS show more prominent distinctions between the different physiological conditions and shown in Figure 9. Hence these indices have been selected as the main inputs to the different ML approaches. Correlation and dependency matrices of different time domain indices on mean HRV for randomly chosen samples from resting Figure 9(a) and working/stressed and Figure 9(b) persons of different age groups and genders.

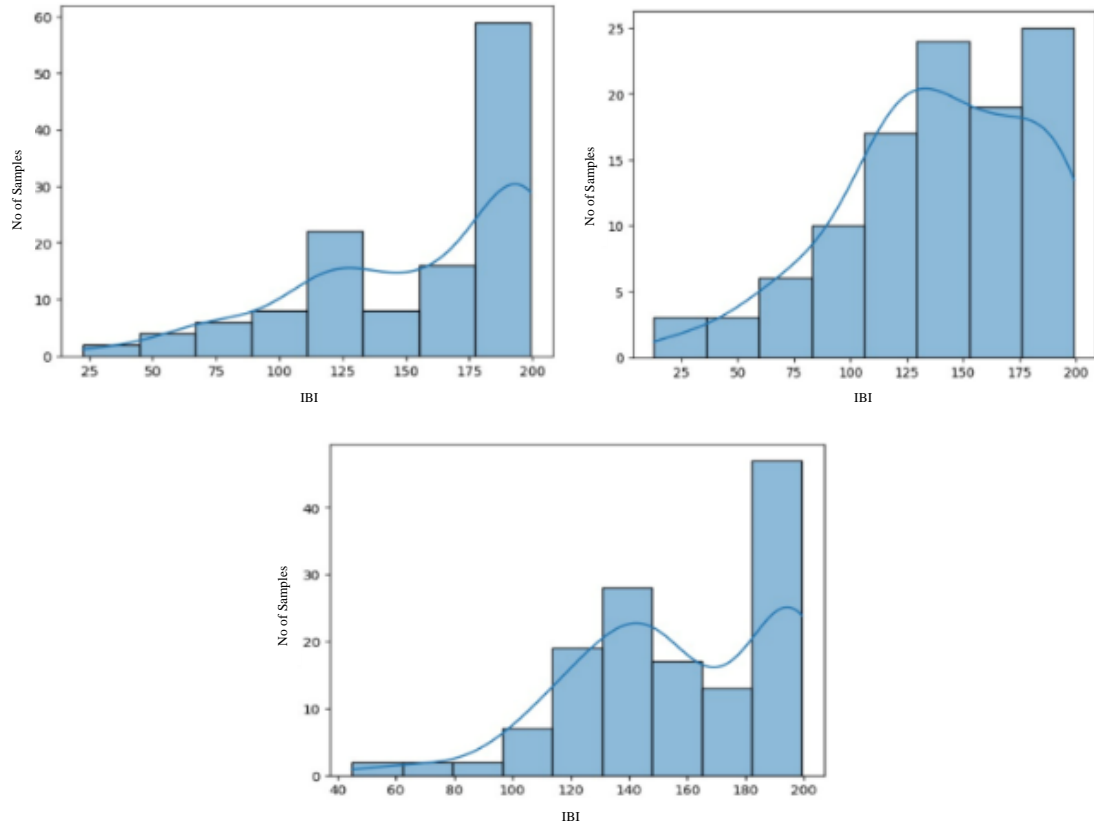


Figure 5. IBI values for resting people vary more in longer intervals

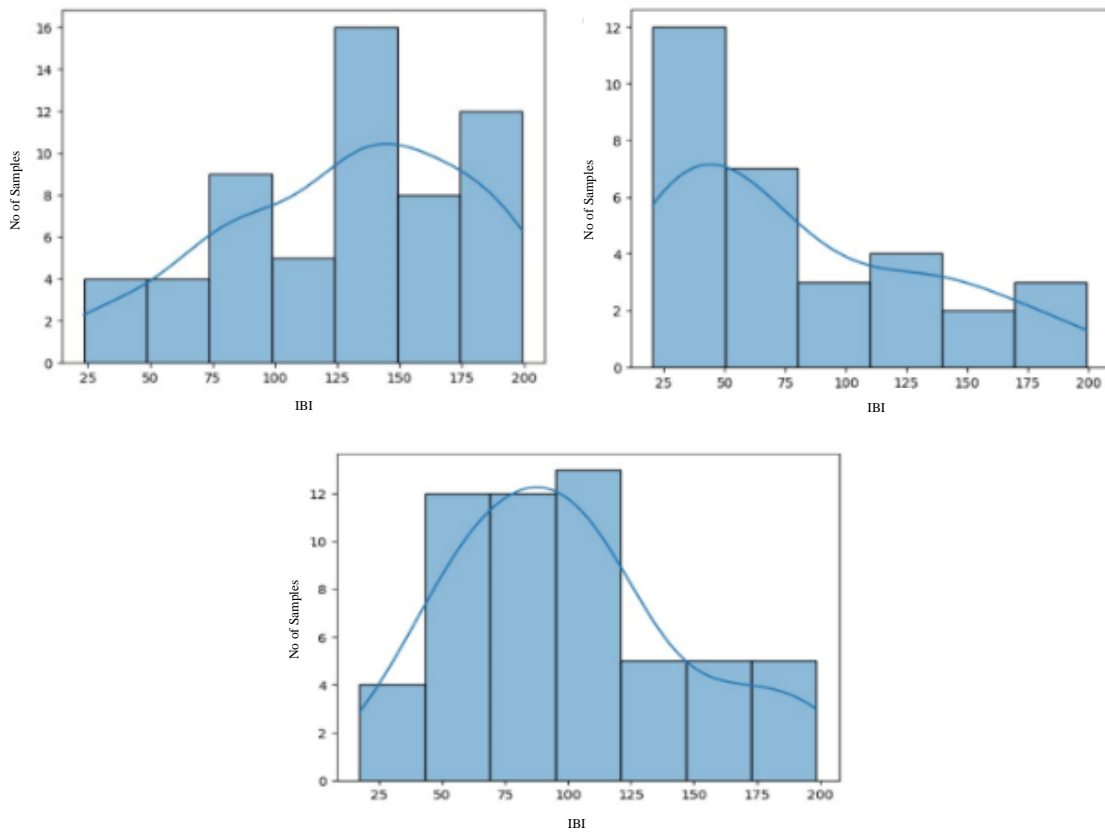


Figure 6. Working/stressed subjects IBI values vary with some lower intervals

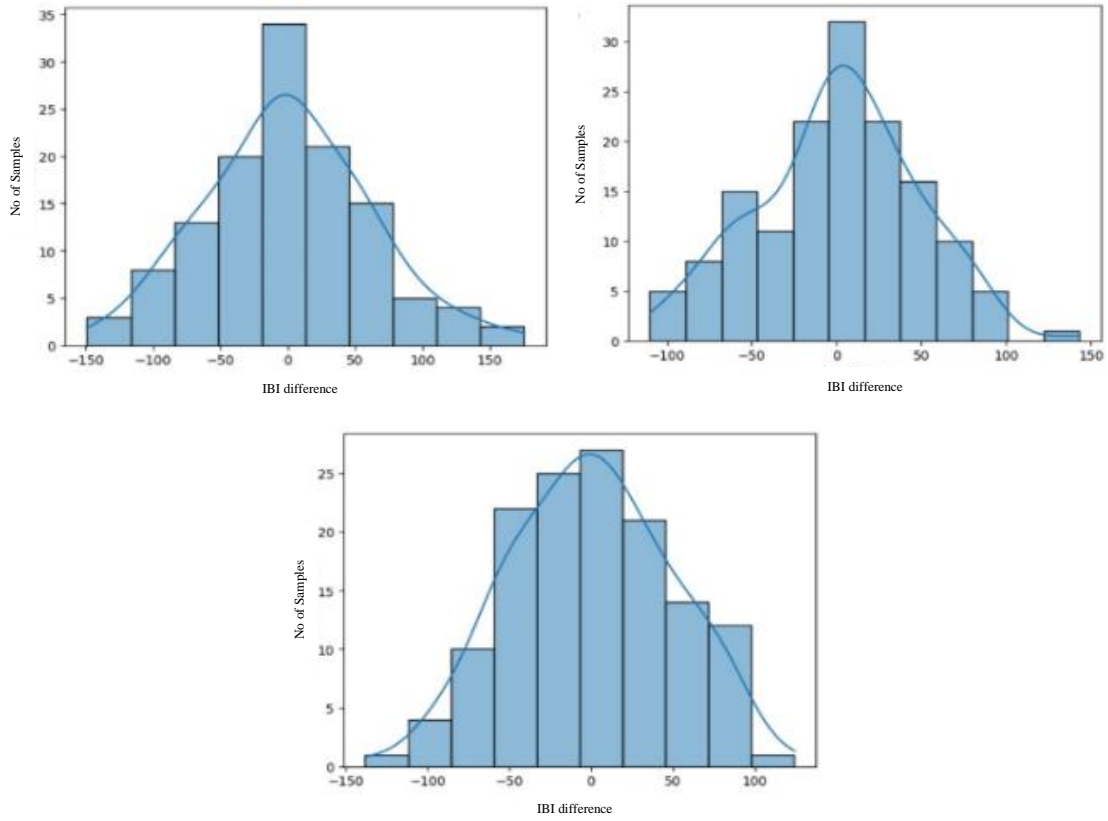
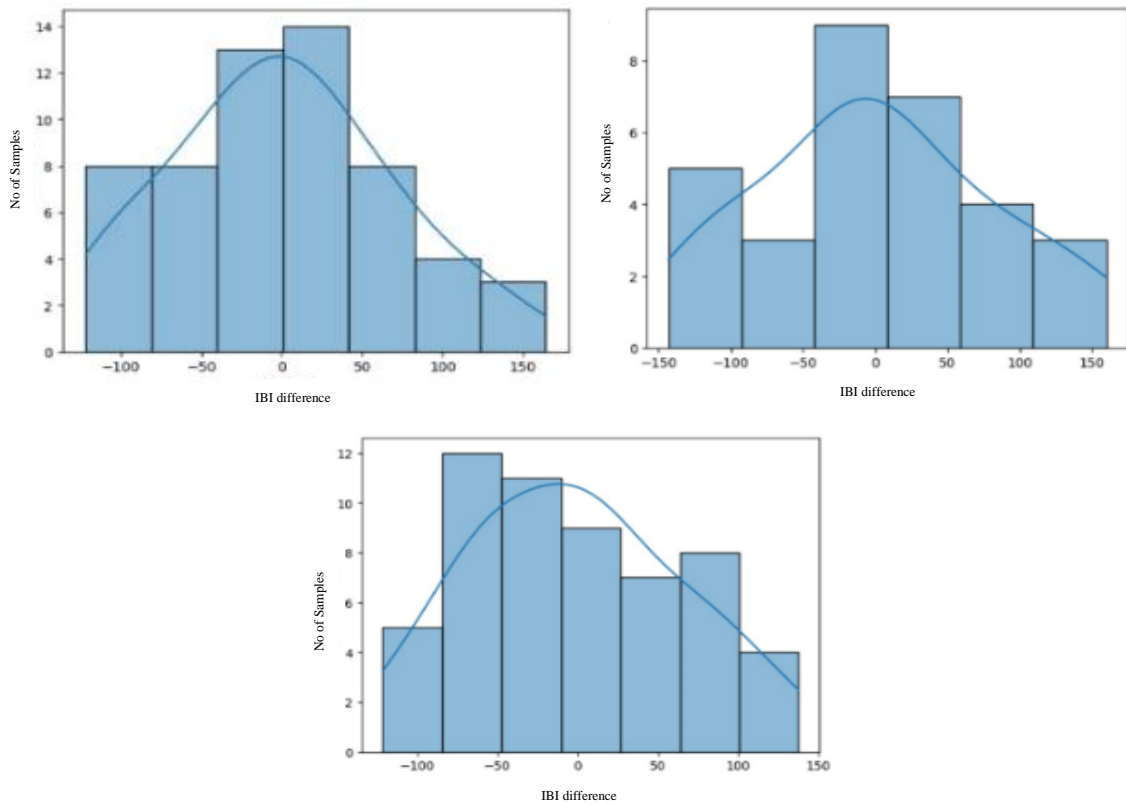


Figure 7. $IBI_{\text{difference}}$ variation occurs less for resting people



(c)

Figure 8. Larger variation in $IBI_{\text{difference}}$ for working/stressed people

Table 2. Equations used for the calculation of time domain parameters

SI No.	Time domain parameters	Equation
1	IBI	$IBI = \frac{60000}{HRV(ms)}$
2	IBI _{difference}	It refers to the difference between consecutive peaks, i.e., the variation in the time interval between successive heartbeats. $HRV_{mean} = \frac{\sum_0^{100}(HRV)}{n}$
3	Mean HRV	$HRV_{mean} = \frac{HRV_1 + HRV_2 + HRV_3 + \dots + HRV_{100}}{100}$
4	Standard deviation of HRV	$SD \text{ of HRV} = \sqrt{\frac{\sum_0^{100}(HRV - HRV_{mean})^2}{(n-1)}}$ Where, n=number of RRs in the specified period
5	SDNN	Standard deviation of normal-to-normal intervals reflects the overall variability of the heart rate and is calculated as the standard deviation of a set of inter-beat intervals measured during a specific period.
6	NN50	$NN50 = \sum RR_{n+1} - RR_n > 50 \text{ ms}$ Where, RR_{n+1} = The RR for the next R wave, RR_n = The RR for the current R wave
7	PNN50	$PNN50 = \frac{NN50 \times 100}{\text{Total Number of NN intervals}}$ Where NN50 = number of pairs of adjacent NN intervals that differ by more than 50 milliseconds
8	NN20	$NN20 = \sum RR_{n+1} - RR_n > 20 \text{ ms}$ Where, RR_{n+1} = The RR for the next R wave, RR_n = The RR for the current R wave
9	PNN20	$PNN20 = \frac{NN20 \times 100}{\text{Total Number of NN intervals}}$ Where NN20 = number of pairs of adjacent NN intervals that differ by more than 20 milliseconds
10	RMSSD	$RMSSD = \sqrt{\frac{\sum_0^{100}(RR_{n+1} - RR_n)^2}{(N+1)}}$ Where, RR_{n+1} = The IBI for the next R wave, RR_n = The IBI for the current R wave N = Total number of NN intervals
11	ARV	$ARV = \frac{\sum_0^{100} RR_{n+1} - RR_n }{(N-1)}$ Where, RR_{n+1} = The RR interval for the next R wave, RR_n = The RR interval for the current R wave
12	SDSD	$SDSD = \sqrt{\frac{\sum_0^{100}(RR_{n+1} - RR_n - \text{Mean RR})^2}{(N-2)}}$
13	AVNN	$AVNN = \frac{\sum NN \text{ Intervals}}{N}$

Table 3. Time domain parameters at resting condition

SI No	Time domain indices	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
1	IBI	127.579	80.4754	99.9172	124.4372	143.7754
2	IBI _{difference}	0.27390	0.39875	0.84423	0.02728	0.02331
3	Mean HRV	575	1210	786	725.5	486.7
4	SD	671.567303	845.565343	535.105	619.86840	256.190006
5	SDNN	50.81212	52.5339	45.5383	59.26368	45.548682
6	NN50	14	8	13	14	30
7	PNN50	0.233333	0.242424	0.22414	0.212121	0.2702703
8	NN20	21	11	20	20	41
9	PNN20	0.35	0.333333	0.34483	0.30303	0.3693694
10	RMSSD	74.87	81.896	67.418	70.799	69.832
11	ARV	60.817	67.764	54.961	54.991	57.464
12	SDSD	68.14811	81.09645	66.1766	70.99478	69.831905
13	AVNN	127.579	80.4754	99.9172	124.4372	143.77545

2.1.2. Frequency domain features

Frequency bins have been identified corresponding to the LF, HF, VHF, and VLF bands and sum the PSD values within these bands to obtain the LF and HF powers, respectively. The sampling rate has been taken as 4 Hz. Ultra-low frequency (ULF) and very low frequency (VLF) power has been obtained from the ULF and VLF power bands from the PSD of the RR interval data which has been shown in Figure 10. Figures 10(a) and 10(b) of a randomly chosen person at different physiological conditions. Less variation in resting subjects and larger variation in working/stressed subjects clearly distinguish different physiological conditions.

Table 4. Time domain parameters at working/stressed condition

Sl No	Time domain indices	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
1	IBI	156.2834	139.6485	126.363	157.0275	169.8324
2	IBI _{difference}	0.08470	0.04729	0.52790	0.19054	0.80402
3	Mean HRV	448.5	528.2	712	409.84	371.26
4	SD	282.33641	473.833657	664.725	136.781495	107.1690608
5	SDNN	44.86444	43.28091	58.8291	35.652218	31.38581
6	NN50	23	18	20	25	23
7	PNN50	0.18254	0.165138	0.25	0.1798561	0.154362
8	NN20	43	40	30	48	46
9	PNN20	0.34127	0.366972	0.375	0.3453237	0.308725
10	RMSSD	61.632	52.286	73.982	49.598	48.064
11	ARV	47.702	41.951	58.739	40.221	37.534
12	SDSD	61.771	52.28456	73.8335	49.547992	47.05842
13	AVNN	155.9443	139.6485	126.363	157.02752	169.8324

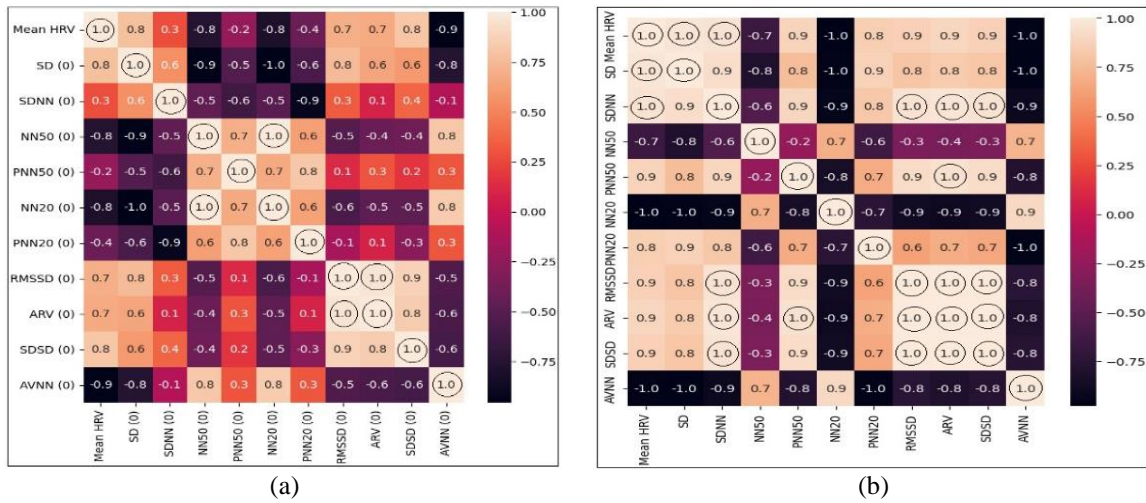


Figure 9. Correlation and dependency matrices of different time domain indices on mean HRV for randomly chosen samples from resting (a) and working/stressed (b) persons of different age groups and genders

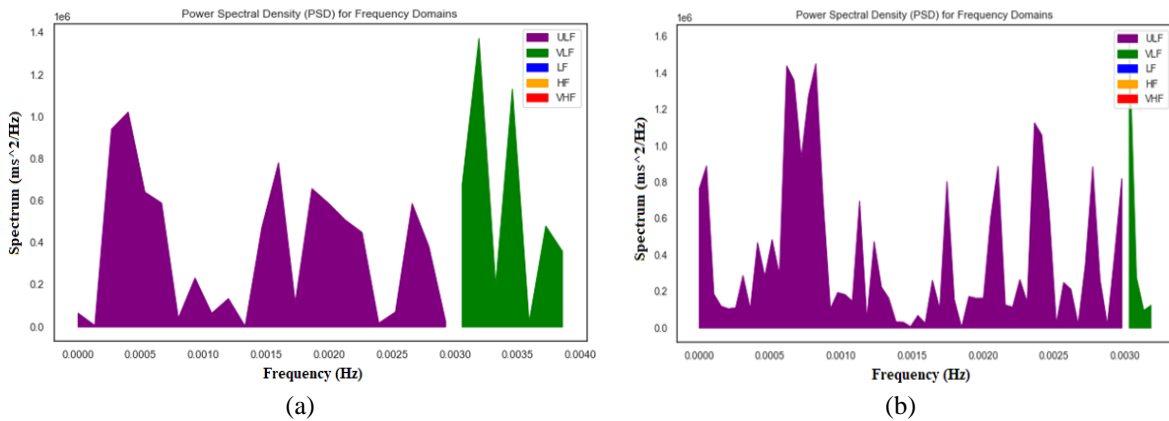


Figure 10. PSD of the RR interval for (a) a resting and (b) a working/stressed person

The VLF and the low frequency (LF) Bands range from 0.0033 to 0.04 Hz (or 3.3 to 40 mHz) and 0.04 to 0.15 Hz (or 40 to 150 mHz) respectively. The high-frequency band ranges from 0.15 to 0.4 Hz (or 150 to 400 mHz) and the Very High-Frequency Band ranges from 0.4 to 0.5 Hz (or 400 to 500 mHz). Frequency spectrum graph from the HRV data have been shown in Figure 11 and Figure 12. The graph displays the PSD, highlighting the VLF, LF, HF, and VHF bands in the background. At resting condition Figures 11 (a)-(c) depicts less variation and at working/stressed condition Figures 12(a)-12(c) depicts more variations in the frequency bands. Different frequency domain parameters have been shown in Tables 5 and 6.

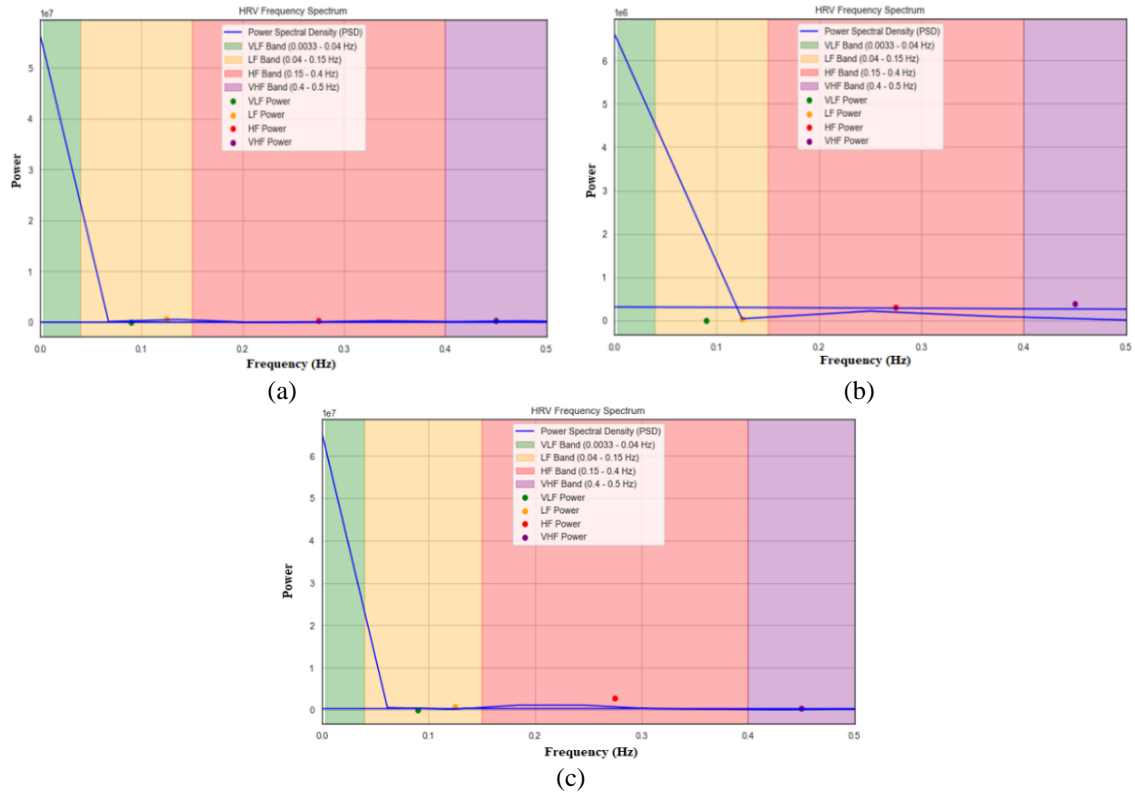


Figure 11. Power spectral density (a)-(c) in the frequency domain for randomly chosen samples from resting persons of different age groups and genders

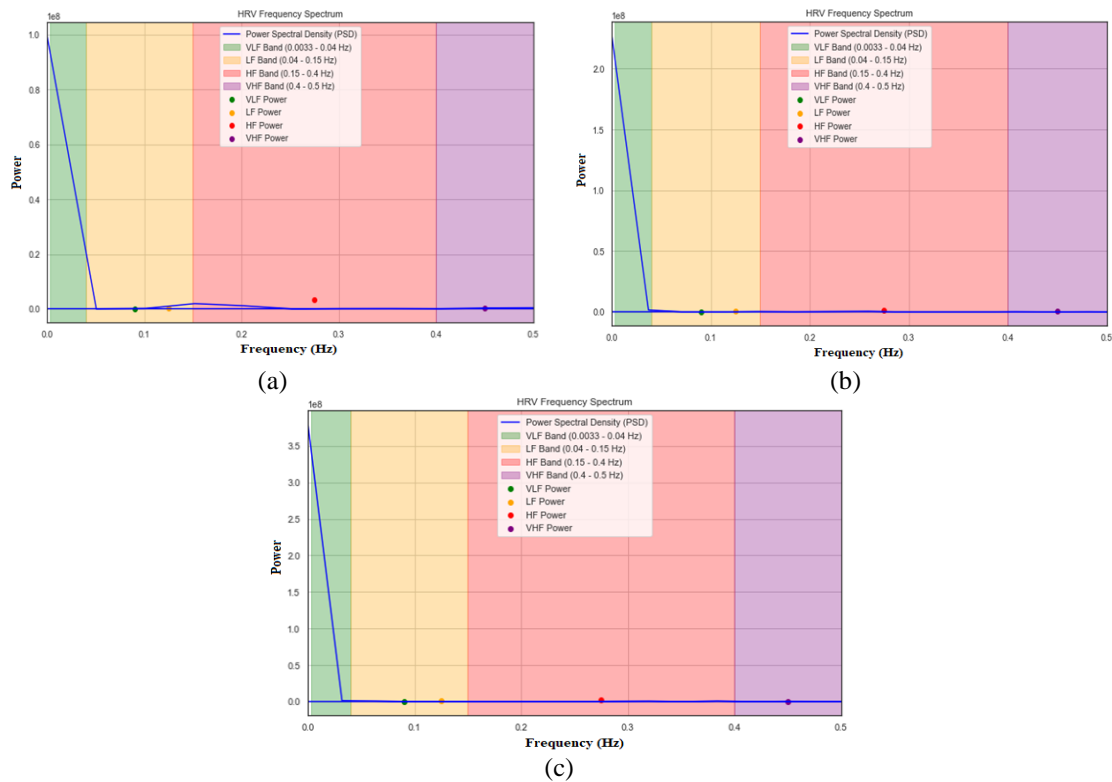


Figure 12. Power spectral density (a)-(c) in the frequency domain for randomly chosen samples from working/stressed persons of different age groups and genders

Table 5. Frequency domain parameters during resting conditions

SI No	Frequency domain indices	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
1	SD ₁	48.18071	57.33519	46.78685	50.19331	49.37116
2	SD ₂	2841.661	2231.304	1957.798	4504.238	1711.117
3	SD Ratio	58.97923	38.91683	41.84504	89.73781	34.65824
4	LF Power	673483.9273687	38860.839793	348084.009	733084.1346571	811986.4265374
5	HF Power	418384.8948564	305832.04258	178545.509	2739136.097402	2006693.112117
6	VLF Power	0	0	0	0	50155.41
7	VHF Power	387312.44	0	94627.35	221619.233	79199.16

Table 6. Frequency domain parameters during work/stress conditions

SI No	Frequency domain indices	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
1	SD ₁	43.67242	36.96519	52.20025	35.03043	33.2703
2	SD ₂	2121.001	2379.637	4196.038	1314.660	862.8903
3	SD Ratio	48.56614	64.37508	80.38348	37.52907	25.93575
4	LF Power	914000.0702291	504977.52335	277942.527	1375636.761779	363341.7327217
5	HF Power	2434037.858509	1309949.9175	3420239.21	1704670.035395	1420910.129393
6	VLF Power	1269613.27	1667310.43	0	421507.1	179543.98
7	VHF Power	704445.55	630481.57	461259.44	525549.02	372151.9

On the other hand, SD₁ represents the standard deviation of the perpendicular distance of each point on the scatter plot to the line of identity (x=y) and SD₂ is the standard deviation of each point from the y = x + average R-R interval specifies the length of the ellipse. Both have been obtained from Poincare plots or calculated using the equations Table 7.

Table 7. Equations used for the calculation of frequency domain parameters

SI No.	Frequency domain parameters	Equation
1	SD ₁	$SD_1 = \sqrt{\frac{\sum_0^{100}(\text{Diff } RR_n - \text{Diff } RR_{n+1})^2}{(2 \times N)}}$ <p>Where N is the total number of R-R interval pairs</p>
2	SD ₂	$SD_2 = \sqrt{\frac{\sum_0^{100}(\text{Diff } RR_n + \text{Diff } RR_{n+1})^2}{(2 \times N)}}$ <p>Where N is the total number of R-R interval pairs</p>

Poincare plot is a graphical representation of two components of the same signal in a 2D graphical plane. The high continuity of the poincare plot indicates low disturbances in the signal which signifies the difference between the consecutive elements is very low. But when discontinuity occurs, the signal generates disturbances in the form of rapid changes in the signal level.

In order to accurate analysis and to increase the classification accuracy of different physiological conditions SD₁ and SD₂ have been chosen for the poincare plots, the distribution of the SD₁ and SD₂ values for different physiological conditions is shown in Figure 13 and Figure 14 respectively. Figures 13 (a)-(d) explain high continuity of the poincare polt which indicates low disturbance at resting conditions of four randomly chosen subjects at different age groups and genders. On the other hand, Figures 14 (a)-(d) shows discontinuity which indicates high disturbances at working/stressed conditions of four randomly chosen subjects of different age groups and genders.

2.1.3. Non-linear features

The Shannon entropy for a discrete random variable X with probability mass-function P(X) is defined using the formula,

$$H(X) = - \sum [P(x) \times \log_2 P(x)]$$

It measures the uncertainty or randomness amongst IBI values. Different Shannon entropy values have been shown in Tables 8 and 9 and it can be concluded that higher values of entropy indicate greater variability and less predictability in the physiological conditions while lower values indicate more regularity and predictability in the analysis of different physiological conditions.

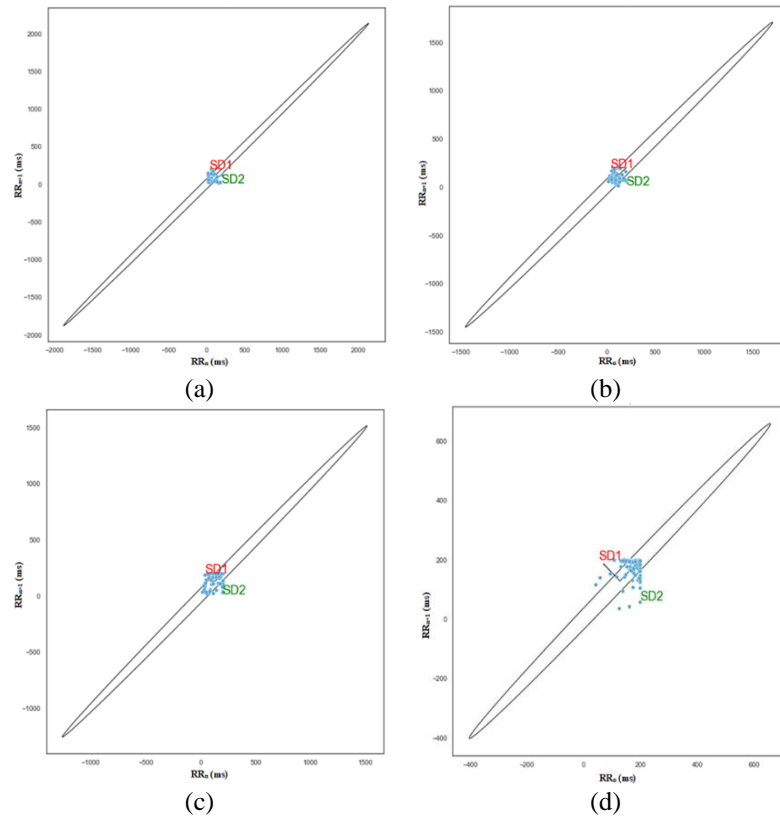


Figure 13. Poincare plot of the different resting person samples considering SD₁ and SD₂ (a)-(d) represents Poincare plot of randomly chosen samples from resting persons of different age groups and genders

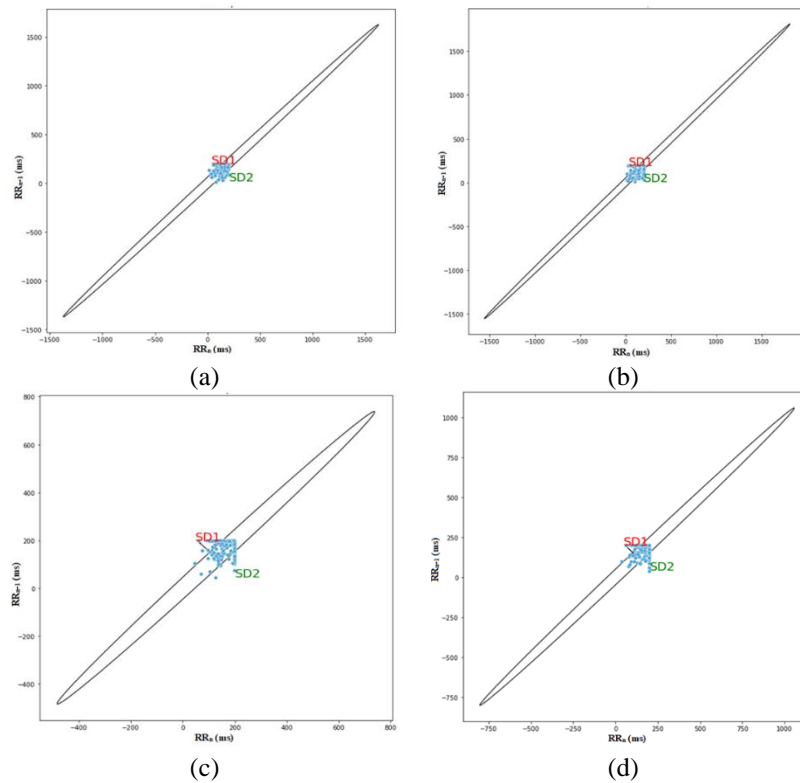


Figure 14. Poincare plot of the different running person samples considering SD₁ and SD₂ (a)-(d) poincare plots of randomly chosen samples from working/stressed persons of different age groups and genders

Table 8. Shannon entropy values at different physiological conditions: resting condition

Non-linear indices	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Shannon entropy	5.598661905257	5	5.72762686	5.658498821422	5.891801565544

Table 9. Shannon entropy values at different physiological conditions: working/stressed

Non-linear indices	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Shannon entropy	5.728621214283	6.3066803544	5.55870981	5.952485198987	5.638952533646

Developing classifier model using machine learning: binary logistic regression, a supervised ML algorithm has been used for the classification problem. All the inputs are taken as independent variables. A probability value between 0 and 1 is produced. Here two different classes have been taken resting as Class 0 and working/stressed as Class 1. The logistic function for input has been set to a threshold value. If it goes beyond the threshold value, then it belongs to Class 1 otherwise it belongs to 0. It has been referred to as binary logistic regression because the dependent variable is binary.

Being a nonparametric algorithm the KNN, a supervised ML method, has been employed to handle classification and regression problems simultaneously. In our case, $K=5$ has been taken. As a lazy learner algorithm, without learning from the training dataset KNN stores the dataset and acts on the dataset directly at the time of classification.

Random forest, a popular ensemble learning algorithm has been employed particularly in large datasets with high dimensionality. Multiple generations of decision trees make the algorithm strong enough. Multiple decision trees have been trained on randomly selected subsets of the dataset. In the initial step, these subsets are sampled randomly from the dataset. Subsequently, decision trees have been constructed based on these subsets, ensuring diversity among the trees. Once the trees are built, predictions are made by aggregating the individual tree outputs through voting for classification problems.

A decision boundary between two classes has been set up in the case of the SVM algorithm which chooses extreme cases or support vectors. For the SVM algorithm, the main goal is to create the decision boundary efficiently that can segregate n -dimensional space into classes so that the new data point can be put in the correct category in the future. For the creation of the best decision boundary, known as a hyperplane, SVM chooses the extreme points/vectors which are called as support vectors, and hence algorithm is termed as SVM.

In case of decision tree, the CART algorithm has been used for classification and regression problems simultaneously. Time domain features have been taken as the dataset and possible solutions have been made based on given conditions.

In this work, out of 300 samples, 100 samples for each class i.e., a total of 200 samples have been considered to develop the training method for the different classifier algorithms, whereas separate 50 samples of each class, i.e., a total of 100 samples have been used to validate each of the proposed ML models. For each HRV sample sex, age, HRV_{mean} , SD, RMSSD, SDSD, ARV, SDNN, PNN50, and PNN20 total of 10 numbers of attributes have been taken as input. Therefore, the size of the input vector is 200×10 .

3. RESULT AND DISCUSSION

In this work, the different classifier models have been used and tested using 100 samples considering 50 samples of each class from a separate dataset. It has been observed that time indices like SD, RMSSD, SDSD, ARV, and SDNN show prominent distinctiveness among the different physiological conditions. Further, it has been observed that PSD shows significant variations in the frequency bands. The Poincare plot shows the SD1 and SD2 values distribution considering low and high signal disturbances. Time indices feature-based ML classifier approaches are simple and fast. As a result, the proposed approach is ideal for a reliable and efficient system for medical diagnosis and to initially predict the physiological conditions of an individual and effective in distinguishing between different medical conditions and states.

The effectiveness of each applied model has also been performed by calculating the metrics such as accuracy, precision, recall, and F1-score. Accuracy measures the ratio of correct predictions to the total case examined making it suitable for balanced classification tasks. Precision addresses the genuinity of the predicted class which gives high confidence in the prediction class. Recall shows the actual proportion of the correct classification task. The F1-score represents the harmonic mean of precision and recall. All the results and calculated metrics of different ML approaches have been compared in Table 10. It has been observed

from Table 10 that the accuracy of the Random Forest ML approach is found to exceed 91% and the remaining classifier gives the accuracy below 90%. Figure 15 shows all the ROC curves of each ML model.

Thus, on an overall analysis, a high accuracy of classification exceeding 91%, with a moderate number of samples for developing the different classifier models, and finally introduction of a very simple, low-cost, and non-invasive scheme in developing a real-time simple time-frequency feature-based ML classifier for the assessment of different physiological states have been the highlights of the proposed work.

Table 10. Comparison of model accuracy

Quality of state Actual state	Predicted state		Total number of correct classifications	Classifier name	Classifier accuracy (%)	Classifier precision	Recall	F1- score
	Resting	Working						
Resting	43	7	43	Logistic regression	83%	0.85	0.8	0.824
Working	10	40	40					
Resting	44	6	44	KNN	83%	0.86	0.78	0.818
Working	11	39	39					
Resting	46	4	46	Random forest	91%	0.92	0.9	0.911
Working	5	45	45					
Resting	42	8	42	SVM	82%	0.83	0.8	0.815
Working	10	40	40					
Resting	41	9	41	Decision tree	81%	0.82	0.8	0.811
Working	10	40	40					

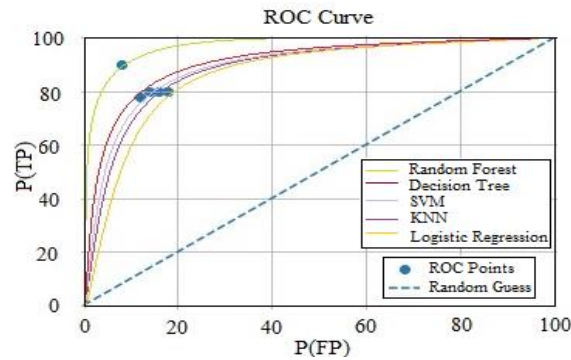


Figure 15. ROC curves of each ML model

4. CONCLUSION

The present work aims to construct accurate and efficient techniques for leveraging the power of ML techniques to enhance the accuracy and efficiency of HRV analysis, leading to improved understanding and diagnosis of various physiological and pathological conditions of human beings. In this proposed approach both time-frequency indices have been considered. It is found that time indices of HRV features produce consistent trends and significant differences during the resting and working/stressed phases. Poincare plot distribution has also been performed to analyze the quadrant. Simultaneously different supervised machine-learning models have been used using time domain indices. After comparison, it has been found that the random forest ML model has achieved an accuracy of 91%, a precision of 0.92, a recall value of 0.9, and an F1-score of 0.911 in the proposed scheme. The other models like logistic regression, KNN, SVM, and decision tree have achieved an accuracy of 83%, 82%, and 81% respectively. Hence the real-time simple time-frequency feature-based classifier makes the ML models' novel and suitable for identifying different medical conditions and states for real-life implementation.

ACKNOWLEDGMENTS

We would like to express our profound gratitude to our educational institution, Techno Main, Salt Lake, for providing us with the infrastructure and facilities to complete our research work with success. We also acknowledge Dr. Alok Mukherjee, Assistant Professor, Government College of Engineering and Ceramic Technology, Kolkata, India for his support.





REFERENCES

- [1] H. Mojtahed, R. R. Rao, C. Paolini, and M. Sarkar, "Temporal modeling of instantaneous interbeat interval based on physical activity," *IEEE Access*, vol. 11, pp. 138279–138291, 2023, doi: 10.1109/ACCESS.2023.3339584.
- [2] C. D. Aswathi, N. A. Mathew, K. S. Riyas, and R. Jose, "Comparison of machine learning algorithms for heart rate variability based driver drowsiness detection," in *2021 2nd Global Conference for Advancement in Technology, GCAT 2021*, Oct. 2021, pp. 1–7, doi: 10.1109/GCAT52182.2021.9587733.
- [3] M. Bussas, "9.1 - Telemetric acquisition of vitality parameters and classification of cognitive condition via machine learning," in *Proceeding - ettc2018*, 2018, pp. 179–184, doi: 10.5162/ettc2018/9.1.
- [4] A. Dolganov and V. Kublanov, "Presentation of the indicative factors of heart rate variability for hypertension swift-diagnostics," in *SIBIRCON 2019 - International Multi-Conference on Engineering, Computer and Information Sciences, Proceedings*, Oct. 2019, pp. 428–431, doi: 10.1109/SIBIRCON48586.2019.8958298.
- [5] A. Dolganov and V. Kublanov, "On some possibilities of the specificity evaluation for the expressing parameters of heart rate variability in task of the arterial hypertension diagnosing," *SIBIRCON 2019 - International Multi-Conference on Engineering, Computer and Information Sciences, Proceedings*, pp. 472–475, 2019, doi: 10.1109/SIBIRCON48586.2019.8958265.
- [6] S. Ishaque, N. Khan, and S. Krishnan, "Trends in heart-rate variability signal analysis," *Frontiers in Digital Health*, vol. 3, no. 2–18, p. 2018, Feb. 2021, doi: 10.3389/fdgh.2021.639444.
- [7] L. Wan-Hua, D. Zheng, G. Li, H. Zhou, and F. Chen, "Investigation on pulse wave forward peak detection and its applications in cardiovascular health," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 2, pp. 700–709, Feb. 2022, doi: 10.1109/TBME.2021.3103552.
- [8] A. Matuz, D. van der Linden, G. Darnai, and Á. Csathó, "Generalisable machine learning models trained on heart rate variability data to predict mental fatigue," *Scientific Reports*, vol. 12, no. 1, May 18, 2022, doi: 10.1038/s41598-022-24415-y.
- [9] J. Morales *et al.*, "Model-based evaluation of methods for respiratory sinus arrhythmia estimation," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 6, pp. 1882–1893, Jun. 2021, doi: 10.1109/TBME.2020.3028204.
- [10] C. W. Sung *et al.*, "Machine learning analysis of heart rate variability for the detection of seizures in comatose cardiac arrest survivors," *IEEE Access*, vol. 8, pp. 160515–160525, 2020, doi: 10.1109/ACCESS.2020.3020742.
- [11] M. F. Jepsen, L. V. Kessing, and K. Munkholm, "Heart rate variability in bipolar disorder: a systematic review and meta-analysis," *Neuroscience and Biobehavioral Reviews*, vol. 73, pp. 68–80, Feb. 2017, doi: 10.1016/j.neubiorev.2016.12.007.
- [12] D. A. Birrenkott, M. A. F. Pimentel, P. J. Watkinson, and D. A. Clifton, "A robust fusion model for estimating respiratory rate from photoplethysmography and electrocardiography," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 2033–2041, Sep. 2018, doi: 10.1109/TBME.2017.2778265.
- [13] G. B. Papini, P. Fonseca, X. L. Aubert, S. Overeem, J. W. M. Bergmans, and R. Vullings, "Photoplethysmography beat detection and pulse morphology quality assessment for signal reliability estimation," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Jul. 2017, pp. 117–120, doi: 10.1109/EMBC.2017.8036776.
- [14] V. H. C. de Albuquerque *et al.*, "Robust automated cardiac arrhythmia detection in ECG beat signals," *Neural Computing and Applications*, vol. 29, no. 3, pp. 679–693, Feb. 2018, doi: 10.1007/s00521-016-2472-8.
- [15] D. Zhao, Y. Sun, S. Wan, and F. Wang, "SFST: A robust framework for heart rate monitoring from photoplethysmography signals during physical activities," *Biomedical Signal Processing and Control*, vol. 33, pp. 316–324, Mar. 2017, doi: 10.1016/j.bspc.2016.12.005.
- [16] N. Pradhan, S. Rajan, A. Adler, and C. Redpath, "Classification of the quality of wristband-based photoplethysmography signals," in *2017 IEEE International Symposium on Medical Measurements and Applications, MeMeA 2017 - Proceedings*, May 2017, pp. 269–274, doi: 10.1109/MeMeA.2017.7985887.
- [17] A. F. Hussein, N. A. Kumar, M. Burbano-Fernandez, G. Ramirez-Gonzalez, E. Abdulhay, and V. H. C. De Albuquerque, "An automated remote cloud-based heart rate variability monitoring system," *IEEE Access*, vol. 6, pp. 77055–77064, 2018, doi: 10.1109/ACCESS.2018.2831209.
- [18] L. Peter, I. Vorek, B. Massot, I. Bryjova, and T. Urbanczyk, "Determination of blood vessels expandability: multichannel photoplethysmography," *IFAC-PapersOnLine*, vol. 49, no. 25, pp. 284–288, 2016, doi: 10.1016/j.ifacol.2016.12.048.
- [19] Q. Fan and K. Li, "Non-contact remote estimation of cardiovascular parameters," *Biomedical Signal Processing and Control*, vol. 40, pp. 192–203, Feb. 2018, doi: 10.1016/j.bspc.2017.09.022.
- [20] F. P. Karegar, A. Fallah, and S. Rashidi, "ECG based human authentication with using generalized hurst exponent," in *2017 25th Iranian Conference on Electrical Engineering, ICEE 2017*, May 2017, pp. 34–38, doi: 10.1109/IranianCEE.2017.7985480.
- [21] T. D. Pham and M. Oyama-Higa, "Photoplethysmography technology and its feature visualization for cognitive stimulation assessment," in *Proceedings of the IEEE International Conference on Industrial Technology*, Mar. 2015, vol. 2015-June, no. June, pp. 1735–1740, doi: 10.1109/ICIT.2015.7125348.
- [22] L. Quintero, P. Papapetrou, and J. E. Munoz, "Open-source physiological computing framework using heart rate variability in mobile virtual reality applications," in *Proceedings - 2019 IEEE International Conference on Artificial Intelligence and Virtual Reality, AIVR 2019*, Dec. 2019, pp. 126–133, doi: 10.1109/AIVR46125.2019.00027.
- [23] E. Mejía-Mejía, J. Allen, K. Budidha, C. El-Hajj, P. A. Kyriacou, and P. H. Charlton, "Photoplethysmography signal processing and synthesis," in *Photoplethysmography: Technology, Signal Analysis and Applications*, Elsevier, 2021, pp. 69–146.
- [24] R. G. Priyadarshini, M. Kalimuthu, S. Nikes, and M. Bhuvaneshwari, "Review of PPG signal using machine learning algorithms for blood pressure and glucose estimation," *IOP Conference Series: Materials Science and Engineering*, vol. 1084, no. 1, p. 012031, Mar. 2021, doi: 10.1088/1757-899x/1084/1/012031.
- [25] P. Mehrgardt, M. Khushi, S. Poon, and A. Withana, "Deep learning fused wearable pressure and PPG data for accurate heart rate monitoring," *IEEE Sensors Journal*, vol. 21, no. 23, pp. 27106–27115, Dec. 2021, doi: 10.1109/JSEN.2021.3123243.





- [26] S. Iqbal, S. Agarwal, I. Purcell, A. Murray, J. Bacardit, and J. Allen, "Deep learning identification of coronary artery disease from bilateral finger photoplethysmography sensing: A proof-of-concept study," *Biomedical Signal Processing and Control*, vol. 86, p. 104993, Sep. 2023, doi: 10.1016/j.bspc.2023.104993.
- [27] X. Liu, X. Yang, R. Song, J. Zhang, and L. Li, "VideoCAD: an uncertainty-driven neural network for coronary artery disease screening from facial videos," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2023, doi: 10.1109/TIM.2022.3229704.
- [28] T. Sadad, S. A. C. Bukhari, A. Munir, A. Ghani, A. M. El-Sherbeeney, and H. T. Rauf, "Detection of cardiovascular disease based on ppg signals using machine learning with cloud computing," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–11, Aug. 2022, doi: 10.1155/2022/1672677.

BIOGRAPHIES OF AUTHORS



Soumyadip Jana     is an assistant professor of the Electrical Engineering Department at Techno Main Salt Lake, Kolkata, West Bengal, India. He obtained his B. Tech and M. Tech Degrees from the West Bengal University of Technology in the year 2010 and 2012 respectively. Currently he is pursuing his Ph.D. degree under the guidance of Prof. Dr. Partha Sarathi Pal. His research interests include Bio-Medical Instrumentation, signal acquisition, signal processing, the application of artificial intelligence, and machine learning in signal analysis. He can be contacted at email: soumyadipjana@gmail.com.



Prof. Dr. Partha Sarathi Pal     is currently holding the position of professor of the Electrical Engineering Department at Netaji Subhash Engineering College, Garia, Kolkata, West Bengal, India. He obtained his B. Tech and M. Tech Degrees from the University of Calcutta in the years 1997 and 1999 respectively and Ph.D. degree from the Indian Institute of Technology (IIT) Kharagpur in the year 2006. His research interests include Photoplethysmography, Bioprocess control, Data acquisition systems, Instrumentation, Measurement and Web Technology. He has several research papers published in national and international journals. He can be contacted at email: pspal2k@gmail.com.