Study of Feature Extraction Algorithms on Photoplethysmography (PPG) Signals to Detect Coronary Heart Disease

Muhammad Fadhil Ihsan
School of Computing
Telkom University
Bandung, Indonesia
fadhilihsn@student.telkomuniversity.ac.id

Satria Mandala
Human Centric (HUMIC) Engineering &
School of Computing
Telkom University Bandung, Indonesia
satriamandala@telkomuniversity.ac.id

Miftah Pramudyo
Department of Cardiology and
Vascular - UNPAD
Bandung, Indonesia
miftah.pramudyo@gmail.com

2022 International Conference on Data Science and Its Applications (ICoDSA) | 978-1-6654-8665-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICoDSA55874.2022.9862855

***Abstract*—Coronary Heart Disease (CHD) is the most dangerous heart disease, this disease occurs, when the blood supply containing oxygen and nutrients to the heart muscle blocked by plaque in the heart blood vessels or coronary arteries. Currently, there are many ways of diagnosing coronary heart disease, starting from using ECG to Cardiac catheterization. However, it has some drawbacks, including the inflexibility of diagnosing quickly and invasive procedures. Heart rate variability (HRV) is a strong indication of cardiovascular diseases; as a result, any change in the normal heart rate (or blood volume) activity is a major marker for a potential cardiovascular malfunction. Through a series of waves and peak detection, photoplethysmography (PPG) detects blood pressure, oxygen saturation, and cardiac output. In recent years, there have been more studies using ECG signals to detect CHD compared to PPG signals, especially those discussing feature extraction on PPG signals in detecting CHD because this greatly affects the accuracy of CHD detection. In this study, proposed a literature study of feature extraction algorithm for detecting coronary heart disease using photoplethysmography. For the feature extraction, three algorithm will be discussed are respiratory rate (RR) interval, HRV Features and Time Domain Features. HRV features, with 94.4% accuracy, 100% sensitivity, and 90.9% specificity, is the best feature extraction approach of the three proposed techniques using decision tree classifier.**

Keywords—photoplethysmography, coronary heart disease, feature extraction, signal

I. INTRODUCTION

Cardiovascular diseases (CVDs) are the most common cause of death around the world. According to the World Health Organization (WHO), 17.9 million people died from cardiovascular disease (CVDs) in 2019, representing for 32% of all deaths worldwide [1]. Heart attacks and strokes were responsible for 85% of these deaths [2], [3] . One of the most common CVDs is coronary heart disease (CHD). When plaque in the heart blood vessels or coronary arteries blocks the blood supply oxygen and nutrients to the heart muscles, CHD occurs [4]. Plaque made up of deposits of cholesterol and other substances in the artery. CHD is a disease that is very dangerous for human health. Considering the risk, universality, complexity and danger of coronary heart disease, taking prompt diagnosis and timely treatment of coronary heart disease is the key.

Currently, there are many ways of diagnosing coronary heart disease, starting from using ECG to Cardiac catheterization [5], [6], [7]. However, it has some drawbacks, including the inflexibility of diagnosing quickly because the procedure is expensive and only in certain hospitals or clinics also an invasive procedures that can only be performed by highly skilled cardiologists [8]. Given the importance of cardiac monitoring and the difficulty of implementing it, it is critical to have a system that is widely accessible, economical, and non-invasive for the public to utilize. PPG satisfies all of the criteria [9].

Photoplethysmopgraphy (PPG) is a non-invasive technique for measuring capillary volumetric blood flow [10]. It does not necessitate expensive equipment and can be develop into a wearable device. PPG also been shown to be able to identify several diseases such as arrhythmia [11], atrial fibrillation [12] and coronary heart disease [13], [8]. Banerjee *et al.* [13] offer a composite feature set for diagnosing CAD patients based on heart rate variability (HRV) PPG features from datasets of MIMIC II dataset and 30 people gathered from an urban hospital. Their algorithm successfully achieved 88% specificity and 87% specificity. Paradkar *et al.* [8] extract distinguishing features from MIMIC-II (v2.6) matched subset database for classifying CAD patients using time domain analysis of PPG signal and its second derivative. The algorithms successfully achieved 78% specificity. In this research, three feature extraction methods will be used are respiratory rate (RR) interval, HRV Features and Time Domain Features. For the evaluation matrix, there are accuracy, sensitivity and specificity.

The following is how the rest the paper is organized: Section II is concerned with related knowledge on feature extraction; Section III is concerned with theory to achieving the objectives; Section IV is concerned with the method to achieve the objectives; Section V is concerned with the experiment and discussion; and Section VI is the conclusion and future works.

II. RELATED WORK

There are not so many research for feature extraction algorithms on PPG to detect coronary heart disease [14], [3]. Related work for identifying coronary heart disease using PPG has done in [13]. For identifying CHD from CHD patients and non-CHD, a composite feature set linked to heart rate variability (HRV) as well as the morphologies of PPG waveform defined in this study. Support Vector Machine (SVM) using Gaussian RBF kernel as a classifier and MIMIC II as a dataset, this results in an overall classification accuracy of 85% when using the proposed methodology. The results reveal that their methodology beats the previous art in terms of specificity with 88% for the MIMIC II dataset and 87% for the urban hospital dataset when compared to the features employed in [15].

In [16], the feature extraction method for classify coronary heart disease get research too, MIMIC II as the dataset for ranking and choosing the optimum features in this work experiment. Time series features, HRV features, and morphological features are the feature extraction methods used. One of the most strongly related features after the feature selection process is HRV features. In detecting CAD patients, the algorithms achieved average sensitivity and specificity of higher than 0.8. Using ANN as a classifier, the RR interval can be successful in classifying persons at variable risk of CHD, according to the results of [17]. Another crucial statistic inferred from the PPG [18] is the RR. All measurement based on the segment's identified and accepted peaks. RIFV: determined using a Fast Fourier Transform (FFT), RIIV: the maximum intensity of the PPG pulses, RR estimation quality, and RR fusion used to estimate RR in [19]. The performance of the RR algorithm evaluated using the un-normalized root mean square (RMS).

Until now, there are several supervised learning models to predict coronary heart disease such as KNN, Decision Tree, Random Forest, SVM and Multinomial Naïve Bayes. Based on the findings of [20], the Decision tree outperforms the Multinomial Naïve Bayes technique of predicting coronary artery disease. Decision Tree had a 99.63% accuracy rate, 100% sensitivity, 99.33% specificity, and 99.23% precision. These results shows that the Decision Tree technique, which incorporates independent data, is suitable for predicting coronary artery disease.

III. THEORY

For the feature extraction algorithms to detect coronary heart disease, the first step is to read all the data for healthy and CHD patients as a dataset. Then, the dataset will be denoised. The next step is to perform feature extraction on a dataset that has been denoised. The last step is to classify the dataset. The following explained theory of denoising method, feature extraction and classifier.

A. Finite Impulse Response (FIR) Filter

Filter with a finite number of impulse responses known as a FIR filter [21]. The filter design includes an initialization coefficient that matches the frequency response of the system specifications. This coefficient determines the response to the filter. Because the output of a finite input is always finite, FIR filters are always stable, and it has a property known as a linear phase [22].

B. RR Interval (RRI)

RR interval is the measurement distance between two consecutive beats. Individuals with variable CHD risk identified using the RR interval [17]. The R wave is the most striking wave because it has the highest peak. All measurements based on the segment identified and accepted peaks. Detecting the RR interval on the PPG signal is the initial process to get the characteristics of each signal [18].

Only the intervals formed by two contiguous, acceptable peaks considered when computing using the RR interval. This ensures that any peaks that rejected do not introduce measurement inaccuracy into subsequent measure calculations.

From identified peaks, the RR interval features calculated. The algorithms measures are Beats Per Minute (BPM), Inter Beat Interval (IBI), Standard Deviation of RR interval (SDNN), Standard Deviation of Successive Differences (SDSD), proportion of successive differences above 20ms (pNN20), proportion of successive differences above 50ms (pNN50), Median Absolute Deviation of RR intervals (MAD), breathing rate (BR) and Root Mean Square of Successive Differences (RMSSD) as detailed in [19].

C. Time Domain Features (TDF)

The analysis of physical signals, mathematical functions, or time series data known as time domain. The Time domain depicts how a signal evolves over time. Each cycle of a PPG waveform has two through points, a systolic peak, and a dicrotic notch point, as in Fig.1 [23]. For feature extraction, accurate detection of this site is necessary [13]. Time domain will be used to get the mean of peak to peak interval (PPI), standard deviation of PPI), mean of pulse signal and standard deviation of pulse signal.

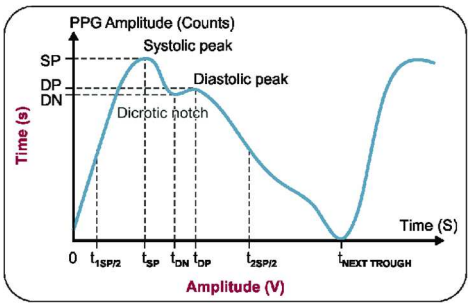

Fig. 1. Sample PPG Signal.

D. HRV Features (HRVF)

HRV is simply a measure of the variation in time between each heartbeat. The successive peak-peak distances in a signal measured to determine HRV-related properties. The Shannon entropy discovered to be a crucial feature for classification [16]. Shannon Entropy, consider as a random variable taking many values with a finite limit, and consider as its distribution of probability.

$$E_{sh} = -\sum_{m=1}^{N} P_m log P_m \quad (1)$$

Pm represents the empirical of probability of each bin. Spectral power normalized to three frequency regions (VLF, LF and HF) [13] is employed. Other features include, mean absolute deviation (MAD), kurtosis and skewness.

E. Decision Tree

Decision tree is a supervised learning technique that is particularly well suited to the solution of classification issues. It is a tree-structured classifier with internal nodes represent dataset features, braches that represent decision rules, and leaf that reflect the outcome [20]. The goal is to learn simple decision rules from data attributes to develop a model that predicts the value of a target variable.

IV. METHODS

The following are some explanations of the methods used in this study.

4.1 Design System

The design System of this study seen in Fig 2. First, input the PPG dataset obtained from an android application that been verified by a doctor. The next step is to design an algorithm consisting of three stages: preprocessing to clean up the noise in the PPG signal; feature extraction to extract features from the PPG signal to improve CHD detection accuracy; and obtain the results of CHD detection accuracy in classification. The last is to test several algorithms to find the best one.

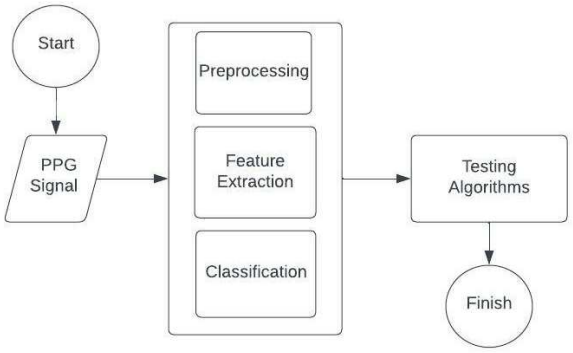

Fig. 2. Flowchart System.

4.2 Data

The study using data that consist of two groups. First group is CHD patients whose data were took at Salamun Hospital. Second group is healthy subjects taken from the surrounding community who are not diagnose with CVD or its risk factors. Total data includes 28 CHD patients and 30 healthy subjects.

4.3 Confusion Matrix

The Confusion Matrix is a matrix that is used to assess a model's performance [16]. The accuracy, sensitivity, and specificity of the model developed in this study measured using a confusion matrix. The following accuracy, sensitivity, specificity are shown in (2), (3) and (4).

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (2)$$

$$Sensitivity = \frac{TP}{(TP + FN)} \quad (3)$$

$$Specificity = \frac{TN}{(TN + FP)} \quad (4)$$

TP represent the value of a positive class in successfully predicts CHD. *TN* represent the value of a negative class in successfully predicts CHD. *FP* represent the value of a positive class in wrongly predicts CHD. *FN* represent the value of a negative class in wrongly predicts CHD [24].

4.4 Scenario of Experiments

4.4.1 Preprocessing

Program will read the PPG signals as a dataset. PPG signals that have been label are denoised in order to get a better signal, which can make it easier to get feature extraction. The process to remove PPG signal noise is using FIR filter.

4.4.2 Feature Extraction

Several feature extraction algorithms tested in this part to produce fiducial points on the denoised PPG data. The following are several scenarios in the feature extraction section.

4.4.2.1 Scenario I

In this scenario, a comparison made between three different algorithms, namely RR interval, time domain features and HRV features. After successfully obtaining the features of each algorithm, each algorithm classified using a decision tree.

4.4.2.2 Scenario II

Combining two different feature extraction algorithms is the idea of scenario II. The first combines the features obtained from each RR interval and HRV features; the second combines the features obtained from each RR interval and time domain features. Furthermore, the decision tree will be the classifier of each combination of features.

4.4.2.3 Scenario III

In scenario III, combines the features of three different algorithms, which obtained from each algorithms, namely, RR interval, time domain features, and HRV features. Then, using decision tree as a classifier.

V. Experiment and Discussion

5.1 Experiment Results (preprocessing)

The following is a sample PPG signal before and after denoising using an FIR filter.

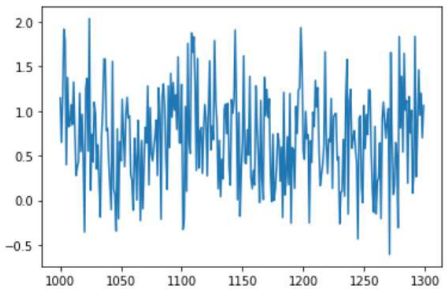

Fig. 3. Noisy PPG Signal Sample.

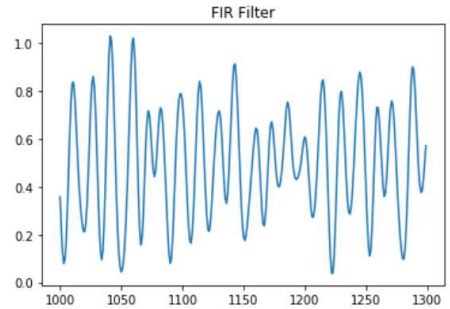

Fig. 4. Sample of Signal PPG Denoised.

5.2 Experiment Results (Feature Extraction)

5.2.1 HRV Features

The following are the features obtained from feature extraction using the HRV features algorithm on the PPG signal. Table I shows the value of each feature on healthy subjects and CHD patient.

TABLE I. HRV FEATURES

Feature	Value	
	Healthy	CHD
ESH	12.67	13.06
MAD	0.132	0.161
Skewness	0.215	-0.426
Kurtosis	-0.012	-0.106
VLF	0.0023	0.0062
LF	0.000064	0.000078
HF	8.816926e-07	1.744376e-07

5.2.2 RR Interval

The following are the features obtained from feature extraction using the RR interval features algorithm on the PPG signal using HeartPy library. Table II shows the value of each feature on healthy subjects and CHD patient

TABLE II. RR INTERVAL

Feature	Value	
	Healthy	CHD
BPM	86	87
IBI	694.17	6688.7
SDNN	162.02	147.18
SDSD	125.51	94.69
pNN20	1	0.875
pNN50	1	0.875
MAD	112.5	112.5
BR	0.22	0.415
RMSSD	246.09	190.78

5.2.3 Time Domain Features

The following are the features obtained from feature extraction using the time domain features algorithm on the PPG signal. Table III shows show the value of each feature on healthy subjects and CHD patient.

TABLE III. TIME DOMAIN FEATURES

Feature	Value	
	Healthy	CHD
Mean PPI	1.145	2.566
STD PPI	0.376	2.96
Mean signal	0.563	0.644
STD signal	0.169	0.131

5.3 Experiment Results (Confusion Matrix)

After feature extraction, performance evaluation is carried out with the confusion matrix mentioned earlier, namely accuracy, sensitivity, and specificity. The following tables VII to IX depict all matrix values. After this, there will be a discussion related to the result of the evaluation matrices.

A. Scenario I

Scenario I depicts a comparison of feature extraction algorithms. Table IV show the results of scenario I.

TABLE IV. COMPARISON BETWEEN DIFFERENT ALGORITHMS

Scenario I	Accuracy	Sensitivity	Specificity
RRI	88.89%	100%	80%
TDF	83.34%	71.42%	90.9%
HRVF	94.44%	100%	90.9%

The table above show that the maximum accuracy obtained at HRV features, with a score of 94.44%. Based on data, the lowest accuracy of the three algorithms is time domain features, at 83.34%. Based on data presented above, HRV features obtained the highest value on other confusion matrix, with 100% sensitivity, 90.9% specificity. The number of features used has no significant impact on result, but the importance of the features is. It is obvious from the time domain features that have less features than the HRV features and RR interval that has more features than HRV features.

B. Scenario II

Scenario II shows a test of combining two different algorithms. The results of Scenario II appear in Table V.

TABLE V. COMBINING TWO DIFFERENT ALGORITHMS

Scenario II	Accuracy	Sensitivity	Specificity
RRI + TDF	88.89%	100%	77.78%
RRI + HRVF	94.4%	100%	85.71%

The table above show that the maximum accuracy obtained from combining RRI and HRVF at 94.44%. Based on data, the lowest accuracy of combining two different algorithms are RRI+TDF at 88.89%. Accuracy results from combining two different algorithms, each combination still dominated by algorithm that have the highest results in scenario I.

C. Scenario III

Scenario III depicts a combining three different algorithms. Table VI shows the result of scenario III.

TABLE VI. COMBINING ALL ALGORITHMS

Scenario III	Accuracy	Sensitivity	Specificity
RRI + TDF + HRVF	94.44%	100%	90%

The table above show that combining three different algorithms obtained 94% accuracy, 100% sensitivity and 90% specificity. The highest result obtained by the combination algorithm that uses the HRV features because it measured from successive peak-peak distances and Shannon entropy as one of the important features in classification [16].

5.4 Discussion

HRV features proved to be strong in influencing the accuracy of CHD detection using PPG signals. Based on the evaluation of the three scenarios that been tested above using

a decision tree as a classifier, it shows that HRV features can produce better accuracy, sensitivity, specificity with 94.44%, 100%, 90.9%. It seen from scenario I, II, and III that each algorithm that uses HRV features in the feature extraction process has the highest accuracy in each test. It can happen because HRV features are obtain from the successive peak-peak distances of the signal. Shannon entropy used in HRV features discovered to be a crucial feature for classification [16]. The number of features possessed by the time domain features and RR interval did not significantly affect the high accuracy in this research.

VI. Conclusion

This research has achieved the objectives. Of the three methods, the best feature extraction algorithm is the HRV features. The accuracy, sensitivity and specificity values of each technique produce high values compared to other methods. The challenges in conducting this research are finding feature extraction algorithms for this kind of data and finding classifiers that are suitable for this kind of algorithms. For further study, researchers may be able to detect CHD by including feature selection method or researched PPG signal denoising to improve detection accuracy.

References

[1] "Cardiovascular Disease (CVDs)," Jun. 11, 2021. www.who.int/news-room/fact-sheets/detail/cardiovascular-disease-(cvds) (accessed May 17, 2022).

[2] M. Farhan, S. Mandala, and M. Pramudyo, "Detecting Heart Valve Disease Using Support Vector Machine Algorithm based on Phonocardiogram Signal," 2021. doi: 10.1109/ICICyTA53712.2021.9689142.

[3] W. R. Putra, S. Mandala, and M. Pramudyo, "Study of Feature Extraction Methods to Detect Valvular Heart Disease (VHD) Using a Phonocardiogram," 2021. doi: 10.1109/ICICyTA53712.2021.9689119.

[4] A. D. Choudhury and A. S. Chowdhury, "CHANGE: Cardiac Health Analysis Using Graph Eigenvalues," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2018, vol. 2018-July. doi: 10.1109/EMBC.2018.8513302.

[5] S. Mandala, Y. N. Fuadah, M. Arzaki, and F. E. Pambudi, "Performance analysis of wavelet-based denoising techniques for ECG signal," 2017. doi: 10.1109/ICoICT.2017.8074701.

[6] K. Husain, M. S. M. Zahid, S. U. Hassan, S. Hasbullah, and S. Mandala, "Advances of ECG sensors from hardware, software and format interoperability perspectives," *Electronics (Switzerland)*, vol. 10, no. 2. 2021. doi: 10.3390/electronics10020105.

[7] S. Mandala, T. Cai Di, M. S. Sunar, and Adiwijaya, "ECG-based prediction algorithm for imminent malignant ventricular arrhythmias using decision tree," *PLoS ONE*, vol. 15, no. 5, 2020, doi: 10.1371/journal.pone.0231635.

[8] N. Paradkar and S. Roy Chowdhury, "Coronary artery disease detection using photoplethysmography," 2017. doi: 10.1109/EMBC.2017.8036772.

[9] J. Kommineni, S. Mandala, M. S. Sunar, and P. M. Chakravarthy, "Advances in computer–human interaction for detecting facial expression using dual tree multi band wavelet transform and Gaussian mixture model," *Neural Computing and Applications*, 2020, doi: 10.1007/s00521-020-05037-9.

[10] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological Measurement*, vol. 28, no. 3, pp. R1–R39, Mar. 2007, doi: 10.1088/0967-3334/28/3/R01.

[11] L. F. Polania, L. K. Mestha, D. T. Huang, and J. P. Couderc, "Method for classifying cardiac arrhythmias using photoplethysmography," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2015, vol. 2015-November. doi: 10.1109/EMBC.2015.7319899.

[12] J. Lee, B. A. Reyes, D. D. McManus, O. Mathias, and K. H. Chon, "Atrial fibrillation detection using a smart phone," 2012. doi: 10.1109/EMBC.2012.6346146.

[13] R. Banerjee, R. Vempada, K. M. Mandana, A. D. Choudhury, and A. Pal, "Identifying coronary artery disease from photoplethysmogram," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, Sep. 2016, pp. 1084–1088. doi: 10.1145/2968219.2972712.

[14] J. Kommineni, S. Mandala, M. S. Sunar, and P. M. Chakravarthy, "Accurate computing of facial expression recognition using a hybrid feature extraction technique," *Journal of Supercomputing*, vol. 77, no. 5, 2021, doi: 10.1007/s11227-020-03468-8.

[15] G. Angius, D. Barcellona, E. Cauli, L. Meloni, and L. Raffo, "Myocardial infarction and Antiphospholipid Syndrome: A first study on finger PPG waveforms effects," in *Computing in Cardiology*, 2012, vol. 39.

[16] R. Banerjee, S. Bhattacharya, and S. Alam, "Time series and morphological feature extraction for classifying coronary artery disease from photoplethysmogram," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018, vol. 2018-April. doi: 10.1109/ICASSP.2018.8462604.

[17] F. Azuaje *et al.*, "Neural network approach to Coronary Heart Disease risk assessment based on short-term measurement of RR intervals," 1997. doi: 10.1109/cic.1997.647828.

[18] K. H. Shelley, "Photoplethysmography: Beyond the calculation of arterial oxygen saturation and heart rate," *Anesthesia and Analgesia*, vol. 105, no. SUPPL. 6, 2007, doi: 10.1213/01.ane.0000269512.82836.c9.

[19] W. Karlen, S. Raman, J. M. Ansermino, and G. A. Dumont, "Multiparameter respiratory rate estimation from the photoplethysmogram," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 7, 2013, doi: 10.1109/TBME.2013.2246160.

[20] Endang S Kresnawati, Yulia Resti, Bambang Suprihatin, M. Rendy Kurniawan, and Widya Ayu Amanda, "Coronary Artery Disease Prediction Using Decision Trees and Multinomial Naïve Bayes with k-Fold Cross Validation," *INOMATIKA*, vol. 3, no. 2, 2021, doi: 10.35438/inomatika.v3i2.266.

[21] Dimurtadha, Melinda, Elizar, and Dewi Meuthia, "Analisis Filter Finite Impulse Response (FIR) pada Sinyal Electroensephalogram (EEG)," *Seminar Nasional dan Expo Teknik Elektro*, pp. 101–104, 2019.

[22] A. Chatterjee and U. K. Roy, "PPG based heart rate algorithm improvement with butterworth IIR filter and Savitzky-Golay FIR filter," 2018. doi: 10.1109/IEMENTECH.2018.8465225.

[23] H. Aggarwal, P. B. Sharma, H. K. Aggarwal, D. Jain, and T. Pahuja, *Biosensors in Hypertension*. 2021. [Online]. Available: https://www.researchgate.net/publication/350072252

[24] D. Krishnani, A. Kumari, A. Dewangan, A. Singh, and N. S. Naik, "Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms," in *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, 2019, vol. 2019-October. doi: 10.1109/TENCON.2019.8929434.