2023 IEEE 4th International Multidisciplinary Conference on Engineering Technology (IMCET) | 979-8-3503-1382-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/IMCET59736.2023.10368256

Detecting Cardiovascular Disease From PPG Signals using Machine Learning

Tamara Sadek
Department of Electrical and Computer Engineering
American University of Beirut
Beirut, Lebanon
tos05@mail.aub.edu

Julia Ahmad
Department of Electrical and Computer Engineering
American University of Beirut
Beirut, Lebanon
jhh16@mail.aub.edu

Fadi Khoury
Biomedical Engineering Department
American University of Beirut
Beirut, Lebanon
fck05@mail.aub.edu

Heba Badawe
Biomedical Engineering Department
American University of Beirut
Beirut, Lebanon
hmb28@mail.aub.edu

Massoud Khraiche
Biomedical Engineering Department
American University of Beirut
Beirut, Lebanon
mkhraiche@aub.edu.lb

***Abstract*— Cardiovascular diseases (CVD) are the leading cause of death worldwide. Early detection of CVD allows for the use of effective and inexpensive interventions. CVD prediction using machine learning has achieved high accuracy using multiple data types such as Photoplethysmography (PPG), Electrocardiogram (ECG), and Blood Pressure (BP). However, these biological signals cannot all be reliably collected in real-time from a single wearable device and necessitate patients to be equipped with several devices. In this work, we trained a machine learning model to predict CVD using only PPG data collected from wearable devices, achieving an accuracy of 88% along with an F1-score of 90%. The information that this system can provide the patients with regarding their health status would enable the early detection and management of CVD, eventually decreasing the risk of morbidity.**

***Keywords*— cardiovascular diseases (CVD), Photoplethysmography (PPG), Machine Learning (ML)**

I. Introduction

Smart healthcare wearables provide affordable and non-invasive means of acquiring real-time vital information outside a controlled clinical setting[1, 2]. Advancements in the fields of electronics, communications and information technology provide access to efficient, lightweight, and compact wearable devices [3, 4], the integration of active recording instrumentation for example, has shown to significantly improve bio-signal fidelity [5, 6]. Enhanced devices like these contribute to the shift towards a more personalized healthcare system, promoting self-monitoring, and emphasizing preventative care and early detection. They enable individuals to take an active role in managing their health status. [1, 7].

Cardiovascular diseases are the leading cause of death worldwide with 17.9 million lives lost in 2019, making up to 32% of all global deaths with 85% of them being from stroke and heart attack. In 2017, 38% of the 17 million premature deaths under 70 years of age due to noncommunicable diseases were linked to CVD, disproportionately affecting low and middle-income countries, accounting for three quarters of them [8, 9]. Low economic budgets exacerbate the burden of CVD by limiting the access to early detection in primary care. Early detection of CVD in the long pre-symptomatic phase is crucial as it allows the use of highly effective and inexpensive interventions, reducing the need for invasive treatment options [10].

Photoplethysmography (PPG), an optical measurement technique that monitors heart rate, employs LEDs and photodetectors to measure the volumetric changes of blood during circulation from the surface of the skin. It is conventionally used for pulse oximetry and heart rate estimation. It possesses important indicators of cardiac health that can be extracted from the signal itself and from the first and second derivative.[11] PPG is inexpensive, widely accessible in wearable devices unlike other heart monitoring techniques like electrocardiogram (ECG), and does not require exact positioning[12].

Wearable devices provide an easy, low cost, and practical means to collect vital data which can be leveraged with machine learning (ML) towards potential early detection of CVD. Previous attempts to detect CVD using ML has been based on ECG, which introduces several limitations like short-term monitoring, higher cost, and possibly limited mobility during acquisition [13-16]. There have also been attempts to detect CVD by combining ECG and PPG, with models trained on features extracted from both signals, such as Pulse Transit Time (PTT) [16]. Systolic (SBP) and Diastolic Blood Pressure (DBP) have been used in all these studies, so traditional BP cuffs [17]. Although efforts have been made to estimate blood pressure without cuffs from wearable devices, satisfactory accuracy levels have not been achieved, rendering it currently unavailable [18, 19]. In this work, we propose using supervised machine learning (ML) to detect CVD solely from PPG data collected from wearable devices, providing individuals with a warning sign. This approach has the potential to assist in the early detection and improved management of CVD. We first performed pre-processing and feature extraction on the chosen dataset and trained ML algorithms on the extracted features and information available in the dataset, including SBP and DBP. After further optimization of the training process, we eliminated SBP and DBP from the training data to rely solely on data that can be acquired through wearables.

II. Methods

A. *Dataset Choice and Preprocessing*

We utilized an open-source dataset (PPG-BP) containing anonymous clinical data of 219 patients from Guilin People's Hospital, China [20]. The dataset includes three 2.1s finger PPG recordings for each patient, along with relevant clinical information including age, sex, BMI, and blood pressure. Patients were labeled by the authors of the dataset into 4 categories: Normal, Prehypertension, Stage 1 Hypertension, and Stage 2 Hypertension as depicted in Fig. 1.

Signal processing and feature extraction were performed using a custom code written in MATLAB R2023a. The Skewness Signal Quality Index (SSQI) was adopted to evaluate the quality of the PPG signals that were recorded after the SSQI was found to be the optimal SQI for PPG signals. The authors of the dataset reported the SSQI values of each of the 3 segments for every patient, allowing the selection of the segment with the highest SSQI for further processing.

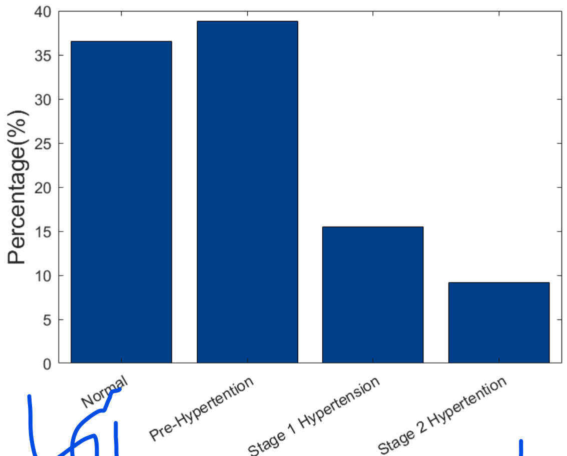

Fig. 1 Patient distribution in the dataset

Liang, Y. *et al* concluded that a 4th order Chebyshev II filter provided the best improvement in the PPG SQI after evaluating the performance of nine filters with 10 different orders for removing noise from the signals of this dataset. Therefore, the signals were filtered using a 4th order Chebyshev II bandpass filter (0.5 Hz-10 Hz) as shown in Fig. 2b [21]. The lower bound of the passband filter was chosen to remove the respiratory noise between 0.1Hz- 0.5 Hz and the DC component from the PPG signal leaving the pulsatile AC component. The upper bound was chosen to include the main frequency components of the PPG signal along with the heart rate related information captured from the fourth harmonics [22].

B. *Feature Extraction*

The supervised Machine Learning Algorithms were trained using features extracted from the chosen PPG signals along with the clinical data on each subject. Features were extracted from the filtered PPG signal, its 1st derivative, and its 2nd derivative after a thorough review of the existing literature:

1. Features extracted from the PPG signal:
   - Photoplethysmogram intensity ratio (PIR): quantifies the variation in the arterial diameter during one cardiac cycle, computed as the ratio of the peak intensity (IH) to the trough intensity (IL). IH was calculated as the average of the systolic peaks, and IL was the average of the troughs. Systolic peaks were detected using the "findpeaks" function in MATLAB with parameters (MinPeakHeight, MinPeakProminence and MinPeakDistance) carefully tuned by iterative testing to ensure accurate detection of systolic peaks as shown in Fig. 2c. The same procedure was employed to detect the troughs (Fig. 2c) by finding the peaks of the inverted signals.
   - AC Component Maximum Amplitude (MEU): represents the viscosity of the blood, calculated as the average of the AC amplitude peaks of the pulse signal [23].
   - Womersley number (alpha): a ratio that characterizes the inertial forces (pulsatile frequency) to viscous effects in blood flow. A low Womersley number indicates that the viscous forces are dominant increasing the resistance to the pulsatile blood flow affecting the Blood Pressure (BP). It is calculated as:

     $$alpha = IL * \left(\frac{\sqrt{1060} * Heart\ rate}{MEU} \right) \quad (1)$$

     where density of blood= 1060 kg/m^3 . [23]
   - Systolic Upstroke Time (SUT): represents the duration of the systolic phase of the cardiac cycle where the heart contracts. It is calculated

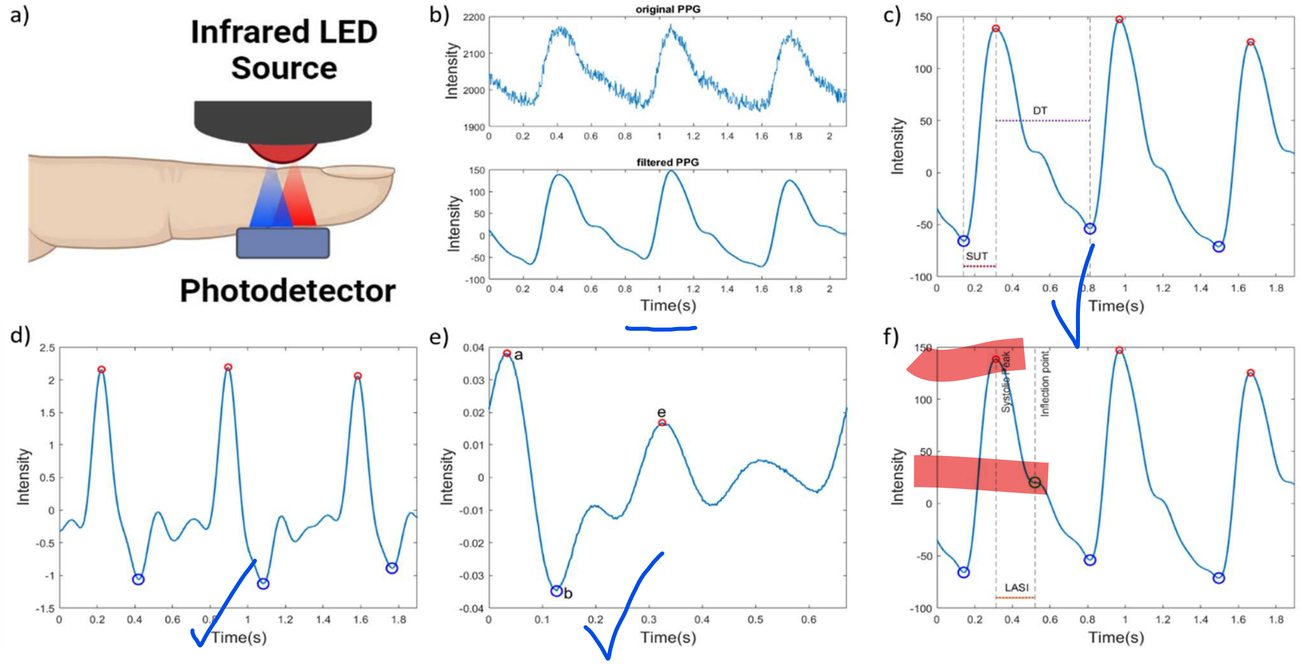

Fig. 2 PPG feature extraction. (a) PPG working principle. (b) original (top) and filtered (bottom) PPG signal sample. (c) peaks and trough detection of PPG signal with SUT and DT marked. (d) peak and trough detection of the first derivative, DPPG. (e) "a", "b", and "e" detection on the second derivative, SDPPG. (f) inflection point and LASI detection

as the average time interval between the systolic peak and the trough preceding it (Fig. 2c). [23]

- Diastolic time cycle (DT): represents the duration of the diastolic phase of the cardiac cycle where the heart relaxes. It is computed as the time interval from systolic peak till the subsequent trough (Fig. 2c) [23].

2. ~~Features~~ from the first derivative of the PPG signal (DPPG):
   - DPPG Height: defined as the average of the highest peaks in the first derivative. The peaks were detected using the "findpeaks" function, employing a suitable 'MinPeakHeight' to capture the highest peaks (Fig. 2d) [24].
   - DPPG Width: defined as the time interval between the highest peaks and the lowest troughs of the DPPG signal (Fig. 2d) [24].

3. Features from the second derivative of the PPG signal (SDPPG):
   - Large Artery Stiffness Index (LASI): an indicator of arterial stiffness. It exhibits an inverse relationship with the time interval between the dicrotic notch (inflection point) and systolic peak preceding it. Inflection points were located by finding the positions within the SDPPG signal where the amplitude is equal or very close to 0. The first inflection point after the systolic peak within a certain threshold determined using an iterative process was identified, and the time between this point and the systolic peak was computed. LASI was calculated as the ratio of the patient's height to this time interval (Fig. 2f) [18, 23].
   - Augmented Index (AI): represents a measure of the wave reflection on the wall of the arteries. It is an indicator of arterial stiffness; a cardiovascular disease marker. It is the ratio of the magnitude at the inflection point to the magnitude of the systolic peak [18, 25].
   - Peaks of SDPPG: The SDPPG signal has 5 peaks labeled as "a", "b", "c", "d", and "e" following the naming proposed by Elgendi, M. et al [26]. While detecting peaks "a" and "b" was straightforward, variations in the PPG signal lead to the appearance of one peak instead of 3 distinct "c", "d", and "e" peaks. Following the recommendation by H. A. Hamid et al., when distinct "c", "d" and "e" peaks are absent, the peak was considered a merged "e" peak and features that rely on "c" and "d" peaks were not computed. The features that were calculated and their corresponding biological significance include:
     - ○ b/a: it is related to the stiffness of blood vessels, and it increases with age.
     - ○ e/a: it is inversely proportional to the vascular stiffness, and it decreases with age.
     - ○ (b-e)/a: it is related to atherosclerosis and vascular aging assessment (Fig. 2e) [22, 27].

Table I. Accuracy results

Algorithm			Accuracy	Precision	Recall	F1-Score
Logistic Regression	19-Features	Normal	0.94	0.92	0.92	0.92
		Diseased		0.95	0.95	0.95
	5-Features	Normal	0.88	0.85	0.85	0.85
		Diseased		0.9	0.9	0.9

C. *Model Selection and Training*

Following feature extraction, we partitioned the dataset into two groups, "diseased" (Prehypertension, Stage 1 and Stage 2 Hypertension), and "normal" to perform binary classification. We attempted to train multiple models like Logistic Regression, Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Adaboost with DT as the base classifier. The Binary Logistic Regression algorithm yielded the most favorable results, so further optimization and testing were focused on this model. The rest of the algorithms yielded an inferior performance with 54% accuracy for DT, 68 % for RF, 75% for SVM, and 63% for Adaboost. The logistic regression model used was from python's scikit-learn library using the "lbfgs" solver with all the other parameters left at their default settings. The inferior performance and overfitting observed from other algorithms may be attributed to the dataset imbalance where normal individuals constituted only 36.6% of the total dataset.

First, we divided the dataset into a 15% testing subset and an 85% training subset. We included all the calculated features excluding those used in more than one feature to avoid redundancy. For example, the calculated MEU was not included because it is naturally used in calculating the Womersley number. The final list of features computed, and those accompanied by the dataset were 19 including: Sex, Age, Height, Weight, SBP, DBP, Heart Rate, BMI, and the features described above.

To determine the most suitable and significant subset of features, we trained the model using every feature combination possible and chose the subset that yielded the best performance and accuracy. This feature combination includes: PPG Intensity Ratio, Diastolic Time, Age, Weight, and b/a.

III. Results and Discussion

The accuracy, precision, recall and F1-scores were reported for the LR models. Precision is defined as the percentage of positive cases predicted by the model that are actually positive cases, with emphasis on minimizing false positives, necessary to avoid misdiagnosis. Recall is defined as the percentage of positive cases correctly predicted as positive cases by the model, with emphasis on decreasing false negatives. This is extremely important in the medical context as it ensures the right detection of positive cases, allowing the patient to seek immediate medical attention. The F1-score combines accuracy and precision providing a middle ground between false positives and false negatives.

A. *Logistic Regression using 19 features*

Initially, our model was trained with a complete set of 19 available features including Systolic (SBP) and Diastolic Blood Pressure (DBP). An accuracy of 94% was obtained along with a recall of 95% indicating minimal occurrence of false negatives and a precision of 95% as shown in table I. The F1-score at 95% demonstrates the ability of our model to obtain a balance, minimizing both false negatives and false positives.

To evaluate the performance of our model, we adopted a 5-fold cross-validation strategy. The validation curve in fig. 3a showed a consistent trend where both the validation and training accuracies increased with the regularization parameter "c". The validation accuracy converged towards the training accuracy, suggesting an equilibrium between bias and variance avoiding both underfitting and overfitting. Consequently, our model showed a robust performance over a range of regularization strengths.

Our learning curve revealed a convergence trend as the training dataset size increased where the training accuracy curve initially dropped, given it was exposed to a larger and more diverse dataset as expected (Fig. 3b), ultimately stabilizing and converging with the validation curve. The validation accuracy curve initially increased, followed by its convergence towards the training accuracy curve, solidifying the reliability of our model while avoiding overfitting.

B. *Logistic Regression using Optimal Features Subset*

Our model was trained using the optimal feature subset that excluded SBP and DBP depending only on data that can be collected from a wearable device. An accuracy of 88% was achieved along with a recall of 90% indicating minimal occurrence of false negatives and a precision of 90% as shown in table I. The F1-score at 90% demonstrated the ability of our model to reach a balance, minimizing both false negatives and false positives.

The same procedure was followed to evaluate the model's performance. The validation curve in Fig. 3c showed a similar behavior to the model trained with 19 features, exhibiting a consistent trend where both the validation and

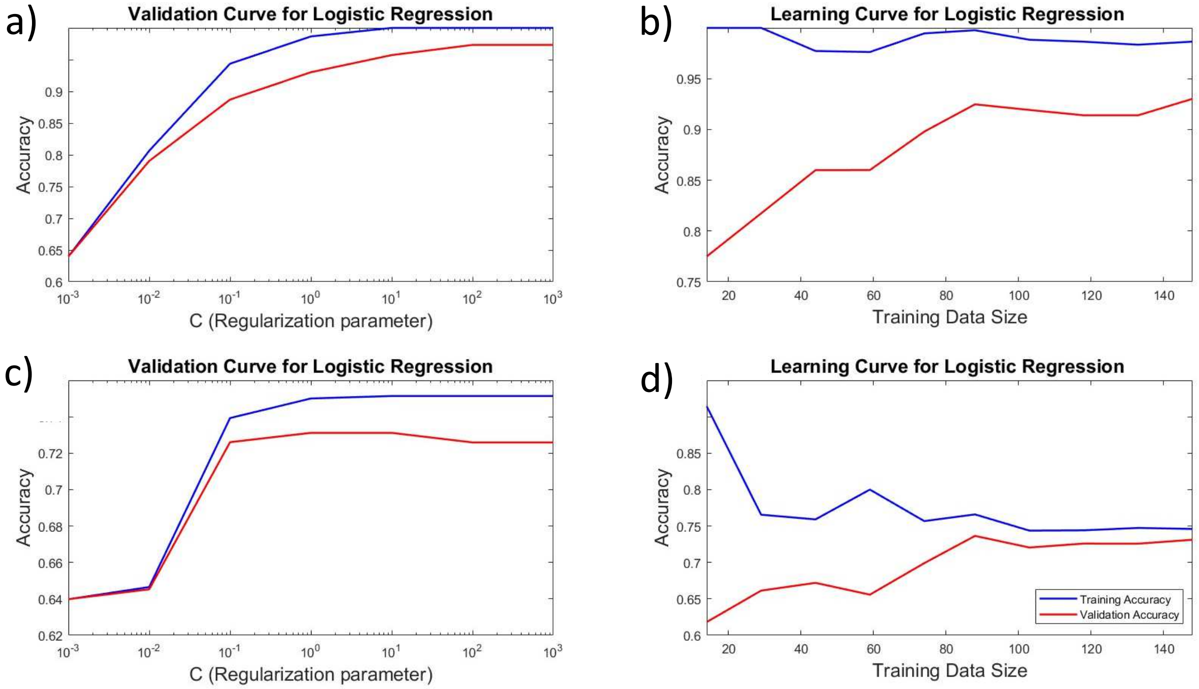

Figure 3. Validation and learning curves for Logistic Regression Algorithms. (a) validation curve for Logistic Regression with 19 features. (b) learning curve for Logistic Regression with 19 features. (c) validation curve for Logistic Regression with 5 features. (d) learning curve for Logistic Regression with 19 features.

training accuracies increased with the regularization parameter “c”, and converged towards each other. This suggested an equilibrium between bias and variance avoiding both underfitting and overfitting, showing a robust performance over a range of regularization strengths.

The learning curve (Fig. 3d) appeared to have a similar behavior to the model trained with the 19 features. The validation and training accuracy curves converged as the training size increased, demonstrating the ability of our model to avoid overfitting.

IV. CONCLUSION

Using solely PPG signals, possessing physiological and pathological indicators, we trained a binary logistic regression model to detect CVD from the features extracted. Using SBP and DBP, we were able to attain an accuracy of 94% along with a high F1-score of 95%. Without employing SBP and DBP that require a BP cuff, we identified the optimal feature combination and were able to attain an accuracy of 88% along with an F1-score of 90%, while also demonstrating our model’s ability to avoid overfitting. Our model was able to accurately identify CVD relying on data that can be acquired easily from wearable devices. This enables the advancement of self-monitoring capabilities that would increase the early detection and proper management of CVD.

REFERENCES

[1] Z. Habli, F. Kobeissy, and M. L. Khraiche, "Advances in point-of-care platforms for traumatic brain injury: recent developments in diagnostics," *Rev Neurosci,* vol. 33, no. 3, pp. 327-345, Apr 26 2022, doi: 10.1515/revneuro-2021-0103.

[2] Z. Habli, W. AlChamaa, R. Saab, H. Kadara, and M. L. Khraiche, "Circulating Tumor Cell Detection Technologies and Clinical Utility: Challenges and Opportunities," *Cancers (Basel),* vol. 12, no. 7, Jul 17 2020, doi: 10.3390/cancers12071930.

[3] M. Moshawrab, M. Adda, A. Bouzouane, H. Ibrahim, and A. Raad, "Smart Wearables for the Detection of Cardiovascular Diseases: A Systematic Literature Review," *Sensors,* vol. 23, no. 2, p. 828, 2023.

[4] Z. Habli, S. Saleh, H. Zaraket, and M. L. Khraiche, "COVID-19 in-vitro Diagnostics: State-of-the-Art and Challenges for Rapid, Scalable, and High-Accuracy Screening," *Front Bioeng Biotechnol,* vol. 8, p. 605702, 2020, doi: 10.3389/fbioe.2020.605702.

[5] F. C. Khoury, S. A. Saleh, Z. S. Habli, and M. L. Khraiche, "Development of Inkjet-Printed PEDOT: PSS-Based Organic Electrochemical Transistor (OECT) for Biopotential Amplification."

[6] S. Ha, M. L. Khraiche, G. A. Silva, and G. Cauwenberghs, "Direct inductive stimulation for energy-efficient wireless neural interfaces," *Annu Int Conf IEEE Eng Med Biol Soc,* vol. 2012, pp. 883-6, 2012, doi: 10.1109/EMBC.2012.6346073.

[7] E. Grönvall and N. Verdezoto, "Beyond self-monitoring: understanding non-functional aspects of home-based healthcare technology," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, 2013, pp. 587-596.

[8] "Cardiovascular diseases (CVDs)." World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed August, 2023).

[9] H. Alawieh, T. E. Chemaly, S. Alam, and M. Khraiche, "Towards Point-of-Care Heart Failure Diagnostic Platforms: BNP and NT-proBNP Biosensors," *Sensors (Basel),* vol. 19, no. 22, Nov 16 2019, doi: 10.3390/s19225003.

[10] H. Eyre *et al.*, "Preventing cancer, cardiovascular disease, and diabetes: a common agenda for the American Cancer Society, the American Diabetes Association, and the American Heart Association," *Circulation,* vol. 109, no. 25, pp. 3244-3255, 2004.

[11] D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur, and H. Nazeran, "A review on wearable photoplethysmography sensors and their potential future applications in health care," *International journal of biosensors & bioelectronics,* vol. 4, no. 4, p. 195, 2018.

[12] Y. Maeda, M. Sekine, and T. Tamura, "The advantages of wearable green reflected photoplethysmography," *Journal of medical systems,* vol. 35, pp. 829-834, 2011.

[13] N. Nazrul Anuar *et al.*, "Cardiovascular Disease Prediction from Electrocardiogram by Using Machine Learning," *International Journal of Online and Biomedical Engineering (iJOE),* vol. 16, no. 07, pp. pp. 34-48, 06/19 2020, doi: 10.3991/ijoe.v16i07.13569.

[14] K. Gowda, N. Makhija, H. Ochani, and S. Sondur, "HealthyHeart: ML based Analysis and Prediction of Cardiovascular Diseases," in *2021 International Conference on Intelligent Technologies (CONIT)*, 25-27 June 2021 2021, pp. 1-5, doi: 10.1109/CONIT51480.2021.9498535.

[15] J. D. Pollard *et al.*, "Electrocardiogram machine learning for detection of cardiovascular disease in African Americans: the Jackson Heart Study," *European Heart Journal - Digital Health,* vol. 2, no. 1, pp. 137-151, 2021, doi: 10.1093/ehjdh/ztab003.

[16] J. Yu, S. Park, S. H. Kwon, K. H. Cho, and H. Lee, "AI-Based Stroke Disease Prediction System Using ECG and PPG Bio-Signals," *IEEE Access,* vol. 10, pp. 43623-43638, 2022, doi: 10.1109/ACCESS.2022.3169284.

[17] T. Sadad, S. A. C. Bukhari, A. Munir, A. Ghani, A. M. El-Sherbeeny, and H. T. Rauf, "Detection of cardiovascular disease based on PPG signals using machine learning with cloud computing," *Computational Intelligence and Neuroscience,* vol. 2022, 2022.

[18] P. Li and T.-M. Laleg-Kirati, "Central blood pressure estimation from distal PPG measurement using semiclassical signal analysis features," *IEEE Access,* vol. 9, pp. 44963-44973, 2021.

[19] M. Kachuee, M. M. Kiani, H. Mohammadzade, and M. Shabany, "Cuffless blood pressure estimation algorithms for continuous health-care monitoring," *IEEE Transactions on Biomedical Engineering,* vol. 64, no. 4, pp. 859-869, 2016.

[20] Y. Liang, Z. Chen, G. Liu, and M. Elgendi, "A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in China," *Scientific data,* vol. 5, no. 1, pp. 1-7, 2018.

[21] Y. Liang, M. Elgendi, Z. Chen, and R. Ward, "An optimal filter for short photoplethysmogram signals," *Scientific data,* vol. 5, no. 1, pp. 1-12, 2018.

[22] J. Park, H. S. Seok, S.-S. Kim, and H. Shin, "Photoplethysmogram analysis and applications: An integrative review," *Frontiers in Physiology,* vol. 12, p. 808451, 2022.

[23] G. Thambiraj, U. Gandhi, U. Mangalanathan, V. J. M. Jose, and M. Anand, "Investigation on the effect of Womersley number, ECG and PPG features for cuff less blood pressure estimation using machine learning," *Biomedical Signal Processing and Control,* vol. 60, p. 101942, 2020.

[24] S. Lee, G. P. Joshi, C.-H. Son, and G. Lee, "Combining Gaussian Process with Hybrid Optimal Feature Decision in Cuffless Blood Pressure Estimation," *Diagnostics,* vol. 13, no. 4, p. 736, 2023.

[25] D. G. Brillante, A. J. O'sullivan, and L. G. Howes, "Arterial stiffness indices in healthy volunteers using non‐invasive digital photoplethysmography," *Blood pressure,* vol. 17, no. 2, pp. 116-123, 2008.

[26] M. Elgendi, Y. Liang, and R. Ward, "Toward generating more diagnostic features from photoplethysmogram waveforms," *Diseases,* vol. 6, no. 1, p. 20, 2018.

[27] H. Ab Hamid, N. A. Nayan, M. Z. Suboh, and A. Aminuddin, "Second Derivatives of Photoplethysmogram for Hyperuricemia Classification using Artificial Neural Network," in *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, 2021: IEEE, pp. 494-498.