TALLINN UNIVERSITY OF TECHNOLOGY

SCHOOL OF ENGINEERING

LINEAR REGRESSION ML MODEL FOR SYSTOLIC BLOOD PRESSURE PREDICTION

LINEAARSE REGRESSIOONI MÕ MUDEL SÜSTOOLSE VERERÕHU ENNUSTAMISEKS

BACHELOR THESIS

Student: Aleksandra Petra Berger

Student code: 176381

Supervisor: Andrei Krivošei , Ph.D.

Tallinn 2024

AUTHOR'S DECLARATION

Hereby I declare, that I have written this thesis independently.

No academic degree has been applied for based on this material. All works, major viewpoints and data of the other authors used in this thesis have been referenced.

"19" May 2024

Author: 

/signature /

Thesis is in accordance with terms and requirements

"....." ..... 20….

Supervisor: ….....

/signature/

Accepted for defence

"....."..... 20… .

Chairman of theses defence commission: .....

/name and signature/

Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I, Aleksandra Petra Berger

1. grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis Linear Regression ML model for Systolic Blood Pressure prediction,
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.

2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.

3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

19.05.2024

supervised by Andrei Krivošei, Ph.D.

[1] *The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.*

School of Engineering

THESIS TASK

Student: Aleksandra Petra Berger 176381

Study programme: MVEB Integrated Engineering

Supervisor: Andrei Krivošei, Ph.D

Consultants: Olev Märtens, Ph.D., Ph.D, Margus Metshein, Ph.D.

Thesis topic:

(in English): Linear Regression ML model for Systolic Blood Pressure prediction

(in Estonian): Lineaarse Regressiooni MÕ Mudel Süstoolse Vererõhu Ennustamiseks

Thesis main objectives:

1. Studying of existing PPG features and Time Series filtering techniques
2. Selection of the relevant features and creation of Python code prototype
3. Creation of ML algorithm

Thesis tasks and time schedule:

No	Task description	Deadline
1.	Discussion of important features	21.10.2024
2.	Design of the working prototype of the code to extract the features	
3.	Training and improving the model	18.04.2024
4.	Thesis Finalisation	24.03.2024

Language: …… **Deadline for submission of thesis:** "06" May 2024.a

Student: Aleksandra Berger "19" May 2024

Supervisor: Andrei Krivošei, Ph.D. */digital signature/* "....."....20….a

Consultant: Olev Märtens, Ph.D. */digital signature/* "....."....20….a

Head of study programme: Tauno Otto */digital signature/* "....."....20…..a

/signature/

CONTENTS

PREFACE

Prior to enrolling in the Integrated Engineering program, my fascination with signal processing was already well established. The idea of conveying information through waveform patterns struck me as both aesthetically pleasing and sentimentally captivating. Now employed as a Data Analyst, I have chosen to merge my passion for data manipulation with my long standing interest in signal processing. This endeavour not only serves as a platform to showcase my data visualisation skills, which were honed during my Project Course work, but also represents a personal journey of growth and validation.

The topic of data analysis and data visualisation also drew inspiration in the prior Course Project with Kristina Vassiljeva, which faced challenges and was quite subpar, however, received a "Passed" mark. Consequently, my upcoming thesis work not only signifies the culmination of my studies in the Integrated Engineering curriculum but also serves as a testament to the dedication and support of Tauno Otto, Ph.D., whom I am particularly grateful to for his unwavering encouragement and permission to explore topics that pique my interest.

I want to express sincere appreciation to Olev Märtens, Ph.D., and Andrei Krivošei, Ph.D., for their guidance in methodological approaches, consistent communication throughout the research process, and inspiration since the very beginning of work. Their time and resource contribution is invaluable, especially considered remote supervision, which is always challenging.

Additionally, I wish to express gratitude to Margus Metshein, Ph.D., for his early support and willingness to assist with sourcing relevant research articles.

I hope that you find reading this work as rewarding as I found working on it.

Keywords: cuffless blood pressure estimation, PPG, support vector regression, linear regression, machine learning

List of abbreviations and symbols

BP – Blood Pressure
BPM – Blood Pressure Monitor
CT – Crest Time
CVD – cardiovascular disease
DBP – Diastolic Blood Pressure
DT – Diastolic Time
ECG – electrocardiogram
EDA – Exploratory Data Analysis
FIR – Finite Impulse Response
IIR – Infinite Impulse Response
LR – Linear Regression
MAP – Mean Arterial Pressure
MNA – Motion Noise Artefacts
ML – Machine Learning
PAT – Pulse Arrival Time
PPG – Photoplethysmogram
PPI - Peak to Peak Interval
PTT – Pulse Transit Time
PWA – Pulse Wave Analysis
PWV – Pulse Wave Velocity
SBP – Systolic Blood Pressure
SQI – Signal Quality Indices
ST – Systolic Time, Upstroke time
SVR – Support Vector Regression

1. STATE OF THE ART

1.1. Introduction

Monitoring blood pressure is essential for assessing heart health [1], [2] since it is a leading cause of death throughout the world and a critical factor for increasing the risk of serious diseases [3], as it serves as a vital biomarker. Traditional blood pressure measurement devices can be uncomfortable for patients, prompting researchers to explore alternative methods like PPG to estimate blood pressure without using a cuff. In this thesis was investigated the correlation between the individual PPG signal features and systolic blood pressure as well is undertaken the attempt to predict the systolic blood pressure with several Machine Learning models: Linear Regression and Support Vector Regression, as well observed the outcome of the models and undertaken the attempts on bettering the ML models performance.

The choice of utilising Photoplethysmogram (PPG) technology for estimating blood pressure was made due to its widespread use in wearable devices, which is increasingly popular and paves the way for future implementation of the research, as well as its potential to provide reliable and accurate measurements.Recent research has shown that PPG signals alone [4] can be effective in estimating blood pressure, especially when the signal quality is high. Even though using both Electrocardiogram (ECG) signal and PPG signal can help estimate blood pressure effectively [5] collecting ECG and PPG simultaneously using a mobile phone [6] remain challenges.

The dataset [1] was chosen based on i) having a good annotation, ii) the dataset has to be published under Creative Commons Attribution, as it has to be in Compliance with privacy laws and regulations, and iii), the last but not the least point was the fact that this dataset was mentioned in several sources [2], [7], [8] during the literature review course of preparing this work.

The data acquisition was conducted at the Guilin People's Hospital, China. The dataset was recorded with a sampling rate set to 1kHz, with 12-bits AD conversion precision, and the hardware filter design is 0.5–12 Hz bandpass.The PPG sensor model was SEP9AF-2 (SMPLUS Company, Korea), which contains dual LED with 660nm (Red light) and 905 nm (Infrared) wavelengths. Each recorded segment of signal includes 2100 sampling points with correspondence to a length of 2.1 seconds, with a number of subjects is 219, aged 21–86 years, with a median age of 58 years. basic physiological

characteristics of individuals presented in the dataset, along with the information on cardiovascular diseases, were extracted from the hospital electronic medical record of the hospital.

In order to extract the PPG signal features accurately, the signals underwent the filtering stage. Before, comparative analysis of the number of digital filters: Chebyshev II, FIR with Hanning window, FIR with Hamming window was done to choose the most appropriate one. To extract the vital features of a signal 12 functions were written in Python programming language.

1.2. Work Structure and Methods Overview

In Chapter 2.1 I review the existing methods to measure blood pressure without actual contact with the body, along with a brief discussion of advantages and disadvantages of the methods. Later, I test and compare results of the different filtering and normalising techniques, and methods of extracting features along with different combinations of features themselves. Firstly, I start with a short Exploratory Data Analysis (EDA) to learn more about the dataset the work is based on, then I move to detrending and denoising signals. Next step is signal components overview along with defining the crucial points of a signal, on which is based the next step of the work – features calculation.

Signal processing, an important stage before any characteristic extraction, is reviewed in Chapter 2.6, where several filtering and preprocessing techniques are explored.

The main characteristics, further called features, which are extracted and later used as features for the ML model, are viewed in Chapter 3.3. Before, in Chapter 2.4, main features of PPG signal are explored. The workload is summarised in the diagram below:

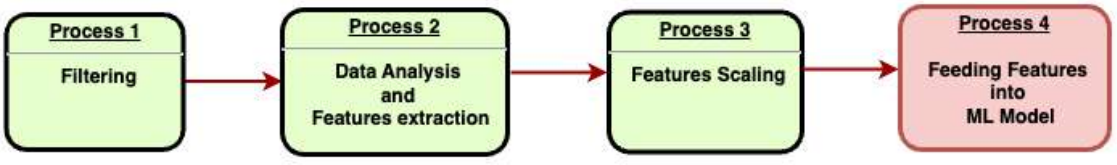


Figure 1. *Principle diagram of Main Pipeline of the work.*

The work is finished with 1.9, where the model is trained and tested, and an attempt to better model performance is made, followed by model result discussion. Limitations, which are restricted model performance, are viewed and discussed in Chapter 4.4.

2. PPG: Origins, Acquisition, and Signal Components

2.1. Existing noninvasive cuffless BP methods

Since the blood pressure can be only measured through either cuff-based or invasive instruments that cannot be continuously used due to their characteristics [9], there are attempts for non invasive and continuous blood pressure measurement based on the photoplethysmographic (PPG) signal alone or integrated with the ECG signal [3], [9], [10]. Several main methods of noninvasive cuffless BP methods are presented below with the corresponding limitations.

- BP estimation using sphygmomanometer

Clinical practice for noninvasive BP measurements, but has limitations, such as requiring an inflatable arm cuff which can be inconvenient and obtrusive, and only providing intermittent measurements [11].

- BP estimation by calculating speed of wave propagation

BP can be estimated from parameters obtained from the two signals, such as pulse transit time (PTT) – the time delay for the pressure waveform to travel between two arterial sites – and pulse arrival time (PAT) [12], [13], which is which is the time difference between the R-peak of the ECG signal and the peak of the PPG signal when measured within the same cardiac cycle [14]. Studies have reported an inverse correlation between BP and pulse transit time (PTT) [15], [16], [17], and PTT based approach has mounting evidence that it can provide cuffless non-invasive BP measurements [3]. However, the approach has several challenges such as requiring two sensors, synchronous signals, and difficulties in acquiring the ECG unobtrusively [11].

- BP estimation from PPG pulse wave shape by extracting time-domain features

Firstly, signals are pre-processed, then features are extracted and later different ML and DL techniques are used for BP prediction. There are several general limitations related to

the way signals were recorded, level of noise, and so on, which will be discussed in chapter 6.2 of this work.

- BP estimation using machine learning and deep learning through a combination of signals such as ECG, PPG

Method based on using PPG solely was chosen from others for its good results shown in different works. As such, El-Hajj and Kyriacou in their "A review of Machine Learning techniques in photoplethysmography for non-invasive cuff-less measurements of blood pressure" [3] provide an overview of the summary of the different ML models, which testifies that the attempts of using waveform features extracted solely from PPG signals has been here for more than a decade and have shown reliable results.

2.2. PPG Physiological Origins and Structure

PPG is a non-invasive method for measuring blood volume changes in a microvascular bed of the skin based on optical properties, such as absorption, scattering, and transmission properties of human body composition under a specific light wavelength. [18].

Photoplethysmography is usually performed by a tissue contact sensing device. A conventional PPG sensing device consists of a light emitter and a highly sensitive photodetector mounted inside a reusable spring-loaded clip. The most commonly used light sources are light-emitting diodes (LEDs).

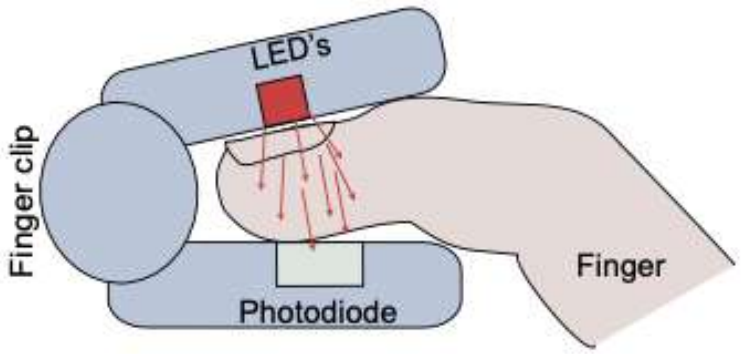


Figure 2. *Device for capturing PPG structure.*

In transmission probes, the LED and photodiode are mounted on opposite sides of a clip, and the light transmitted by the LED passes through the tissue sample to the photodiode

(see Fig. 2). Transmission probes are the most widely used probes in healthcare settings, used to measure arterial oxygen saturation [19]. Their use is limited to peripheral sites to which a clip can be attached, such as the finger, earlobe, or toe.

The volume of blood in the measurement site, arterial diameter, haemoglobin concentration, and haemoglobin direction according to the cardiac cycle are also major factors that affect the detected light intensity [20]. For example, during the diastolic phase, blood volume, arterial diameter, and haemoglobin concentration in the measurement site are minimised. Thus, absorbance is minimised, while the amount of light detected by the photodetector is maximised. Conversely, in the systolic phase, the light intensity detected by the photodetector becomes minimum [21], [22].

Major blood vessels and arteries with strong pulsation are mainly located in the skin dermis or subcutaneous tissue. Thus, light with a red wavelength of 640–660 nm and infrared wavelength of 880–940 nm is mainly used for PPG measurement . PPG is mainly obtained at the extremities of the human body, such as fingers, toes, and earlobes that are advantageous for measuring changes in blood volume, because the vascular bed is shallow and widely spread [23], [24], [25]. PPG can also be obtained from the forehead, oesophagus, and nose [26], [27], [28].

2.3. PPG Signal Acquisition

Signal acquisition in PPG involves placing a sensor on the skin, which emits light into the tissue and detects the amount of light that is absorbed or reflected by the blood vessels. The sensor converts this information into an electrical signal that represents the changes in blood volume over time. At first, the light particles from the LED interact with the finger skin, where the sensor is placed (see Figure 2), arteries in the finger let the blood, ejected from the heart, through, which results in the diameter of the artery changing. The diameter of the artery dynamically adjusts to the volume of the blood being let through. As such, in the phase of the cardiac activity called diastole, diameter shrinks to the minimum diameter, whereas at a systolic phase a big pressure wave of the blood ejected from the heart propagates through the artery, the thickness of the artery diameter changes to maximum. The change of diameter follows the AC curve. There is a second component, photodetector, which transmits an oscillating curve to an actual PPG signal: it receives the light emitted from LED after their interaction with the finger tissues, including the oscillating diameter, and reflects it back, converting the optical

particles into electrical signal. However, PPG signals can be affected by various sources of noise, including motion artefacts, ambient light interference, sensor misplacement, and physiological factors such as skin pigmentation and perfusion.

Sensors and electronics play a significant role in obtaining the signal and in ensuring the signal has adequate quality. The optoelectronics of the PPG sensor and their tissue attachment mechanism are fundamental building blocks in a PPG measurement system. PPG sensors are designed in two different distinct forms: transmission mode and reflectance mode. In transmission mode, the light source and detector are separated by the tissue, whereas in reflectance mode, the photodetector is positioned along the light source on the same side of the tissue to measure the reflected light [29]. Their design will impact on signal quality, subsequent pulse wave analysis techniques, and the need for skilful denoising.

The quality of a PPG signal is determined by both the instrumentation and the sensor setup. Noise in PPG signals can lead to inaccuracies in measurements and affect the reliability of the data. It can obscure important features of the signal, making it difficult to extract meaningful information. The consequences of noisy signals include incorrect heart rate readings, inaccurate blood pressure estimates, and unreliable assessment of oxygen saturation levels.

The important point to consider when acquiring pulse wave fronts for analysis is a sampling frequency – the number of samples per second (or per other unit) taken from a continuous signal to make a discrete or digital signal [30] – while recording. For instance, J. Allen, when analysing The morphological features of individual PPG pulse waves, required a relatively high sampling rate (e.g., > 500 Hz) [31]. The frequency in the investigated dataset is set to 1kHz.

2.4. Main PPG Signal Components

Each PPG pulse wave consists of two distinct phases: the anacrotic and catacrotic phases, corresponding to the rising and falling limbs respectively [31], whose projection on the time axis corresponds to systolic and diastolic phase respectively. The morphology of the PPG pulse wave is influenced by:

- the heart (characteristics of cardiac ejection including heart rate, heart rhythm, and stroke volume);

- the circulation (including cardiovascular properties such as arterial stiffness and blood pressure);
- additional physiological processes including respiration and the autonomic nervous system (which can be affected by stress);
- disease [31], [32]

The main components of a PPG waveform include the systolic peak, the diastolic peak, and the dicrotic notch. The systolic peak describes the maximum arterial blood volume during the systolic phase when the heart contracts and ejects blood into the arteries. When the heart relaxes and refills with blood, what happens in the diastolic phase, the arterial blood volume is minimal. This state corresponds to the diastolic peak. The dicrotic notch is a small bump in the waveform that occurs when the aortic valve closes, causing a brief interruption in blood flow. Cardiovascular health and function is described with the morphology of a PPG signal.

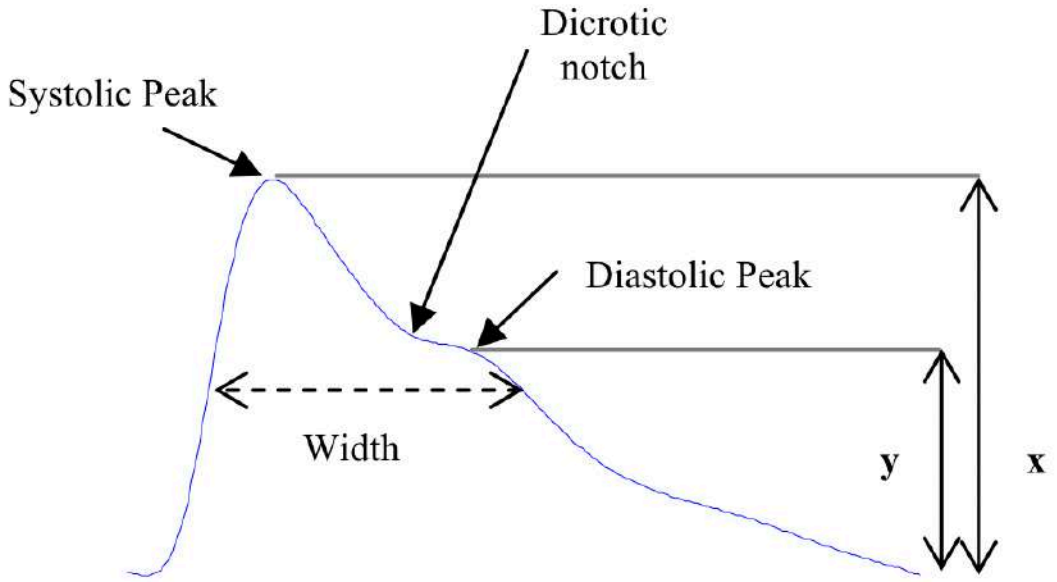


Figure 3. A typical waveform of the PPG and its characteristic parameters, whereas the amplitude of the systolic peaks is x while y is the amplitude of the diastolic peak [33].

1) SNR

A distinctive attribute extracted from an individual signal prior to undergoing signal processing is the Signal-To-Noise Ratio (SNR). SNR quantifies the ratio of signal power to

noise power within the PPG waveform. Elevated SNR values signify a clearer and more dependable signal quality, while diminished SNR levels can impede the extraction of significant information from the signal.

2) Systolic Amplitude

As shown Fig. 3 the systolic amplitude (x) is an indicator of the pulsatile changes in blood volume caused by arterial blood flow around the measurement site [34], [35]. Systolic amplitude has been related to stroke volume [36]. Dorlas and Nijboer found that systolic amplitude is directly proportional to local vascular distensibility over a remarkably wide range of cardiac output [37]. It has also been suggested that systolic amplitude is potentially a more suitable measure than pulse arrival time for estimating continuous blood pressure [38].

3) Area under curve

The pulse area is measured as the total area under the PPG curve. Seitsonen et al. [39] found the PPG area response to skin incision to differ between movers and non-movers.

4) Peak to Peak Interval and Pulse Interval

The distance between two consecutive systolic peaks will be referred to as Peak-Peak interval, as shown below:

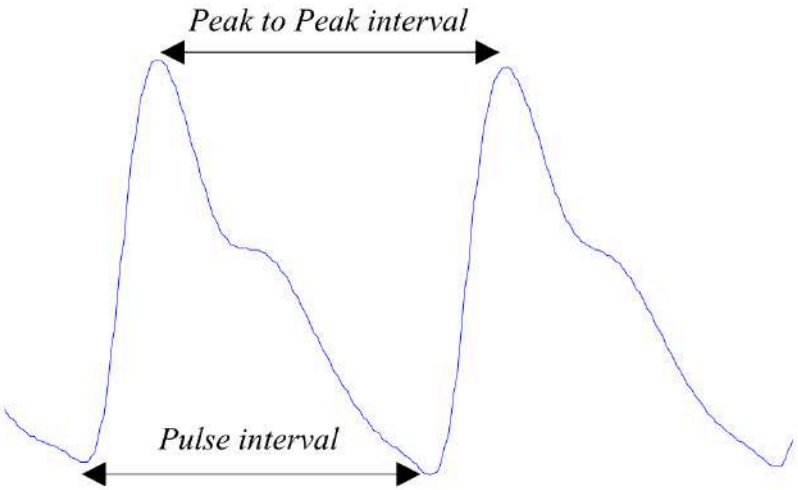


Figure 4. Two consecutive cycles of a PPG signal [33].

The Peak-to-peak-interval (PPI) represents cardiac beat-to-beat interval of PPG signal. The distance between the beginning and the end of the PPG waveform, as shown in Fig. 4. The Pulse interval is usually used instead of the Pulse interval when the diastolic peaks are more clear and easier to detect compared to the systolic peak. Poon et al. [40] suggested that the ratio of Pulse interval to its systolic amplitude could provide an understanding of the properties of a person's cardiovascular system. In 2008, Lu et al. [41] compared the heart rate variability (HRV) using the Pulse interval in PPG signals with the HRV using Peak-to-peak intervals in ECG signals. Their results demonstrated that HRV in PPG and ECG signals are highly correlated. They strongly suggested that PPG signals could be used as an alternative measurement of HRV.

Certain physiological characteristics of a signal may present challenges in accurately identifying specific fiducial points, such as the dicrotic notch or diastolic peak, within the pulse wave. These unique features of the signal can impact the visibility and reliability of these fiducial points, requiring careful consideration and specialised techniques to enhance their detection and analysis. In this case the first and second derivatives might serve as a great help. Points and features , requiring derivation, are listed below:

5) Diastolic point and Dicrotic Notch

Diastolic point definition relies on finding the first derivative: as such, Millasseau et al [42] defined the diastolic point as the point at which the first derivative of the waveform is closest to zero.

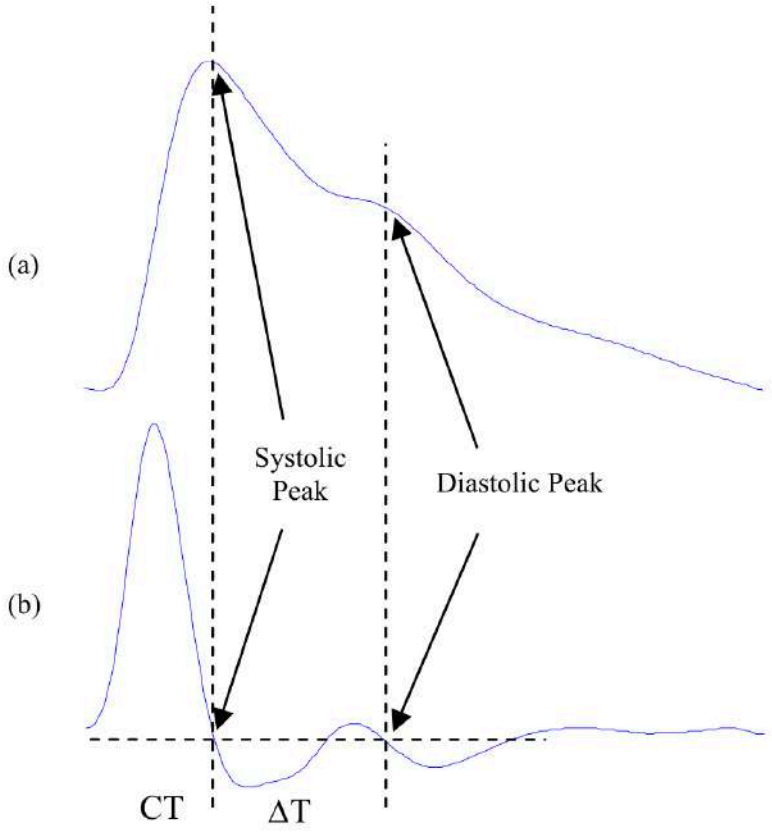


Figure 5. Signal measurements: a) Original fingertip PPG and b) first derivative of the PPG [33].

The second derivative is more commonly used than the first derivative. In literature, the second derivative of photoplethysmogram is also called the acceleration plethysmogram because it is an indicator of the acceleration of the blood in the finger [33].

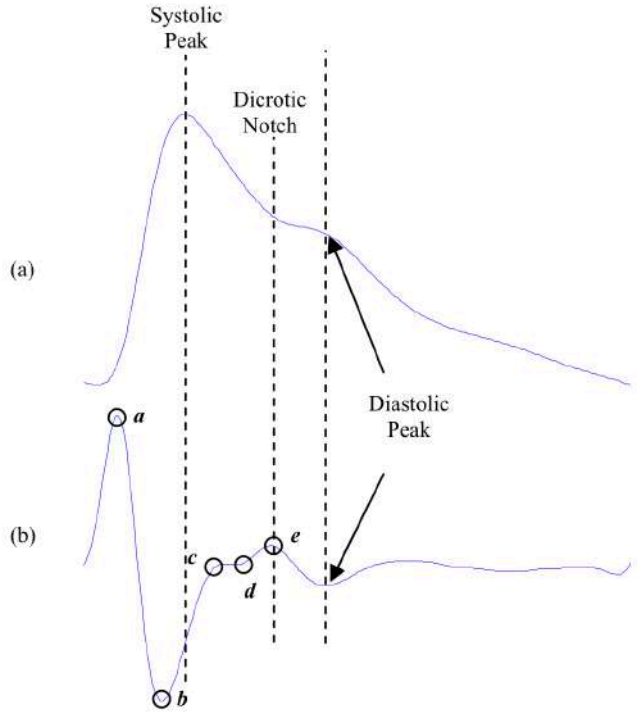
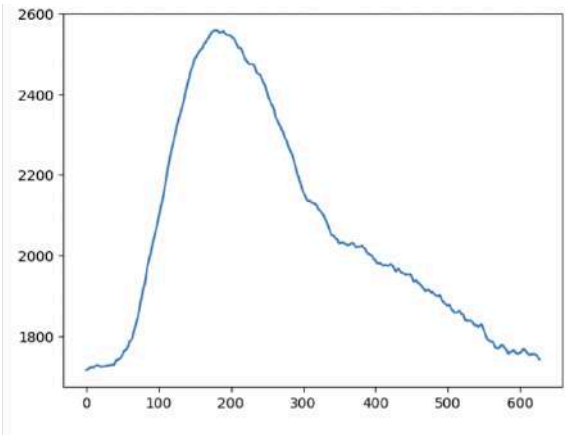
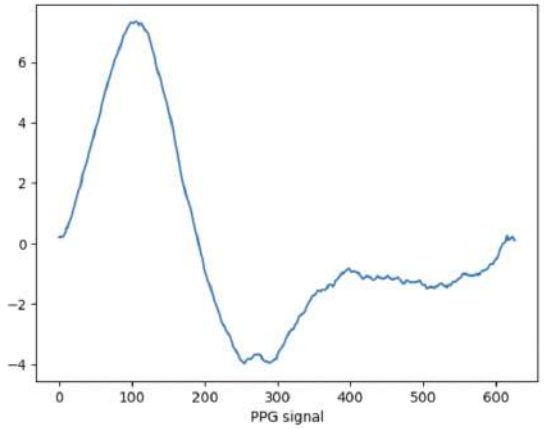


Figure 6. Signal measurements: a) Original fingertip PPG and b)second derivative of the PPG [33].

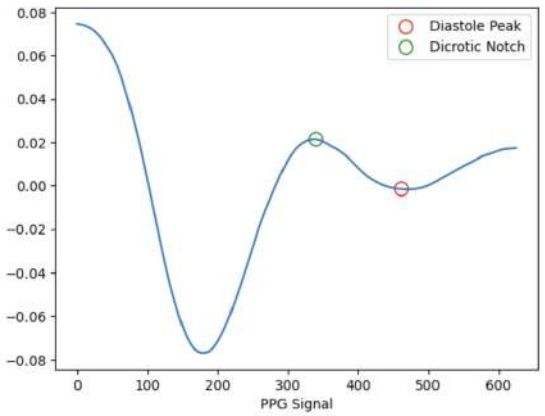
As described in chapter 2.4, The dicrotic notch is a small bump in the waveform that occurs when the aortic valve closes, causing a brief interruption in blood flow. In the diagram below I used a smoothed second derivative to identify dicrotic notch and diastolic peak. Dicrotic notch and diastolic peak are proved to provide valuable information about cardiovascular activity, including vascular compliance, peripheral resistance, and cardiac function. To define a dicrotic notch, the initial cycle was isolated, then its first derivative was calculated, and subsequently its second derivative. A moving average smoothing function with a larger window was then applied between the first and second derivatives stages to mitigate the amplification of small noises present in the derivative results:



a) cycle of PPG signal



b) first derivative of PPG signal



c) second derivative of PPG signal

Figure 7. Fiducial points: Diastolic Peak and Dicrotic notch locations found with the help of second derivative: a) cycle of PPG signal, b) first derivative of PPG signal, c) second derivative of PPG signal.

3. PPG Signals Processing And Features Determination

3.1. Dataset Overview

The basis for the research served the dataset for the non-invasive detection of cardiovascular disease (CVD), containing 657 data segments from 219 subjects [1], each including 3 samples of signals, each, in turn, containing from 2 to 3 cardiac cycles. The dataset covers an age range of 20–89 years, from which males account for 48% of participants, and records of diseases including hypertension and diabetes. The dataset is annotated, each signal has corresponding Systolic and Diastolic Blood Pressure values.

Cardiovascular Dataset Information File									Hospital Electronic Medical Record			
subject_ID	Sex(M/F)	Age (year)	Height (cm)	Weight (kg)	Systolic Blood Pressure(mmHg)	Diastolic Blood Pressure(mmHg)	Heart Rate(b/m)	BMI(kg/m^2)	Hypertension	Diabetes	cerebral infarction	cerebrovascular disease
84	Female	76	150	50	106	53	69	22.22	Normal		cerebral infarction	
85	Female	58	164	53	158	89	73	19.71	Stage 1 hypertension		cerebral infarction	
86	Female	71	159	80	170	87	74	31.64	Stage 2 hypertension		cerebral infarction	
87	Female	78	160	55	164	73	85	21.48	Stage 2 hypertension		cerebral infarction	
88	Female	77	153	60	120	69	76	25.63	Prehypertension		cerebral infarction	
89	Female	74	152	56	100	55	63	24.24	Normal			insufficiency of cerebral blood supply
90	Male	63	166	90	128	77	85	32.66	Prehypertension			insufficiency of cerebral blood supply
91	Female	53	160	64	116	58	71	25.00	Normal			cerebrovascular disease
92	Female	53	160	64	128	76	73	25.00	Prehypertension			
93	Male	54	169	64	161	95	78	22.41	Stage 2 hypertension			cerebrovascular disease
95	Female	64	163	72	153	71	59	27.10	Stage 1 hypertension			insufficiency of cerebral blood supply
96	Female	76	158	62	136	71	72	24.84	Prehypertension			
97	Male	76	168	75	143	65	66	26.57	Stage 1 hypertension		cerebral infarction	

Figure 8. Snippet of the dataset.

All the physical parameters close to Gaussian normal distribution, including both of the BP values:

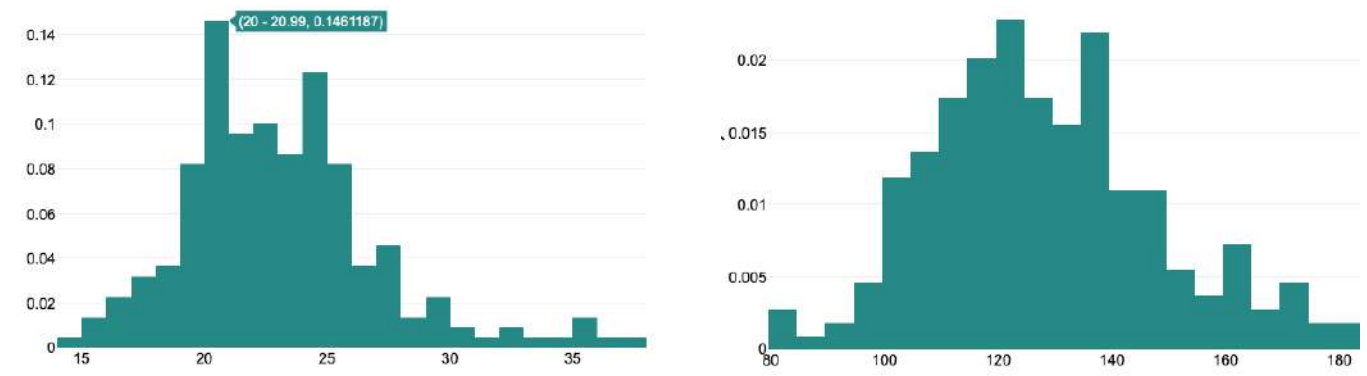


Figure 9: *BMI and Systolic BP distribution histograms.*

Information of heart diseases and their stages follow eye-catching pattern in both cases, be them distributed either by age or weight:

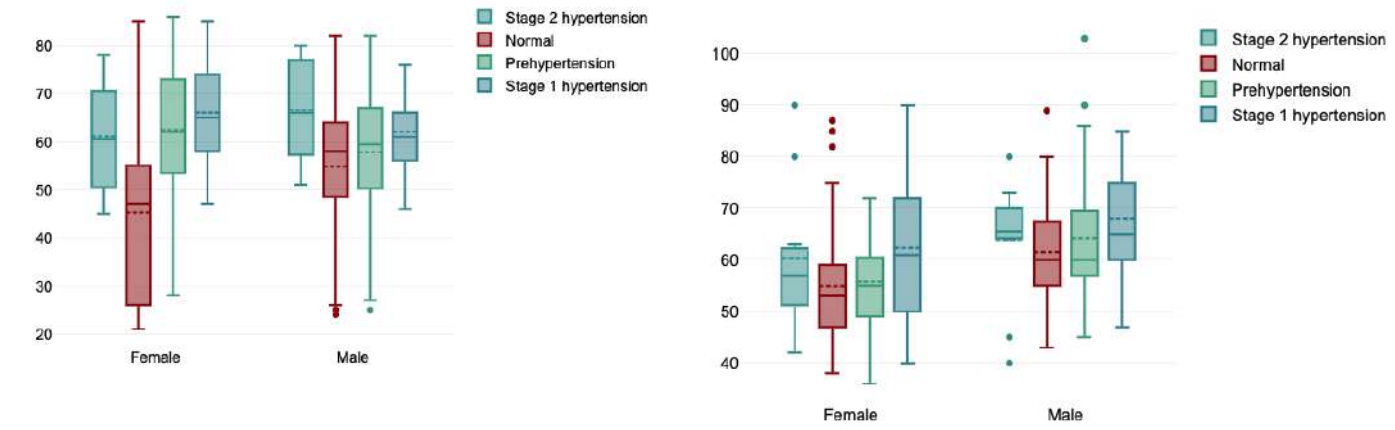


Figure 10: Hypertension by Age and Weight distribution boxplots.

As for BP, both of them, SBP and DBP, follow almost the same trend for Hypertension:

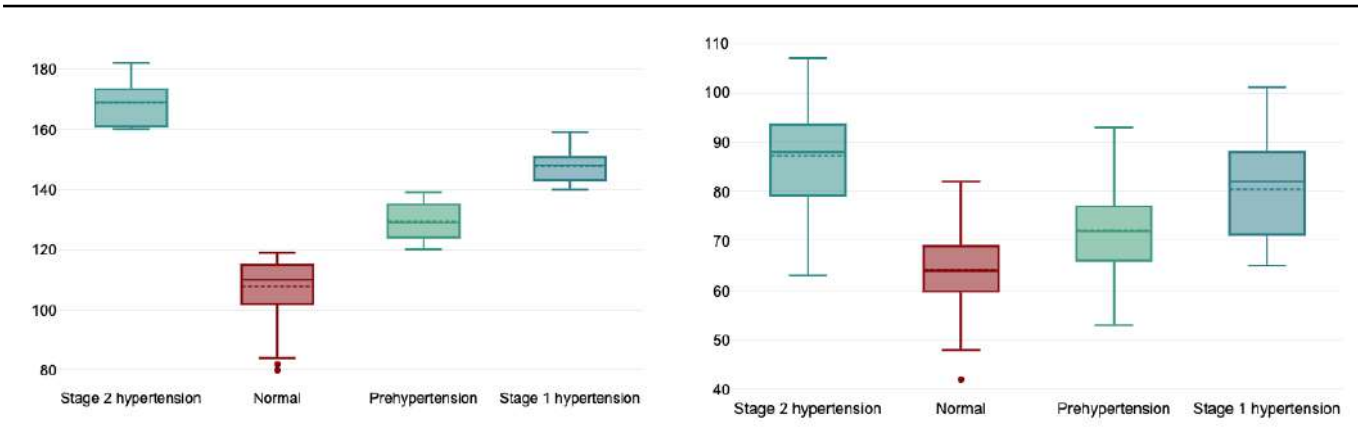


Figure 11. Systolic and DIastolic BP by hypertension distribution boxplots

The flow diagram pipeline already was summarised in the introduction, here I further compartmentalise it. The work pipeline comprised four main parts: preprocessing and

analysing the data of the signals, extracting the features, teaching ML models and post-processing the data (i.e. tuning the ML model):

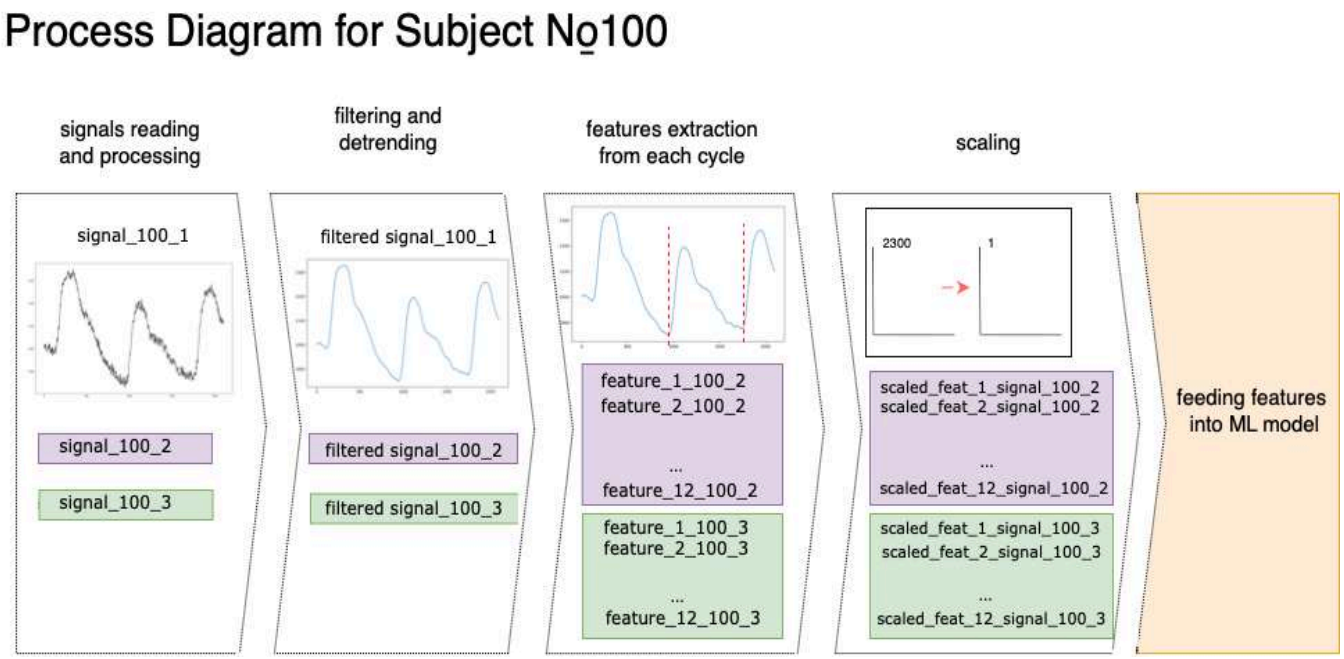


Figure 12. Diagram of the working process based on an example of a signal of a patient n. 100.

Each file, containing a signal, has a name, corresponding to the person's id number. Dataset with given physical characteristics (Fig. 8) contains physical characteristics of each person along with their id number. The first step was to annotate each person's data provided in the dataset (Fig. 8) with the signals belonging to the same persons by pairing their id numbers. Later, the dataset was augmented with the features of each patient, derived from their PPG signals. Subsequently, various filtering and normalisation methods are further explored.

Detrending is utilised to standardise the signal along the Y-axis by eliminating the mean value (0th order polynomial), slope (1st order polynomial), and baseline drift (polynomial of order greater than 1, such as 2, 3, etc.) (stage 2 on Fig. 12). Examination of the dataset during these stages revealed significant variations in feature scales, prompting the use of scaling techniques prior to inputting the features into a Machine Learning algorithm.

3.2. Signal Preprocessing and Filtering Techniques

Big role in assessing signals plays a signal preprocessing stage. It is responsible for cleaning a waveform from oscillations which do not contain essential information, and rather distort accurate assessment of the signal morphology. There is a great number of studies, books and manuscripts devoted to signal processing, as well as it being a separate study field. Different filters employ various techniques, and in the work several filters were explored such as FIR filters with Hanning and Hamming windows, Butterworth, Chebyshev II, wavelet. In the end Savitzky-Golay and Moving Average filters were chosen due to their versatility. Following filters along with detrending method are overviewed:

a) Detrend

Detrend approach for Time Series analysis could be defined as :

$$\hat{x}_t = x_t - T_t \text{, where} \tag{3.1}$$

$\hat{x}_t$ - detrended time series

x_t – time series,

T_t - trend [43]

Signals in the work were detrended as follows: detrended_signal = y - baseline, where baseline was found as a polynomial of degree 1 (a line) fitted into data.

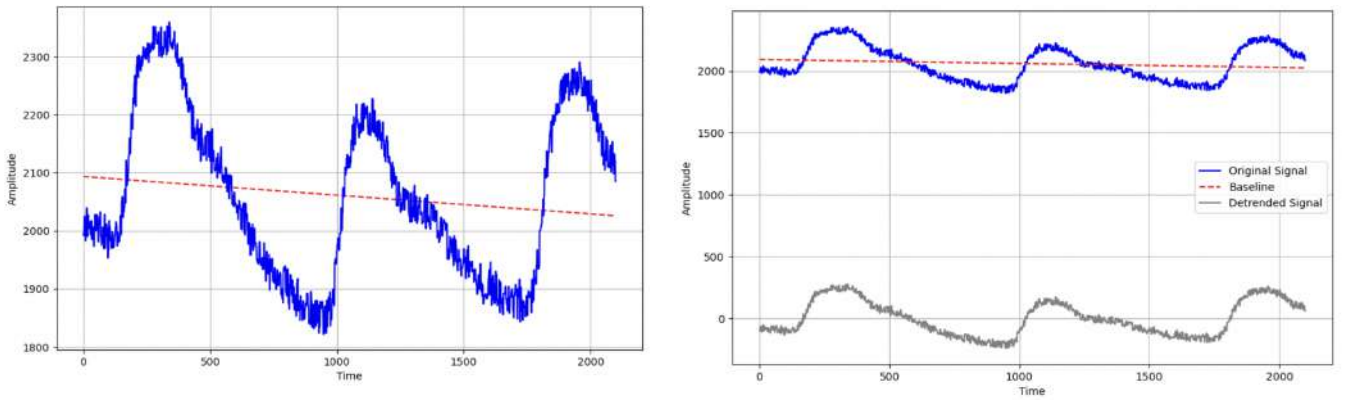


Figure 13. Example of detrended signal.

b) Savitzky-Golay filter

The Savitzky-Golay filter is a least-squares polynomial algorithm, i.e it fits a polynomial to a small window of data points and uses this polynomial to estimate the smoothed value at the centre point of the window. The algorithm involves convolving the data with a set of coefficients derived from a least-squares fit of a polynomial within the window. Moving polynomial fit can be numerically handled in exactly the same way as a weighted moving average, since the coefficients of the smoothing procedure are constant for all y values [44].

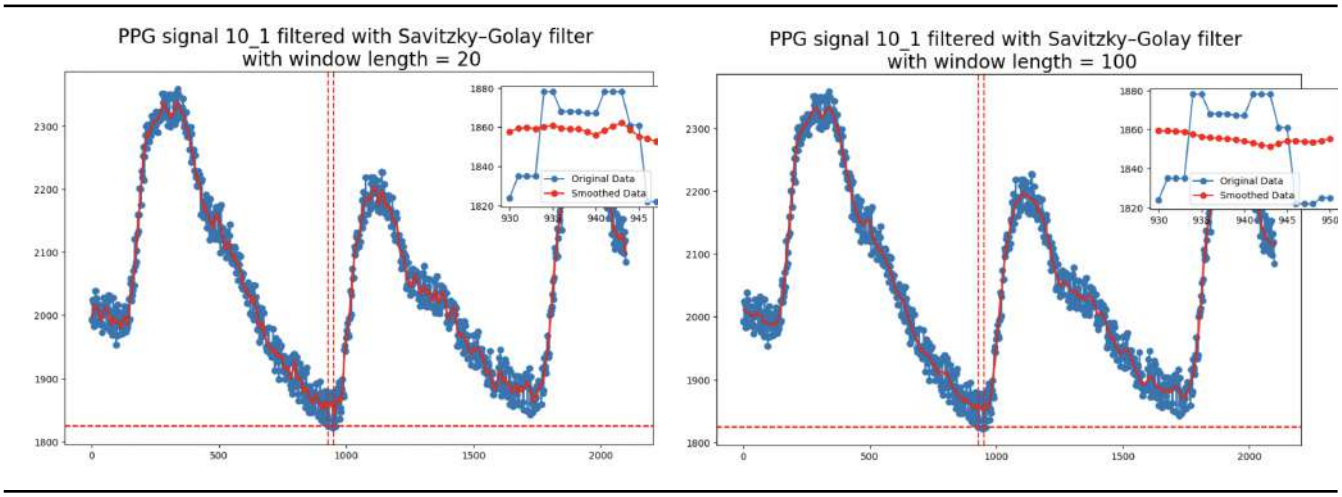


Figure 14. Example of PPG signal filtered with Savitzky-Golay filter with a polyorder=2 and window = 20 and window = 100 correspondingly.

c) Moving Average

The computationally efficient moving average filter averages a subset of data points values within a window centred around the current point. This average value replaces the original data point, resulting in a smoothed signal.

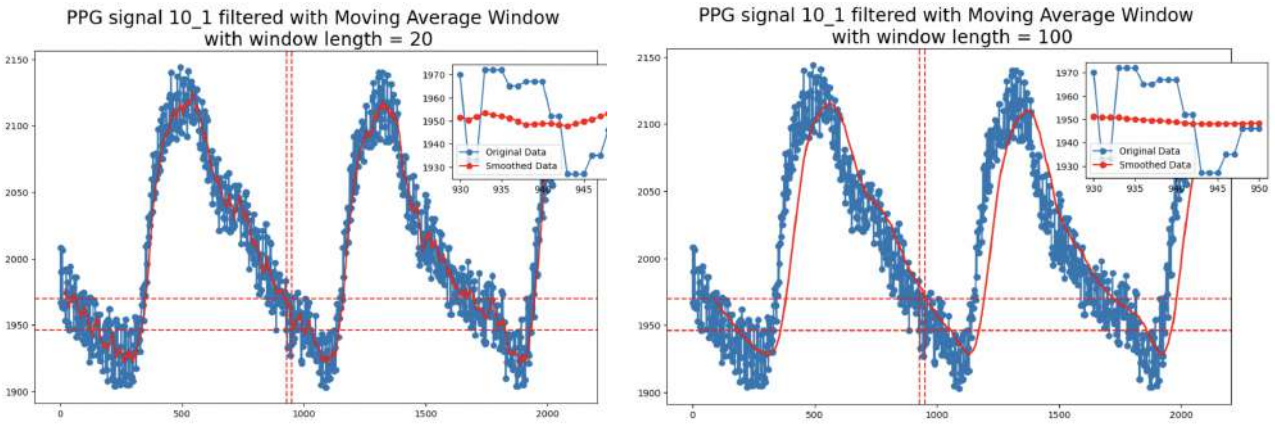


Figure 15. Example of PPG signal filtered with Moving Average Window filter with a window = 20 and window = 100 correspondingly.

4. Features Calculation

Methodology	Algorithm
Detection of a Peak Coordinates	Description: According to the signal processing algorithm represented in Fig. 12, there are several stages of preprocessing algorithms, including filtering and smoothing techniques, taken prior to any feature detection. After the preprocessing is done, peak coordinate is found with the help of find_peaks module of scipy.signal Python library. Since there are other peaks, for instance, the peak of the diastolic phase, the first peak is taken. Reason why the feature is considered a significant aspect of cardiovascular function: peak coordinate provides information about the strength of the heart's contractions and the efficiency of blood flow through the arteries. Changes in the peak amplitude or shape of the

	PPG signal can indicate abnormalities in blood pressure regulation, heart function, or arterial stiffness. Is a detector of diseases such as hypertension, atherosclerosis, and heart failure. Why the feature is important The peak in the PPG signal is directly related to the systolic blood pressure because it reflects the force with which blood is being ejected from the heart and the pressure exerted on the arterial walls during each heartbeat.
Systolic Peak Amplitude	Description: Peak was calculated using the find_peaks module of scipy.signal Python library, but pulse width was used instead, as Awad *et al.* suggested [45]. Why the feature is important: Low systolic peak amplitude signifies decreased blood volume pulsations, high Increased blood volume pulsations [33].
Identification of the cycle width at 25%, 50%, 75% of the cycle height	Description: After the preprocessing and filtering stage the part of the wavefront which is higher than a given level is identified. The latter allows us to find pertinent time domain coordinates at a given level of the amplitude. Why the feature is important: It can provide information about the vascular tone and the elasticity of blood vessels.
Area Under Curve	Description: Taken from a signal after it has gone through preprocessing techniques, the window for moving average is enlarged to 100. AUC for the first cycle is calculated with the formula: auc = simpson(first_cycle, dx = 5), where simpson is an integrating function along the given axis module imported from scipy.signal library. Why the feature is important: The area under the curve represents the total volume of blood that is pulsating through the vessels during this cycle. A larger area under the PPG curve may be a contributing factor towards the increasing of arterial stiffness [46], [47].

SNR	Description: SNR was calculated as SNR (dB) = 10 * log10(P_signal / P_noise), where P_signal is the power of the PPG signal, P_noise is the power of background noise Why the feature is important: Feature helps to understand if a signal passes the quality threshold.
Perfusion Index	Description: PI was calculated as a ratio of ac and dc components, where ac_component = np.max(signal) - np.min(signal) dc_component = np.mean(moving_average_window(signal)) Why the feature is important: represents the ratio of pulsatile on non-pulsatile light absorbance or reflectance of the PPG signal. PI determinants are complex and interlinked, involving and reflecting the interaction between peripheral and central haemodynamic characteristics, such as vascular tone and stroke volume [48].

Table 1. PPG signal main features calculation.

5. ML Model description

For Linear Regression model LinearRegression library was imported from sklearn.linear_model. All the parameters excluding the SBP value were used as training data, whereas SBP was used as a predicted value.

The 219 patients' data was divided into train and test subsets with the test_size=0.3 and random_state=5, i.e. 70% of the initial data was assigned to train data and 30% of the initial data was assigned to the test data. This way, the train subset contains 153 values, along with the test subset containing 66 values. The split was done by train_test_split library imported from sklearn.model_selection.

After dividing the train subset X_train, y_train was fit to the model with the .fit() function from LinearRegression, and test data X_test, y_test was reserved for testing. Then two

values of y were obtained: y_train_predict and y_test_predict, to compare how the model behaves on the training data, which it was trained on, and on the test data, which the model has not "seen".

```
train, predict = split_predict(df_selected)
X_train, X_test, y_train, y_test = train_test_split(train, predict, test_size=0.3, random_state=5)
```

Figure 16. Example of LR model fitting with test_size = 0.3, random_state = 5.

```
def y_train_predict (X_tr, y_tr, X_te, y_te):

    lm = LinearRegression()
    lm.fit(X_tr, y_tr)
    y_train_predict = lm.predict(X_tr)
    y_test_predict = lm.predict(X_te)

    return y_train_predict, y_test_predict
```

Figure 17. Example of function which creates an object of LR model by calling LinearRegression module from the imported sklearn.linear_model, then fits the training subset of selected columns for X along with corresponding values of SBP. Next, the function predicts results are obtained for the train subset and for the test subset.

6. Results

6.1. Metric

As for the metric to assess the model performance, the coefficient of determination, denoted as R2 [49] , was selected. as the primary metric for the model evaluation due to its simplicity and well representation of the model fit for such a dataset. R2 measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s), i.e. variance of Systolic Blood Pressure from features values. The R2 metric library was imported from sklearn.metrics as r2_score.

Comparison of the results shown in the chapters **No Scaler**, **Standard Scaler** and **MinMax Scaler** (see ML_BP_Prediction file in Appendices) shows that neither the Standard Scaler nor the MinMaxScaler did not show any significant improvement of the R2 score on the test data, and I remain with R2 = **0.9196**, which indicates that approximately **91.96%** of the variance in the dependent variable can be explained by the independent variables in the given LR model.

However, on the actual train dataset R2 has improved to **0.93**:

R2 train for systolic BP	R2 test for systolic BP
0.919619384996917	0.9300379226236517

Table 2. Coefficient of determination value for the training data subset and for the test ("unseen") data subset.

These results indicate that approximately **91.96%** and **93.00%** correspondingly of the variance in the dependent variable can be explained by the independent variables in the given LR model.

Below the comparison of the the values given (actual train value in the working dataset), and the values predicted is demonstrated for both delimited parts, train and test subsets:

Actual train value	Predicted train value
137	136.0652901
110	108.0541972
161	171.8487959
115	114.2235993
131	124.8318383
126	125.330447
120	119.8555417
120	127.3502206
137	128.1257165
126	131.462402
128	134.522672
123	117.4409788

Table 3. Comparison of actual SBP values and SBP values predicted by ML model, for train subset.

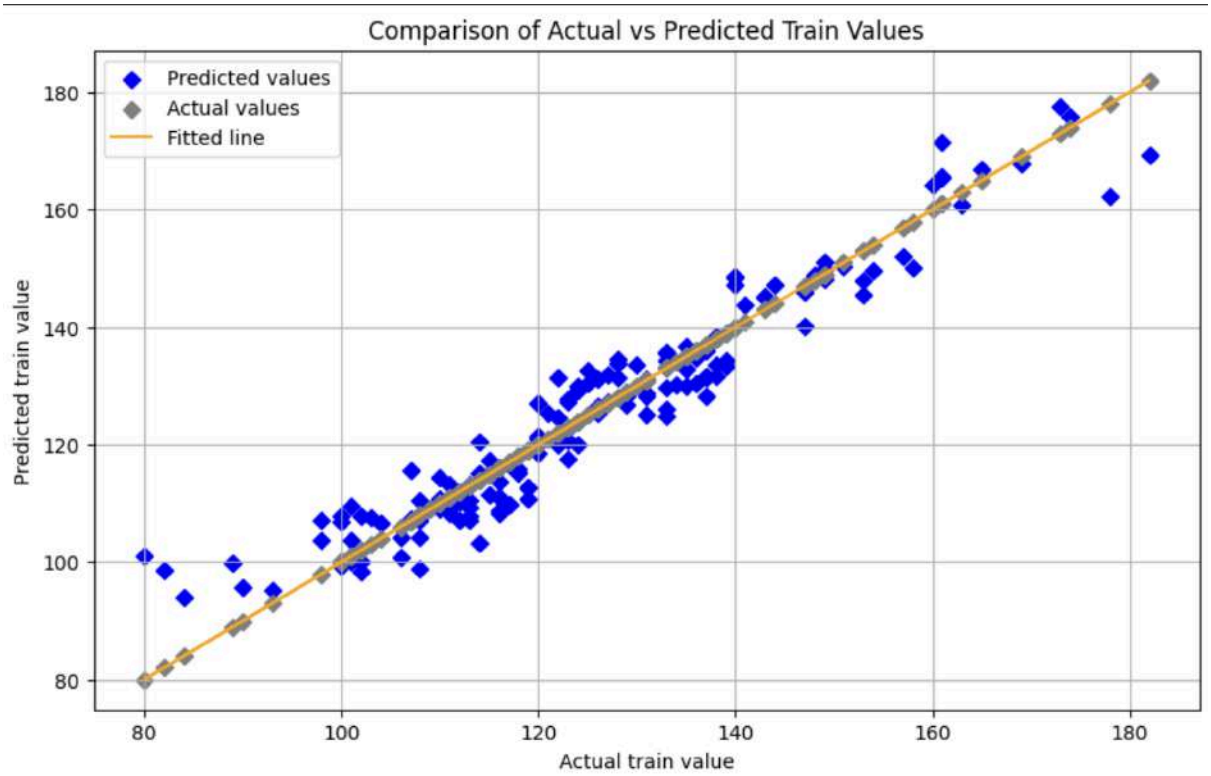


Figure 18. Comparison of Actual and Predicted values, for the train dataset.

Actual test value	Predicted test value
115	109.9703995
170	171.4327976
109	100.6130494
106	109.3346939
139	136.2794009
96	100.4342055
147	147.8300893
150	149.6146633
126	125.6441725
100	106.6982589
138	132.7249691
110	105.6246273

Table 4. Comparison of actual SBP and SBP values predicted by ML model, for test subset.

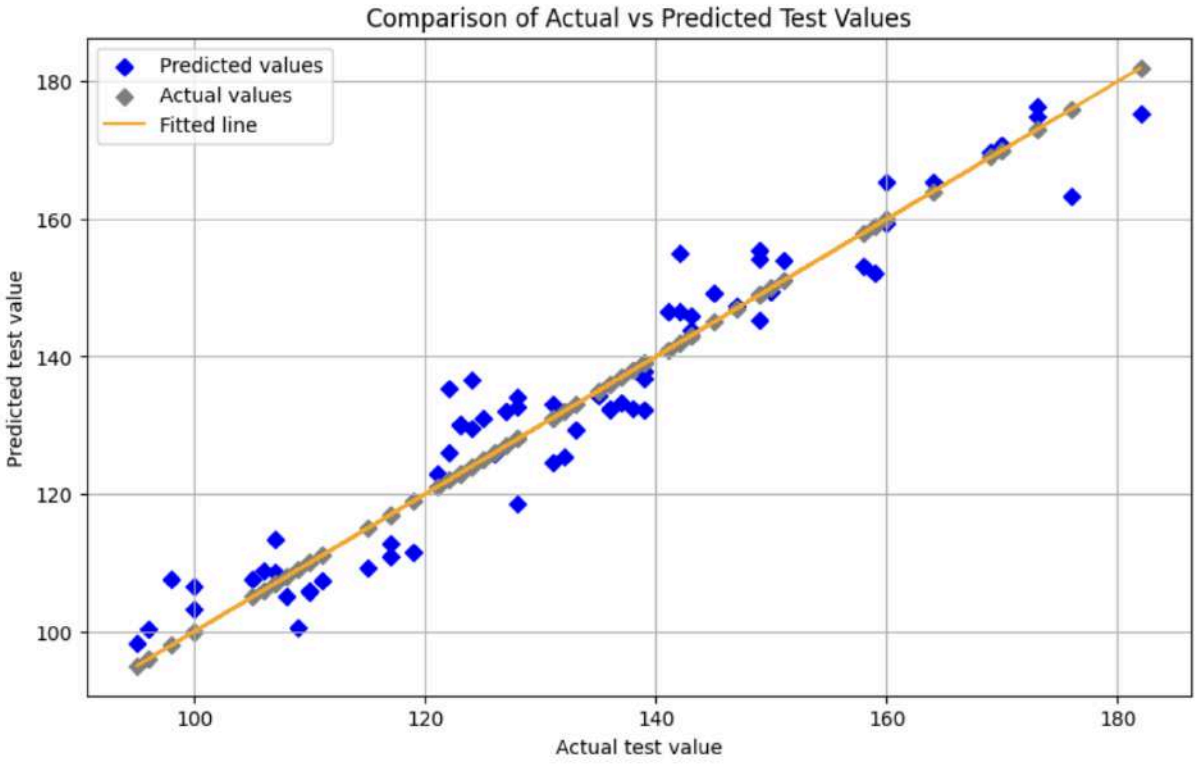


Figure 19. Comparison of Actual and Predicted values, for the test dataset.

6.2. Features Engineering

As a part of EDA heatmap was displayed to show correlation of the dataset attributes, which now include both given physical attributes of individuals along with the calculated features. It is interesting to see how some individual physical attributes have higher correlation index, whereas other physical attributes have less impact on Systolic BP.

As such, sex and height from the individual physical attributes have the minimal impact: *0.09* and *0* correspondingly, BMI has more of an impact (*0.23*); heart rate (*0.14*), hypertension (all states: normal (*0.75*), first stage (*0.42*), second stage(*0.64*)), diabetes (*0.18*), cerebral infarction (*0.13*) and cerebrovascular disease (*0.14*) have higher impact than prehypertension (*0.06*) and diabetes type 2 (*0.04*).

For calculated features the situation is as follows: peak coordinates for the first peak *(0.11)* and for the third peak *(0.06)* are less correlated with SPB than the second peak *(0.15)*. This is due to the fact that the second cycle, being a well-mediated cycle, always was completely captured. Therefore, correct determination of the second cycle peak

never causes problems, while the first and third cycles may not be completely recorded, hence the not quite correct determination of their peaks, and hence peak 1 and peak 3 are not so related to SBP as peak number 2.

Width of a cycle at different amplitudes, at 25% (*0.08*), at 50% (*0.18*) and at 75% (*0.22*) have different correlation index with the SBP: the close to the peak (100%) the width at a given amplitude is, the more the width at the given amplitude is related to the SBP. The start of the diastole phase has surprisingly low correlation index: *0.03*, calculated average pulse rate has index of *0.05*, area under cycle has index of *0.01*, which is explainable, given low correlation between the width at the 25% of the cycle amplitude: the width closer to the baseline defines the area under curve value.

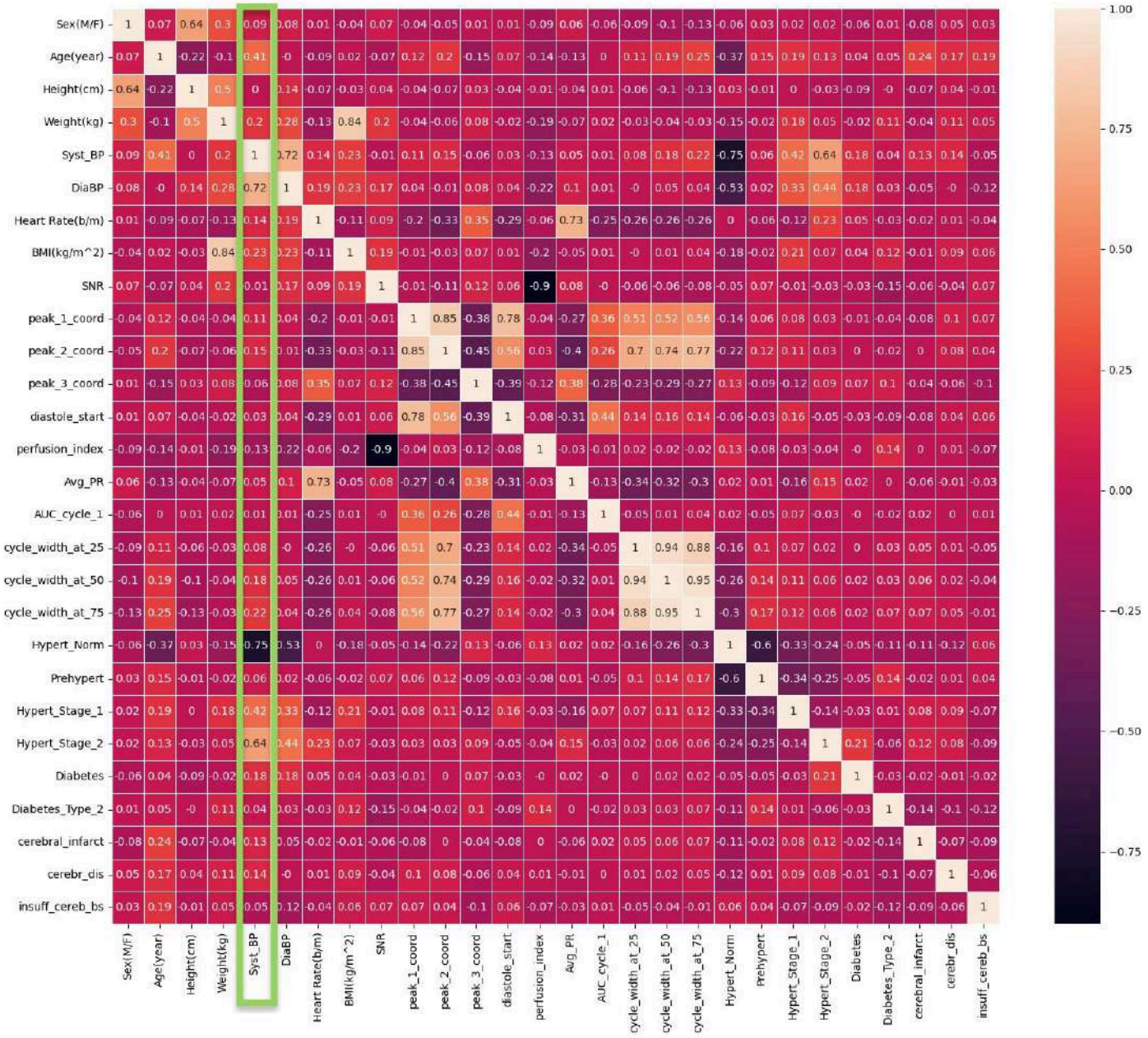


Figure 20. Heatmap of all the dataset features.

In Data Engineering stage it was decided to select features by 3 criteria:

- avoid negative values, which appeared on peak_3 feature due to incomplete capture of the last cycle of the signal on the record, and, consequently, incorrect determination of the peak value of the third cycle. Instead, a feature of distance between the peaks has been added
- pass the threshold of correlation coefficient of 0.10 on the heatmap
- not have outliers.

This way, features ‘peaks distance 1’, ‘peaks distance 2’ have been added, features

Age, Heart_Rate, BMI, peak_1_coord, peak_2_coord, perfusion_index, cycle_width_at_50, cycle_width_at_75, Hypertension_Normal, Hypert_Stage_1, Hypert_Stage_2, Diabetes, cerebral_infraction, cerebrovascular_disease were selected as passed correlation index threshold. ‘Peaks distance 2’ was excluded from the final selection as one having a number of outliers. The same applies to BMI:

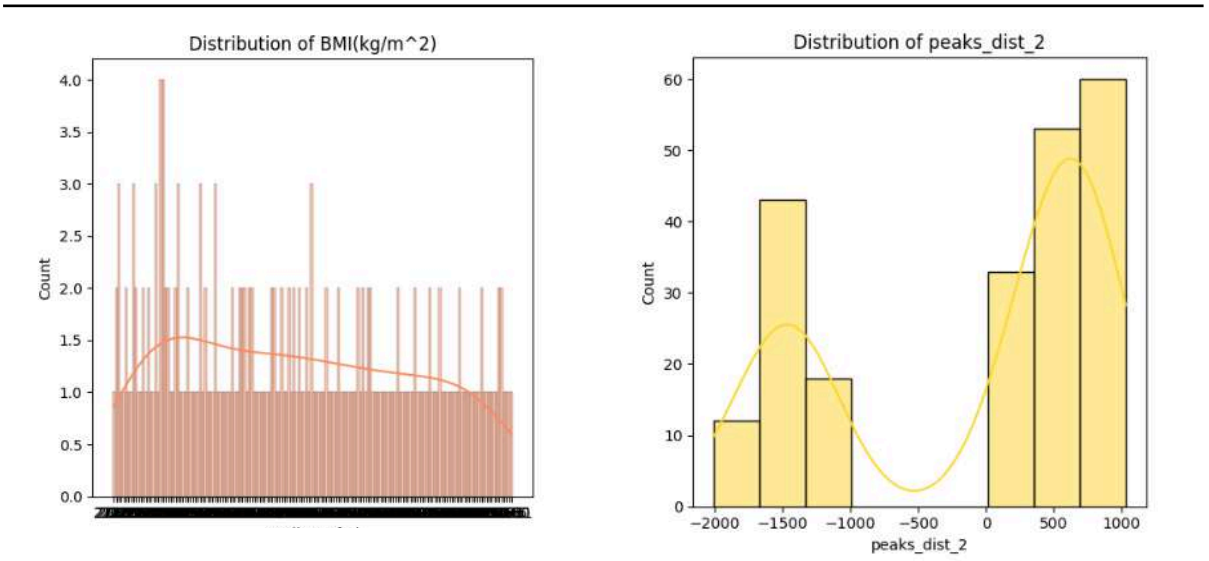


Figure 21. Not normal distributions of the features of BMI and peaks_distance_2.

This way, the final set of the selected features is presented on the heatmap:

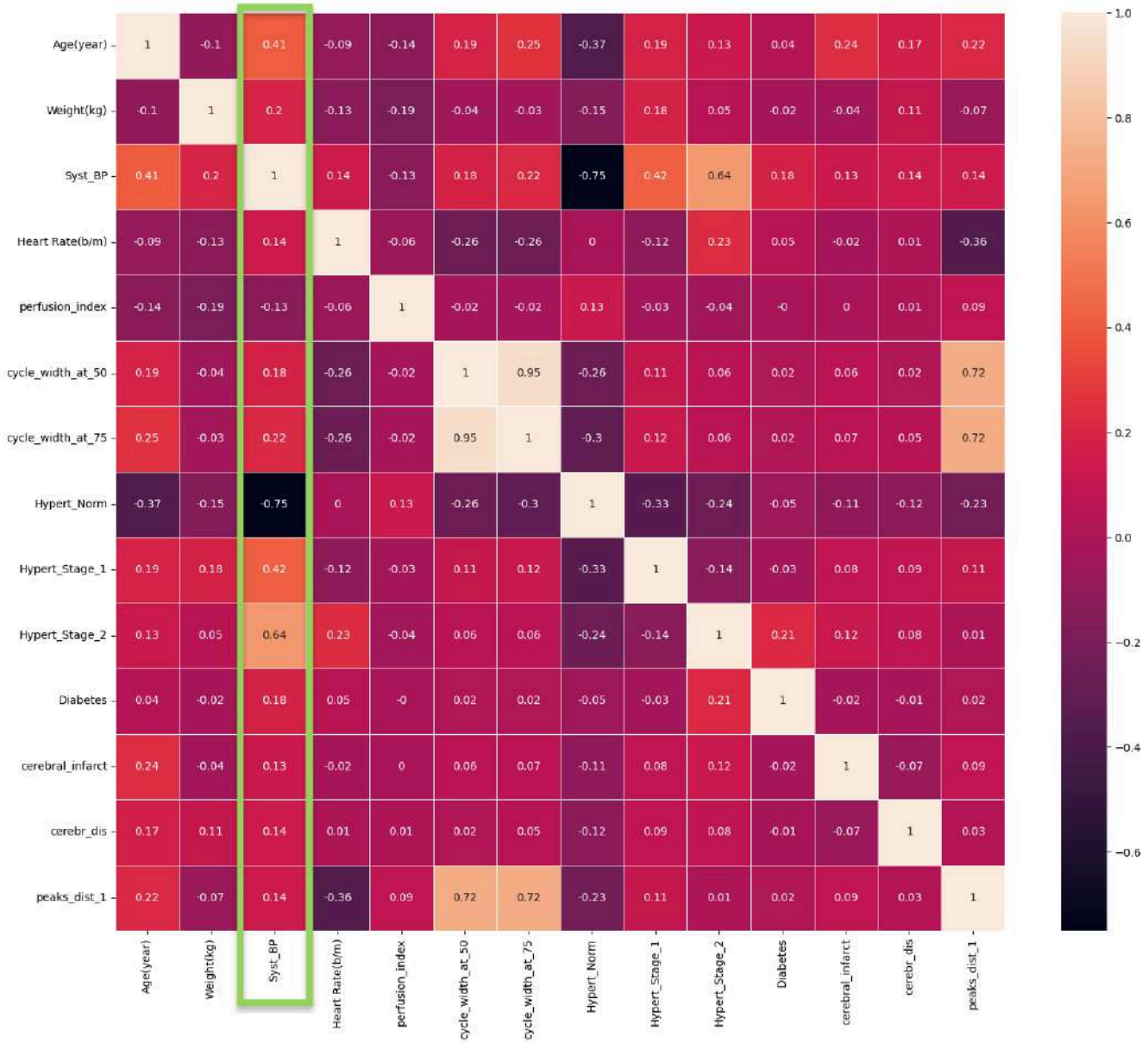


Figure 22. Features, selected for ML model training.

All column for model training and testing	columns selected for model training and testing
['Sex(M/F)', 'Age(year)', 'Height(cm)', 'Weight(kg)', 'Systolic Blood Pressure(mmHg)', 'Diastolic Blood Pressure(mmHg)', 'Heart Rate(b/m)', 'BMI(kg/m^2)', 'SNR', 'peak_1_coord', 'peak_2_coord','peak_3_coord', 'diastole_start', 'perfusion_index', 'average_Pulse_Rate', 'area_under_first_cycle','cycle_width_at_25', 'cycle_width_at_50', 'cycle_width_at_75',	['Age(year)',Weight(kg), Heart Rate (b/m), 'perfusion_index', 'cycle_width_at_50', 'cycle_width_at_75', ,'Hypertension_Normal', 'Hypertension_Stage 1', 'Hypertension_Stage 2', 'Diabetes', 'cerebral infarction', 'cerebrovascular disease', 'peaks_dist_1',]

'Hypertension_Normal', 'Hypertension_Prehypertension', 'Hypertension_Stage 1 hypertension', 'Hypertension_Stage 2 hypertension', 'Diabetes_Diabetes','Diabetes_Type 2 Diabetes', 'cerebral infarction_cerebral infarction','cerebrovascular disease_cerebrovascular disease', 'cerebrovascular disease_insufficiency of cerebral blood supply']	
R2 for model with all columns	**R2 for model with the selected columns**
R2 for systolic BP: 0.9196	R2 for systolic BP: 0.9016

Table 5. Comparison of the results from all all features and with selected features.

As shown in Table 5 above, filtering out the features with lower correlation ration and keeping only the ones with the higher correlation ratio, along with adding an additional complex feature of peaks_dist_1 rather decreased the model performance to **0.9016 , hence it is decided to use the initial ML model without feature engineering stage with the coefficient of determination = 0.9196.**

7. Discussions

7.1. Limitations

Obtaining well annotated datasets can be challenging due to the need for participants' consent to publish their private data, which must be protected. There are some more challenges to it:

- Diversity Issues

  Data collected from homogeneous groups may lack diversity, limiting the generalizability of findings to broader populations and potentially introducing bias. Moreover, some features, such as BMI or blood type distribution or disease susceptibility could be ethnicity-specific, and hence have limitations in application.

- Data Completeness

  Medical datasets often contain incomplete or missing data due to various reasons, posing challenges in data analysis and interpretation. Filling missing values is complex and can impact the reliability and validity of results.

- Data Annotation and Interpretability Challenges

  Datasets created by non-data experts, like clinicians, may lack clear descriptions or feature names, hindering interpretation and potentially introducing biases in analyses. Inadequate explanations of dataset features, sensor inaccuracies, doctor errors, can lead to misinterpretations and inaccuracies, affecting research validity.

Preprocessing is crucial to address outliers and missing values, ensuring data quality.

7.2. Future Prospects

Given the findings and scope of limitations, it is seen that the study could be continued in finding MAP, instead of SBP and DBP [20]. It is a calculated value that represents the average pressure in a persons' arteries during one cardiac cycle. It is considered a better indicator of perfusion to vital organs than SBP.

There are further improvements which were gained while working with ML algorithm:

- Choosing the best features with KNN algorithm

These days there are another type of ML models presented in both business and academic domains, which include several ML models presented in different stages of the main ML model, i.e., ML model 1 is used for classifying the features of for choosing the right features for stage 2, where ML model 2 is trained on the features selected in ML model 1. This could be an option for further project development.

- Encoding values

  There are models which are able to encode categorical parameters better. For instance, catboost [50] is a gradient boosting library, which is able to handle categorical features directly without the need for manual encoding by using an efficient method called "Ordered Target Encoding" or "Bayesian Target Encoding".

- Prior decomposition of a signal into Fourier series.
  Fourier series is a representation of the signal components as a sum of trigonometric functions. Decomposition of a signal into a Fourier series could help filtering out noise artefacts and extracting features such as dominant frequencies, harmonics, or spectral power that may be indicative of certain health conditions or physiological states, however, this is of an advanced research.

- Prior data classification of a data Blood pressure: normal, prehypertension, hypertension i and hypertension ii and assigning some weights accordingly.

- Dividing AUC into the systolic part of AUC and the diastolic part of AUC, along with using the ratio between the two areas, as done by Mejʹıa-Mejʹıa et al.

- Cutting off all the parts of a signal irrelevant to the analysis: the part before the first (or second) cycle upstroke phase and after the end of the third cycle downstroke phase, if presented. Next, averaging all three (or two) signals into a merged averaged cycle, hoping the individual cycle wavefront remains still,

- Using all the cycles for extracting the features, with the features from each cycle stored in separate columns, filling the missing values for a partially captured cycle with NANs, and feeding the model with the features columns for all cycles.

8. SUMMARY

The attempt to predict blood pressure based on a dataset of photoplethysmography signals is motivated by the belief that PPG data alone can effectively forecast blood pressure, as indicated by research papers [4].

This potential has sparked interest in developing new PPG sensors and enhancing existing ones, as well as integrating machine learning models into cuffless wearable devices.

To predict blood pressure accurately, an annotated dataset is carefully selected, explored, and analysed. Various filtering techniques are evaluated, and the most suitable ones are chosen based on specific requirements. The literature review guides the selection of PPG signal features for extracting and utilising in predicting systolic blood pressure. These features are integrated into the dataset, complementing the existing physical characteristics of individuals. The initial model demonstrates a coefficient of determination of **0.9196**, indicating a strong performance though not perfect. To enhance the model further, diverse data engineering techniques are employed, followed by a comparative analysis.

The application of wavelet mathematical transformations shows significant promise and has generated considerable interest, prompting further investigation and potential integration into future model enhancements. The continuous improvement of the model encompasses both hardware aspects, such as sensor enhancements, and machine learning techniques including principal component analysis, various encoding methods, and scaling techniques. This multifaceted approach underscores the ongoing quest for refining predictive models in the realm of blood pressure estimation.

LIST OF REFERENCES

[1] Y. Liang, Z. Chen, G. Liu, and M. Elgendi, "A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in China," *Sci. Data*, vol. 5, no. 1, p. 180020, Feb. 2018, doi: 10.1038/sdata.2018.20.

[2] S. Kilickaya, A. Guner, and B. Dal, "Comparison of Different Machine Learning Techniques for the Cuffless Estimation of Blood Pressure using PPG Signals," in *2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, Ankara, Turkey: IEEE, Jun. 2020, pp. 1–6. doi: 10.1109/HORA49412.2020.9152602.

[3] C. El-Hajj and P. A. Kyriacou, "A review of machine learning techniques in photoplethysmography for the non-invasive cuff-less measurement of blood pressure," *Biomed. Signal Process. Control*, vol. 58, p. 101870, Apr. 2020, doi: 10.1016/j.bspc.2020.101870.

[4] P. S. Addison, "Slope Transit Time (STT): A Pulse Transit Time Proxy requiring Only a Single Signal Fiducial Point," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 11, pp. 2441–2444, Nov. 2016, doi: 10.1109/TBME.2016.2528507.

[5] D. Zheng and A. Murray, "Non-invasive quantification of peripheral arterial volume distensibility and its non-linear relationship with arterial pressure," *J. Biomech.*, vol. 42, no. 8, pp. 1032–1037, May 2009, doi: 10.1016/j.jbiomech.2009.02.011.

[6] "Wireless Machine-to-Machine Healthcare Solution Using Android Mobile Devices in Global Networks | Request PDF." Accessed: Apr. 23, 2024. [Online]. Available: https://www.researchgate.net/publication/260636637_Wireless_Machine-to-Machine_Healthcare_Solution_Using_Android_Mobile_Devices_in_Global_Networks

[7] H. Tjahjadi, K. Ramli, and H. Murfi, "Noninvasive Classification of Blood Pressure Based on Photoplethysmography Signals Using Bidirectional Long Short-Term Memory and Time-Frequency Analysis," *IEEE Access*, vol. 8, pp. 20735–20748, 2020, doi: 10.1109/ACCESS.2020.2968967.

[8] Y. Liang, Z. Chen, R. Ward, and M. Elgendi, "Hypertension Assessment via ECG and PPG Signals: An Evaluation Using MIMIC Database," *Diagn. Basel Switz.*, vol. 8, no. 3, p. E65, Sep. 2018, doi: 10.3390/diagnostics8030065.

[9] G. Fortino and V. Giampa, "PPG-based methods for non invasive and continuous blood pressure measurement: an overview and development issues in body sensor networks," in *2010 IEEE International Workshop on Medical Measurements and Applications*, Ottawa, ON: IEEE, Apr. 2010, pp. 10–13. doi: 10.1109/MEMEA.2010.5480201.

[10] S. Rao M., N. K. C, A. K., and S. C. Bangera, "An experimental investigation on pulse transit time and pulse arrival time using ecg, pressure and ppg sensors," *Med. Nov. Technol. Devices*, vol. 17, p. 100214, Mar. 2023, doi: 10.1016/j.medntd.2023.100214.

[11] E. Mejía-Mejía, J. Allen, K. Budidha, C. El-Hajj, P. A. Kyriacou, and P. H. Charlton, "Photoplethysmography signal processing and synthesis," in *Photoplethysmography*, Elsevier, 2022, pp. 69–146. doi: 10.1016/B978-0-12-823374-0.00015-3.

[12] H. Gesche, D. Grosskurth, G. Küchler, and A. Patzak, "Continuous blood pressure measurement by using the pulse transit time: comparison to a cuff-based method," *Eur. J. Appl. Physiol.*, vol. 112, no. 1, pp. 309–315, Jan. 2012, doi: 10.1007/s00421-011-1983-3.

[13] W. Chen, T. Kobayashi, S. Ichikawa, Y. Takeuchi, and T. Togawa, "Continuous estimation of systolic blood pressure using the pulse arrival time and intermittent calibration," *Med. Biol. Eng. Comput.*, vol. 38, no. 5, pp. 569–574, Sep. 2000, doi: 10.1007/BF02345755.

[14] M. Sharma *et al.*, "Cuff-Less and Continuous Blood Pressure Monitoring: A Methodological Review," *Technologies*, vol. 5, no. 2, p. 21, May 2017, doi:

10.3390/technologies5020021.
[15] L. A. Geddes, M. H. Voelz, C. F. Babbs, J. D. Bourland, and W. A. Tacker, "Pulse Transit Time as an Indicator of Arterial Blood Pressure," *Psychophysiology*, vol. 18, no. 1, pp. 71–74, Jan. 1981, doi: 10.1111/j.1469-8986.1981.tb01545.x.
[16] Y. Choi, Q. Zhang, and S. Ko, "Noninvasive cuffless blood pressure estimation using pulse transit time and Hilbert–Huang transform," *Comput. Electr. Eng.*, vol. 39, no. 1, pp. 103–111, Jan. 2013, doi: 10.1016/j.compeleceng.2012.09.005.
[17] R. Mukkamala *et al.*, "Toward Ubiquitous Blood Pressure Monitoring via Pulse Transit Time: Theory and Practice," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 8, pp. 1879–1901, Aug. 2015, doi: 10.1109/TBME.2015.2441951.
[18] A V J Challoner and C A Ramsay, "A photoelectric plethysmograph for the measurement of cutaneous blood flow," *Phys. Med. Biol.*, vol. 19, no. 3, pp. 317–328, May 1974, doi: 10.1088/0031-9155/19/3/003.
[19] J. G. Webster, Ed., *Design of Pulse Oximeters*, 0 ed. CRC Press, 1997. doi: 10.1201/9780367802592.
[20] J. Park, H. S. Seok, S.-S. Kim, and H. Shin, "Photoplethysmogram Analysis and Applications: An Integrative Review," *Front. Physiol.*, vol. 12, p. 808451, Mar. 2022, doi: 10.3389/fphys.2021.808451.
[21] Xiao-Rong Ding and Yuan-Ting Zhang, "Photoplethysmogram intensity ratio: A potential indicator for improving the accuracy of PTT-based cuffless blood pressure estimation," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Milan: IEEE, Aug. 2015, pp. 398–401. doi: 10.1109/EMBC.2015.7318383.
[22] X. Ding, B. P. Yan, Y.-T. Zhang, J. Liu, N. Zhao, and H. K. Tsang, "Pulse Transit Time Based Continuous Cuffless Blood Pressure Estimation: A New Extension and A Comprehensive Evaluation," *Sci. Rep.*, vol. 7, no. 1, p. 11554, Sep. 2017, doi: 10.1038/s41598-017-11507-3.
[23] M. Nitzan, F. Glikberg, J. Fidel, C. Gross, H. Bar-On, and Y. Mahler, "The Relationship between Systolic Blood Pressure and Microvascular Resistance in Non-Diabetic and Diabetic Subjects," *J. Basic Clin. Physiol. Pharmacol.*, vol. 3, no. 3, pp. 193–206, Jul. 1992, doi: 10.1515/JBCPP.1992.3.3.193.
[24] R. M. Stern, "Ear Lobe Photoplethysmography," *Psychophysiology*, vol. 11, no. 1, pp. 73–75, Jan. 1974, doi: 10.1111/j.1469-8986.1974.tb00824.x.
[25] J. Allen and A. Murray, "Age-related changes in peripheral pulse timing characteristics at the ears, fingers and toes," *J. Hum. Hypertens.*, vol. 16, no. 10, pp. 711–717, Oct. 2002, doi: 10.1038/sj.jhh.1001478.
[26] R. W. Barnes, J. M. Clayton, G. E. Bone, E. E. Slaymaker, and J. Reinertson, "Supraorbital photoplethysmography. Simple, accurate screening for carotid occlusive disease," *J. Surg. Res.*, vol. 22, no. 4, pp. 319–327, Apr. 1977, doi: 10.1016/0022-4804(77)90150-0.
[27] P. A. Kyriacou, S. Powell, R. M. Langford, and D. P. Jones, "Investigation of oesophageal photoplethysmographic signals and blood oxygen saturation measurements in cardiothoracic surgery patients," *Physiol. Meas.*, vol. 23, no. 3, pp. 533–545, Aug. 2002, doi: 10.1088/0967-3334/23/3/305.
[28] B. M. Choi *et al.*, "Development of a new analgesic index using nasal photoplethysmography," *Anaesthesia*, vol. 73, no. 9, pp. 1123–1130, Sep. 2018, doi: 10.1111/anae.14327.
[29] 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," *Int. J. Biosens. Bioelectron.*, vol. 4, no. 4, pp. 195–202, 2018, doi: 10.15406/ijbsbe.2018.04.00125.
[30] "Sampling rate (audio) - Glossary - Federal Agencies Digitization Guidelines Initiative." Accessed: May 25, 2024. [Online]. Available: https://www.digitizationguidelines.gov/term.php?term=samplingrateaudio
[31] J. Allen, "Quantifying the Delays Between Multi-Site Photoplethysmography Pulse and Electrocardiogram R-R Interval Changes Under Slow-Paced Breathing," *Front. Physiol.*, vol. 10, p. 1190, Sep. 2019, doi: 10.3389/fphys.2019.01190.

[32] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiol. Meas.*, vol. 28, no. 3, pp. R1–R39, Mar. 2007, doi: 10.1088/0967-3334/28/3/R01.
[33] M. Elgendi, "On the Analysis of Fingertip Photoplethysmogram Signals," *Curr. Cardiol. Rev.*, vol. 8, no. 1, pp. 14–25, Jun. 2012, doi: 10.2174/157340312801215782.
[34] H. H. Asada, P. Shaltis, A. Reisner, Sokwoo Rhee, and R. C. Hutchinson, "Mobile monitoring with wearable photoplethysmographic biosensors," *IEEE Eng. Med. Biol. Mag.*, vol. 22, no. 3, pp. 28–40, May 2003, doi: 10.1109/MEMB.2003.1213624.
[35] C. P. Chua and C. Heneghan, "Continuous Blood Pressure Monitoring using ECG and Finger Photoplethysmogram," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, New York, NY: IEEE, Aug. 2006, pp. 5117–5120. doi: 10.1109/IEMBS.2006.259612.
[36] W. B. Murray and P. A. Foster, "The peripheral pulse wave: Information overlooked," *J. Clin. Monit.*, vol. 12, no. 5, pp. 365–377, Sep. 1996, doi: 10.1007/BF02077634.
[37] J. C. Dorlas and J. A. Nijboer, "PHOTO-ELECTRIC PLETHYSMOGRAPHY AS A MONITORING DEVICE IN ANAESTHESIA," *Br. J. Anaesth.*, vol. 57, no. 5, pp. 524–530, May 1985, doi: 10.1093/bja/57.5.524.
[38] E. C.-P. Chua, S. J. Redmond, G. McDarby, and C. Heneghan, "Towards Using Photo-Plethysmogram Amplitude to Measure Blood Pressure During Sleep," *Ann. Biomed. Eng.*, vol. 38, no. 3, pp. 945–954, Mar. 2010, doi: 10.1007/s10439-009-9882-z.
[39] E. R. J. Seitsonen *et al.*, "EEG spectral entropy, heart rate, photoplethysmography and motor responses to skin incision during sevoflurane anaesthesia," *Acta Anaesthesiol. Scand.*, vol. 49, no. 3, pp. 284–292, Mar. 2005, doi: 10.1111/j.1399-6576.2005.00654.x.
[40] C. C. Y. Poon, X. F. Teng, Y. M. Wong, C. Zhang, and Y. T. Zhang, "Changes in the Photoplethysmogram Waveform After Exercise," in *2004 2nd IEEE/EMBS International Summer School on Medical Devices and Biosensors*, Hong Kong, China: IEEE, 2004, pp. 115–118. doi: 10.1109/ISSMD.2004.1689576.
[41] S. Lu *et al.*, "Can Photoplethysmography Variability Serve as an Alternative Approach to Obtain Heart Rate Variability Information?," *J. Clin. Monit. Comput.*, vol. 22, no. 1, pp. 23–29, Jan. 2008, doi: 10.1007/s10877-007-9103-y.
[42] S. C. Millasseau, R. P. Kelly, J. M. Ritter, and P. J. Chowienczyk, "Determination of age-related increases in large artery stiffness by digital pulse contour analysis," *Clin. Sci.*, vol. 103, no. 4, pp. 371–377, Oct. 2002, doi: 10.1042/cs1030371.
[43] "Time Series Analysis." Accessed: May 25, 2024. [Online]. Available: https://sherbold.github.io/intro-to-data-science/intro-to-data-science/09_Time-Series-Analysis.html
[44] "Study of smoothing filters – Savitzky-Golay filters | Bart Wronski." Accessed: May 01, 2024. [Online]. Available: https://bartwronski.com/2021/11/03/study-of-smoothing-filters-savitzky-golay-filters/
[45] A. A. Awad *et al.*, "The relationship between the photoplethysmographic waveform and systemic vascular resistance," *J. Clin. Monit. Comput.*, vol. 21, no. 6, pp. 365–372, Nov. 2007, doi: 10.1007/s10877-007-9097-5.
[46] R. Ferizoli, P. Karimpour, J. M. May, and P. A. Kyriacou, "Arterial stiffness assessment using PPG feature extraction and significance testing in an in vitro cardiovascular system," *Sci. Rep.*, vol. 14, no. 1, p. 2024, Jan. 2024, doi: 10.1038/s41598-024-51395-y.
[47] S. Usman, R. Mohamad Rozi, M. B. I. Reaz, and M. A. Mohd Ali, "Analysis of area under curve of PPG and its relation with HbA1c," in *2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences*, Langkawi, Malaysia: IEEE, Dec. 2012, pp. 260–263. doi: 10.1109/IECBES.2012.6498065.
[48] M. Coutrot, E. Dudoignon, J. Joachim, E. Gayat, F. Vallée, and F. Dépret, "Perfusion index: Physical principles, physiological meanings and clinical implications

in anaesthesia and critical care," *Anaesth. Crit. Care Pain Med.*, vol. 40, no. 6, p. 100964, Dec. 2021, doi: 10.1016/j.accpm.2021.100964.
[49] "Coefficient of Determination (R^2) | Calculation & Interpretation." Accessed: May 25, 2024. [Online]. Available: https://www.scribbr.com/statistics/coefficient-of-determination/
[50] "CatBoost - open-source gradient boosting library." Accessed: May 01, 2024. [Online]. Available: https://catboost.ai/

APPENDICES