TESIS TE142599

IDENTIFIKASIPERSONAL BIOMETRIKBERDASARKAN SINYAL PHOTOPLETHYSMOGRAPHY DARI DETAK JANTUNG

RAUDHATUL JANNAH
2213 205025

PEMBIMBING
Prof. Shi-Jinn Horng
Dr. I Ketut Eddy Purnama,S.T., M.T
Prof.Ir. Mauridhi Hery Purnomo, M.Eng, Ph.D

PROGRAM MAGISTER
BIDANG KEAHLIAN JARINGAN CERDAS MULTIMEDIA
JURUSAN TEKNIK ELEKTRO
FAKULTAS TEKNOLOGI INDUSTRU
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2015

THESIS TE142599

BIOMETRIC PERSONAL IDENTIFICATION BASED ON PHOTOPLETHYSMOGRAPHY SIGNAL BY HEART RATE

RAUDHATUL JANNAH
2213 205 025

SUPERVISOR
Prof. Shi-Jinn Horng
Dr. I Ketut Eddy Purnama,S.T., M.T
Prof. Ir. Mauridhi Hery Purnomo, M.Eng, Ph.D

MASTER PROGRAM
MULTIMEDIA INTELLIGENT NETWORK
DEPARTMENT OF ELECTRICAL ENGINEERING
FACULTY OF INDUSTRIAL TECHNOLOGY
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2015

Thesis has been done to satisfy one of the requirements to obtain the degree of
Magister Teknik (MT)
in
Institut Teknologi Sepuluh Nopember

By :
Raudhatul Jannah
NRP. 2213205025

Examination Date : 29 June 2015
Graduation Period : September 2015

Approved by:

1. Dr. I Ketut Eddy Purnama, ST, MT. (Advisor I)
NIP. 19690730 199512 1 001

2. Prof. Ir. Mauridhi Hery Purnomo, M.Eng, Ph.D. (Advisor II)
NIP. 19580916 198601 1 001

Chairman of Graduate Program,

KEMENTERIAN RISET, TEKNOLOGI, DAN PENDIDIKAN TINGGI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
PROGRAM PASCASARJANA

Prof. Dr. Ir. Adi Soeprijanto, MT.
NIP. 19640405 199002 1 001

IDENTIFIKASI PERSONAL BIOMETRIK BERDASARKAN SINYAL PHOTOPLETHYSMOGRAPHY DARI DETAK JANTUNG

Nama	: Raudhatul Jannah
NRP	: 2213205025
Promotor	: Prof. Shi-Jinn Horng
Co-Promotor 1	: Dr. I Ketut Eddy Purnama, S.T., M.T
Co-Promotor 2	: Prof. Ir. Mauridhi Hery Purnomo, M.Eng., Ph.D

Sistem biometrik sangat berguna untuk membedakan karakteristik individu seseorang. Sistem identifikasi yang paling banyak digunakan diantaranya berdasarkan metode fingerprint, face detection, iris atu hand geometry. Penelitian ini mencoba untuk meningkatkan sistem biometrik menggunakan sinyal Photoplethysmography dari detak jantung. Algoritma yang diusulkan menggunakan seluruh ektraksi fitur yang didapatkan melalui sistem untuk pengenalan biometrik. Efesiensi dari algoritma yang diusulkan didemonstrasikan oleh hasil percobaan yang didapatkan menggunakan metode klasifikasi Multilayer Perceptron, Naïve Bayes dan Random Forest berdasarkan fitur ekstraksi yang didapatkan dari proses sinyal prosesing. Didapatkan 51 subjek pada penelitian ini; sinyal PPG signals dari setiap individu didapatkan melalui sensor pada dua rentang waktu yang berbeda. 30 fitur karakteristik didapatkan dari setiap periode dan kemudian digunakan untuk proses klasifikasi. Sistem klasifikasi menggunakan metode Multilayer Perceptron, Naïve Bayes dan Random Forest; nilai true positive dari masing-masing metode adalah 94.6078 %, 92.1569 % dan 90.3922 %. Hasil yang didapatkan menunjukkan bahwa seluruh algoritma yang diusulkan dan sistem identifikasi biometrik dari pengembangan sinyal PPG ini sangat menjanjikan untuk sistem pengenalan individu manusia.

BIOMETRIC PERSONAL IDENTIFICATION BASED ON PHOTOPLETHYSMOGRAPHY SIGNAL BY HEART RATE

Name : Raudhatul Jannah
NRP : 2213205025
Promotor : Prof. Shi-Jinn Horng
Co-Promotor 1 : Dr. I Ketut Eddy Purnama, S.T., M.T
Co-Promotor 2 : Prof. Ir. Mauridhi Hery Purnomo, M.Eng., Ph.D

The importance of biometric system can distinguish the uniqueness of personal characteristics. The most popular identification systems have concerned the method based on fingerprint, face detection, iris or hand geometry. This study is trying to improve the biometric system using Photoplethysmography signal by heart rate. The proposed algorithm calculates the contribution of all extracted features to biometric recognition. The efficiency of the proposed algorithms is demonstrated by the experiment results obtained from the Multilayer Perceptron, Naïve Bayes and Random Forest classifier applications based on the extracted features. There are fifty one persons joined for the experiments; the PPG signals of each person were recorded for two different time spans. 30 characteristic features were extracted for each period and these characteristic features are used for the purpose of classification. The results were evaluated via the Multilayer Perceptron, Naïve Bayes and Random Forest classifier models; the true positive rates are then 94.6078 %, 92.1569 % and 90.3922 %, respectively. The obtained results showed that both the proposed algorithm and the biometric identification model based on this developed PPG signal are very promising for contact less recognizing systems.

Acknowledgment

In the name of Allah, the Most Gracious, the Most Merciful. All praise to Allah as with His will and blessing this work could be finished. People say it is not that important to get degree if we can make our millions without it, while they do not realize that the function of education is to teach one to think intensively and to think critically. Intelligence plus character. At a time like this, I must thank those who have supported me as I choose to go on the less popular (and much less financially-rewarding) path of personal intellectual satisfaction. These include my father, the most handsome, dr. H. Atmawiyat and prettiest mom Tuti Haryanti thanks for all the support given. Also for my two sisters, I hope we can make our parents proud of us. For someone who so highly values intellectual satisfaction, I am very lucky to have chosen Professor Shi-Jinn Horng as my supervisor. I've also had the privilege to receive the same kind of guidance on occasion from Dr. I Ketut Eddy Purnama and Prof. Mauridhi Hery Purnomo. They gave me the exact kind of guidance that I need and feel comfortable with. For my favorite person, thank you for seeing me through the bad times as well as the good. I have been thankful for being the member of Information Security and Parallel Processing Laboratory RB304-3 and meeting friendly labmates. Also for "keluarga JCM 2013". Thank you for your hospitality and great helps along the way. And I would like to thank DIKTI and National Taiwan University of Science and Technology for the financial support provided to those involved in this work. The last but not least, I would like to thank the anonymous reviewers for their helpful suggestion on this research.

Table of contents

List of Tables

Chapter 1 Introduction

Several studies have been developed for many biometric identification using fingerprint, face detection, iris, hand geometry, lip-movements, walkingstyle, electroencephalography (EEG) and electrocardiography(ECG). This biometric system plays an important role in network security issues. The uniqueness of each individual was needed to improve the security system applications. Biometric identification methods greatly affect the security of information technology. Instead of using password for access control, biometric identification can be used for authentication. Biometric information is hard to be duplicated, lost, forgotten, shared or transferred because it is a part of human body. Unfortunately, hackers can possibly get into system via counterfeit biometric information. For example, the security system using fingerprint, face or iris recognitions can let the hackers use the duplicated images stored in the network database do to the authentication. Fingerprints can be affected by chemical reactions for the people who are working in the industry. Furthermore, biometric identification system using voice can be changed seriously due to aging and health condition. Finally, EEG and ECG-based methods are impractical as various electrodes are required in order to acquire the bio-signals.

In this study, the photoplethysmography (PPG) signal was used for data input that is capable of implementing identification functionality. PPG is a non-invasive electro-optical method which gives information about the volume of blood flowing through a testing zone of the body, close to the skin. PPG device receives and responds to a signal or stimulus from pulse oximeter technology to capture changes in blood volume based Light Emitting Diodes (LEDs). Due to changes in blood volume corresponding to the number (synchronous) of the heartbeat, PPG technique can be used to measure the beat of heart rate. In order to acquire PPG signal, a source of light, the wavelength of which is λ, is placed on one side of a jut of the body (e.g. a finger) and on the other side, a photo-detector is placed right across the source to see the transmitted light. A typical PPG signal consists of a large DC component passing through the skin, muscle and bone without passing through the blood vessels, a small AC component passing directly through the blood vessels by detaching it self from the skin, muscle and bone and also a light passing through the arterial blood vessels. Shortly after the systole, the amount of blood in the arteries increase, thus the intensity of light received decreases. During the diastole, the amount of blood in the arteries decrease and an increase in the light transmittance is observed. The advantage of using the PPG signal is widely used, it is easier and more affordable price sensors.

The method proposed for biometric recognition in this study is composed by data acquisition, pre-processing, PPG signaling and the feature extraction of PPG signal using smoothing PPG signal and its first and second derivatives. The process of data acquisition was provided by 51 volunteers through a PPG data acquisition-card. Various artifacts like analog circuit noises to be found in the signal acquired by pre-processing, medium illuminance change, respiration and base deviation arising from movement are eliminated. 30 features of PPG signal in the time domain, such as systolic peak, diastolic peak, augmentation index, and peak-to-peak interval were found using the PPG signal and its derivatives. For each of these features, feature ranking process was performed by separately calculating their contribution to biometric recognition.

1.1 Related Work

Gu et. al. [1] provided a new approach of human verification using the PPG signals acquired easily from the fingertips. For the group consisting of 17 healthy subjects, they performed experimental studies by obtaining four feature parameters from digitized PPG signals. A feature vector template was formulated using the recorded signals, and later on, the discriminant function was applied in order to verify the data. This promising method of human identification finally achieved a 90% success.

Yao et. al. [2], two important conclusions like the derivatives of PPG signals, and the consistency of subjects within themselves and the distinguishability among different subject are examined. Data taken using Pulse Oximeter Sensor, statistically the results of same subject have a constant time interval against the generated maximum/minimum points, and the derivatives can certainly indicate the features of one's PPG signal and can be used as biometrics for recognition.

Spachos et. al. [3], on the other hand, the feasibility of the application of PPG signal as a single biometrical feature along with the signal-processing methods for the matter involved is being researched. The PPG signals were acquired from the fingertips of 29 healthy subjects using BvpPLUX System from OpenSignal PPG Dataset and also using NONIN pulse oximeter from BioSec PPG Dataset. The classification was applied using the Nearest Neighbor and Majority Voting for the data to match the input signal. The accuracy results of identification depend on the dataset used. This can occur because of the influence of the circuit, the sensor and the current state of data collection. The experimental results suggesting biometrics for identification can be used when PPG signals come under a controlled environment with infallible sensors.

Wei et. al. [4] addressed that PPG signals could reflect numerous physiological parameters, such as heart functions, blood vascular elasticity and blood viscosity. This is a new non-invasive method with the advantages like smoothness and accuracy. It is important to find out efficient pre-processing and feature extraction algorithms in order to deal with the original PPG signal that could be affected by many other factors. Most of the practical methods include median and FIR (finite infinite response) filtration. In this study, a new algorithm is recommended in order to eliminate the wavelet transform-based baseline deviation. The inference of feature spots is another important issue. A sophisticated differential algorithm is used to solve this problem. All these practical algorithms have created an effective platform for determining the physiological parameters.

Gu et. al. [5], had showed a fuzzy-logic approach to examine the feasibility of the application of PPG signals as a new method in the identification of humans. The PPG signals were acquired from the fingertips of 17 healthy subjects and were used as fuzzy entries for the classification of four distinctive features such as the peak number, the upward slope, the downward slope, and the time interval. This fuzzy-based decision-making can reach up to 94% in the same testing and 82.3% for two different trials. This outcome suggests that this new PPG-based biometry is potentially feasible in the verification of humans.

In Wan et. al. [6], the design of an amplifier circuit intended for extracting the DC component of the signal is being negotiated for PPG signals. Consequently, a high AC signal with SNR (signal-to-noise ratio) is acquired from a raw PPG signal, adding a bias-adjusted circuit to the amplifier. This hardware development resulted in acquiring a better signal quality and a data handling convenience in recognition (identification).

In Singh et. al. [7], the fingerprint of someone could be imitated by placing a thin film or using the artificial copy of that print in a biometric system operating via finger scanning. The uniqueness of a finger impact profile was approved in the preliminary studies. This study creates researches into the possibility to utilize the PPG signal as an additional parameter along with the fingerprint.

Kavsaoglu et. al. [8]got data using a microcontroller and sensors DCM03 then using k-NN (k-Nearest Neighbor) to do classification. Forty different features were used for feature extraction stage, including augmentation index, systolic and diastolic peak, pulse width, and peak-to-peak interval. When the results were evaluated for the k-NN classifier model created along with the proposed algorithm, an identification of 90.44% for the 1st configuration, 94.44% for the 2nd configuration, and 87.22% for the 3rd configuration has successfully been attained.

1.2 Organization of the Thesis

The following chapters describe the organization of this thesis. For the second chapter, the basic theories used are explained. For the third chapter, the data acquisition, signal processing and features extraction are described. Experiments to evaluate the performance of the identification system for the proposed approach are described in the fourth chapter. Finally, the conclusion remarks are included in the last chapter.

Chapter 2 Overview

2.1 Photoplethysmoraphy Signal

Heart is a vital organ of the human body that have a function to circulate or pump blood throughout the body. Owing to the pumping of blood in an organ volume will change, Photoplethysmograph (PPG) is a device that can be used to detect changes in the blood volume. Although it works to detect changes in blood volume, commonly the use of PPG is to calculate and show the heart rate per minute, while the changes in blood volume information is not displayed. It would be helpful if the PPG, which shows the cardiac signal of the inspected person, displays the data in graphical form for blood volume changes continuously. This study does researches into the possibility to display blood volume information. PPG graph continuous observation can be used to detect heart problems, for example, the heart contraction premature, observing the observation cycle, and so on.

PPG is an instrument used to measure changes in blood volume within an organ or the whole body. Usually the result of fluctuations in the volume of blood or air contained therein. Photoplethysmograph (PPG) is an instrument that works plesthysmograph using optical sensors [16]. In PPG technique known two kinds of sensor mounting configuration modes: 1. Mode of transmission: a light source (LED) mounted dealing with light sensor (LDR). LDR detects changes in the light emitted by the LED due to absorption by organs (blood, skin, and meat / muscle) directly. 2. Mode reflection: In reflection mode LED and LDR fitted lined. The signal or light changes detected by LDR is reflected signal or reflection.

PPG signal was generated by periodic ejection of the heart, so it has a close relationship with the ejection period, from which the heart rate could be extracted. In another aspect, the blood flowing in the vessels was affected by the vessel elasticity and blood viscosity. Hence, much cardiovascular information, like the degree of angiosclerosis, could also be picked up from the PPG signal. All the physiological parameters could be reflected in PPG signal feature points. In Figures 2.1 and 2.2, the relationships between the original PPG signals, its first and second derivatives are shown, respectively. Feature points in PPG signals are determined using these correlations. Systolic peak is the main peak of the signal. At this point, the blood pressure (BP) is highest in the whole period. Another crest is called diastolic peak, which could reflect the compliance of the arteriola. Heart rate could be obtained from the interval of two main peaks (TPP); the time interval of systolic-diastolic peak was an index of arteriosclerosis. However, due to the presence of various factors, the feature points for real PPG signals were always hard to get directly. So there is great meaning to find efficient and practical pre-processing and feature extraction methods to pinpoint the PPG signal.

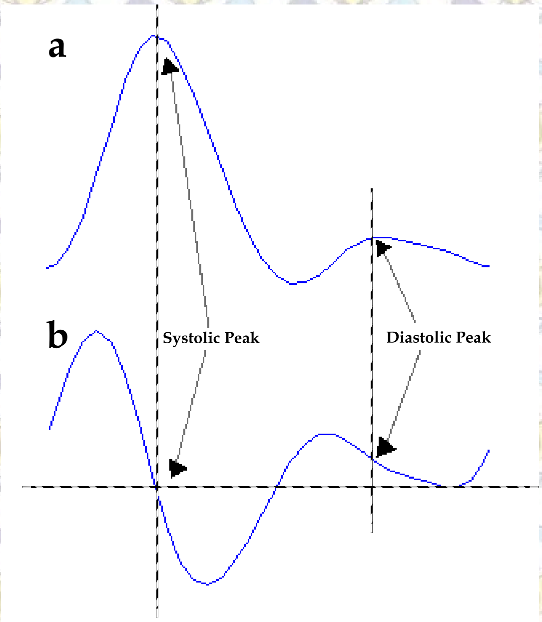


Figure 2.1: Signal measurements.(a) original fingertip PPG and (b) the first derivative of PPG

2.2 Signal Processing

The signal is measured by electrodes attached to the skin and is sensitive to disturbances such as power source interference and noises due to movement artifacts. Segmentation signal is part of signal processing that aims to remove motion artifact and frequency noise contained in PPG signal. For object recognition system, the

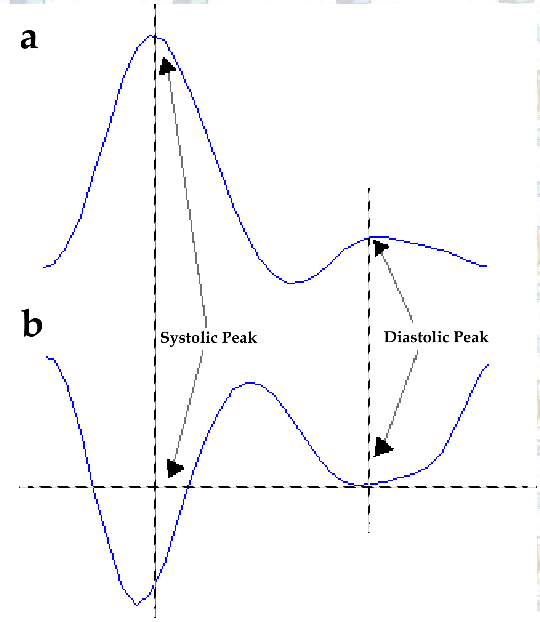


Figure 2.2: Signal measurements.(a) original fingertip PPG and (b) the second derivative of PPG

signal will be divided into several sections with a limit of two different signals. In this study, the signal is divided as a wave where the wave has only one systolic peak and diastolic peak then be normalized. Using the Savitzky-Golay filter, the unwanted peaks are removed via a given threshold and only the systolic peak and diastolic peak are determined and located. We perform the peak detection on the smooth signals and use the logical indexing to find the locations of the peaks.

2.2.1 Savitzky-Golay Filter

Savitzky-Golay smoothing filters (also called digital smoothing polynomial filters or least-squares smoothing filters) are typically used to "smooth out" a noisy signal whose frequency span (without noise) is large. In this type of application, Savitzky-Golay smoothing filters perform much better than standard averaging FIR filters, which tend to filter out a significant portion of the signal's high frequency content along with the noise. Although Savitzky-Golay filters are more effective at preserving the pertinent high frequency components of the signal, they are less successful than standard averaging FIR filters at rejecting noise.

Savitzky-Golay filters are optimal in the sense that they minimize the least-squares error in fitting a polynomial to frames of noisy data. An examples of Savitzky-Golay smoothing filters is shown in Figure 2.3

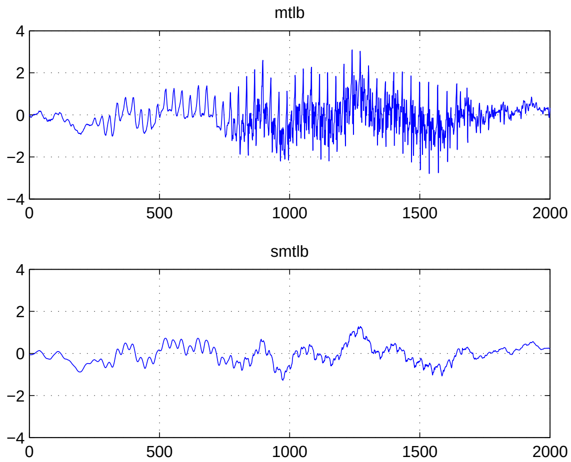


Figure 2.3: An example of Savitzky-Golay smoothing filters: the upper diagram shows the original signals and the lower diagram shows the result after filtering.

2.3 *Classification*

2.3.1 Random Forest

Random forests is a novel ensemble classifier; it uses a similar but improved method of bootstrap as bagging. It uses the strategy of a random selection of a subset of m predictors to grow each tree, where each tree is grown on a bootstrap sample of the training set. This number, m, is used to split the nodes and is much smaller than the total number of variables available for analysis. For further information, please see Breiman [9], in detail.

2.3.2 Multilayer Perceptron

A neural network is an interconnected group of artificial neurons that uses a computational model for information processing. The neural network selected for this study is a multilayer perceptron [10]. The model of a neuron shown in Figure 2.4 indicates that q input signals are received by a neuron. These inputs are weighted and summed together. The threshold, which is treated as an extra connection weight, is then applied to the weighted-sum result. Thus, the linear combiner output (z) or input to the activation function is given by Equation 2.1.

$$z = \sum_i w_i u_i \tag{2.1}$$

where u_i is the i th input to the neuron and wi is the connection weight for the the input u_i. In addition, $u_0 = -1$ and w_0 is the threshold. The neuron output ($h(z)$) is the output from the activation function and is denoted by Equation 2.2. As a result, the output signal from each neuron is limited by a logistic sigmoid function. The neuron model described above is used throughout the multilayer feed-forward network.

$$h(z) = \frac{1}{1 + exp(-z)}. \tag{2.2}$$

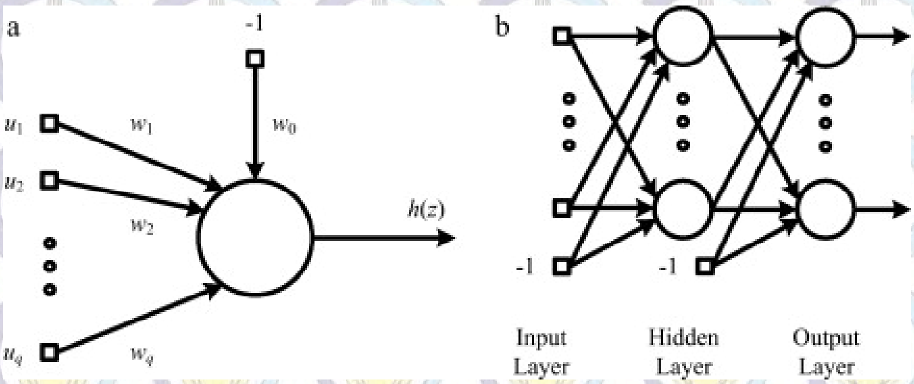
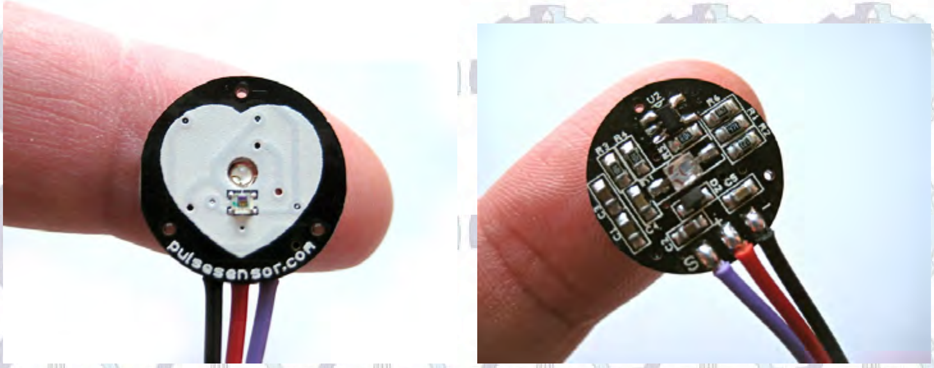


Figure 2.4: Schematic diagram of a multilayer perceptron: (a) computational model of a neuron and (b) feed-forward network with one hidden layer.



(a) Front side (b) Back side

Figure 2.5: The pulse sensor

2.3.3 Naïve Bayes

Bayesian classification is a statistical classification that is able to predict the probability of a class. Bayesian classification is calculated based on Bayes' Theorem described in Equation 2.3.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{2.3}$$

Based on the formula described above, H and X represent a class of events and an attribute, respectively. P (H) is called the prior probability of a class H. P (X) is the prior probability of an attribute X. P (X|H) is the posterior probability that reflects the probability of a class H on the attribute X. P (X|H) indicates the possibility of predictors X in a class H.

2.4 Pulse Sensor

[11]Pulse Sensor is a well-designed plug-and-play heart-rate sensor for Arduino. It can be used by students, artists, athletes, makers, and game & mobile developers who want to easily incorporate live heartrate data into their projects. The sensor clips onto a fingertip or earlobe and plugs right into Arduino. It also includes an open-source monitoring app that graphs pulses obtained in real time.

As shown in Figure 2.5, the front of the sensor with the Heart logo is the side that makes contact with the skin. On the front, a small round hole is where the LED shines through from the back, and there is also a little square just under the LED. The square is an ambient light sensor, exactly like the one used in cellphones, tablets, and laptops, to adjust the screen brightness in different light conditions. The LED shines light into the fingertip or earlobe, or other capillary tissue, and sensor reads the amount of light that bounces back. The other side of the sensor is where the rest of the parts are mounted. The cable is a 24" flat color coded ribbon cable with 3 male header connectors. Red wire = +3V to +5V, black wire = GND and purple wire = Signal as seen in 2.6.

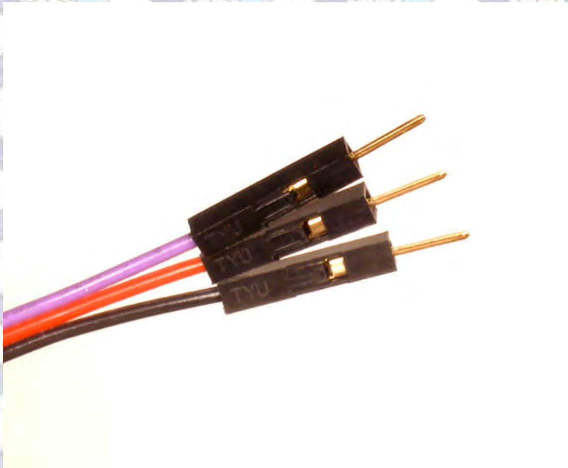


Figure 2.6: Pulse sensor wire

In 2.7., the schematic diagram of the pulse sensor used for acquiring the PPG signal is shown.

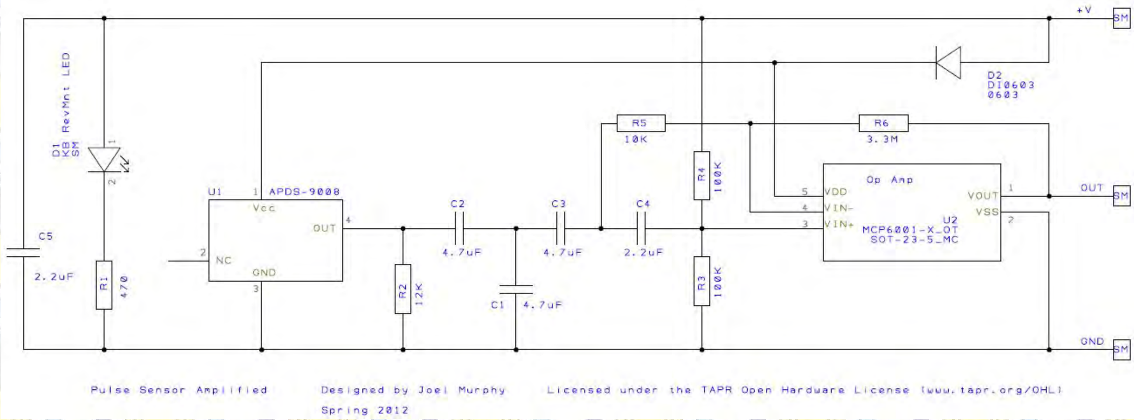


Figure 2.7: Schematic of Pulse sensor

Chapter 3 Approach

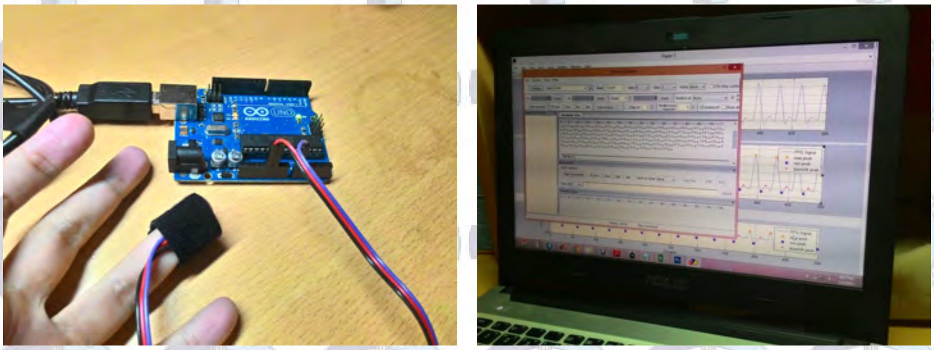
3.1 Data Acquisition

In this study, PPG signals are acquired from a total of fifty one healthy volunteers, and twenty one of them are male and the remaining persons are female. The statistics for age, weight and height of the volunteers are shown in Table 3.1. The data were obtained from their right index fingers while they were seated in a calm position. Total 90 period-signal was acquired from each individual at two different time spans. 30 characteristic features were extracted for each period and these characteristic features are used for the purpose of classification.

Groups	Age (years)	Height (cm)	Weight (kg)
Mean ± SD (range) for males	25±4.36 (18-41)	175±7.62 (157-185)	67.81±13.06 (48-95)
Mean ± SD (range) for females	24.46±2.44 (21-32)	157.87±3.75 (150-165)	55.37±9.59 (43-95)
Mean ± SD (range) for total	24.68±3.35 (18-41)	162.75±7.92 (150-185)	60.44±12.63 (43-95)

Table 3.1: Descriptive statistics for male and female groups

Pulse sensor is a heart rate detection sensor consisting of LEDs and An Avalanche Photodiodes (APDs) and it is connected to the arduino. LEDs serve as the light source while the LDR serves as the light receiver which receives the passing or reflected light by the skin. APDs can be considered as photodetectors which are electronic semiconductor devices that utilize highly sensitive photoelectric effect to convert light into electricity. APDs receive a light that changes according to changes in blood flow in the skin. The Pulse Sensor connected to the Arduino catching the heart beats in digital signals is shown in Figure 3.1. The Arduino is then connected to a computer and the heart beat signals can be transferred to the computer for



(a) Arduino and pulse sensor series (b) Data digital from sensor

Figure 3.1: The acquisition of PPG signal from the volunteers.

further analysis.

3.2 Identification System of PPG Signal

In this study, the block diagram of the identification system using PPG signals is shown in Figure 3.2. In this system were the PPG signals acquired by an arduino and a pulse sensor with a 5 Hz sampling frequency. In order to debug the noises in the PPG signal as a pre-processing, a band-pass filter out of 3rd order Butterworth low-pass and high-pass filters with cutoff frequencies of .8Hz and 5Hz is utilized. Low order polynomial POLYFIT and the polynomial POLYVAL are used to detrend the signal with obvious baseline drift. This drift was mainly caused by the breathe signal and the motion artifact. A Polynomial method is proposed to eliminate the influence of the breathe signal. The result of the polynomial reconstruction method is shown in Figure 3.4.

The method for detecting the peaks was implemented in MATLAB®. Peaks detection was implemented using the function FINDPEAKS, in which, the data of heartbeat have to be input and there are two input arguments have to be defined as shown in Figure 3.5. The MinPeakDistance defines the function used to specify the small peaks distance, or minimum separation between peaks as a positive integer. We can use the MinPeakDistance option to specify that the algorithm

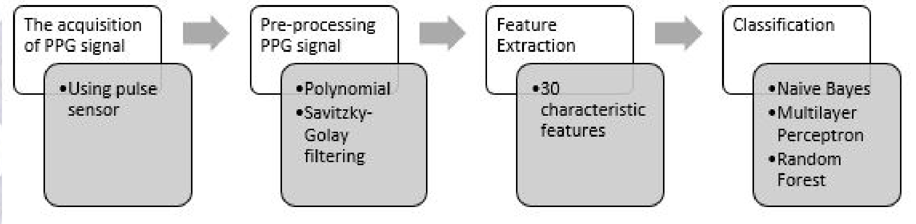
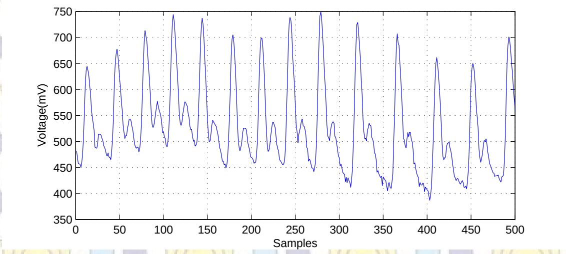
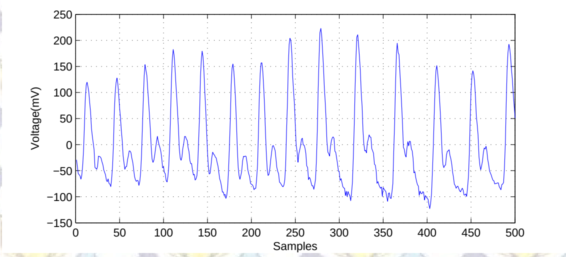


Figure 3.2: The block diagram of an identification system using Heartbeat PPG Signals.



(a) PPG signal with a trend



(b) PPG signal after trend removing

Figure 3.3: Detrend signal

ignore small peaks that occur in the neighborhood of a larger peak. When we specify a value for MinPeakDistance, the algorithm initially identifies all the peaks in the input data and sorts those peaks in descending order. Beginning with the largest peak, the algorithm ignores all identified peaks not separated by more than the value of MinPeakDistance. The MinPeakHeight function finds only those peaks that are greater than the value of MinPeakHeight. findpeaks only returns peaks that exceed the MinPeakHeight.

Figure 3.6 (a) shows the original signal which results from the sensor with a baseline shift and therefore does not represent the true amplitude. In order to remove the trend, fit a low order polynomial to the signal and use the polynomial to remove the trend. Figure 3.6(b) shows the detrend PPG signal. Before peak detection step, determine the locations of the each proposed peak. Thresholding the peaks to locate the proposed peak can remove unwanted peaks caused by noise and it can be done by filtering. In this paper, Savitzky-Golay filtering is used to remove the noise in the signal and the result is shown in Figure 3.6(c). After detrending as shown in Figure 3.6(b), find the main-complex peaks, which are the most prominent repeating peaks in the PPG signal, such as systolic peak (maximum value), diastolic peak and minimum peak. Figure 3.6(d) shows the obtained peaks. The derivative for one-dimensional signals can be calculated by Equations 3.1 and 3.2. The graphs of the 1st and 2nd derivatives are shown in Figures 3.6(e) and 3.6(f), respectively.

$$1st\ derivative: \frac{\partial f}{\partial x} = f(x+1) - f(x) \tag{3.1}$$

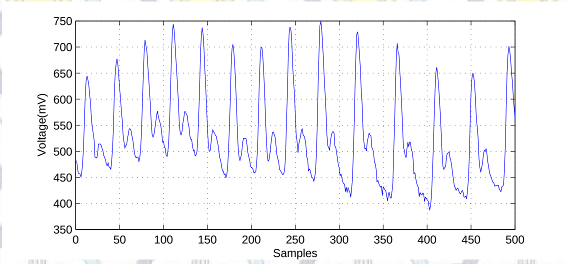
$$2nd\ derivative: \frac{\partial^2 f}{\partial x^2} = f(x+1) + f(x-1) - 2f(x) \tag{3.2}$$

In Figures 3.7, we do some labels on the PPG signal and its corresponding first derivative and second derivative. For example, x, y, z labeling from smoothing signal means systolic peak, diastolic peak and minimum peak, respectively, with their corresponding times are labeled as t1, t2, and t3, respectively. The other features can be calculated by these major feature spots detected in the time domain. Time between two systolic peak referred to as Time Peak to Peak (tpp). The distance between the beginning and the end of the PPG waveform labeled as Time Pulse Interval (tpi). Time between to diastolic peak labeled as Time between Middle Peak (tmp). Augmentation index (AI) is defined for a ratio calculated from the blood pressure waveform as the ratio of systolic peak to diastolic peak (diastolic peak/systolic peak). Takazawa et. al. [12] defined the augmentation index (AI) as the ratio of y to x as Equation 3.3. Rubins et. al. [13] used the reflection index as in Equation 3.4 and introduced an alternative augmentation index. The initial peak point for the first derivative and second derivative are a1 and a2 respectively. Then comes b1 and e1 points for the first derivative and b2 for the second derivative, following the position of systolic peak point. Corresponding times of each feature from both first derivative and second derivative signals are labeled as a1 time, b1 time, c1 time, a2 time, and b2 time, respectively.

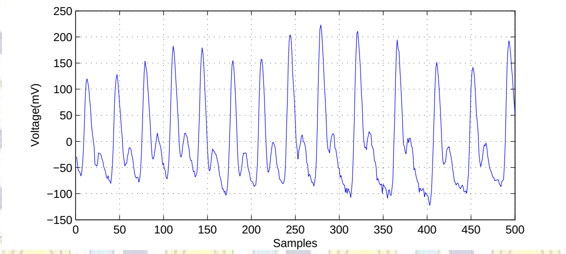
A total of 30 characteristic features are calculated. Table 3.2 shows all 30 features defined for the system. In second column from Table 3.2 are shown the values of each labels from the signal of Figure 3.7.

$$AI : \frac{diastolic\ peak(y)}{systolic\ peak(x)} \tag{3.3}$$

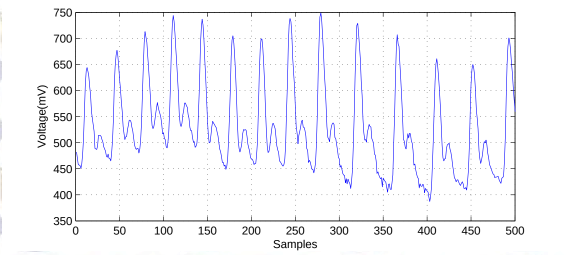
$$Alternative AI : \frac{systolic\ peak(x) - diastolic\ peak(y)}{systolic\ peak(x)} \tag{3.4}$$



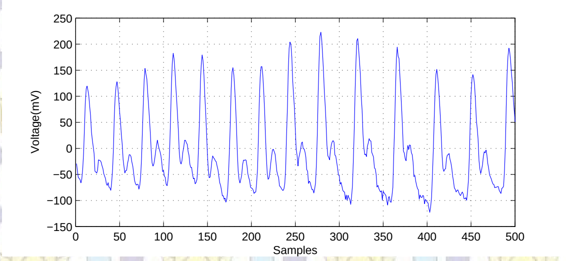
(a) Peak Detection in Original PPG Signal



(b) Peak Detection in Median Filtering PPG Signal



(c) Peak Detection in Savitzky-Golay Filtering PPG Signal



(d) Peak Detection in 3-point moving average Filtering PPG Signal

Figure 3.4: Detrend signal

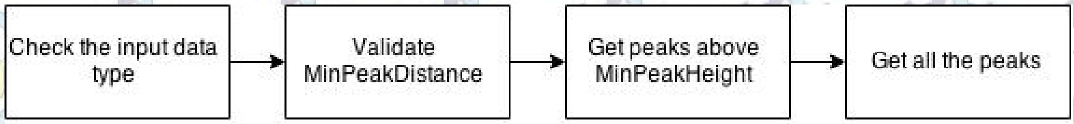
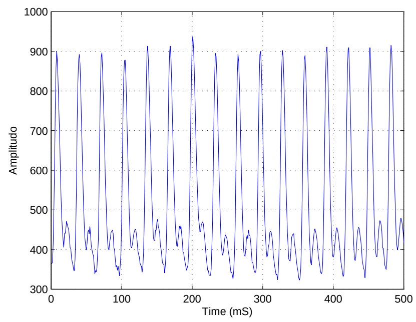
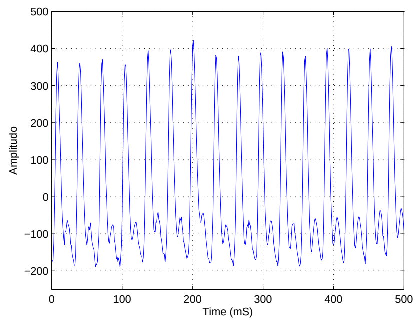


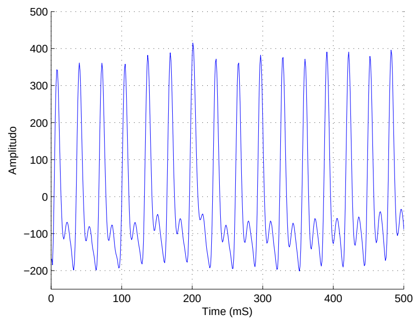
Figure 3.5: Block Diagram of FINDPEAKS function



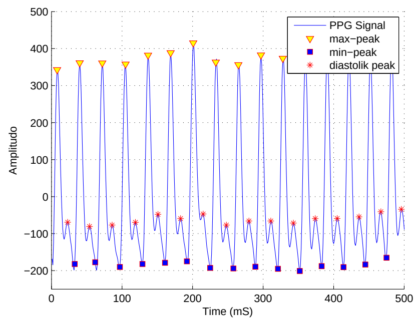
(a) Original signal



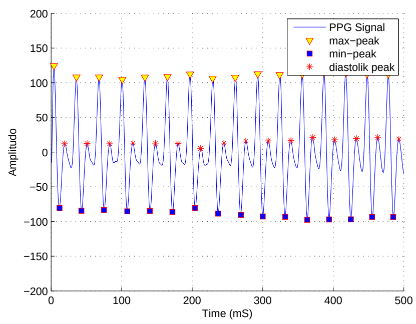
(b) Detrended PPG signal



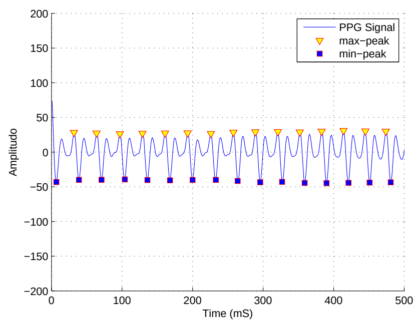
(c) Savitzky-Golay Filtering in PPG Signal



(d) Peak detection in Smoothing PPG signal



(e) Peak detection in 1-st derivative PPG signal



(f) Peak detection in 2-nd derivative PPG signal

Figure 3.6: Peak detection result of Subject 1

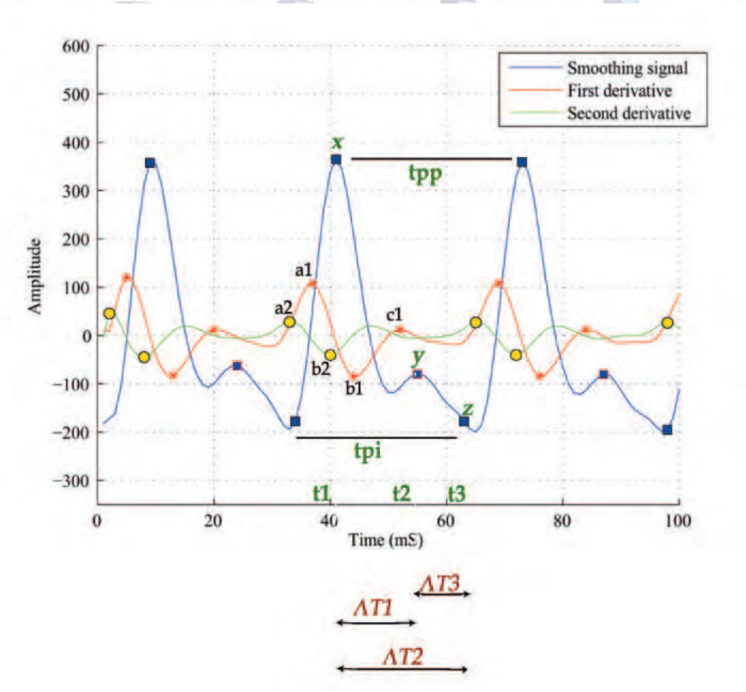


Figure 3.7: The specified parameters used to derive the characteristic features from the PPG signal.

No.	Features	The values of each label
1	Systolic peak (x)	363.907
2	Systolic peak time (t1)	41
3	Time peak to peak (tpp)	32
4	Minimum peak (z)	186.3494
5	Minimum peak time (t3)	63
6	Time Pulse interval (tpi)	35
7	Diastolic peak (y)	-78.5053
8	Diastolic peak time (t3)	55
9	Time between diastolic peaks (tdp)	32
10	y/x (augmentation index)	-0.21573
11	(x-z)/x (alternative augmentation index)	1.215729
12	t1/x (systolic peak output curve)	0.112666
13	y/(tpi-t3) (diastolic peak downward curve)	2.80376
14	t1/tpp	1.28125
15	t2/tpp	1.96875
16	t3/tpp	1.71875
17	ΔT1 (time between diastolic and systolic peaks)	14
18	ΔT2 (time between minimum and systolic peaks)	22
19	ΔT3 (time between minimum and diastolic peaks)	8
20	ΔT1/tpp	0.4375
21	a1	108.1641
22	a1 time	37
23	b1	84.50754
24	b1 time	44
25	c1	12.24569
26	c1 time	52
27	a2	28.34428
28	a2 time	33
29	b2	40.01237
30	b2 time	40

Table 3.2: All 30 features defined for the system

Chapter 4 Experimental Results

4.1 Classification Result

After processing each subject in the dataset, the extracted features are used as an input data in the classification stage. Four scenarios are used to test the accuracy of the system: training set, supplying test set, cross validation and percentage split. In the training set option, testing is performed by using the training data itself. In the cross-validation option, 10-fold is used. As for the percentage split option, from 90-period-signal, 66% is used for the training data and the rest is used for testing data. The data used for training and testing is chosen by the system itself. For the supplying test set option, from 90-period-signal taken, 70-period-signal of the initial data are used for training data and the remaining data are used for testing data. Furthermore, classification methods using Naïve Bayes, Multilayer Perceptron and Random Forest are proposed, respectively.

A feature ranking algorithm is proposed for the 30 features calculated during this study and the result is shown in Table 4.1. The first 5, 10, 15, 20, 25 and 30 features from the ranked ones are selected and used as the classification input.

The percentages of classification success using Multilayer Perceptron are shown in Table 4.2 and are graphically shown in Figure 4.1. The best classification success rate is achieved as 98.6928% for using the training set option and 94.6078% for supplying test set option, where all 30 features are used. As for the supplying test set option, it is seen that the classification success having the same accuracy even if there is feature selection and ranking process. The performance measurement values calculated from Multilayer Perceptron, such as True Positive, False Positive, Precision, Recall, F-measure and ROC Area, are shown in Table 4.5.

The percentages of classification success using Naïve Bayes are shown in Table 4.3 and are graphically shown in Figure 4.2. The best classification success rate is achieved as 92.7451 % for supplied test set option, where the first 15 features are used. Besides, it is seen that when there is no feature selection and ranking process, 92.1569% of classification success at most could be achieved for the classification algorithm in the event that all the features are used. In this case, a 0.58 % of increase in the classification success is attained through the feature ranking and selection process. The performance measurement values calculated from Nave Bayes, such as True Positive, False Positive, Precision, Recall, F-measure and ROC Area, are shown in Table 4.7.

The percentages of classification success using Random Forest are shown in Table4.4 and are graphically shown in Figure 4.3. The best classification success rate is achieved as 99.9346% for use training set option of the classification algorithm, while 91.9164% is achieved for supplied test set option, where the first 25 features are used. Besides, it is seen when there is no feature selection and ranking process, 90.3922% of classification success at most could be achieved for the classification algorithm in the event that all the features are used. In this case, a 1.52% of increase in the classification success is attained through the feature ranking and selection process. The performance measurement values calculated from Random Forest, such as True Positive, False Positive, Precision, Recall, F-measure and ROC Area, are shown in Table 4.6.

Table 4.5, 4.6 and 4.7 the results of the accuracy of the classification method implemented in where the 30 features used are shown. There are four options to classify the data: use training set, supplied test set, cross validation and percentage split. From those options, use training set is slightly higher than the rate for other options, in particular the accuracy is higher for the Multilayer Perceptron. The value indicating the accuracy of identification shows that Multilayer Perceptron method has a slightly higher accuracy rate supplied test set, but it takes the more time of computation.

True positive (TP) indicates that the type of signal from the PPG signal is appropriately identified according to the class. False positive (FP) is a type of signal from the PPG signal which should be identified correctly in class classification wrong turns in the process of identification. The percentage of correct classified instances is often called accuracy or sample accuracy. Kappa is a chance corrected measure of agreement between the classifications and the true classes. It is calculated by taking the agreement, expected by chance away from the observed agreement and dividing by the maximum possible agreement. A value greater than 0 means that the classifier is doing better than chance. The error rates are used for numeric prediction rather than classification. In numeric prediction, predictions are not just right or wrong, the error has a magnitude, and these measures reflect that.

Table 4.8 the comparison of different classification methods to see how the performance varies across different architectures are shown. The performance results show accuracy (ACC), for the following architectures: Fuzzy, K-nn, Nave Bayes, Random Forest and Multilayer Perceptron. The first column of Table 4.8 the performance for fuzzy logic by Gu et al. and the second column the performance for k-nn by Kasgaovlu et al. are shown for seventeen subjects and thirty subjects, respectively. In our study, fifty one subjects joined for the experiments, this method can achieve good performance for more data samples. In this case, a 0.6% increase in the classification success is attained, compared with Fuzzy logic method and a 0.16% increase compared with K-nn method.

Clustering the data with poincare plots was used to clarify the PPG signal characteristics possessed by each individual. Poincare plots are drawn for the three major features: systolic peak, diastolic peak and minimum peak. While a poincare plot with all the data points clustered together produces a good quality of the signal, the poincare plot with the scattered data points produces the corrupted signals. This clustering encompassed the data testing in five subjects is shown in Figure 4.4. As shown in the graph clustering results, the value of the feature extraction on each individual has their own characteristics. In some subjects, signal value changes in heart rate is unstable, which is causes the data matching process task cannot be optimal.

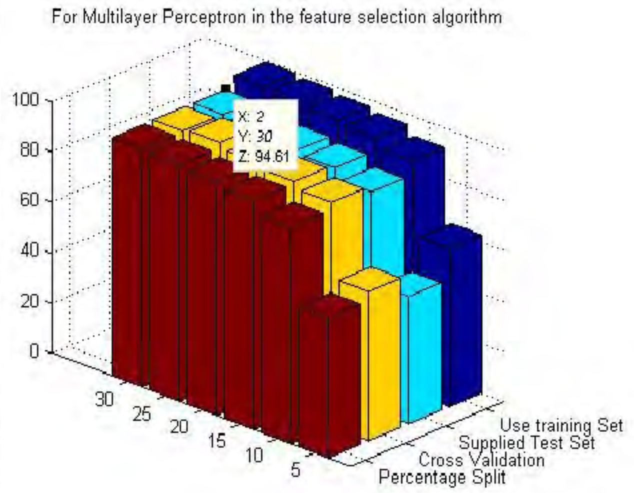


Figure 4.1: The selected feature numbers and the percentages of classification success using Multilayer Perceptron

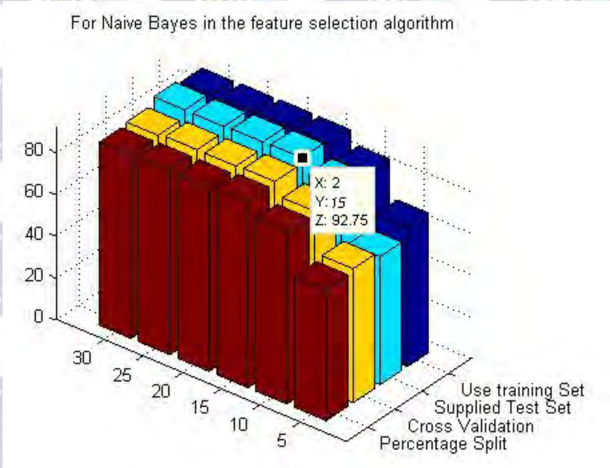


Figure 4.2: The selected feature numbers and the percentages of classification success using Naïve Bayes

Ranked attributes	% Accuracy	features
1	2.8615	feature 19
2	2.81223	feature 18
3	2.42353	feature 17
4	2.14412	feature 20
5	1.91653	feature 11
6	1.7881	feature 10
7	1.71755	feature 23
8	1.68597	feature 29
9	1.63417	feature 21
10	1.57064	feature 1
11	1.49101	feature 4
12	1.4482	feature 27
13	1.21659	feature 25
14	1.08117	feature 7
15	1.0099	feature 3
16	0.76155	feature 9
17	0.61778	feature 6
18	0.59481	feature 13
19	0.54094	feature 12
20	0.018798	feature 8
21	0.017113	feature 26
22	0.09972	feature 2
23	0.09625	feature 28
24	0.09597	feature 22
25	0.08143	feature 16
26	0.0807	feature 30
27	0.07054	feature 15
28	0	feature 14
29	0	feature 5
30	0	feature 24

Table 4.1: Attribute Evaluator (supervised, Class (nominal): 30 class): Information Gain Ranking Filter

Multilayer Perceptron						
The selected feature numbers						
	5	10	15	20	25	30
Using training Set	62.7015%	91.7429%	95.3595%	96.7974%	97.7996%	**98.6928%**
Supplying Test Set	50.4995%	86.376%	90.2941%	91.644 %	93.3697%	**94.6078%**
Cross Validation	59.1503%	89.0196%	91.6993%	93.5294%	**95.8606%**	94.7277%
Percentage Split	55.6054%	85.9705%	90.7111%	92.5048%	93.9142%	**94.4266%**

Table 4.2: The selected feature numbers and the percentages of classification success using Multilayer Perceptron

Naïve Bayes						
The selected feature numbers						
	5	10	15	20	25	30
Using training Set	66.6449%	84.6623%	90.7625%	**90.8932%**	90.6536%	90.3486%
Supplying Test Set	61.2745%	87.0588%	**92.7451%**	92.4614 %	92.0073%	92.1569%
Cross Validation	64.0087%	83.8126%	**90%**	89.5425%	89.5425%	88.5839%
Percentage Split	62.9084%	82.319%	**88.9814%**	88.0846%	88.0205%	87.7002%

Table 4.3: The selected feature numbers and the percentages of classification success using Naïve Bayes

Random Forest						
The selected feature numbers						
	5	10	15	20	25	30
Using training Set	98.8889%	99.7386%	99.9129%	99.8039%	**99.9346%**	99.8257%
Supplying Test Set	69.2157%	85.4902%	89.6078%	91.7348 %	**91.9164%**	90.3922%
Cross Validation	82.1786%	92.7669%	94.4227%	94.8802%	**95.3377%**	95.207%
Percentage Split	80.8456%	91.4158%	92.6329%	93.5298%	93.4017%	**93.5939%**

Table 4.4: The selected feature numbers and the percentages of classification success using Random Forest

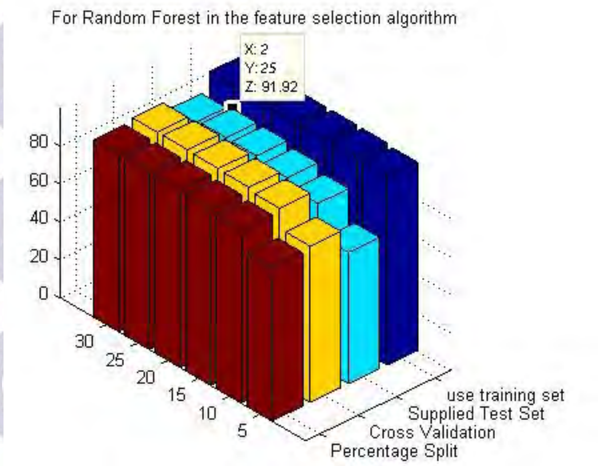


Figure 4.3: The selected feature numbers and the percentages of classification success using Random Forest

Multilayer Perceptron				
	Use training Set	Supplied Test Set	Cross Validation	Percentage Split
Correctly Classified Instances	98.6928%	94.6078%	94.7277%	94.4266%
Incorrectly Classified Instances	1.3072%	5.3922%	5.2723%	5.5734%
Kappa statistic	0.9867	0.945	0.9462	0.9431
Mean absolute error	0.0015	0.0045	0.0036	0.0042
Root mean squared error	0.0225	0.0456	0.0414	0.0435
Relative absolute error	4.0163%	11.6916%	9.3319%	10.8539%
Root relative squared error	16.2607%	32.8716%	29.8865%	31.3847%
TP Rate	0.987	0.946	0.947	0.944
FP Rate	0	0.001	0.001	0.001
Precision	0.987	0.951	0.948	0.947
Recall	0.987	0.946	0.947	0.944
F-Measure	0.987	0.945	0.947	0.944
ROC Area	0.99	0.998	0.997	0.998
Calculation Time	754.75 s	954.31 s	330.42	376.27 s

Table 4.5: The results of the testing process with the target output using Multilayer Perceptron with 30 features are used

Naïve Bayes				
	Use training Set	Supplied Test Set	Cross Validation	Percentage Split
Correctly Classified Instances	90.3486%	92.1569%	88.5839%	87.7002%
Incorrectly Classified Instances	9.6514%	7.8431%	11.4161%	12.2998%
Kappa statistic	0.9.016	0.92	0.8836	0.8745
Mean absolute error	0.0038	0.0033	0.0046	0.005
Root mean squared error	0.0572	0.051	0.0626	0.0657
Relative absolute error	10.0124%	8.5873%	11.8542%	12.9764%
Root relative squared error	41.2403%	36.8115%	45.1387%	47.371%
TP Rate	0.903	0.922	0.886	0.877
FP Rate	0.002	0.002	0.002	0.002
Precision	0.911	0.931	0.894	0.893
Recall	0.903	0.922	0.886	0.877
F-Measure	0.905	0.923	0.887	0.88
ROC Area	0.99	0.997	0.99	0.995
Calculation Time	0.39 s	0.11 s	0.15 s	0.05 s

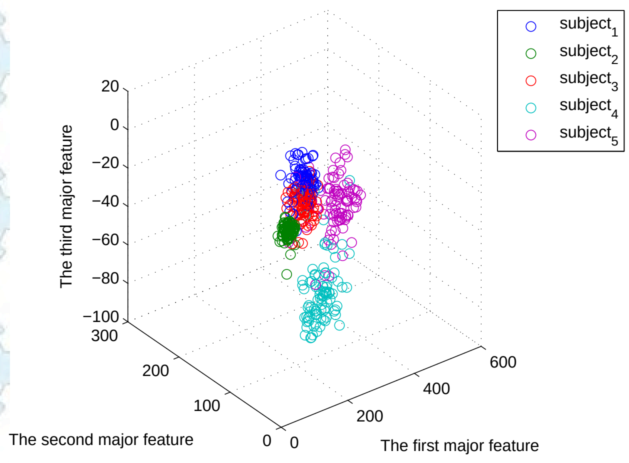
Table 4.6: The results of the testing process with the target output using Naïve Bayes with 30 features are used

Random Forest				
	Use training Set	Supplied Test Set	Cross Validation	Percentage Split
Correctly Classified Instances	99.8257%	90.3922%	95.3595%	93.5939%
Incorrectly Classified Instances	0.1743%	9.6078%	4.6405%	6.4061%
Kappa statistic	0.9982	0.902	0.9527	0.9346
Mean absolute error	0.0028	0.0125	0.0077	0.0086
Root mean squared error	0.0225	0.0665	0.0493	0.053
Relative absolute error	7.3489%	32.39%	20.1467%	22.4212%
Root relative squared error	16.2029%	47.9917%	35.5478%	38.2248%
TP Rate	0.998	0.904	0.954	0.936
FP Rate	0	0.002	0.001	0.001
Precision	0.998	0.911	0.954	0.937
Recall	0.998	0.904	0.954	0.936
F-Measure	0.998	0.902	0.953	0.935
ROC Area	1	0.993	0.99	0.996
Calculation Time	1.51 s	1.25 s	1.53 s	2.01 s

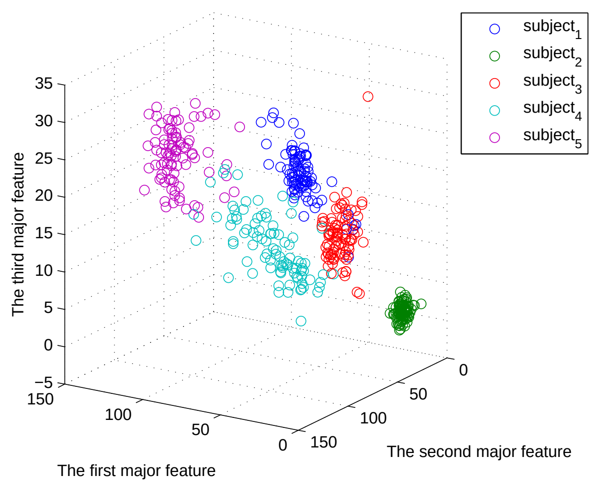
Table 4.7: The results of the testing process with the target output using Random Forest with 30 features are used

	Fuzzy	K-nn	Naive Bayes	Random Forest	MLP
ACC	94%	94.44%	92.15%	90.39%	94.6%

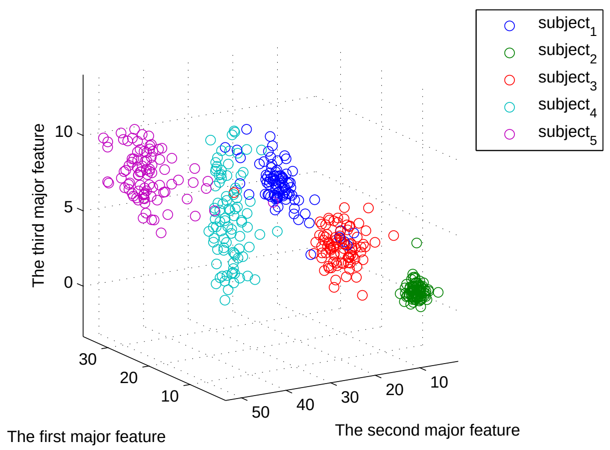
Table 4.8: Comparison with different classification methods.



(a) The class distribution for the smoothing signal.



(b) The class distribution for the 1st derivative signal.



(c) The class distribution for the 2nd derivative signal.

Figure 4.4: The class distributions of five people according to the major features.

Chapter 5 Conclusion

This study has tested the ability of PPG signals for biometric identification system. Based on the research that has been done can be concluded that:

- Designed system can identify the heartbeat of each individual.
- Feature extraction based on the three major peaks value of the photoplethysmography signal.
- The results were evaluated via the Multilayer Perceptron, Nave Bayes and Random Forest classifier models; the true positive rates are then 94.6078%, 92.1569% and 90.3922%, respectively. The obtained results showed that the proposed algorithm and the biometric identification model based on this developed PPG signal are very promising for contact less recognizing systems.

References

[1] Y. Gu, Y. Zhang, and Y. Zhang, "A novel biometric approach in human verification by photoplethysmographic signals," in *Information Technology Applications in Biomedicine, 2003. 4th International IEEE EMBS Special Topic Conference on*, pp. 13–14, April 2003.

[2] J. Yao, X. Sun, and Y. Wan, "A pilot study on using derivatives of photoplethysmographic signals as a biometric identifier," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pp. 4576–4579, Aug 2007.

[3] P. Spachos, J. Gao, and D. Hatzinakos, "Feasibility study of photoplethysmographic signals for biometric identification," in *Digital Signal Processing (DSP), 2011 17th International Conference on*, pp. 1–5, July 2011.

[4] C. Wei, L. Sheng, G. Lihua, C. Yuquan, and P. Min, "Study on conditioning and feature extraction algorithm of photoplethysmography signal for physiological parameters detection," in *Image and Signal Processing (CISP), 2011 4th International Congress on*, vol. 4, pp. 2194–2197, Oct 2011.

[5] Y. Gu and Y. Zhang, "Photoplethysmographic authentication through fuzzy logic," in *Biomedical Engineering, 2003. IEEE EMBS Asian-Pacific Conference on*, pp. 136–137, Oct 2003.

[6] Y. Wan, X. Sun, and J. Yao, "Design of a photoplethysmographic sensor for biometric identification," in *Control, Automation and Systems, 2007. ICCAS '07. International Conference on*, pp. 1897–1900, Oct 2007.

[7] Y. Singh and P. Gupta, "Correlation-based classification of heartbeats for individual identification," *Soft Computing*, vol. 15, no. 3, pp. 449–460, 2011.

[8] A. R. Kavsaoglu, K. Polat, and M. R. Bozkurt, "A novel feature ranking algorithm for biometric recognition with {PPG} signals," *Computers in Biology and Medicine*, vol. 49, no. 0, pp. 1 – 14, 2014.

[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[10] D. Setsirichok, T. Piroonratana, W. Wongseree, T. Usavanarong, N. Paulkhaolarn, C. Kanjanakorn, M. Sirikong, C. Limwongse, and N. Chaiyaratana, "Classification of complete blood count and haemoglobin typing data by a c4.5 decision tree, a nave bayes classifier and a multilayer perceptron for thalassaemia screening," *Biomedical Signal Processing and Control*, vol. 7, no. 2, pp. 202 – 212, 2012.

[11] M. Joel and G. Yury, "http://pulsesensor.com/."

[12] K. Takazawa, N. Tanaka, M. Fujita, O. Matsuoka, T. Saiki, M. Aikawa, S. Tamura, and C. Ibukiyama, "Assessment of vasoactive agents and vascular aging by the second derivative of photoplethysmogram waveform.," in *Hypertension*, p. 32(2): 365—370, August 1998.

[13] U. Rubins, A. Grabovskis, J. Grube, and I. Kukulis, "Photoplethysmography analysis of artery properties in patients with cardiovascular diseases," in *14th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics* (A. Katashev, Y. Dekhtyar, and J. Spigulis, eds.), vol. 20 of *IFMBE Proceedings*, pp. 319–322, Springer Berlin Heidelberg, 2008.

BIOGRAPHY

Name: Raudhatul Jannah (Rana)

Birthday: Cianjur, 13 Juli 1991

Religion: Islam

Address: Kp Rancagoong No. 21, RT.02 RW.05, Cianjur (43285) West Java, Indonesia

Email: ranaraudhatul@gmail.com

Educational Background:

2014 – 2015 **Dual-Degree Master**
Department of Computer Science and Information Engineering
in National Taiwan University of Science and Technology (NTUST), Taiwan

2013 – 2015 **Master Degree**
Department of Intelligent Multimedia Network
in Institut Teknologi Sepuluh Nopember (ITS), Indonesia

2009 – 2013 **Bachelor Degree**
Electronic Engineering Polytechnic Institute of Surabaya (EEPIS)

2006 – 2009 **Senior High School**,
SMA Negeri 1 Cianjur

2003 – 2006 **Junior High School**,
SMP Negeri 1 Cianjur

1997 – 2003 **Elementery School**,
SD Ibu Jenab 1 Cianjur

Biometric Personal Identification Based on PPG Signal by Heart Rate

Raudhatul Jannah[1], I Ketut Edi Purnama[2], Mauridhi Hery Purnomo[3]
[1,2,3)]Jurusan Teknik Elektro, Fakultas Teknologi Industri, Institut Teknologi Sepuluh Nopember
Kampus ITS Sukolilo, Surabaya
Telp. (031) 5947274 ext. 1206, Faks. (031) 5931237
Email: [1]ranaraudhatul@gmail.com, [2]ketut@ee.its.ac.id, [3]hery@ee.its.ac.id

***Abstract*—The importance of biometric system can distinguish the uniqueness of personal characteristics. The most popular identification systems have concerned the method based on fingerprint, face detection, or iris. This study is trying to improve the biometric system using the Photoplethysmography signal by heart rate. The proposed algorithm calculates the contribution of all extracted features to biometric recognition. The efficiency of the proposed algorithms is demonstrated by the experiment results obtained from the Multilayer Perceptron, Nave Bayes and Random Forest classifier applications based on the extracted features. There are fifty one persons joined for the experiments; the PPG signals of each person were recorded for two different time spans. The first half data are used for training and the rest are used for testing. 30 characteristic features were extracted for each period and these characteristic features are used for the purpose of classification. The results were evaluated via the Multilayer Perceptron, Nave Bayes and Random Forest classifier models; the true positive rates are then 94.6078%, 92.1569% and 90.3922%, respectively. The obtained results showed that both the proposed algorithm and the biometric identification model based on this developed PPG signal are very promising for contact less recognizing systems.**

Keywords*—Bakteri Tuberkulosis, Segmentasi, Klasifikasi, Counting,Probabilistic Neural Network.

I. Introduction

SEVERAL studies have been developed for many biometric identification using fingerprint, face detection, iris, hand geometry, lip-movements, walkingstyle, electroencephalography (EEG) and electrocardiography(ECG). This biometric system plays an important role in network security issues. The uniqueness of each individual was needed to improve the security system applications. Biometric identification methods greatly affect the security of information technology. Instead of using password for access control, biometric identification can be used for authentication. Biometric information is hard to be duplicated, lost, forgotten, shared or transferred because it is a part of human body. Unfortunately, hackers can possibly get into system via counterfeit biometric information.

In this study, the photoplethysmography (PPG) signal was used for data input that is capable of implementing identification functionality. PPG is a non-invasive electro-optical method which gives information about the volume of blood flowing through a testing zone of the body, close to the skin. PPG device receives and responds to a signal or stimulus from pulse oximeter technology to capture changes in blood volume based Light Emitting Diodes (LEDs). Due to changes in blood volume corresponding to the number (synchronous) of the heartbeat, PPG technique can be used to measure the beat of heart rate.

Gu et. al. [3], had showed a fuzzy-logic approach to examine the feasibility of the application of PPG signals as a new method in the identification of humans. The PPG signals were acquired from the fingertips of 17 healthy subjects and were used as fuzzy entries for the classification of four distinctive features such as the peak number, the upward slope, the downward slope, and the time interval. This fuzzy-based decision-making can reach up to 94% in the same testing and 82.3% for two different trials. This outcome suggests that this new PPG-based biometry is potentially feasible in the verification of humans.

Spachos et. al. [2], on the other hand, the feasibility of the application of PPG signal as a single biometrical feature along with the signal-processing methods for the matter involved is being researched. The PPG signals were acquired from the fingertips of 29 healthy subjects using BvpPLUX System from OpenSignal PPG Dataset and also using NONIN pulse oximeter from BioSec PPG Dataset. The classification was applied using the Nearest Neighbor and Majority Voting for the data to match the input signal. The accuracy results of identification depend on the dataset used. This can occur because of the influence of the circuit, the sensor and the current state of data collection. The experimental results suggesting biometrics for identification can be used when PPG signals come under a controlled environment with infallible sensors.

Kavsaoglu et. al. [4] got data using a microcontroller and sensors DCM03 then using k-NN (k-Nearest Neighbor) to do classification. Forty different features were used for feature extraction stage, including augmentation index, systolic and diastolic peak, pulse width, and peak-to-peak interval. When the results were evaluated for the k-NN classifier model created along with the proposed algorithm, an identification of 90.44% for the 1st configuration, 94.44% for the 2nd configuration, and 87.22% for the 3rd configuration has successfully been attained.

The method proposed for biometric recognition in this study is composed by data acquisition, pre-processing, PPG signaling and the feature extraction of PPG signal using smoothing PPG signal and its first and second derivatives. The process of data acquisition was provided by 51 volunteers through a PPG data acquisitioncard. Various artifacts like analog circuit noises to be found in the signal acquired by pre-processing, medium illuminance change, respiration and base deviation arising from movement are eliminated. 30 features of PPG signal in the time domain, such as systolic peak, diastolic peak, augmentation index, and peak-to-peak interval were found using the PPG signal and its derivatives. For each of these features, feature ranking process was performed by separately calculating their contribution to biometric recognition.

II. Data Acquisition and Methodology

A. Data Acquisition

Pulse sensor is a heart rate detection sensor consisting of LEDs and An Avalanche Photodiodes (APDs) and it is connected to the arduino. LEDs serve as the light source while the LDR serves as the light receiver which receives the passing or reflected light by the skin. APDs can be considered as photodetectors which are electronic semiconductor devices that utilize highly sensitive photoelectric effect to convert light into electricity. APDs receive a light that changes according to changes in blood flow in the skin. The Pulse Sensor connected to the Arduino catching the heart beats in digital signals is shown in Figure 1. The Arduino is then connected to a computer and the heart beat signals can be transferred to the computer for further analysis.

Fig. 1. The acquisition of PPG signal from the volunteers.

B. Identification System of PPG Signal

In this study, the block diagram of the identification system using PPG signals is shown in Figure 2. The signal is measured by electrodes attached to the skin and is sensitive to disturbances such as power source interference and noises due to movement artifacts. Segmentation signal is part of signal processing that aims to remove motion artifact and frequency noise contained in PPG signal. For object recognition system, the signal will be divided into several sections with a limit of two different signals. In this study, the signal is divided as a wave where the wave has only one systolic peak and diastolic peak then be normalized. Using the Savitzky-Golay filter, the unwanted peaks are removed via a given threshold and only the systolic peak and diastolic peak are determined and located. We perform the peak detection on the smooth signals and use the logical indexing to find the locations of the peaks.

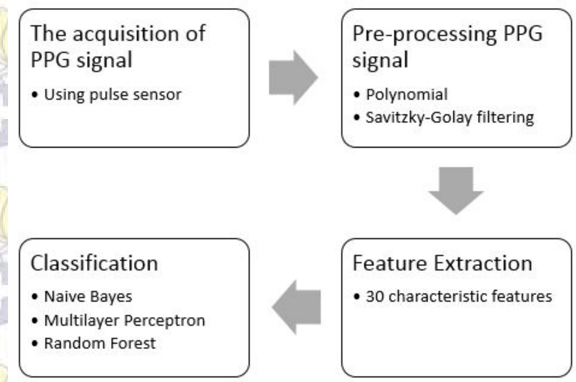


Fig. 2. The block diagram of an identification system using Heartbeat PPG Signals.

After detrending and filtering, find the main-complex peaks, which are the most prominent repeating peaks in the PPG signal, such as systolic peak (maximum value), diastolic peak and minimum peak. Figure 3 shows the obtained peaks. The derivative for one-dimensional signals can be calculated by Equations 1 and 2. The graphs of the 1st and 2nd derivatives are shown in Figures 4 and 5, respectively.

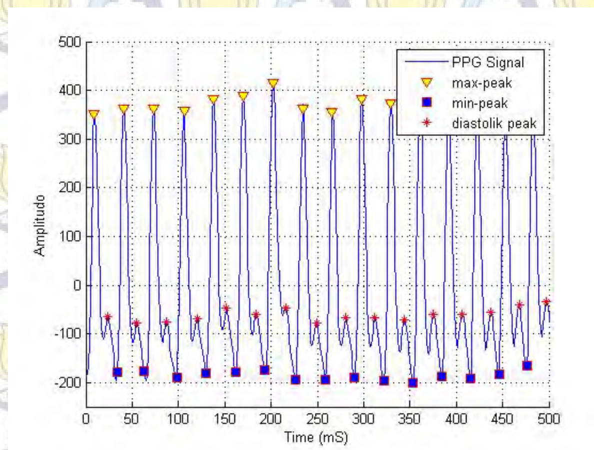


Fig. 3. Peak detection in Smoothing PPG signal.

$$1st\ derivative : \frac{\partial f}{\partial x} = f(x+1) - f(x) \tag{1}$$

$$2nd\ derivative : \frac{\partial^2 f}{\partial x^2} = f(x+1) + f(x-1) - 2f(x) \tag{2}$$

In Figures 6, we do some labels on the PPG signal and its corresponding first derivative and second derivative. For example, x, y, z labeling from smoothing signal means systolic peak, diastolic peak and minimum peak, respectively, with their corresponding times are labeled as t1, t2, and t3, respectively. The other features can be calculated by these major feature spots detected in the time domain. Time between two systolic peak referred to as Time Peak to Peak (tpp).

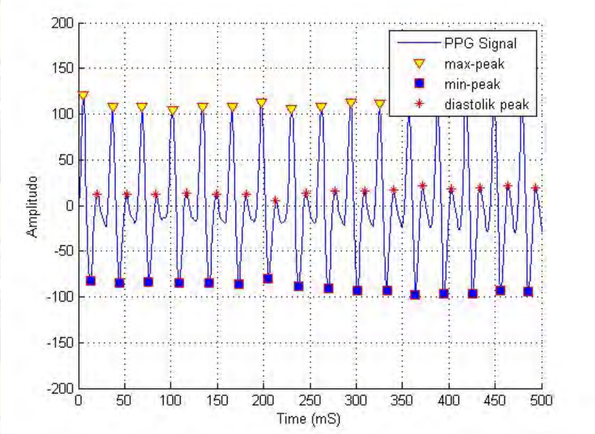


Fig. 4. Peak detection in 1-st derivative PPG signal.

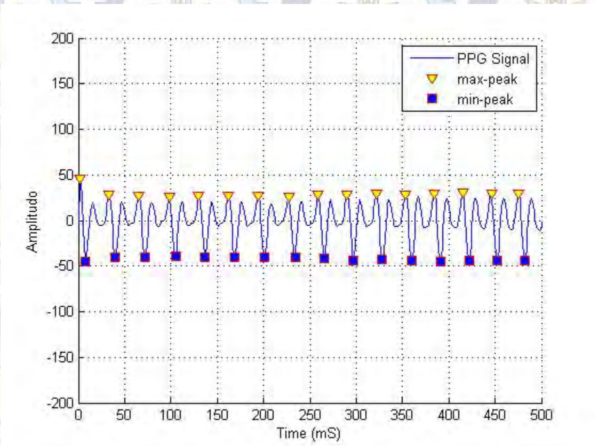


Fig. 5. Peak detection in 2-nd derivative PPG signal.

The distance between the beginning and the end of the PPG waveform labeled as Time Pulse Interval (tpi). Time between to diastolic peak labeled as Time between Middle Peak (tmp). Augmentation index (AI) is defined for a ratio calculated from the blood pressure waveform as the ratio of systolic peak to diastolic peak (diastolic peak/systolic peak). Takazawa et. al. [6] defined the augmentation index (AI) as the ratio of y to x as Equation 3. Rubins et. al. [7] used the reflection index as in Equation 4 and introduced an alternative augmentation index. The initial peak point for the first derivative and second derivative are a1 and a2 respectively. Then comes b1 and e1 points for the first derivative and b2 for the second derivative, following the position of systolic peak point. Corresponding times of each feature from both first derivative and second derivative signals are labeled as a1 time, b1 time, c1 time, a2 time, and b2 time, respectively.

A total of 30 characteristic features are calculated. Table I shows all 30 features defined for the system. In second column from Table I are shown the values of each labels from the signal of Figure 6.

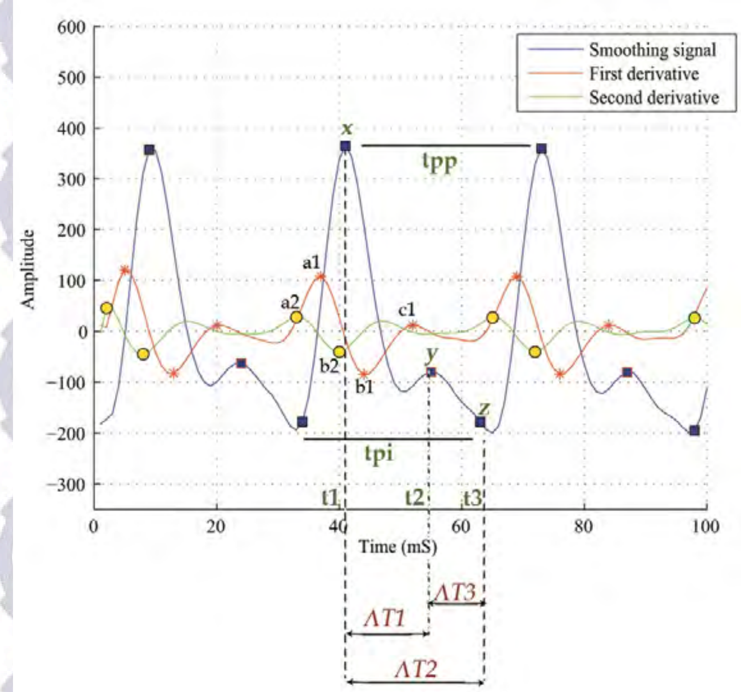


Fig. 6. The specified parameters used to derive the characteristic features from the PPG signal.

$$AI : \frac{y}{x} \tag{3}$$

$$Alternative\ AI : \frac{x-y}{x} \tag{4}$$

C. Classification Result

After processing each subject in the dataset, the extracted features are used as an input data in the classification stage. Four scenarios are used to test the accuracy of the system: training set, supplying test set, cross validation and percentage split. In the training set option, testing is performed by using the training data itself. In the cross-validation option, 10-fold is used. As for the percentage split option, from 90-period-signal, 66% is used for the training data and the rest is used for testing data. The data used for training and testing is chosen by the system itself. For the supplying test set option, from 90-period-signal taken, 70-period-signal of the initial data are used for training data and the remaining data are used for testing data. Furthermore, classification methods using Naïve Bayes, Multilayer Perceptron and Random Forest are proposed, respectively.

A feature ranking algorithm is proposed for the 30 features calculated during this study and the result is shown in Table II. The first 5, 10, 15, 20, 25 and 30 features from the ranked ones are selected and used as the classification input.

The percentages of classification success using Multilayer Perceptron are shown in Table III and are graphically shown

TABLE I. ALL 30 FEATURES DEFINED FOR THE SYSTEM

No.	Features	The values of each label
1	Minimum peak (z)	363.907
2	Minimum peak time (t1)	41
3	Distance between minimum peak	32
4	Systolic peak (x)	186.3494
5	Systolic peak time (t2)	63
6	peak to peak (tpp)	35
7	Diastolic peak (y)	-78.5053
8	Diastolic peak time (t3)	55
9	pulse interval (tpi)	32
10	x/z (augmentation index)	-0.21573
11	(x-z)/x (alternative augmentation index)	1.215729
12	∆T (time between systolic and diastolic peaks)	0.112666
13	∆T2 (time between systolic and minimum peaks)	2.80376
14	∆T3 (time between minimum and diastolic peaks)	1.28125
15	t2/x (systolic peak output curve)	1.96875
16	y/(tpi-t3) (diastolic peak downward curve)	1.71875
17	t1/tpp	14
18	t2/tpp	22
19	t3/tpp	8
20	∆T/tpp	0.4375
21	a1	108.1641
22	a1 time	37
23	b1	84.50754
24	b1 time	44
25	c1	12.24569
26	c1 time	52
27	a2	28.34428
28	a2 time	33
29	b2	40.01237
30	b2 time	40

TABLE II. ATTRIBUTE EVALUATOR (SUPERVISED, CLASS (NOMINAL): 30 CLASS): INFORMATION GAIN RANKING FILTER

Ranked attributes	% Accuracy	features
1	2.8615	feature 19
2	2.81223	feature 18
3	2.42353	feature 17
4	2.14412	feature 20
5	1.91653	feature 11
6	1.7881	feature 10
7	1.71755	feature 23
8	1.68597	feature 29
9	1.63417	feature 21
10	1.57064	feature 1
11	1.49101	feature 4
12	1.4482	feature 27
13	1.21659	feature 25
14	1.08117	feature 7
15	1.0099	feature 3
16	0.76155	feature 9
17	0.61778	feature 6
18	0.59481	feature 13
19	0.54094	feature 12
20	0.018798	feature 8
21	0.017113	feature 26
22	0.09972	feature 2
23	0.09625	feature 28
24	0.09597	feature 22
25	0.08143	feature 16
26	0.0807	feature 30
27	0.07054	feature 15
28	0	feature 14
29	0	feature 5
30	0	feature 24

in Figure **??**. The best classification success rate is achieved as 98.6928% for using the training set option and 94.6078% for supplying test set option, where all 30 features are used. As for the supplying test set option, it is seen that the classification success having the same accuracy even if there is feature selection and ranking process. The performance measurement values calculated from Multilayer Perceptron, such as True Positive, False Positive, Precision, Recall, F-measure and ROC Area, are shown in Table VI.

The percentages of classification success using Naïve Bayes are shown in Table IV and are graphically shown in Figure **??**. The best classification success rate is achieved as 92.7451 % for supplied test set option, where the first 15 features are used. Besides, it is seen that when there is no feature selection and ranking process, 92.1569% of classification success at most could be achieved for the classification algorithm in the event that all the features are used. In this case, a 0.58 % of increase in the classification success is attained through the feature ranking and selection process. The performance measurement values calculated from Nave Bayes, such as True Positive, False Positive, Precision, Recall, F-measure and ROC Area, are shown in Table VIII.

The percentages of classification success using Random Forest are shown in TableV and are graphically shown in Figure **??**. The best classification success rate is achieved as 99.9346% for use training set option of the classification algorithm, while 91.9164% is achieved for supplied test set option, where the first 25 features are used. Besides, it is seen when there is no feature selection and ranking process, 90.3922% of classification success at most could be achieved for the classification algorithm in the event that all the features are used. In this case, a 1.52% of increase in the classification success is attained through the feature ranking and selection process. The performance measurement values calculated from Random Forest, such as True Positive, False Positive, Precision, Recall, F-measure and ROC Area, are shown in Table VII.

Table VI, VII and VIII the results of the accuracy of the classification method implemented in where the 30 features used are shown. There are four options to classify the data: use training set, supplied test set, cross validation and percentage split. From those options, use training set is slightly higher than the rate for other options, in particular the accuracy is higher for the Multilayer Perceptron. The value indicating the accuracy of identification shows that Multilayer Perceptron method has a slightly higher accuracy rate supplied test set, but it takes the more time of computation.

True positive (TP) indicates that the type of signal from the PPG signal is appropriately identified according to the class. False positive (FP) is a type of signal from the PPG signal which should be identified correctly in class classification wrong turns in the process of identification. The percentage of correct classified instances is often called accuracy or sample accuracy. Kappa is a chance corrected measure of agreement between the classifications and the true classes. It is calculated by taking the agreement, expected by chance away from the observed agreement and dividing by the maximum possible agreement. A value greater than 0 means that the classifier is doing better than chance. The error rates are used for numeric prediction rather than classification. In numeric prediction, predictions are not just right or wrong, the error has a magnitude, and these measures reflect that.

Table IX the comparison of different classification methods to see how the performance varies across different architectures are shown. The performance results show accuracy (ACC), for the following architectures: Fuzzy, K-nn, Nave Bayes, Random Forest and Multilayer Perceptron. The first column of Table 4.8 the performance for fuzzy logic by Gu et al. and the second column the performance for k-nn by Kasgaovlu et al. are shown for seventeen subjects and thirty subjects, respectively. In our study, fifty one subjects joined for the experiments, this method can achieve good performance for more data samples. In this case, a 0.6% increase in the classification success is attained, compared with Fuzzy logic method and a 0.16% increase compared with K-nn method.

III. Conclusion

This study has tested the ability of PPG signals for biometric identification system. Based on the research that has been done can be concluded that:

- Designed system can identify the heartbeat of each individual.
- Feature extraction based on the three major peaks value of the photoplethysmography signal.
- The results were evaluated via the Multilayer Perceptron, Nave Bayes and Random Forest classifier models; the true positive rates are then 94.6078%, 92.1569% and 90.3922%, respectively. The obtained results showed that the proposed algorithm and the biometric identification model based on this developed PPG signal are very promising for contact less recognizing systems.

References

[1] Y. Gu, Y. Zhang, and Y. Zhang, "A novel biometric approach in human veri
cation by photoplethysmographic signals," in Information Technology Applications in Biomedicine, 2003. 4th International IEEE EMBS Special Topic Conference on, pp. 13-14, April 2003.

[2] P. Spachos, J. Gao, and D. Hatzinakos, "Feasibility study of photoplethysmographic signals for biometric identi
cation," in Digital Signal Processing (DSP), 2011 17th International Conference on, pp. 1-5, July 2011.

[3] Y. Gu and Y. Zhang, "Photoplethysmographic authentication through fuzzy logic," in Biomedical Engineering, 2003. IEEE EMBS Asian-Paci c Conference on, pp. 136-137, Oct 2003.

[4] A. R. Kavsaoglu, K. Polat, and M. R. Bozkurt, "A novel feature ranking algorithm for biometric recognition with fPPGg signals," Computers in Biology and Medicine, vol. 49, no. 0, pp. 1-14, 2014.

[5] M. Joel and G. Yury, "http://pulsesensor.com/."

[6] K. Takazawa, N. Tanaka, M. Fujita, O. Matsuoka, T. Saiki, M. Aikawa, S. Tamura, and C. Ibukiyama, "Assessment of vasoactive agents and vascular aging by the second derivative of photoplethysmogram waveform," in Hypertension, p. 32(2): 365370, August 1998.

[7] U. Rubins, A. Grabovskis, J. Grube, and I. Kukulis, "Photoplethysmography analysis of artery properties in patients with cardiovascular diseases," in 14th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (A. Katashev, Y. Dekhtyar, and J. Spigulis, eds.), vol. 20 of IFMBE Proceedings, pp. 319-322, Springer Berlin Heidelberg, 2008.

TABLE III. The selected feature numbers and the percentages of classification success using Multilayer Perceptron

Multilayer Perceptron						
	The selected feature numbers					
	5	10	15	20	25	30
Using training Set	62.70%	91.74%	95.36%	96.80%	97.80%	**98.69%**
Supplying Test Set	50.50%	86.38%	90.30%	91.64%	93.37%	**94.61%**
Cross Validation	59.15%	89.02%	91.70%	93.53%	**95.86%**	94.73%
Percentage Split	55.60%	85.97%	90.71%	92.50%	93.91%	**94.42%**

TABLE IV. The selected feature numbers and the percentages of classification success using Naïve Bayes

Naive Bayes						
	The selected feature numbers					
	5	10	15	20	25	30
Using training Set	66.65%	84.66%	90.76%	**90.89%**	90.65%	90.34%
Supplying Test Set	61.27%	87.06%	**92.74%**	92.46%	92.01%	92.16%
Cross Validation	64.01%	83.8%	**90%**	89.54%	89.54%	88.58%
Percentage Split	62.91%	82.32%	**88.98%**	88.08%	88.02%	87.70%

TABLE V. The selected feature numbers and the percentages of classification success using Random Forest

Random Forest						
	The selected feature numbers					
	5	10	15	20	25	30
Using training Set	98.89%	99.74%	99.91%	99.80%	**99.93%**	99.82%
Supplying Test Set	69.21%	85.49%	89.61%	91.73%	**91.91%**	90.39%
Cross Validation	82.18%	92.76%	94.42%	94.88%	**95.34%**	95.21%
Percentage Split	80.85%	91.42%	92.63%	93.52%	93.40%	**93.59%**

TABLE VI. The results of the testing process with the target output using Multilayer Perceptron with 30 features are used

Multilayer Perceptron				
	Use training Set	Supplied Test Set	Cross Validation	Percentage Split
Correctly Classified Instances	98.6928%	94.6078%	94.7277%	94.4266%
Incorrectly Classified Instances	1.3072%	5.3922%	5.2723%	5.5734%
Kappa statistic	0.9867	0.945	0.9462	0.9431
Mean absolute error	0.0015	0.0045	0.0036	0.0042
Root mean squared error	0.0225	0.0456	0.0414	0.0435
Relative absolute error	4.0163%	11.6916%	9.3319%	10.8539%
Root relative squared error	16.2607%	32.8716%	29.8865%	31.3847%
TP Rate	0.987	0.946	0.947	0.944
FP Rate	0	0.001	0.001	0.001
Precision	0.987	0.951	0.948	0.947
Recall	0.987	0.946	0.947	0.944
F-Measure	0.987	0.945	0.947	0.944
ROC Area	0.99	0.998	0.997	0.998
Calculation Time	754.75 s	954.31 s	330.42	376.27 s

TABLE VII. THE RESULTS OF THE TESTING PROCESS WITH THE TARGET OUTPUT USING NAÏVE BAYES WITH 30 FEATURES ARE USED

Naïve Bayes				
	Use training Set	Supplied Test Set	Cross Validation	Percentage Split
Correctly Classified Instances	90.3486%	92.1569%	88.5839%	87.7002%
Incorrectly Classified Instances	9.6514%	7.8431%	11.4161%	12.2998%
Kappa statistic	0.9.016	0.92	0.8836	0.8745
Mean absolute error	0.0038	0.0033	0.0046	0.005
Root mean squared error	0.0572	0.051	0.0626	0.0657
Relative absolute error	10.0124%	8.5873%	11.8542%	12.9764%
Root relative squared error	41.2403%	36.8115%	45.1387%	47.371%
TP Rate	0.903	0.922	0.886	0.877
FP Rate	0.002	0.002	0.002	0.002
Precision	0.911	0.931	0.894	0.893
Recall	0.903	0.922	0.886	0.877
F-Measure	0.905	0.923	0.887	0.88
ROC Area	0.99	0.997	0.99	0.995
Calculation Time	0.39 s	0.11 s	0.15 s	0.05 s

TABLE VIII. THE RESULTS OF THE TESTING PROCESS WITH THE TARGET OUTPUT USING RANDOM FOREST WITH 30 FEATURES ARE USED

Random Forest				
	Use training Set	Supplied Test Set	Cross Validation	Percentage Split
Correctly Classified Instances	99.8257%	90.3922%	95.3595%	93.5939%
Incorrectly Classified Instances	0.1743%	9.6078%	4.6405%	6.4061%
Kappa statistic	0.9982	0.902	0.9527	0.9346
Mean absolute error	0.0028	0.0125	0.0077	0.0086
Root mean squared error	0.0225	0.0665	0.0493	0.053
Relative absolute error	7.3489%	32.39%	20.1467%	22.4212%
Root relative squared error	16.2029%	47.9917%	35.5478%	38.2248%
TP Rate	0.998	0.904	0.954	0.936
FP Rate	0	0.002	0.001	0.001
Precision	0.998	0.911	0.954	0.937
Recall	0.998	0.904	0.954	0.936
F-Measure	0.998	0.902	0.953	0.935
ROC Area	1	0.993	0.99	0.996
Calculation Time	1.51 s	1.25 s	1.53 s	2.01 s

TABLE IX. COMPARISON WITH DIFFERENT CLASSIFICATION METHODS.

	Fuzzy	K-nn	Naive Bayes	Random Forest	MLP
ACC	94%	94.44%	92.15%	90.39%	94.6%

List of Figures

Biometric Personal Identification Based on Photoplethysmography Signal by Heart Rate

Supervisor : Prof. Shi-Jinn Horng
Dr. I Ketut Eddy Purnama,S.T., M.T
Prof. Ir. Mauridhi Hery Purnomo, M.Eng, Ph.D

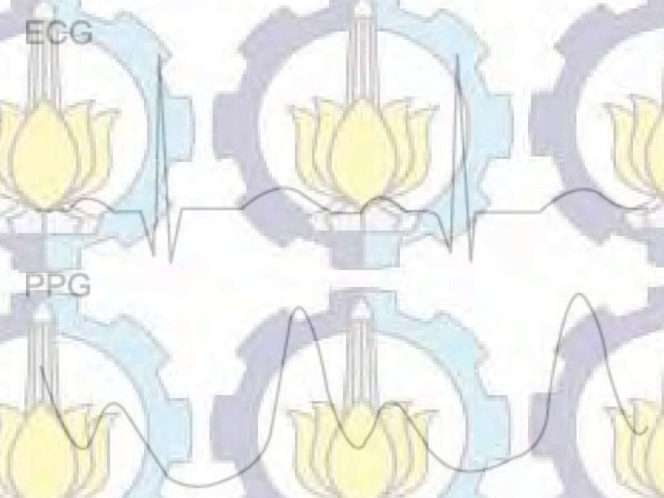
Presented by:
RAUDHATUL JANNAH
2213205025

Outline

- Introduction
- Related Works
- Methodology
  - Data acquisition
  - Signal Processing
  - Feature Extraction
  - Classification Result
- Conclusion

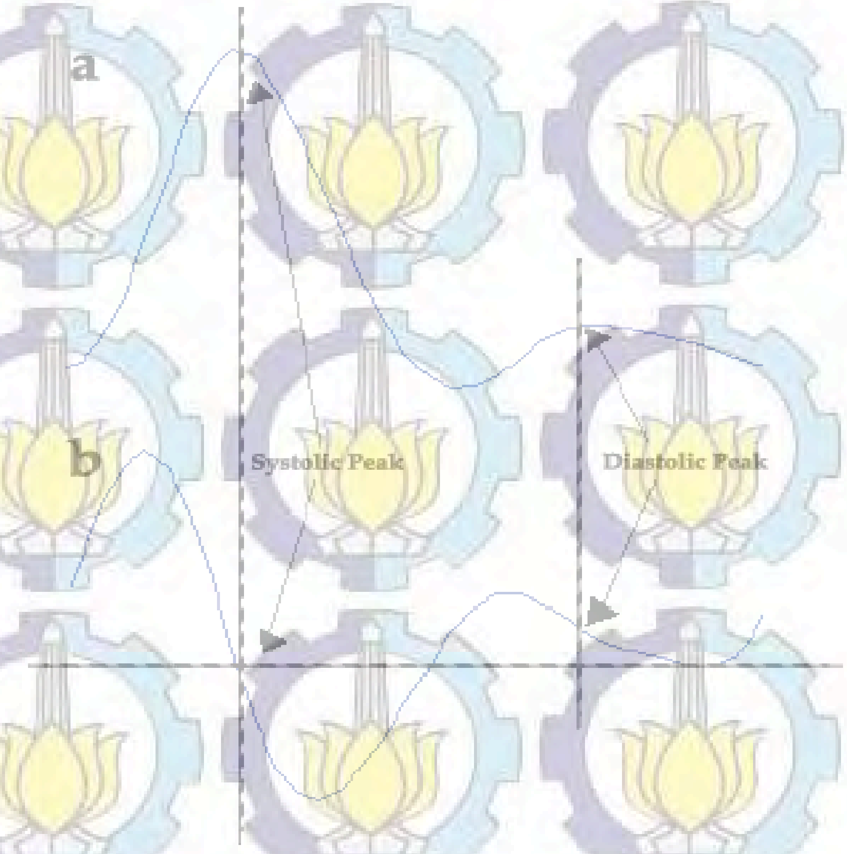
Introduction

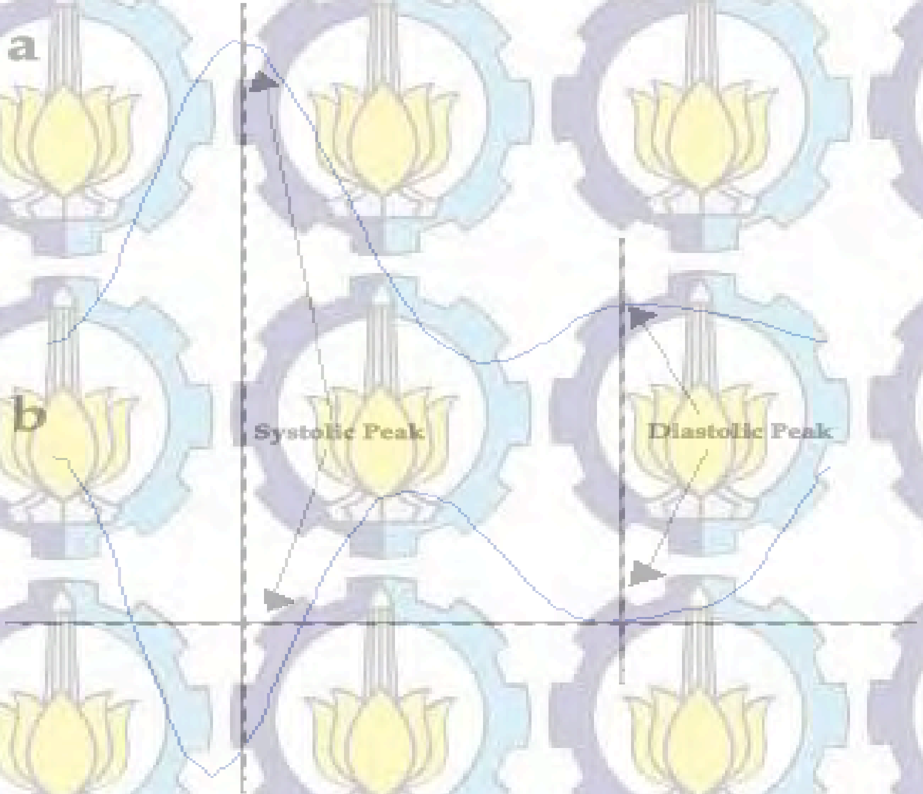
- Several studies have been developed for many biometric identification using fingerprint, face detection, iris, hand geometry, lip-movements, walking style, electroencephalography (EEG) and electrocardiography(ECG).
- In this study, the Photoplethysmography (PPG) signal was used for data input that is capable of implementing identification functionality.
- PPG is a non-invasive electro-optical method which gives information about the volume of blood owing through a testing zone of the body, close to the skin.



1^{st} *derivative:* $\frac{\partial f}{\partial x} = f(x+1) - f(x)$

2^{nd} *derivative:* $\frac{\partial^2 f}{\partial x^2} = f(x+1) + f(x-1) - 2f(x)$





The relationships between the original fingertip PPG, its first derivative and second derivative

Related Works

No.	Authors	Year	Proposed Works
1	Gu et. al.	2003	• 17 subjects • 4 features • 94% accuracy using Fuzzy-logic
2	Yao et. al.	2007	• 2 subjects • Proved the consistency of subjects within themselves and the distinguish ability among different subject
3	Spachos et. al.	2011	• 29 subjects • Nearest Neighbor and Majority Voting
4	Wei et. al.	2011	• Purposed : to find efficient pre-processing and feature extraction algorithms • Preprocessing : median filter and FIR • Feature Extraction : interpolation, differentiation and extreme point extraction
5	Kavsaoglu et. al.	2014	• 30 subjects • 40 features • 94.44% accuracy using K-nn

Methodology

The acquisition of PPG signal
- Using pulse sensor

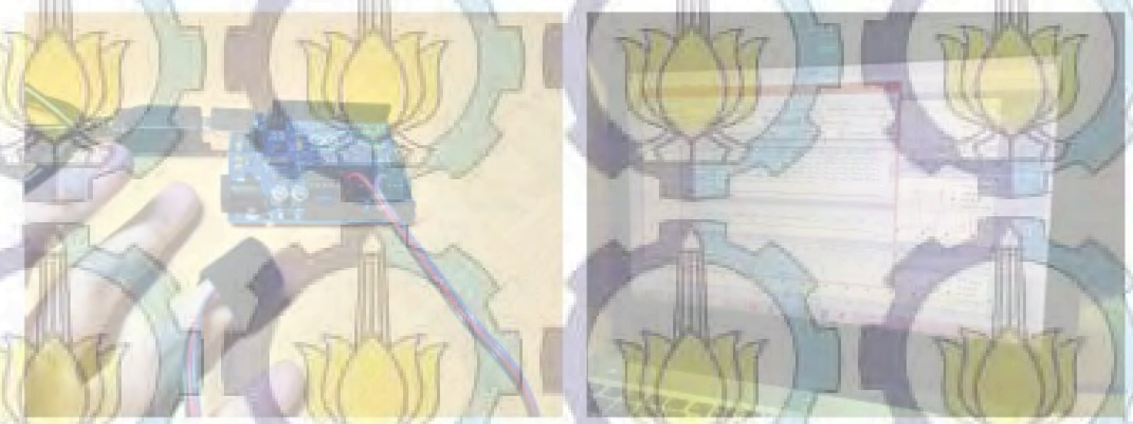
Pre-processing PPG signal
- Polynomial
- Savitzky-Golay filtering

Feature Extraction
- 30 characteristic features

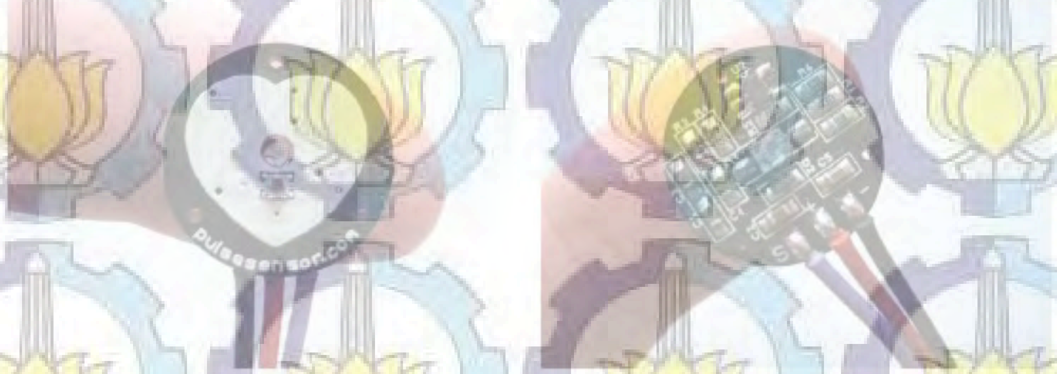
Classification
- Naive Bayes
- Multilayer Perceptron
- Random Forest

The block diagram of an identification system using Heartbeat PPG Signals

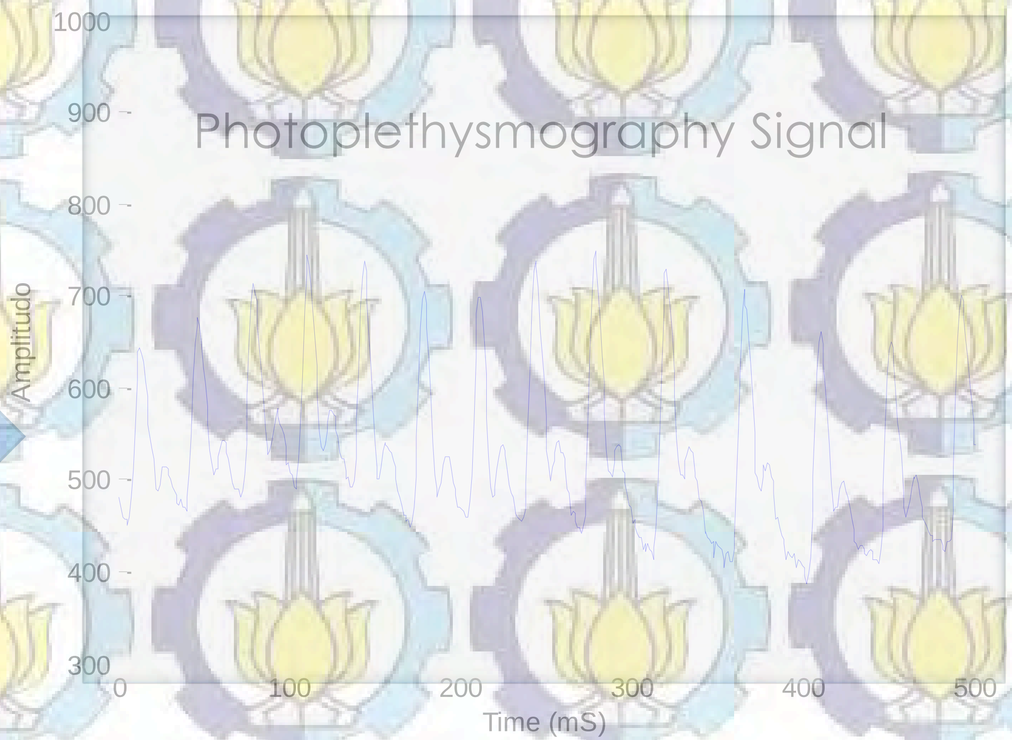
Data Acquisition



(a) Arduino and pulse sensor series (b) Data digital from sensor



(a) Front side (b) Back side



Descriptive statistics for subjects

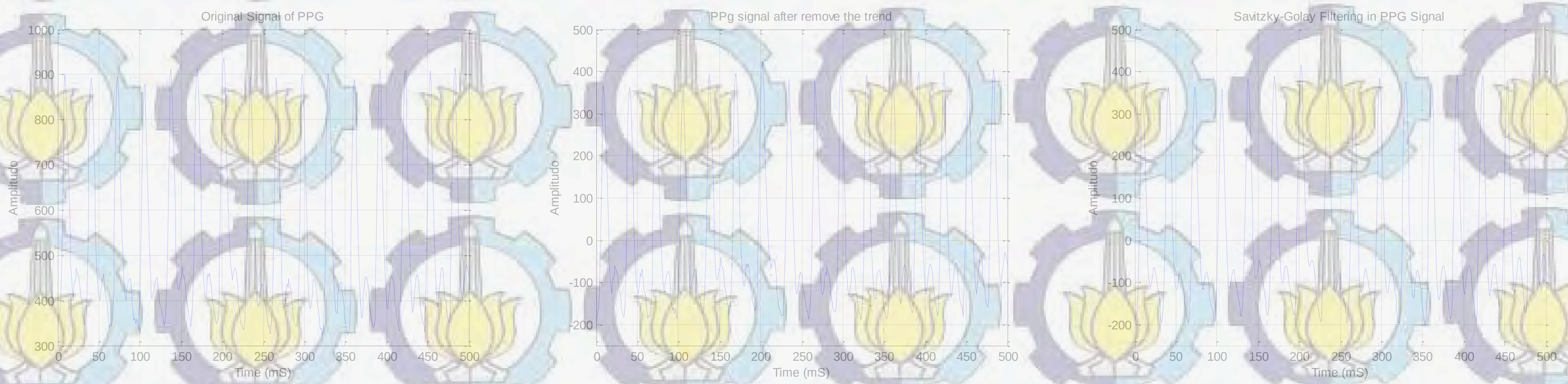
- In this study, PPG signals are acquired from a total of 51 healthy subjects.
- Twenty one of them are male and the remaining persons are female.

Groups	Age (years)	Height (cm)	Weight (kg)
Mean ± SD (range) for males	25±4.36 (18-41)	175±7.62 (157-185)	67.81±13.06 (48-95)
Mean ± SD (range) for females	24.46±2.44 (21-32)	157.87±3.75 (150-165)	55.37±9.59 (43-95)
Mean ± SD (range) for total	24.68±3.35 (18-41)	162.75±7.92 (150-185)	60.44±12.63 (43-95)

Block Diagram of findpeaks function

- The MinPeakDistance defines the function used to specify the minimum peak distance, or minimum separation between peaks as a positive integer.
- The MinPeakHeight function finds only those peaks that are greater than the value of MinPeakHeight.

Signal Preprocessing

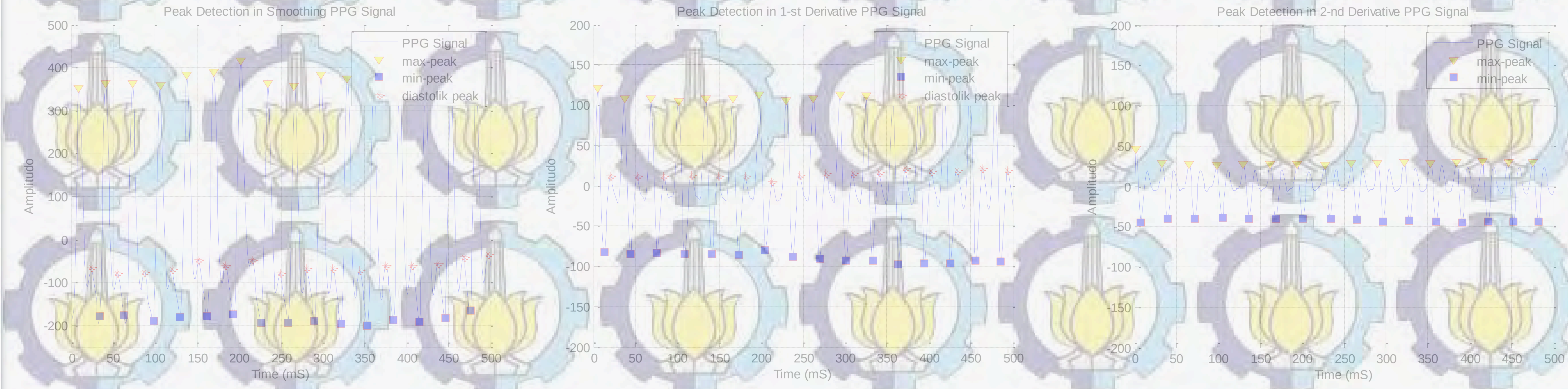


(a) Original signal of PPG

(b) PPG signal after removing the Trend

(c) Savitzky-Golay filtering in PPG signal

Feature Extraction

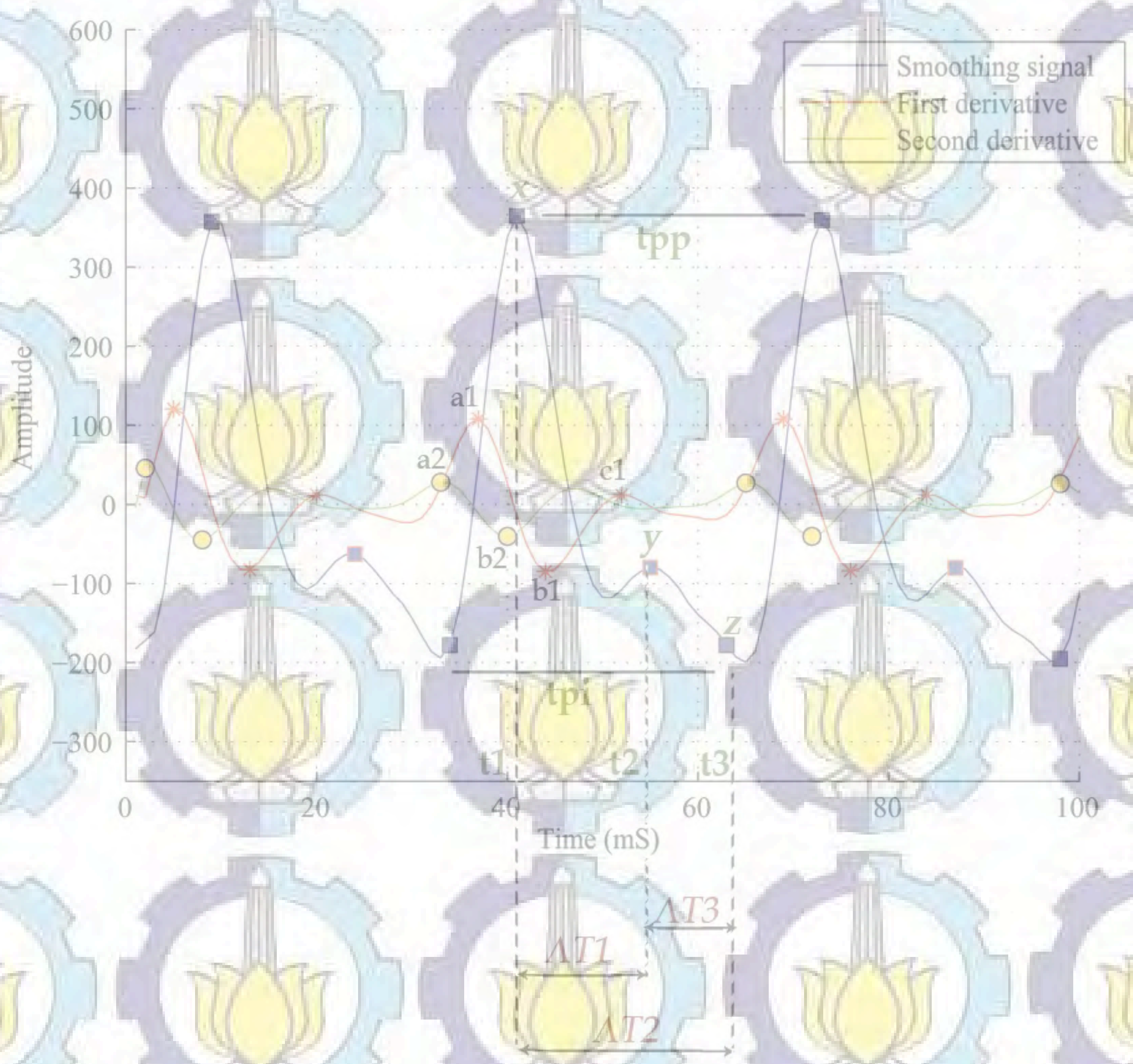


(a) Peak detection in smoothing PPG signal

(b) Peak detection in 1-st derivative PPG signal

(c) Peak detection in 2-nd derivative PPG signal

Feature Extraction



The specified parameters are used to derive the characteristic features from the PPG signal.

No.	Features	The values of each label
1	Systolic peak (x)	363.907
2	Systolic peak time (t1)	41
3	Time peak to peak (tpp)	32
4	Minimum peak (z)	186.3494
5	Minimum peak time (t3)	63
6	Time Pulse interval (tpi)	35
7	Diastolic peak (y)	-78.5053
8	Diastolic peak time (t2)	55
9	Time between diastolic peaks (tdp)	32
10	y/x (augmentation index)	-0.21573
11	(x-z)/x (alternative augmentation index)	1.215729
12	t1/x (systolic peak output curve)	0.112666
13	y/(tpi-t3) (diastolic peak downward curve)	2.80376
14	t1/tpp	1.28125
15	t2/tpp	1.96875

No.	Features	The values of each label
16	t3/tpp	1.71875
17	ΛT1 (time between diastolic and systolic peaks)	14
18	ΛT2 (time between minimum and systolic peaks)	22
19	ΛT3 (time between minimum and diastolic peaks)	8
20	ΛT1/tpp	0.4375
21	a1	108.1641
22	a1 time	37
23	b1	84.50754
24	b1 time	44
25	c1	12.24569
26	c1 time	52
27	a2	28.34428
28	a2 time	33
29	b2	40.01237
30	b2 time	40

Experimental Results

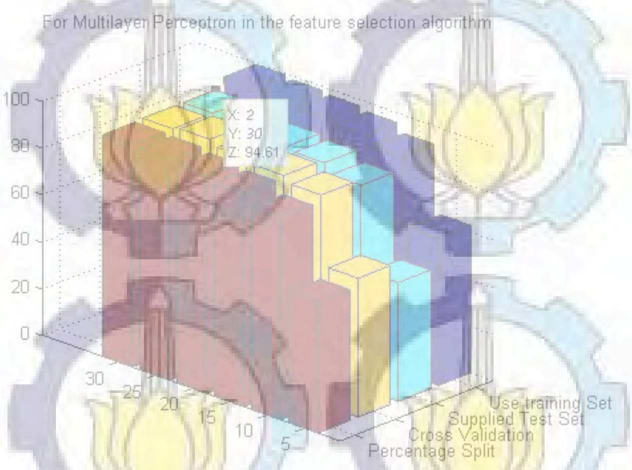
- After processing each subject in the dataset, the results of the process of feature extraction are used to input data in the classification stage.
- Classification methods using Naïve Bayes, Multilayer Perceptron and Random Forest are proposed, respectively.
- Four scenarios are used to test the accuracy of the system:
  - use training set : Testing is performed by using the training data itself
  - supply test set : From 90-period-signal taken, 70-period-signal are used for training data and the remaining data are used for testing data
  - cross validation : 10-fold is used
  - percentage split : Training data 66% and Testing data 34%
- A features ranking algorithm is proposed for the 30 features calculated.

Features Ranking Algorithm

Ranked attributes	% Accuracy	features	Ranked attributes	% Accuracy	features
1	2.8615	feature_19	16	0.76155	feature_9
2	2.81223	feature_18	17	0.61778	feature_6
3	2.42353	feature_17	18	0.59481	feature_13
4	2.14412	feature_20	19	0.54094	feature_12
5	1.91653	feature_11	20	0.18798	feature_8
6	1.7881	feature_10	21	0.17113	feature_26
7	1.71755	feature_23	22	0.09972	feature_2
8	1.68597	feature_29	23	0.09625	feature_28
9	1.63417	feature_21	24	0.09597	feature_22
10	1.57064	feature_1	25	0.08143	feature_16
11	1.49101	feature_4	26	0.0807	feature_30
12	1.4482	feature_27	27	0.07054	feature_15
13	1.21659	feature_25	28	0	feature_14
14	1.08117	feature_7	29	0	feature_5
15	1.0099	feature_3	30	0	feature_24

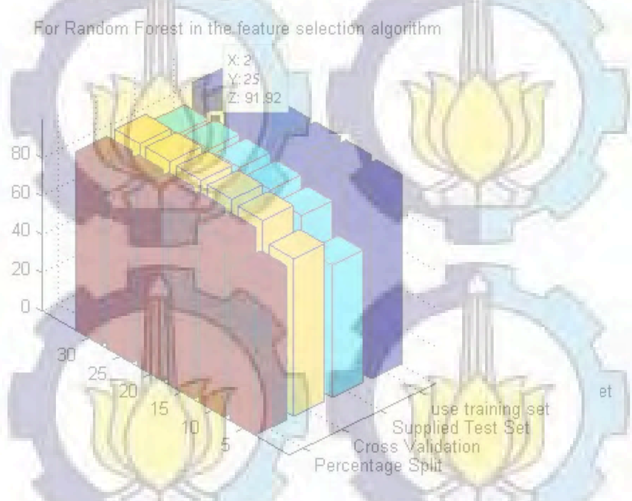
The selected feature numbers and the percentages of classification success using Multilayer Perceptron

Multilayer Perceptron						
	The selected feature numbers					
	5	10	15	20	25	30
Use training Set	62.7015 %	91.7429 %	95.3595 %	96.7974 %	97.7996 %	**98.6928 %**
Supplied Test Set	50.4995 %	86.376 %	90.2941 %	91.644 %	93.3697 %	**94.6078 %**
Cross Validation	59.1503 %	89.0196 %	91.6993 %	93.5294 %	**95.8606 %**	94.7277 %
Percentage Split	55.6054 %	85.9705 %	90.7111 %	92.5048 %	93.9142 %	**94.4266 %**



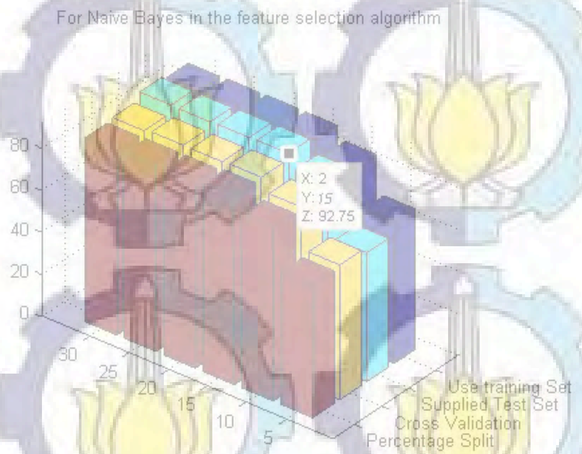
The selected feature numbers and the percentages of classification success using Random Forest

Random Forest						
	The selected feature numbers					
	5	10	15	20	25	30
Use training Set	98.8889 %	99.7386 %	99.9129 %	99.8039 %	99.9346 %	99.8257 %
Supplied Test Set	69.2157 %	85.4902 %	89.6078 %	91.7348 %	91.9164 %	90.3922 %
Cross Validation	82.1786 %	92.7669 %	94.4227 %	94.8802 %	95.3377 %	95.207 %
Percentage Split	80.8456 %	91.4158 %	92.6329 %	93.5298 %	93.4017 %	93.5939 %



The selected feature numbers and the percentages of classification success using Naïve Bayes

Naïve Bayes						
	The selected feature numbers					
	5	10	15	20	25	30
Use training Set	66.6449 %	84.6623 %	90.7625 %	**90.8932 %**	90.6536 %	90.3486 %
Supplied Test Set	61.2745 %	87.0588 %	**92.745 %**	92.4614 %	92.0073 %	92.1569 %
Cross Validation	64.0087 %	83.8126 %	**90 %**	89.5425 %	89.5425 %	88.5839 %
Percentage Split	62.9084 %	82.319 %	**88.9814 %**	88.0846 %	88.0205 %	87.7002 %



Multilayer Perceptron				
	Use training Set	Supplied Test Set	Cross Validation	Percentage Split
Correctly Classified Instances	98.6928%	94.6078%	94.7277%	94.4266%
Incorrectly Classified Instances	1.3072%	5.3922%	5.2723%	5.5734%
Kappa statistic	0.9867	0.945	0.9462	0.9431
Mean absolute error	0.0015	0.0045	0.0036	0.0042
Root mean squared error	0.0225	0.0456	0.0414	0.0435
Relative absolute error	4.0163%	11.6916%	9.3319%	10.8539%
Root relative squared error	16.2607%	32.8716%	29.8865%	31.3847%
TP Rate	0.987	0.946	0.947	0.944
FP Rate	0	0.001	0.001	0.001
Precision	0.987	0.951	0.948	0.947
Recall	0.987	0.946	0.947	0.944
F-Measure	0.987	0.945	0.947	0.944
ROC Area	0.99	0.998	0.997	0.998
Calculation Time	754.75 s	954.31 s	330.42	376.27 s

The results of the testing process with the target output using Multilayer Perceptron with 30 features are used

Naive Bayes				
	Use training Set	Supplied Test Set	Cross Validation	Percentage Split
Correctly Classified Instances	90.3486%	92.1569%	88.5839%	87.7002%
Incorrectly Classified Instances	9.6514%	7.8431%	11.4161%	12.2998%
Kappa statistic	0.9016	0.92	0.8836	0.8745
Mean absolute error	0.0038	0.0033	0.0046	0.005
Root mean squared error	0.0572	0.051	0.0626	0.0657
Relative absolute error	10.0124%	8.5873%	11.8542%	12.9764%
Root relative squared error	41.2403%	36.8115%	45.1387%	47.371%
TP Rate	0.903	0.922	0.886	0.877
FP Rate	0.002	0.002	0.002	0.002
Precision	0.911	0.931	0.894	0.893
Recall	0.903	0.922	0.886	0.877
F-Measure	0.905	0.923	0.887	0.88
ROC Area	0.99	0.997	0.99	0.995
Calculation Time	0.39 s	0.11 s	0.15 s	0.05 s

The results of the testing process with the target output using Naïve Bayes with 30 features are used

Random Forest				
	Use training Set	Supplied Test Set	Cross Validation	Percentage Split
Correctly Classified Instances	99.8257%	90.3922%	95.3595%	93.5939%
Incorrectly Classified Instances	0.1743%	9.6078%	4.6405%	6.4061%
Kappa statistic	0.9982	0.902	0.9527	0.9346
Mean absolute error	0.0028	0.0125	0.0077	0.0086
Root mean squared error	0.0225	0.0665	0.0493	0.053
Relative absolute error	7.3489%	32.39%	20.1467%	22.4212%
Root relative squared error	16.2929%	47.9917%	35.5478%	38.2248%
TP Rate	0.998	0.904	0.954	0.936
FP Rate	0	0.002	0.001	0.001
Precision	0.998	0.911	0.954	0.937
Recall	0.998	0.904	0.954	0.936
F-Measure	0.998	0.902	0.953	0.935
ROC Area	1	0.993	0.99	0.996
Calculation Time	1.51 s	1.25 s	1.53 s	2.01 s

The results of the testing process with the target output using Random Forest with 30 features are used

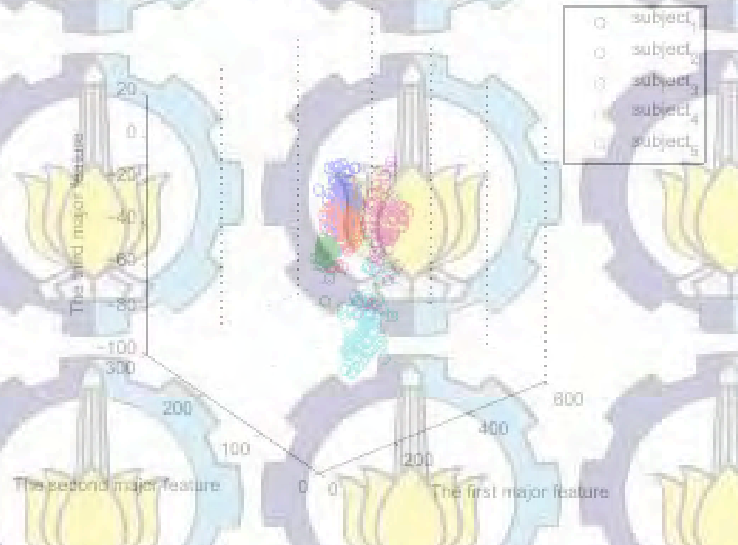
Comparison with other studies

	Fuzzy [5]	K-nn [8]	Naïve Bayes	Random Forest	Multilayer Perceptron
ACC	94%	94.44%	92.15%	90.39%	94.6%

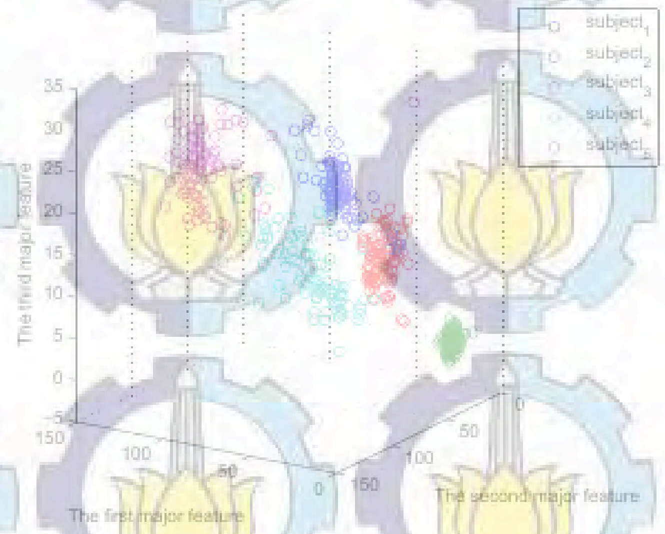
[5] Y. Gu and Y. Zhang, "Photoplethysmographic authentication through fuzzy logic, "in Biomedical Engineering, 2003. IEEE EMBS Asian-Pacific Conference on, pp. 136-137, Oct 2003.

[8] A. R. Kavsaoglu, K. Polat, and M. R. Bozkurt, "A novel feature ranking algorithm for biometric recognition with {PPG} signals, "Computers in Biology and Medicine, vol. 49, no. 0, pp. 1-14, 2014.

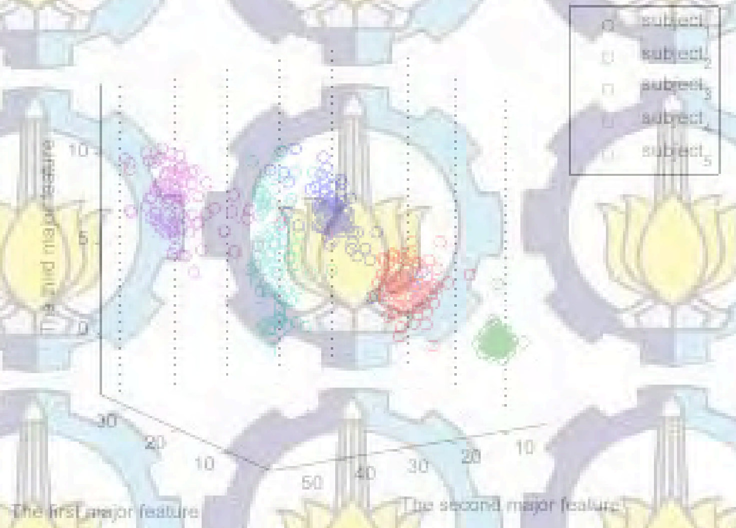
The class distributions



The class distribution for the smoothing signal



The class distribution for the 1st derivative signal



The class distribution for the 2nd derivative signal

Conclusion

This study has tested the ability of PPG signals for biometric identification system. Based on the research that has been done can be concluded that:

- The designed system can identify the heartbeat of each individual.
- Feature extraction based on the three major peaks value of the photoplethysmography signal.
- The results were evaluated via the Multilayer Perceptron, Naïve Bayes and Random Forest classifier models; the true positive rates are then 94.6078%, 92.1569% and 90.3922%, respectively. The obtained results showed that the proposed algorithm and the biometric identification model based on this developed PPG signal are very promising for contact less recognizing systems.

THANK YOU