Feature Selection of Photoplethysmograph Data in Machine Learning

Faris Atoil Haq
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
6025211011@mhs.its.ac.id

Riyanarto Sarno
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

Rifqi Abdillah
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
6025211012@gmail.com

Taufiq Choirul Amri
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
6025211009@mhs.its.ac.id

Abdullah Faqih Septiyanto
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
7025221022@mhs.its.ac.id

Kelly Rossa Sungkono
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
kelly@its.ac.id

2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC) | 978-1-6654-5645-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAIIC57133.2023.10067116

Abstract— Photoplethysmography signals are more responsive to changes in blood volume, not vascular pressure. Nowadays, more and more research is being developed for medical purposes, one of which is to diagnose diseases through fingertip pulse waves. This study proposes a new approach to optimize the statistical parameters of regression produced by PPG signals. The fingertip pulse wave device samples the PPG signal in humans and obtains the value of the signal. By taking the following samples, through processing using machine learning to process PPG signal data. machine learning is built to process PPG signal parameter data by the proposed method. The machine learning of feature selection algorithm that used are Forward Feature Selection Algorithm (FFS) and Sequential Input Selection Algorithm (SISAL). And for Machine learning using several methods is expected to obtain processing from the statistical parameters generated by Random Forest Regressor after going through feature selection by feature extraction with accuracy for FFS is 90% and for SISAL is 89%

Keywords— Feature Selection, PPG, Fast Forward Selection, Sequential Input Selection Algorithm

I. INTRODUCTION

Photoplethysmography (PPG) is a non-invasive technology with a waveform morphology similar to arterial blood pressure waveforms, therefore drawing importance on BP estimation. PPG is a method to measure the amount of light absorbed or reflected by blood vessels. Since the amount of optical absorption or reflection depends on the amount of blood in the optic pathway, the PPG signal is more responsive to changes in blood volume, not vascular pressure. In other words, PPG detects changes in blood volume by photoelectric techniques, both transmissive and reflective, to record blood volume in the sensor coverage area to form a PPG signal.

The photoplethysmography itself is a low-cost measurement technique that measures changes in blood volume in blood vessels by using optical method for monitoring vital signs of the body, such as heart rate, heart rate variability, and blood oxygenation [1].

PPG, Hertzman and Spelman described two significant PPG features; ascending edge (systolic phase), anthropic phase and descending edge (diastolic phase) [2].

The dicrotic phase is usually seen in the falling signal. A typical PPG signal and its characteristic point. The placement of the PPG sensor is an important factor during measurement. Meant. Generally, the sensor is placed either at the fingertips or earlobes or toe tips. Simultaneous acquisition of PPG from different sites. However, its measurement is entirely independent of the color of the skin, and the thickness and volume of the individual's blood. In recent times, PPG signals have been used widely to monitor clinical and physiological parameters such as blood oxygen saturation (SpO2), heart rate, heart rate variability, blood pressure, cardiac output and respiration; vascular peculiarities such as arterial disease, arterial adherence and ageing, endothelial function, venous assessment, vasospastic conditions and autonomic functions, such as vasomotor function and thermoregulation, orthostatic intolerance, neurology and other evaluations of cardiovascular variability.

PPG waveforms represent variations in blood volume and contain essential characteristics and are useful for analyzing cycle periods, baselines, and amplitudes. Fortunately, wearable health monitoring devices, including smartwatches and fitness trackers, can now capture PPG signals and enable monitoring of cardiac activity by lowering the R interval. The same R of PPG derived from an electrocardiogram (ECG). People have used a variety of non-invasive cardiovascular signals such as PPG, electrocardiograph (ECG), and phonocardiograph (PCG), using machine learning algorithms [3]. Several studies have been on PPG, one of which is on the approach to direct estimation of blood pressure. A simple but efficient method is to use a regression model to correlate PPG amplitude with blood pressure, which has proven to be a suitable method [4].

In several studies on PPG that have been applied, namely to diagnose diseases [5]. ECG and PPG are widely used to measure the heart cycle, and heart rate monitors are ECG and Photoplethysmography (PPG). Then the ECG signal provides an essential clinical physiological signal and is highly recommended for continuous monitoring and ensuring international standards for the safe practice of anesthesia. In other studies, PPG signals were also developed to detect a person's emotions [6]. The performance of photoplethysmography (PPG) signals through these devices has yet to be studied enough. In addition, the details of the phase space geometry of this signal have not been investigated in the emotional state to date. In connection with this problem, the Poincare section can measure the geometric pattern of trajectories in high-dimensional phase spaces. Therefore, this work attempts to propose an automatic emotional identifier that can characterize PPG signals during the exposure of emotional music videos. The PPG signal is obtained from the thumb. The data were recorded at a sampling frequency of 512 Hz. Then, the PPG signal was down sampling. Finally, each movement is smoothed out with an average filter.

This PPG data retrieval can be done in several ways and a variety of tools. One of them is by using a Fingertip pulse wave. Fingertip Wave Pulse that has been applied in several studies, such as in oximeters by utilizing pulsatile physiological signal [7]. It is often a noninvasive recording of blood-related physiological measurements used in health monitoring. The quality of these recordings is a major concern in healthcare, as many vital physiological measurements (e.g., respiratory rate, heart rate, and oxygen saturation) are extracted from these signals.

The processing of PPG is applied to several things such as what is done by monitoring heart rate through the average heart rate (beats per minute or bpm) [8] is one of the oldest but most widely used methods for the diagnosis of cardiovascular diseases (CVDs). Modalities commonly used for this purpose include Electrocardiography (ECG), Phonocardiography (PCG) and Photoplethysmography (PPG). ECG and PPG have one different peak during the heart cycle while PCG has two peaks referring to the two main sounds that are heard when the heart beats. Traditionally, ECG signals or PCG electrodes, collected in noise-free environments are common modalities for CVD diagnosis as they are the main source of death in the world. However, in recent years, heart rate monitors (HRM) that can be used in real time have gained prominent popularity. Generally, these monitors are based on ECG and PPG technologies.

Feature selection is algorithm that can select the best extracted feature

II. Research Method

A. *Photoplethysmography*

photoplethysmography (PPG) is a method used to determine the condition of the cardiovascular system by measuring changes in blood volume in skin tissues. In its application, this method uses optical sensors to capture electrical signals coming from passing or reflected light sources.

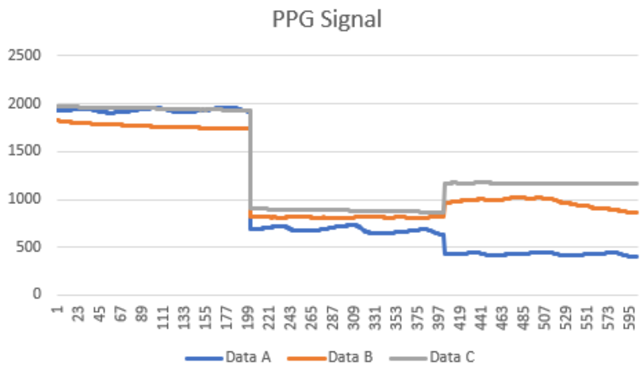

Fig 1. Forms of PPG signals across various light spectrum

Several ways of sampling data with the destination object are in the form of a photoplethysmography (PPG) whose signal form is as shown in Figure 2. 1, it can be done in different ways. The results obtained are quite good, namely the readings by the laser are close to the actual value. However, in this study, the use of lasers is quite expensive and complex in their use. On the other hand, PPG signal data collection can also be widely studied by various countries to be used as a diagnosis of blood sugar levels based on the blood volume detected [9]. In another study that examines the diagnosis of blood sugar levels, namely by utilizing data on light signals absorbed by fingers such as in an oximeter. Diagnosis is carried out by utilizing the results of light readings captured by photodiode sensors and used regression methods to obtain measurement results. However, the error value obtained is quite high, namely in the average error number.

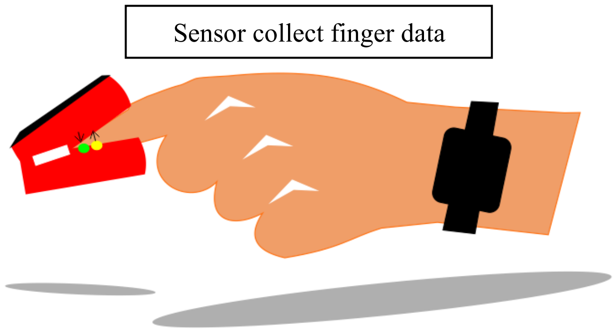

Fig 2. Forms of PPG signals in general

Figure 2. 2 is a s.il obtained Photoplethysmography which is formed and developed for various medical research needs, in addition to the fact that this PPG uses a non-invasive and low-cost method that uses optical sensors with two different wavelengths to detect changes in blood volume in micro vessels, usually from vessels. finger. The PPG waveform shows pulsatile waveforms resulting from the beat-to-beat heart cycle and slow frequency baseline, which vary due to the response of the sympathetic nervous system, respiration, and thermoregulation. In several studies that have been studied, namely in the detection of blood pressure. Both blood oxygenation and changes in blood volume can affect the PPG pulsatile waveform while the most dominant factor shown is a change in blood volume in peripheral blood vessels rather than blood oxygenation. As a result, it was found that the PPG signal would contain important information about blood pressure [10]. In another study, PPG signal processing was also carried out to monitor heart rate [11].

B. *Random Forest Regressor*

Random Forest Regressor (RFR) is an approach for data prediction using decision trees is called). By merging trees (trees) and training them on sample data, RFR prediction is accomplished [15]. More trees will improve the accuracy that is obtained. Decision-making with Random Forest is used in Fig. 2 and is based on the outcomes of each decision tree's voting. The outcomes of each tree's vote are used to make decisions.

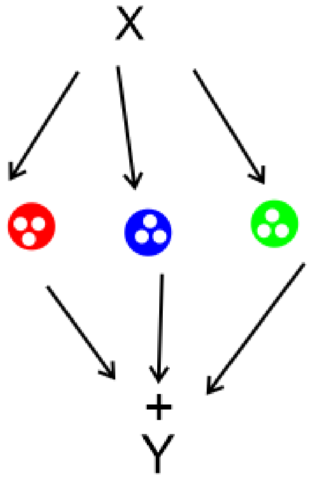

Fig 3. Random Forest tree

With many assessments, the random forest can be configured to create the right number of trees for the best accuracy. Same-class data distribution inevitably includes a tree. After the tree has been formed, the decision-making process is completed.

C. *Sequential Input Selection Algorithm (SISAL)*

Identifies relevant inputs using a backward selection type algorithm with optional branching. Choices are made by assessing linear models estimated with ordinary least squares or ridge regression in a cross-validation setting uses a backward selection technique with optional branching to find pertinent inputs. In a cross-validation setting, decisions are determined by evaluating linear models generated with ordinary least squares or ridge regression [16].

Sequential input selection algorithm that applies cross-validation on a linear model. One variable is eliminated at a time based on the regression coefficients, starting with the entire model. By selecting the model with the fewest variables from this group of models where the validation error is below a certain threshold, a sparse model is discovered. The formula can be expressed in equation (3) below.

$$\hat{y} = \sum_{i=1}^{L} \beta_i x_i + \varepsilon \tag{3}$$

With:
y = Output
B = Collect Variable
X = Sum Variable
E = Epsilon

D. *Forward Feature Selection Algorithm*

Forward Feature Selection Algorithms are often used for the purpose of selecting relevant features associated with output labels and minimizing computational time during process of classification. There are mainly two goals in the selection of features: the first is to minimize the cost of classification by reducing the number of features and thus, forming a more compact classification model; and the second is to improve classification performance [14].

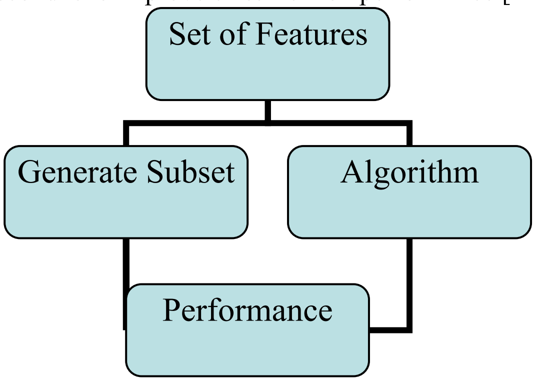

Figure 4. Feature Selection Chart

The chart form of the feature selection is shown in Figure 2. 10, where feature selection aims to identify a subset of the relevant features for the construction of statistical models (e.g., for classification or regression purposes), so that redundant and irrelevant features can be discarded. This reduces model complexity, improves computational efficiency, facilitates model interpretation, and avoids frequent performance degradation when redundant and irrelevant features are included in existing statistical tasks [15].

Here, this study uses several different feature methods including feature selection as follows: Forward Feature Selection Algorithm (FFS) and Sequential Input Selection (SIS).

E. *Mean Square Error (MSE)*

MSE are on one of the most common metrics used to calculate model prediction errors. Prediction error from a single row of data, as in equation (4).

$$PredictionError = ActualValue - PredictedValue \tag{4}$$

with:
PredictionError = Error value on prediction
ActualValue = Actual value
PredictionValuer = Predicted value

F. *Mean Absolute Error (MAE)*

MAE by taking the mean squared difference between the target and prediction values. This value is widely used for many regression problems and larger errors have a greater quadratic contribution to the mean error as in equation (5).

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{5}$$

With:
y' = Predicted Value
y= Actual Value
n= Total Data

The line plot created shows a straight line or linear increase in absolute error values because the difference between the expected and predicted increasing values is seen in Fig 4 below.

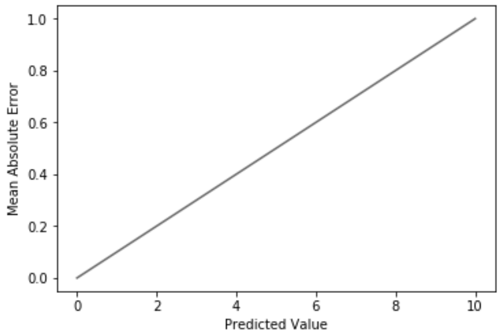

Fig 4. MAE plot on the chart

MAE itself is generally used for measuring error predictions in *time series* analysis, by showing the average error value that is the error from the actual value to the prediction value.

G. Root Mean Squared Error *(RMSE)*

Root Mean Squared Error (RMSE), is an extension of the average squared error. The sum of squared errors or the difference between the actual value and the predetermined predicted value. Where, the square root of the error is calculated, which means that the RMSE unit is equal to the original unit of the predicted target value as in equation (6)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}} \tag{6}$$

Predicted = Error value on prediction
Actual = Actual value
n = amount of data

III. Proposed Method

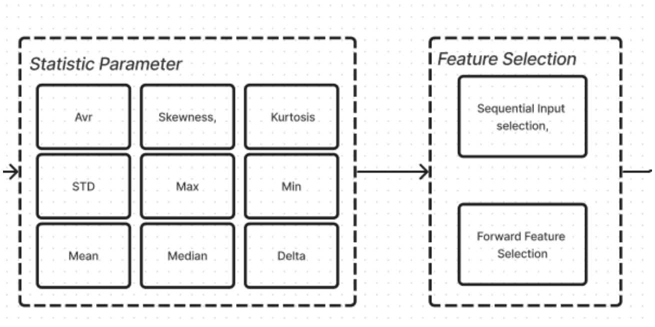

Fig 5 Feature Selection Process

From the proposed method, System testing resulting from statistical features is used several evaluation features, namely in the machine learning process that is proposed to be known for each accuracy, compared to several error checks such as MAE, MSE, RMSE to find out how much accuracy is generated from the system that has been carried out, and the results of each machine learning algorithm will be compared.

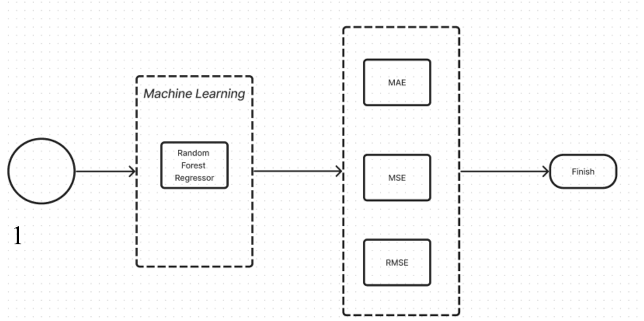

Fig 6. Proposed method

IV. Experimental Result

A. *Feature Extraction*

In this research, there are many features are Mean, Mean-Median, Median, Min, Max and Std are to be learn and by the evaluation model from the features is to be compared with another features is shown below

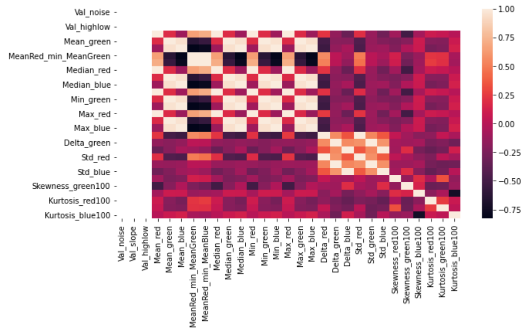

Fig 7. Correlation Heatmap of All Features

And for the correlation result show in table below

Table 1 Testing all Features

Feature	MAE	RMSE	MSE
Kurtosis_blue100	26,279	32,062	1027,982
Kurtosis_green100	26,621	32,054	1027,468
Kurtosis_red100	25,414	31,444	988,718
Skewness_blue100	26,477	32,147	1033,428
Skewness_red100	26,260	31,843	1014,004
Skewness_green101	25,516	31,398	985,829
std rgb	24,352	30,848	951,581
std mean rgb	24,517	30,636	938,549
std meandelta rgb	25,799	31,619	999,765
std median rgb	25,349	31,235	975,652
std min rgb	25,239	31,450	989,081
std max rgb	25,612	31,459	989,683
std delta rgb	25,103	31,690	1004,272
std skew rgb	25,743	32,350	1046,503
std kurt rgb	25,175	31,415	986,914
std mean meandelta rgb	24,975	30,847	951,529

std mean median rgb	25,317	31,147	970,164
std mean min rgb	25,000	31,019	962,186
std mean max rgb	25,294	31,410	986,591
std mean delta rgb	24,525	31,871	1015,754
std mean skew rgb	25,378	31,512	993,008
std mean kurt rgb	24,245	30,763	946,366
std mean kurt meandelta rgb	26,061	32,141	1033,030
std mean kurt median rgb	24,971	31,117	968,245
std mean kurt min rgb	26,834	32,579	1061,390
std mean kurt max rgb	24,483	31,043	963,655
std mean kurt delta rgb	24,675	30,756	945,950
std mean kurt skew rgb	24,578	31,154	970,542
std mean kurt max meandelta rgb	25,161	31,172	971,724
std mean kurt max median rgb	24,345	30,744	945,205
std mean kurt max min rgb	25,121	30,999	960,956
std mean kurt max delta rgb	24,031	29,826	889,565
std mean kurt max skwe rgb	24,630	31,445	988,774
std mean kurt max delta meandelta rgb	22,495	28,529	813,929
std mean kurt max delta median rgb	23,766	29,280	857,344
std mean kurt max delta min rgb	25,012	30,772	946,898
std mean kurt max delta skew rgb	25,012	30,772	946,898
std mean kurt max delta meandelta median rgb	23,861	29,789	887,370
std mean kurt max delta meandelta min rgb	23,983	29,850	891,039
std mean kurt max delta meandelta skew rgb	26,795	33,644	1131,923

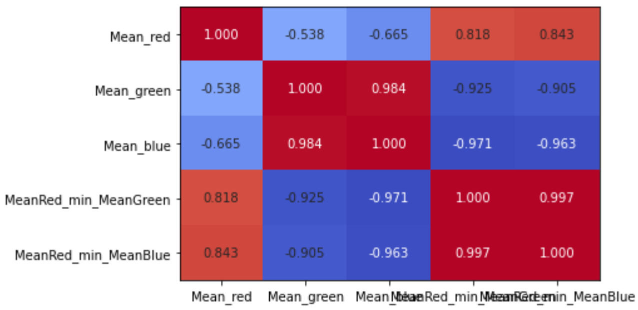

Fig 8. Correlation Heatmap of selected features.

Table 2 Testing 1

Feature	MAE	RMSE	MSE
Mean	27,745	34,146	1165,973
Mean-Mean	25,599	33,039	1091,554
Median	27,158	34,103	1162,983
Min	27,247	33,523	1123,77
Max	27,134	34,035	1158,385
Delta	28,313	34,335	1178,907
Std	26,444	34,093	1162,339
Skew	27,373	33,667	1133,462
Kurtosis	25,196	34,015	1157,021

Table 3 Testing 2

Feature	MAE	RMSE	MSE
Mean	25,487	32,174	1035,174
Mean-Mean	24,716	31,657	1002,185
Median	25,255	32,291	1042,736
Min	25,501	32,759	1073,182
Max	25,562	32,701	1069,326
Delta	24,742	31,073	965,532
Std	25,782	32,337	1045,708
Skew	24,391	30,493	929,833

Table 4 Testing 3

Feature	MAE	RMSE	MSE
Mean	25,727	31,36	983,438
Mean-Mean	24,949	31,371	984,14
Median	25,471	31,503	992,466
Min	24,938	31,319	980,864
Max	25,566	31,539	994,698
Delta	24,696	30,78	947,413
Std	25,782	31,903	1017,782

B. *Evaluation*

The performance evaluation used in this study to evaluate the best model for checking accuracy with machine learning is using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE). For the result of this testing is the accuracy of data get 90% for FFS and 89% for SISAL.

V. CONCLUSION

This research can be concluded that from the PPG signal to ensure that the chosen features are produced, effective feature selection and evaluation should be carried out. Extracted multiple features from feature selection can be processed with machine learning to produce more diversified features for use.

Additionally, the accuracy level created will depend on the outcomes of various accuracy values and how they compare with various features. By choosing different features, accuracy can be increased.

VI. REFERENCES

[1] Alishahi, M., Moghtadaiee, V., & Navidan, H. (2022). Add Noise to Remove Noise: Local Differential Privacy for Feature Selection. *Computers & Security*, *123*, 102934. https://doi.org/10.1016/j.cose.2022.102934

[2] Azar, J., Makhoul, A., Couturier, R., & Demerjian, J. (2021). Deep recurrent neural network-based autoencoder for photoplethysmogram artifacts filtering. *Computers and Electrical Engineering*, *92*(February), 107065. https://doi.org/10.1016/j.compeleceng.2021.107065

[3] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-January*, 1800–1807. https://doi.org/10.1109/CVPR.2017.195

[4] Dash, A., Jain, K., Ghosh, N., & Patra, A. (2022). Non-invasive detection of coronary artery disease from photoplethysmograph using lumped parameter modelling. *Biomedical Signal Processing and Control*, *77*(April), 103781. https://doi.org/10.1016/j.bspc.2022.103781

[5] Dhar, S., Mukhopadhyay, S. K., Pal, S., & Mitra, M. (2018). An efficient data compression and encryption technique for PPG signal. *Measurement: Journal of the International Measurement Confederation*, *116*(November 2017), 533–542. https://doi.org/10.1016/j.measurement.2017.11.006

[6] Goshvarpour, A., & Goshvarpour, A. (2018). Poincaré's section analysis for PPG-based automatic emotion recognition. *Chaos, Solitons and Fractals*, *114*, 400–407. https://doi.org/10.1016/j.chaos.2018.07.035

[7] Harsono, W., Sarno, R., & Sabilla, S. I. (2020). Recognition of original arabica civet coffee based on odor using electronic nose and machine learning. *Proceedings - 2020 International Seminar on Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, ISemantic 2020*, 333–339. https://doi.org/10.1109/iSemantic50169.2020.9234234

[8] Hasan, M. A., Sarno, R., & Sabilla, S. I. (2020). Optimizing machine learning parameters for classifying the sweetness of pineapple aroma using electronic nose. *International Journal of Intelligent Engineering and Systems*, *13*(5), 122–132. https://doi.org/10.22266/ijies2020.1031.12

[9] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[10] Ismail, S., Siddiqi, I., & Akram, U. (2022). Heart rate estimation in PPG signals using Convolutional-Recurrent Regressor. *Computers in Biology and Medicine*, *145*(March), 105470. https://doi.org/10.1016/j.compbiomed.2022.105470

[11] Kavsaoılu, A. R., Polat, K., & Hariharan, M. (2015). Non-invasive prediction of hemoglobin level using machine learning techniques with the PPG signal's characteristics features. *Applied Soft Computing Journal*, *37*, 983–991. https://doi.org/10.1016/j.asoc.2015.04.008

[12] Macedo, F., Valadas, R., Carrasquinha, E., Oliveira, M. R., & Pacheco, A. (2022). Feature selection using Decomposed Mutual Information Maximization. *Neurocomputing*, *513*, 215–232. https://doi.org/10.1016/j.neucom.2022.09.101

[13] Rossi, E., Aliani, C., Francia, P., Deodati, R., Calamai, I., Luchini, M., Spina, R., & Bocchi, L. (2022). COVID-19 detection using a model of photoplethysmography (PPG) signals. *Medical Engineering and Physics*, 103904. https://doi.org/10.1016/j.medengphy.2022.103904

[14] Roy Chowdhury, M., Madanu, R., Abbod, M. F., Fan, S. Z., & Shieh, J. S. (2021). Deep learning via ECG and PPG signals for prediction of depth of anesthesia. *Biomedical Signal Processing and Control*, *68*(May), 102663. https://doi.org/10.1016/j.bspc.2021.102663

[15] Sabilla, S. I., Sarno, R., & Siswantoro, J. (2017). Estimating Gas Concentration using Artificial Neural Network for Electronic Nose. *Procedia Computer Science*, *124*, 181–188. https://doi.org/10.1016/j.procs.2017.12.145

[16] Rodríguez-Rodríguez I, Rodríguez J-V, González-Vidal A, Zamora M-Á. Feature Selection for Blood Glucose Level Prediction in Type 1 Diabetes Mellitus by Using the Sequential Input Selection Algorithm (SISAL). *Symmetry*. 2019; 11(9):1164. https://doi.org/10.3390/sym11091164 5