Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 1018−1026
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2259

Study of Feature Selection Method to Detect Coronary Heart Disease (CHD) on Photoplethysmography (PPG) Signals

Faizal Akbari Putra[1], Satria Mandala[1,*], Miftah Pramudyo[2]

[1] Informatics, Informatics, Telkom University, Bandung, Indonesia
[2] Medicine, Cardiology and Vascular Department, Universitas Padjadjaran, Bandung, Indonesia
Email: [1]faizalakbari@student.telkomuniversity.ac.id, [2,*]satriamandala@telkomuniversity.ac.id, [3]miftah.pramudyo@gmail.com
Correspondence Author Email: satriamandala@telkomuniversity.ac.id
Submitted:**10/09/2022**; Accepted:**24/09/2022**; Published: **30/09/2022**

Abstract−Coronary Heart Disease (CHD) is a condition in which the heart's blood supply is blocked or disrupted by fat in the coronary arteries. This disease is the most significant cause of death in Indonesia. CHD can be detected based on the Heart Rate Variability (HRV) index of the Photophletysmograph (PPG) signal taken from a smartphone's camera. However, the use of PPG from smartphone to detect CHD is still rare in real-world applications. Moreover, studies on CHD detection based on PPG signal are also difficult to be found in the scientific literature, especially those discussing feature selection on PPG signals in detecting CHD. Currently, the Electrocardiogram (ECG) signal still dominates as a signal for detecting CHD. This research fills this research gap by proposing a study on the feature selection of PPG signal taken from smartphone's camera to detect CHD. The detection performance value obtained is highly dependent on the features used. Using irrelevant features can reduce detection performance in the classification model. Therefore, feature selection methods are essential to select optimal features to improve detection performance. There are three feature selection methods studied in this research, i.e., Analysis of Variance (Anova), Pearson Correlation, and Recursive Feature Elimination (RFE). Furthermore, a classification algorithm, called as K-Nearest Neighbors, has also been chosen to create a machine learning model based on the PPG features. The experimental results show that the Pearson Correlation feature selection method produces better CHD detection performance compared to the other two methods (Anova and RFE). CHD detection performance using the Pearson Correlation produces an accuracy of 90.9%, sensitivity of 75%, and specificity of 100%.

Keywords: Coronary Heart Disease, Heart Rate Variability, Photophletysmograph, Electrocardiogram, Feature Selection

1. INTRODUCTION

According to the World Health Organization, in 2019, it was estimated that there were 17.9 million deaths due to heart disease [1]. Coronary Heart Disease (CHD) is Indonesia's most significant cause of death. In 2016, the death rate from this disease reached 122 people per 100,000 population. Furthermore, CHD is a leading cause of mortality in the UK and across the world [2].

CHD is a condition in which the blood supply to the heart is obstructed or disrupted by fat in the coronary arteries [2]. Symptoms of this disease can be known if a person experiences chest pain, pain throughout the body, hard to breath, feeling weak and sick, and nausea. This disease can be caused by an unhealthy lifestyle, age, smoking, diabetes, obesity, or a family history of CHD. Heart Rate Variability (HRV) can be used as a valuable index to predict the occurrence of CHD and other heart diseases by accessing and analyzing the response of the Autonomic Nervous System (ANS) [3]. HRV is a computation of the time difference between each heartbeat. This variation is controlled by the ANS, a primitive part of the nervous system. It operates by automatically controlling heart rate, blood pressure, respiration, and digestion, among other important functions. [4].

Photophletysmograph (PPG) is a signal that can be used to determine the condition of the heart in a certain period. PPG is rapidly being used not only through medical equipment but also in wearable devices and smartphones to monitor oxygen saturation, cardiovascular disease symptoms, and HRV [5]. Currently, the HRV index can be known through the PPG signal by utilizing the camera from a smartphone so that CHD detection can be carried out. PPG uses a light source and a photodetector at the skin's surface to measure the volumetric variations of blood circulation [6]. The utilization of PPG signals to detect CHD is still little implemented. Electrocardiogram (ECG) signals are more often used to detect CHD [5]. Research [7] [8] [9] proposed a study to detect CHD using ECG signals. Therefore, developing a PPG signal for CHD detection is necessary as an alternative to the ECG signal.

In general, the detection of CHD using machine learning has four stages, namely denoising, feature extraction, feature selection, and classification. The denoising stage is the initial stage to remove noise in the signal but still retain important information in the signal. The feature extraction stage is the stage of grouping and distinguishing features based on the information on the signal. The feature extraction results are then evaluated to select the most influential features in the feature selection stage. The selected features are then used as input variables for CHD detection at the classification stage. The feature selection stage is essential because it can improve the detection performance of the built classification model. Moreover, using irrelevant features can reduce detection performance in the classification model. This can be proven in research [10] which tried to detect cardiovascular disease (CVD) using the Anova feature selection method with SVC classification. The datasets used are CVD and Framingham datasets. The highest accuracy achieved using the full feature set is 0.73 for the CVD dataset and 0.66 for the Framingham dataset. However, after using the features from the feature selection method, the accuracy increased to 0.75 and 0.71 for each dataset.

Research [7] conducted a study on feature selection methods and proposed a feature selection method, namely an ensemble algorithm based on multiple feature selection (EA-MFS). To find the best method for CHD detection,

Copyright © 2022 **Faizal Akbari Putra**, Page 1018
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 1018−1026
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2259

they compared feature selection methods. These methods are analysis of variance (Anova), mutual information (MutualInfo) and chi-square test (CHI). Classification is performed using a variety of classification algorithms using CHD datasets from demographics, symptoms and examination, ECG, and laboratory and echo features. As a result, the proposed method outperforms other methods in terms of classification performance and robustness on the CHD dataset.

Research [8] detect CHD using a feature selection Fisher-based Hybrid Extreme Learning Machine. The dataset used is from Z-Alizadeh sani, containing demographics, symptoms and examinations, ECG, laboratory and echo. Classification is done with selected features from the results of the Fisher method, which is the best ten features out of 15 features. The result achieved is accuracy of 97.6%.

Research [9] conducted a CHD prediction experiment using the Supervised algorithm with feature selection using an analytical approach. The data used comes from the FHS dataset in which there is a Heart Rate that comes from the ECG signal. The classification algorithm used is Random Forest, Decision Tree, and KNN. Classification using Random Forest has the highest accuracy of 96.80%.

Research [11] tried to detect Atrial Fibrillation (AF) using PPG signals with SVM classification with the Sequential Forward Selection (SFS) feature selection algorithm to find the best features heuristically. The experiment yielded an accuracy value of 95.7%, sensitivity of 94.2%, and specificity of 96.2%.

Research [12] identified fundamental features of HRV data which can be used for biometric recognition. HRV data is extracted from the PPG signal obtained from the IR-based pulse detection sensor. A Genetic Algorithm (GA) was applied and used an adaptive search technique. The experiment resulted in 15 dominant features out of 101 features.

This research studied feature selection methods to detect CHD based on PPG signal. Three methods have been considered to this study, i.e., Analysis of Variance (Anova), Pearson Correlation, and Recursive Feature Elimination (RFE). The three methods will be compared to determine the best method in this study to detect CHD based on PPG signals. In this study, a comparison of the results of the accuracy, sensitivity, and specificity performance of the three proposed feature selection methods will be carried out.

2. RESEARCH METHODOLOGY

2.1 Research Stages

The stages of this research can be described through the flowchart in Figure 1.

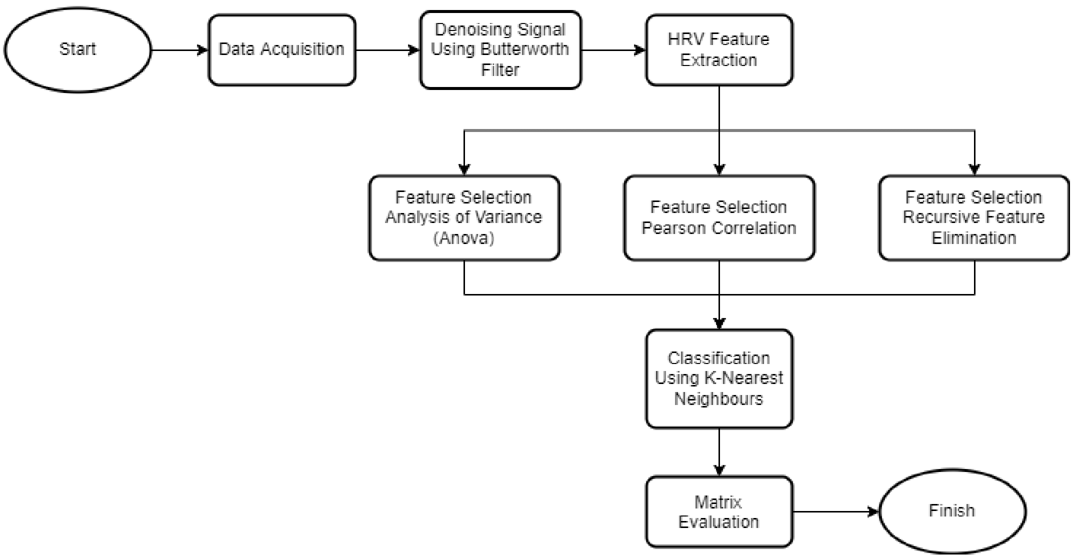

Figure 1. Research Stages Flowchart

It can be seen in Figure 1 that the initial stage of this research is Data Acquisition, which is collecting PPG signal datasets from subjects who have CHD taken from Dr. RSAU. M. SALAMUN and non-CHD subjects taken from the surrounding environment. Total data included 28 CHD patients and 30 healthy subjects. The next stage is to remove the signal using the Butterworth filter method to remove noise in the signal. Then feature extraction is performed to extract HRV-related features. The features that have been successfully extracted are then analyzed using feature selection methods to select the most optimal features for the classification model. This study has considered three methods: Analysis of Variance (Anova), Pearson Correlation, and Recursive Feature Elimination (RFE). The next stage is to create a classification model using the K-Nearest Neighbors classification algorithm to detect CHD based on the selected features. The last stage is evaluating the performance of CHD detection with Matrix based on accuracy, sensitivity, and specificity.

Copyright © 2022 **Faizal Akbari Putra**, Page 1019
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 1018−1026
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2259

2.2 Data Acquisition

Data acquisition is carried out using an application designed on an Android-based smartphone that are connected to the MQTT broker. To secure data, the broker is installed with the IDS as in paper [13] [14]. The application takes PPG signal data from patients by placing the index finger on the smartphone's back camera and covering the flashlight on the camera so that the application display will show a red color in the small box. This utility works by measuring the amount of red it sees through the camera and trying to detect fluctuations. It works better in a dark room and when the pressure on the camera is constant. The PPG signal data acquisition process can be seen in Figure 2.

Figure 2. Signal PPG Acquisition

The result is a fluctuation blood or air contained in the blood [15] that can be seen on the application display, as shown in the Figure 2. The PPG signal data can then be retrieved through the MySQL database connected to the application for further processing.

This research was conducted using data consisting of two groups. The first group is CHD patients whose data were taken at Dr. RSAU. M. SALAMUN. The second group is Healthy subjects from the surrounding community who were not diagnosed with CHD. Total data included 28 CHD patients and 30 healthy subjects.

2.3 Denoising Signal Using Butterworth Filter

The signal that has been successfully obtained still has interference in the form of noise in the signal because the retrieval is done using a camera on a smartphone. Due to these noises, the signal does not accurately reflect the heart's true condition [16]. To overcome this, a denoising technique is needed to remove noise and retain critical information contained in the signal. The denoising method used in this research is Butterworth Filter. The Butterworth filter attempts to maintain the frequency response of its pass-band as flat as possible [17]. Butterworth Low-Pass eliminates high frequencies' noise, while Butterworth High-Pass eliminates low frequencies' noise. [18].

2.4 HRV Feature Extraction

The dataset that has been successfully denoised is then carried out by taking features from the signal to classify and distinguish features depending on the signal's qualities and information. The feature extraction method used in this study is the HRV Feature extraction. HRV-related features are calculated from sequential peak-to-peak distances (NN intervals) in a signal [19]. Features that can be extracted are Shannon Entropy (ESH), Mean Absolute Deviation (MAD), Skewness (Skew), Kurtosis (Kurt), Very Low-Frequency (VLF), Low-Frequency (LF), and High-Frequency (HF).

2.5 Feature Selection

At this stage, the features that have been successfully extracted from the PPG signal dataset will be evaluated to select which features to use. The aim is to determine which features are the most significant to the classification process. Feature selection refers to obtaining a subset of the original feature set according to specific selection criteria and selecting the relevant dataset features [20]. Performing feature selection has the advantage of getting the potential for higher accuracy, less computation, and finding dominant features for the detection algorithm. The experiment in this study used the Analysis of Variance (Anova), Pearson Correlation, and Recursive Feature Elimination (RFE) feature selection methods.

a. Analysis of Variance (Anova)

Analysis of Variance (Anova) was used to compare the mean scores of the dataset and visualize whether the mean scores of different groups (classes) are significantly different. The statistics for Anova are called F-statistics [21]. This method can be used to perform feature selection by considering the distribution of the dataset and then

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 1018−1026
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2259

selecting the feature that is most related to the classification target. In this study, the target of the classification is whether the subject of the data is a person have CHD or not. In classification problems, Anova can be used to determine how well features influence on several classes. Anova can be used for classification problems that use continuous features. If the features used are categorical, the calculation of the mean scores and variance cannot be carried out, so in this case, Anova is unsuitable for feature selection.

b. Pearson Correlation
The Pearson correlation method is the most commonly used method for numerical variables. This method returns a value between -1 and 1. Value 0 means no correlation, 1 means total positive correlation, -1 means total negative correlation [22]. Each feature in the dataset is calculated correlation coefficient to determine how closely the relationship between each feature is.

c. Recursive Feature Elimination (RFE)
Recursive feature elimination (RFE) is a feature selection method that reduces unrelated or weak features and is carried out continuously until the best features are found to be used in building the model [23]. RFE repeatedly searches for a feature subset of all features in the dataset and removes the least significant features until the desired number of features is reached. Features will be ranked based on the level of significance of the feature against the classification target. RFE can use machine learning algorithms as the core model to recursively calculate feature importance.

2.6 Classification Using K-Nearest Neighbors

The K-Nearest Neighbors (KNN) classification algorithm is a supervised machine learning algorithm commonly used for classification. This algorithm classifies data points based on how their neighbors are classified. Each adjacent object is considered to have the same characteristics [24]. The euclidean formula is used to define the distance between two points, namely the point on the training data (x) and the point on the testing data (y) [25]. The formula of Euclidean can be defined as follows:

$$D(x,y) = \sqrt{\sum_{k-1}^{n}(x_k - y_k)^2} \quad (1)$$

2.7 Matrix Evaluation

The confusion matrix [26] is used to measure the performance of the classification models used in this study. Accuracy, sensitivity, and specificity will be used to test the performance of the model. Accuracy measures the number of correct predictions out of the total number of predictions made by the model, specificity measures the number of correct negative predictions out of the total number of true negative samples, sensitivity measures the number of correct positive predictions out of the total number of true positive samples [27]. The formula of accuracy, specificity, and sensitivity can be defined as follows:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + False\ Negative + True\ Negative}\ x\ 100\% \quad (2)$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}\ x\ 100\% \quad (3)$$

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}\ x\ 100\% \quad (4)$$

True Positive is the value of the number of correct predictions of subjects with CHD, and True Negative is the value of the number of correct predictions of subjects who do not have CHD. Meanwhile, False Positive is the value of the number of wrong predictions from subjects with CHD, and False Negative is the value of the number of wrong predictions from subjects who do not have CHD.

2.8 Scenario

Four kinds of test scenarios are carried out in this study to determine the differences in the metric evaluation of the feature selection method used. The aim is to find out which feature selection method produces the best metric evaluation of this study. The following are several scenarios that will be tested:

a. Scenario I : K-NN classification using features extracted from HRV without using feature selection method.
b. Scenario II : K-NN classification using features from the Anova feature selection method.
c. Scenario III : K-NN classification using features from the Pearson Correlation feature selection method.
d. Scenario IV : K-NN classification using features from the RFE feature selection method.

3. RESULTS AND DISCUSSION

Copyright © 2022 **Faizal Akbari Putra**, Page 1021
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 1018−1026
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2259

3.1 Implementation

3.1.1 Signal Denoising

Figure 3 shows one sample of PPG signal data before denoised, and Figure 4 shows PPG signal data after denoised using the Butterworth Filter method.

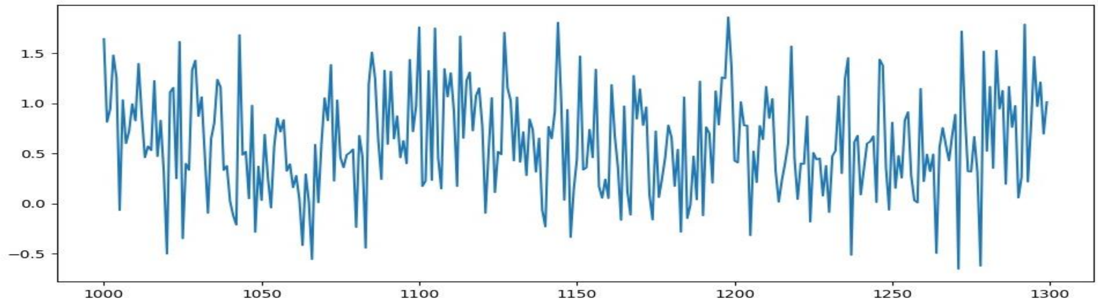

Figure 3. Sample of Signal PPG

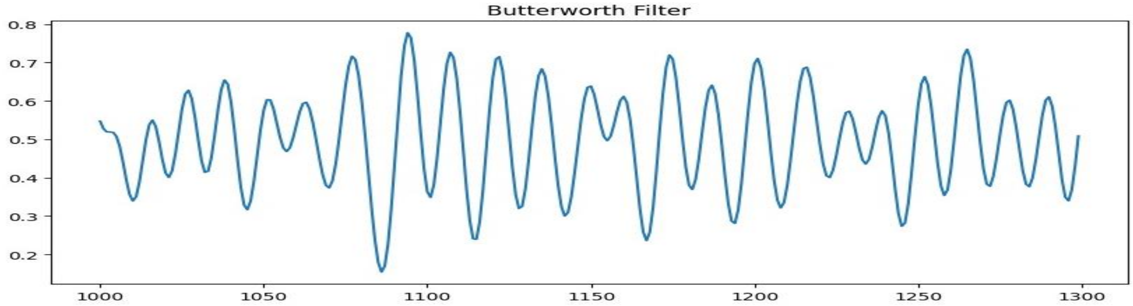

Figure 4. Sample of Signal PPG Denoised

It can be seen in Figure 3 that the PPG signal plotted using Python program using the *matplotlib* library shows a coarser signal than the PPG signal resulting from the denoising process shown in Figure 4. The denoising signal produces a smoother signal and maintains a signal shape that is not too different from the original signal PPG.

3.1.2 Feature Extraction

The power spectrum of the HRV signal is divided into three bands: very low frequency (VLF), low frequency (LF), and high frequency (HF) [28]. These frequency bands are used as features. Other features are shannon entropy (ESH), mean absolute deviation (MAD), kurtosis (Kurt), and skewness (Skew) from NN intervals. These features are obtained from Python program using the Statistical functions module (*scipy.stats*). The following are the results of feature extraction using the HRV Features from the denoised PPG signal dataset. Table 1 is a sample of healthy subject, and Table 2 is a sample of CHD patient.

Table 1. HRV Features on Healthy Subject

Feature	Value
ESH	12.667555
MAD	0.079907
Skew	0.261786
Kurt	-0.725236
VLF	0.000525
LF	0.000101
HF	1.304105e-06

Table 2. HRV Features on CHD Patient

Feature	Value
ESH	12.685187
MAD	0.087452
Skew	-0.248031
Kurt	-0.418951
VLF	0.002040
LF	0.000067
HF	7.951989e-07

Copyright © 2022 **Faizal Akbari Putra**, Page 1022
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 1018−1026
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2259

The tables above shows that seven features have been successfully extracted from the PPG signal. These features are ESH, MAD, Skew, Kurt, VLF, LF, and HF. Each feature will be evaluated at the feature selection stage using the feature selection methods proposed in this study.

3.1.3 Feature Selection

a. Analysis of Variance (Anova)

The implementation of the feature selection Anova uses the *scikit-learn* library from Python, namely the *f_classif()* function. This function is used in conjunction with the *SelectKBest* class to select the *k* most relevant features. The optimal number of features is generated using a grid search, namely the *GridSearchCV()* function, where the *k* argument of the *SeleckKBest* class can be set to produce the best-performing model. The results of the feature selection Anova can be seen in Table 3.

Table 3. Results of Anova Feature Selection

Optimal Number of Features (*k*) : 5	
Feature	**Score**
Mad	6.051159
LF	0.854830
Kurt	0.851622
Skew	0.604537
VLF	0.083199
ESH	0.038505
HF	0.036356

The table above shows that the Mad feature is the most dominant compared to other features. The Mad feature far outperforms other features with a total score of statistical values of 6.051159 (larger is better). The result of the search for k to determine the optimal number of features can also be seen in Table 3. Based on the results of the grid search, it was found that the most optimal features are five features. Therefore, the selected features of this method are the features of Mad, LF, Kurt,Skew, and VLF.

b. Pearson Correlation

The correlation between features is used to find out how strong the relationship between features is. Two features that have a strong relationship will be removed one of them to prevent duplication of features in the classification process. The correlation between features can be visualized using a Python program using *heatmap* from data visualization library, namely *Seaborn*. The results of the Pearson Correlation features can be seen in Figure 5.

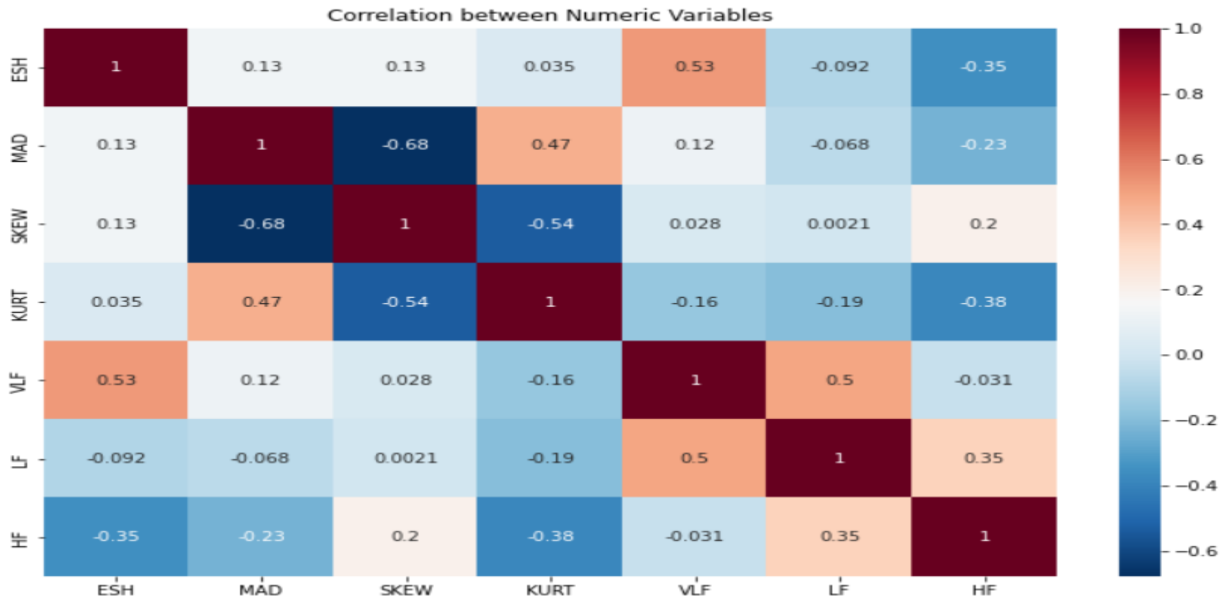

Figure 5. Correlation between Numeric Variables

The figure above shows the correlation between each feature. Each feature is compared to its correlation coefficient with other features. After that, the threshold is determined to select features with a high correlation. The threshold used is 0.5, where features that have a correlation above 0.5 will be removed. In Figure 5, it can be seen that the correlation value between the ESH feature and the VLF feature is 0.53. It exceeds the value of the specified threshold, so either ESH or VLF will be removed. This process is repeated for each feature. The result is that Kurt, VLF, and Skew features will be removed. Therefore, the selected features of this method are Mad, ESH, LF, and HF.

Copyright © 2022 **Faizal Akbari Putra**, Page 1023
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 1018−1026
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2259

c. Recursive Feature Elimination (RFE)

The RFE implementation uses the *sci-kit learn* library from Python, namely the *RFE()* class. Initially, the estimator parameter is set as the core model, and then the *n_features_to_select* parameter is set to determine the number of selected features. An *estimator* algorithm is needed to calculate the important scores. In this study, *RandomForestClassifier* is set as an estimator for RFE. Meanwhile, to find the optimal number of features, the *RFECV()* class is used to perform the cross-validation of several features to automatically select the number of features that produce the best mean score. The results of the RFE can be seen in Table 4.

Table 4. Results of RFE

Optimal Number of Features : 3		
Feature	**Selected**	**Rank**
ESH	TRUE	1
MAD	TRUE	1
Skew	TRUE	1
Kurt	FALSE	4
VLF	FALSE	3
LF	FALSE	2
HF	FALSE	5

The table above shows that the optimal number of features used for classification is three features. It can be seen that each feature that the RFE algorithm has processed has its ranking. The higher the ranking, the more relevant it is to use in predicting the target variable. The features that rank the highest are the ESH, Mad, and Skew features. Therefore, the selected features of this method are the ESH, Mad, and Skew.

3.2 Scenario Results

The scenario tested in this research using the Python program uses the KNN classification algorithm from the machine learning library *scikit-learn*, namely *KNeighborsClassifier*. The following are the results of scenarios of this study:

a. **Scenario I** : K-NN classification using features extracted from HRV features without using feature selection method:

Table 5. Scenario I Results

Scenario I	Accuracy	Sensitivity	Specificity
All HRV Features	72,7%	25%	100%

The results of the scenario I can be seen at Table 5. The results show that the accuracy and sensitivity are pretty low; it shows the accuracy of 72.7% and the sensitivity of 25%. In contrast, the result of specificity is very high, reaching 100%. By using the feature selection method in the next scenarios, it is expected that the results achieved can exceed the results of this test.

b. **Scenario II** : K-NN classification using features from the Anova feature selection method:

Table 6. Scenario II Results

Scenario II	Accuracy	Sensitivity	Specificity
Anova (Mad, LF, Kurt, Skew, VLF)	81,8%	50%	100%

The results of the scenario II can be seen at Table 6. The results show that the accuracy and sensitivity are better than in scenario I. The obtained accuracy reaches 81.8%, and the sensitivity obtained reaches 50%. The specificity obtained is also as high as in scenario I, which is 100%.

c. **Scenario III** : K-NN classification using features from the Pearson Correlation feature selection method:

Table 7. Scenario III Results

Scenario III	Accuracy	Sensitivity	Specificity
Pearson Correlation (Mad, ESH, LF, HF)	90,9%	75%	100%

The results of the scenario III can be seen at Table 7. The results show better accuracy and sensitivity than scenario I and scenario II. The accuracy obtained reached 90.9%, and the sensitivity obtained reached 75%. The specificity obtained is also as high as in scenario I, which is 100%.

d. **Scenario IV**: K-NN classification using features from the RFE feature selection method:

Table 8. Scenario IV Results

Scenario IV	Accuracy	Sensitivity	Specificity
Recursive Feature Elimination (ESH, Mad, Skew)	81.8%	50%	100%

Copyright © 2022 **Faizal Akbari Putra**, Page 1024
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 1018−1026
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2259

The results of the scenario IV can be seen at Table 8. The results show that the accuracy and sensitivity are better than in scenario I. The obtained accuracy reaches 81.8%, and the sensitivity obtained reaches 50%. The specificity obtained is also very high, the same as the results obtained in the scenario I, which is 100%.

3.3 Performance Comparison

Based on the results of the test scenarios, the three feature selection methods are able to improve the performance of CHD detection in the classification model using the KNN algorithm. All scenario results can be seen in Table 9.

Table 9. All Scenarios Results

Feature Selection Method	Accuracy	Sensitivity	Specificity
Without Feature Selection	72.70%	25%	100%
Anova	81.80%	50%	100%
Pearson Correlation	90.90%	75%	100%
RFE	81.80%	50%	100%

The table above shows the comparison of CHD detection performance between the classification model without feature selection and the classification model using feature selection methods. The classification model without using the feature selection method has the lowest detection performance compared to other models, the results are an accuracy of 72.70%, a sensitivity of 25%, and a specificity of 100%. The classification model using the proposed feature selection method has a higher CHD detection performance than the classification model without using feature selection. The classification model using the Anova method achieves an accuracy of 81.80%, a sensitivity of 25%, and a specificity of 100%. Then the classification model using the Pearson Correlation method achieves an accuracy of 90.90%, a sensitivity of 75%, and a specificity of 100%. At the same time, the classification model with the RFE method achieves an accuracy of 81.80%, a sensitivity of 25%, and a specificity of 100%. A comparison of the detection performance improvements can be seen in Figure 6.

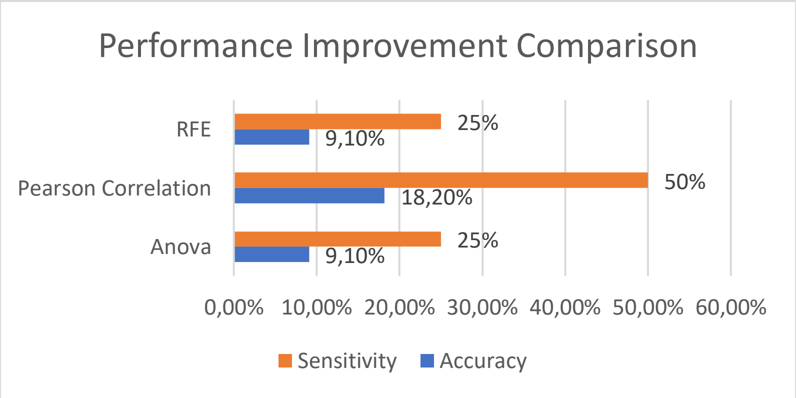

Figure 6. Performance Improvement Comparison

The Figure 6 shows that the three proposed feature selection methods can improve the accuracy and sensitivity performance of the classification model for CHD detection. It can be seen in Figure 6 that the Anova method can increase accuracy by 9.10% and sensitivity by 25%. Then the Pearson Correlation method can increase accuracy by 18.20% and sensitivity by 50%. Finally, the RFE method can increase accuracy by 9.10% and sensitivity by 25%. From these results, the Pearson Correlation method is the best feature selection method in this study because it can improve CHD detection performance more than the Anova and RFE methods.

4. CONCLUSION

This research has achieved the goal of creating a machine learning model to detect CHD with feature selection method based on PPG signals taken from smartphone's cameras. Based on the tests performed in this study, the implementation of three feature selection methods was able to improve the evaluation performance of CHD detection using K-NN classification algorithms. Before using the feature selection method, the results of the CHD detection test in this study were the accuracy of 72.7%, the sensitivity of 25%, and the specificity of 100%. After using the feature selection method, the performance results are improved for accuracy and sensitivity. The Anova method can increase accuracy by 9.1% and sensitivity by 25%, the Pearson Correlation method can increase accuracy by 18.2% and sensitivity by 50%, and the RFE method can increase accuracy by 9.1% and sensitivity by 25%. The best feature selection method in this study is the Pearson Correlation method because it can produce higher CHD detection performance than other methods. The overall results are the accuracy of 90.9%, the sensitivity of 75%, and the specificity of 100%.

Copyright © 2022 **Faizal Akbari Putra**, Page 1025
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 1018−1026
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2259

REFERENCES

[1] World Health Organization, "Cardiovascular diseases (CVDs)," World Health Organization, 11 June 2021. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). [Accessed 20 June 2022].
[2] NHS, "Coronary Heart Disease," NHS, 10 March 2020. [Online]. Available: https://www.nhs.uk/conditions/coronary-heart-disease/. [Accessed 20 June 2022].
[3] N. Mangatharayu, B. P. Rani, V. Janaki, L. S. Kotturi, M. Vallabhapurapu and V. G., "Heart Rate Variability for Predicting Coronary," in Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2020.
[4] Harvard Health Publishing Staff, "Heart rate variability: How it might indicate well-being," Harvard Health, 1 December 2021. [Online]. Available: https://www.health.harvard.edu/blog/heart-rate-variability-new-way-track-well-2017112212789. [Accessed 22 June 2022].
[5] J. K. Kim and J. M. Ahn, "Design of an Optimal Preamplifier of Photoplethysmogram for Age-Related," International Journal of Engineering Research and Technology, vol. 12, no. 10, pp. 1639-1646, 2019.
[6] D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur and H. Nazeran, "A review on wearable photoplethysmography sensors and their potential future applications in health care," International Journal of Biosensors & Bioelectronics, vol. 4, pp. 195-202, 2018.
[7] C.-J. Qin, Q. Guan and X.-P. Wang, "Application of ensemble algorithm integrating multiple criteria feature selection in coronary heart disease detection," Biomedical Engineering: Applications, Basis and Communications, vol. 29, p. 1750043, 2017.
[8] A. Shahid, B. Roy, M. P. Singh and A. Aadarsh, "Coronary Artery Disease Diagnosis Using Feature Selection Based Hybrid Extreme Learning Machine," in 2020 3rd International Conference on Information and Computer Technologies (ICICT), 2020.
[9] D. Krishnani, A. Kumari, A. Dewangan, A. Singh and N. S. Naik, "Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms," in TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, 2019.
[10] M. S. Pathan, A. Nag, M. M. Pathan and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," ArXiv, vol. abs/2206.03239, 2022.
[11] S.-M. Shan, S.-C. Tang, P.-W. Huang, Y.-M. Lin, W.-H. Huang, D.-M. Lai and A.-Y. Wu, "Reliable PPG-based Algorithm in Atrial Fibrillation," in 2016 IEEE Biomedical Circuits and Systems Conference (BioCAS), China, 2016.
[12] N. Akhter, S. Dabhade, N. Bansod and K. Kale, "Feature Selection for Heart Rate Variability Based Biometric Recognition Using Genetic Algorithm," in Intelligent Systems Technologies and Applications, 2016.
[13] C. Yahaya, G. Rouskas, M. S. Abd Latiff, M. A. Razzaque and S. Mandala, "QoS-aware ant-based route, wavelength and timeslot assignment algorithm for optical burst switched networks," Transactions on Emerging Telecommunications Technologies, vol. 26, pp. 1265-1277, 2015.
[14] S. Mandala, K. Jenni, M. A. Ngadi, M. Kamat and Y. Coulibaly, "Quantifying the Severity of Blackhole Attack in Wireless Mobile Adhoc Networks," in Security in Computing and Communications, 2014.
[15] A. R. Fikar, "Monitoring Sinyal ECG,PPG,Suhu Tubuh dan Kondisi cairan infus berbasis Zigbee (pada Bagian Pengirim dan Multiplexing)," 2013.
[16] S. Mandala, Y. N. Fuadah, M. Arzaki and F. E. Pambudi, "Performance analysis of wavelet-based denoising techniques for ECG signal," in 2017 5th International Conference on Information and Communication Technology (ICoIC7), Melaka, 2017.
[17] A. Chatterjee and U. K. Roy, "PPG Based Heart Rate Algorithm Improvement with Butterworth IIR Filter And Savitzky-Golay FIR Filter," in 2018 2nd International Confrence on Electronics, Material Engineering & Nano-Technology (IEMENTech), 2018.
[18] A. J. Parsaoran, S. Mandala and M. Pramudyo, "Study of Denoising Algorithms on Photoplethysmograph (PPG) Signals," in 2022 International Conference on Data Science and Its Applications (ICoDSA), Bandung, 2022.
[19] R. Banerjee, S. Bhattacharya and S. Alam, "TIME SERIES AND MORPHOLOGICAL FEATURE EXTRACTION FOR CLASSIFYING," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018.
[20] J. Cai, J. Luo, S. Wang and S. Yang, "Feature selection in machine learning: A new perspective," Neurocomputing, vol. 300, 2018.
[21] M. Kumar, N. K. Rath, A. Swain and S. K. Rath, "Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor," Procedia Computer Science, vol. 54, pp. 301-310, 2015.
[22] D. Nettleton, Chapter 6. Selection of Variables and Factor Derivation, 2014.
[23] A. R. I. Pratama, S. A. Latipah and B. N. Sari, "Optimasi Klasifikasi Curah Hujan Menggunakan Support Vector Machine (SVM) dan Recursive Feature Elimination (RFE)," JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), vol. 07, no. 02, pp. 314-324, 2022.
[24] H. Tjahjadi and K. Ramli, "Noninvasive Blood Pressure Classification Based on Photoplethysmography Using K-Nearest Neighbors Algorithm: A Feasibility Study," Information, vol. 11, p. 93, 2020.
[25] R. Adi, W. Laksito and P. Harsadi, "IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOR UNTUK IDENTIFIKASI KUALITAS AIR (STUDI KASUS : PDAM KOTA SURAKARTA)," Jurnal Teknologi Informasi dan Komunikasi (TIKomSiN), vol. 6, 2018.
[26] K. Jenni and S. Mandala, "Pre-processing image database for efficient Content Based Image Retrieval," in 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Delhi, 2014.
[27] A. H. Gonsalves, F. Thabtah, R. M. A. Mohammad and G. Singh, "Prediction of Coronary Heart Disease using Machine : An Experimental Analysis," in Proceedings of the 2019 3rd International Conference on Deep Learning Technologies - ICLDT 2019, 2019.
[28] R. Banarjee, R. Vempada, K. M. Mandana, A. D. Choudhury and A. Pal, "Identifying coronary artery disease from photoplethysmogram," in Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, 2016.

Copyright © 2022 **Faizal Akbari Putra**, Page 1026
This Journal is licensed under a Creative Commons Attribution 4.0 International License