e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v6i2.11326

Prediction of Sudden Cardiac Death with Feature Selection Using Particle Swarm Optimization

David[1*], Sani Muhamad Isa[2]

[1, 2] Computer Science Department, BINUS Graduate Program – Master of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
david038@binus.ac.id; sani.m.isa@binus.ac.id

*Correspondence: david038@binus.ac.id

Abstract – *The heart, a vital organ responsible for pumping oxygenated blood through blood vessels, is susceptible to disturbances in heart rate that can have adverse effects. According to data from the World Health Organization (WHO) since 2000, this disease has experienced the most significant increase in fatalities, rising from over 2 million to 8.9 million deaths. The prediction of Sudden Cardiac Death (SCD) continues to gain attention as a promising approach to saving millions of lives threatened by the occurrence of the disease. In this study, we propose the utilization of Particle Swarm Optimization (PSO) as a feature selection method to train the Support Vector Machine (SVM) and Logistic Regression. By employing the proposed algorithm, SCD can be predicted up to 30 minutes before the onset with an accuracy of 92.5%, by using PSO and SVM. Features are extracted from Heart Rate Variability (HRV) analysis and Discrete Wavelet Transform (DWT) obtained from ECG records of MIT-BIH normal sinus rhythm database & MIT-BIH Sudden Cardiac Death Holter database dataset. This paper also compares feature selection algorithm of PSO and Analysis of Variance (ANOVA) and found that PSO is better in accuracy, recall, and F1-score.*

Keywords*: Feature Selection; Sudden Cardiac Death; Particle Swarm Optimization; Cardiovascular Diseases; Support Vector Machine*

I. INTRODUCTION

The heart, a crucial organ responsible for circulating oxygenated blood through the body's blood vessels by maintaining a specific rhythm of contractions. Disturbance in the heart rate could be disastrous. Although cardiac arrhythmia can lead to death, it can be managed if diagnosed promptly (Tavassoli et al., 2012). According to World Health Organization (WHO) data since 2000, this disease has witnessed a significant surge in fatalities, escalating from over 2 million to 8.9 million deaths by 2019 (World Health Organization, 2020). Sudden Cardiac Death (SCD) in medical terms refers to sudden and unexpected death caused by heart failure in a short time, mostly around less than an hour to an individual without observable symptoms (Wong et al., 2019). Atrial fibrillation (AF) and Ventricular Tachycardia (VT) are related to increased risk of cardiovascular and mortality, including SCD (Behnes et al., 2019).

Our understanding of SCD has improved. One of the non-invasive methods that can be used to predict SCD is by observing ECG records. Long term observation of ECG records is a criterion to diagnose Ventricular Arrhythmia (VA). Even though this method is proven in detecting VA, it still lacks the ability to accurately differentiate between normal and abnormal ECG records (Bayasi et al., 2015). Previous research tried to detect SCD using features of Heart Rate Variability (HRV) extracted from electrocardiogram (EKG) (Ebrahimzadeh et al., 2019; Lai et al., 2019). A similar topic by Ashtiyani et al. (Ashtiyani et al., 2018) extracted HRV features using Discrete Wavelet Transform (DWT) method. One of the most used mathematical methods for analyzing HRV is the Fourier transform, which is primarily applicable to stationary signals. However, when dealing with non-stationary conditions, wavelet transform analysis becomes a suitable alternative for quantifying HRV. Wavelet transform (WT) is a mathematical technique employed to examine non-stationary signals (Rhif et al., 2019). Unlike the Fourier transform, which dissects the signal into sine and cosine functions, the wavelet transform employs functions localized in both the Fourier and real spaces, rendering it an apt method for processing medical signals (Al Bassam et al., 2021).

Research performed by (Ebrahimzadeh et al., 2019) uses features generated by HRV and uses a time local subset as feature selection. The features are sent to multilayer perceptron to predict SCD with 83% accuracy, 12 minutes before the SCD. Other previous research by (Lopez-Caracheo et al., 2018) uses linear and non-linear features from fractal dimension. In this research, ANOVA analysis is used as feature selection combined with Multilayer Perceptron to predict SCD 91.4% accuracy, 14 minutes before SCD. (Devi et al., 2019) uses feature extraction from HRV and Continuous Wavelet Transform (CWT). Features generated will be selected using Kruskal-Wallis to train K-Nearest Neighbors (KNN) model, SVM, and decision tree that produce 83.33% accuracy, 10 minutes before SCD.

In feature selection problems, the quest for an efficient global search technique is paramount. Particle Swarm Optimization (PSO) emerges as a relatively recent evolutionary computing (EC) approach grounded in swarm intelligence principles. In contrast to other EC algorithms like genetic programming (GP) and genetic algorithms (GA), PSO uses less computation power (Almufti et al., 2019), rendering it a favored choice across various domains, including feature selection endeavors (Wei et al., 2019). Based on that fact, this research will implement PSO as feature selection. The purpose of this research is to compare accuracy of 3 methods in predicting SCD: one using all features generated from HRV and DWT, another using selected features by PSO, and another using selected features by ANOVA analysis. The classification model used are SVM and Logistic regression.

The structure of this paper is arranged as follows: Section 2 explains the methodology of the research and elaborates the steps of the experiment; Section 3 discusses results of the study; Section 4; closes the study with conclusion and suggestion for future work.

II. METHODS

The illustration of the experimental steps can be seen in Figure 1. The experiment begins with data collection, preprocessing, feature extraction, feature selection, classification model, and evaluation model that uses all the features and models that use the selected feature by PSO.

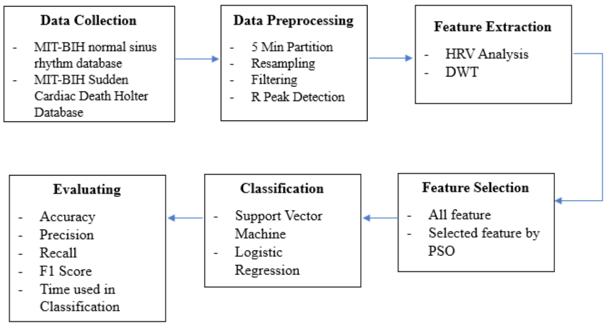


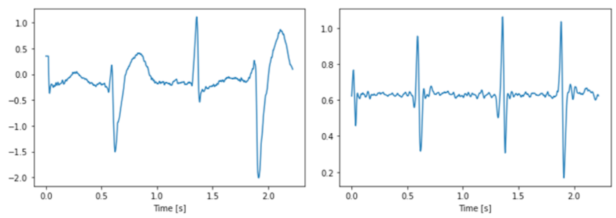
Figure 1. Steps of experiment

2.1 Data acquisition

There were two datasets used in this study, collected from PhysioNet. The first dataset is from MIT-BIH normal sinus rhythm database consisting of 18 electrocardiogram (ECG) recordings with frequency sampling of 128 Hz. The second dataset was sourced from the MIT-BIH Sudden Cardiac Death Holter Database, encompassing 23 electrocardiogram (ECG) recordings with a sampling frequency of 250 Hz. The age range of SCD patients spanned from 18 to 89 years, while that of normal subjects ranged from 20 to 50 years.

2.2 Preprocessing.

5 minutes duration will be sampled from every healthy ECG record randomly for analysis. For each SCD ECG record, we take the first 5 minutes duration in the last 35 minutes before the occurrence of SCD as sample for analysis. In this study, in order to maintain consistency between the SCD and normal groups, the normal ECG signals were uniformly sampled at a rate of 250 Hz. Out of the total 20 SCD ECG signals used for analysis, 3 were excluded due to the absence of Ventricular Fibrillation (VF) episodes. Prior to ECG signal analysis, it is imperative to filter out interference noise stemming from power lines and baseline wander (Kher & others, 2019) can be seen in Figure 2. Signal processing, specifically bandpass filtering, is employed to allow only a specific frequency range of the ECG signal to pass through, effectively reducing signal noise. In the next step of the process, feature extraction requires the location of R-peaks within the ECG signal to obtain HRV (Heart Rate Variability) features. The location of the R-peaks will be detected using the Pan-Tompkins algorithm (Abd Al-Jabbar et al., 2023).



ECG Before Filter

ECG After Filter

Figure 2. Signal ECG

2.3 Feature Extraction

In this study, the features will be extracted using Heart Rate Variability (HRV) (Panday & Panday, 2018) analysis and Discrete Wavelet Transform (DWT) (Chashmi & Amirani, 2019; Dar et al., 2015). According to (Panday & Panday, 2018), There are various methods to analyze HRV. Generally, the analysis of HRV produces time-domain, frequency-domain, and non-linear domain.

Time-domain:

- MeanNN (The mean of the RR intervals)
- SDNN The standard deviation of the RR intervals)
- RMSSD (Root mean square of successive NN interval differences)
- SDSD (Standard deviation of successive NN interval differences)
- MadNN (The median absolute deviation of the RR intervals)
- pNN50 (Proportion of successive NN interval differences larger than 50 ms)
- pNN20 (Proportion of successive NN interval differences larger than 20 ms)
- MinNN (The minimum of the RR interval)
- MaxNN (The maximum of the RR intervals)

Among them, the most used time-domain measures are SDNN and RMSSD.

Frequency-domain:

- VLF (Power spectrum in the frequency range of 0.0033–0.04 Hz)
- LF (Power spectrum in the frequency range of 0.04–0.15 Hz)
- HF (Power spectrum in the frequency range of 0.15–0.4 Hz)
- VHF (Power spectrum in the frequency range of 0.4–0.5Hz)
- LF/HF (Ratio of LF to HF power)
- LFn (Normalized LF)
- HFn (Normalized HF)
- LnHF (Natural logarithm of HF)

Frequency domain measures utilize frequency bands to count the number of RR intervals that fall within each band.

Non-linear domain:

Non-linear methods are utilized because commonly employed moment statistics of HRV may not be capable of detecting subtle yet significant changes in HR within time series. Among these methods, the Poincaré plot stands out as the most widely utilized for HRV analysis. In this plot, each data point denotes a pair of consecutive beats, with the x-axis denoting the current NN interval and the y-axis representing the previous NN interval (Henriques et al., 2020).

DWT (Discrete Wavelet Transform)

is a proven effective method for processing digital signals, including ECG signals. With the discrete wavelet transform, the signal can be decomposed into multiple frequency levels, where each level represents approximation coefficients (discrete-time low filter) and detail coefficients (discrete-time high filter), as shown in Figure 3 (Ashtiyani et al., 2018). In this study, the signal is decomposed using db6 (Bota et al., 2019; Murugappan et al., 2013) wavelet into 6 levels and five statistical features (Min, Max, Mean, STD, Energy) (Chashmi & Amirani, 2019) are extracted from each sub-band obtained through DWT decomposition.

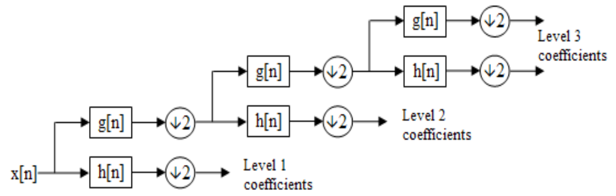


Figure 3. Discrete Wavelet Transform Decomposition Tree (Ashtiyani et al., 2018)

2.4 Feature Selection

In this study, PSO (Particle Swarm Optimization) is used for feature selection, which is an effective algorithm in solving problems, and can be used to find an optimal feature subset (Wei et al., 2019). In PSO, it is supposed that the quality of each candidate solution can be assessed using a fitness function. In this study, the accuracy of classification is considered as a fitness function. PSO parameters are adjusted as follows: Particle size: 10 and number of iterations: 100. The Selected subset is based on PSO results is SDNN, RMSSD, SD2, cA6_Maximum, cD6_Energy, cD5_Maximum, cD4_Std, cD3_Minimum, cD3_Std, cD3_Energy, cD2_Std, cD2_Energy, cD1_Minimum. The selected features will be used to train the final model and will be evaluated with the model without selecting features.

2.5 Classification

To classify normal and SCD subjects, Logistic Regression (LaValley, 2008) and Support Vector Machine (SVM) (Cortes & Vapnik, 1995) are used. The classification model will be trained with the features selected by PSO and all features to compare the results. In this study, the five times five-fold cross-validation method is used to test the performance of all classifiers.

2.6 Evaluation

Classification accuracy is typically assessed based on performance indicators such as Accuracy, Precision, Recall, and F1-score (1). The Confusion Matrix (Krstinić et al., 2020) shown in Figure 4 is used for classification evaluation. It represents the comparison between the model's classification results and the actual classification results.

$$F1 - Score = 2\,\frac{Precision\cdot Recall}{Precision+Recall'},$$

where $Precision = \frac{TP}{(TP+FP)}$ and $Recall = \frac{TP}{(TP+FN)}$ (1)

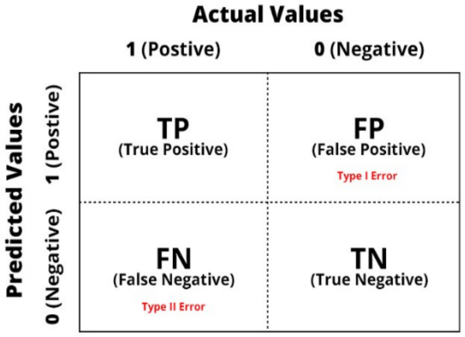


Figure 4. Confusion Matrix

III. RESULTS AND DISCUSSION

In this study, Particle Swarm Optimization (PSO) is utilized as a feature selection technique to obtain the most optimal set of features for training Logistic Regression and SVM models. As a comparative analysis, the models are also trained using the entire set of available features to evaluate their performance. This research aims to compare the evaluation results between the models that incorporate all features and the models that exclusively utilize the features selected by PSO. The validation results obtained from using the K-Fold method, as shown in Table I, were used to assess the performance of the classification model. K-Fold was executed with 5 iterations to ensure the model's accuracy was maximized.

Table I. The result comparison of SVM and Logistic Regression

Method	Feature	Accuracy	Precision	Recall	F1-Score	Time (s)
Support Vector Machine	all features	0.8	0.83	0.8	79.44%	0.0068
	selected features PSO	0.925	0.909	0.95	92.77%	0.0024
Logistic Regression	all features	0.875	0.96	0.8	84.44%	0.0006
	selected features PSO	0.9	0.909	0.9	89.92%	0.0002

From the evaluation results in Table 1, it can be concluded that training the model requires the use of relevant features. In this study, both the SVM and Logistic Regression methods achieved better results when utilizing features selected by PSO. When comparing the SVM and Logistic Regression models using the selected features, it is evident that SVM achieves the highest values for Accuracy, Recall, and F1-Score compared to Logistic Regression. In addition, proper feature selection can also accelerate the model training time. The analysis results of this study indicate that support vector machine (SVM) performs well in identifying SCD patients and normal patients when trained with appropriate features. SVM is considered superior to logistic regression based on higher accuracy, recall, and F1-score. SVM achieved the highest accuracy of 92.5%, highest recall of 0.95, and the highest F1 score of 92.77%.

In this study, a comparison was made between feature selection using the PSO method and ANOVA analysis used by (Lopez-Caracheo et al., 2018) in Table II. PSO proved to be effective in providing suitable features for training the model to identify SCD and normal patients, outperforming ANOVA in terms of Accuracy, Recall, and F1-Score. However, it should be noted that the feature selection process using PSO requires more time compared to ANOVA due to the iterations involved in achieving the optimal solution.

Table II. Evaluation Results between PSO and ANOVA as feature selection

Model	Feature Selection	Accuracy	Precision	Recall	F1-Score
SVM	PSO	0.925	0.909	0.95	92.77%
	ANOVA Analysis	0.9	0.95	0.85	88.75%

IV. CONCLUSION

This study illustrates how proper feature selection can enhance the accuracy of the model, and PSO can be utilized as a feature selection method to improve the performance of the model in predicting SCD, with higher accuracy and faster processing time compared to ANOVA analysis and not using any feature selection. We propose the utilization of PSO as a feature selection method to train the Support Vector Machine model. By employing the proposed algorithm, SCD can be predicted up to 30 minutes before the onset with an accuracy of 92.5%, by using features extracted from HRV and DWT. Furthermore, similar experiments can be conducted with longer prediction intervals and larger datasets, considering the limitations of the dataset used in this study.

REFERENCES

Abd Al-Jabbar, E. Y., Al-Hatab, M. M. M., Qasim, M. A., Fathel, W. R., Fadhil, M. A., & others. (2023). Clinical Fusion for Real-Time Complex QRS Pattern Detection in Wearable ECG Using the Pan-Tompkins Algorithm. *Fusion: Practice and Applications*, *12*(2), 172.

Al Bassam, N., Ramachandran, V., & Parameswaran, S. E. (2021). Wavelet theory and application in communication and signal processing. In *Wavelet Theory* (p. 45). IntechOpen.

Almufti, S. M., Zebari, A. Y., Omer, H. K., & others. (2019). A comparative study of particle swarm optimization and genetic algorithm. *Journal of Advanced Computer Science & Technology*, *8*(2), 40.

Ashtiyani, M., Lavasani, S. N., Alvar, A. A., & Deevband, M. R. (2018). Heart rate variability classification using support vector machine and genetic algorithm. *Journal of Biomedical Physics & Engineering*, *8*(4), 423.

Bayasi, N., Tekeste, T., Saleh, H., Mohammad, B., Khandoker, A., & Ismail, M. (2015). Low-power

ECG-based processor for predicting ventricular arrhythmia. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, *24*(5), 1962–1974.

Behnes, M., Rusnak, J., Taton, G., Schupp, T., Reiser, L., Bollow, A., Reichelt, T., Engelke, N., Ellguth, D., Kuche, P., & others. (2019). Atrial fibrillation is associated with increased mortality in patients presenting with ventricular tachyarrhythmias. *Scientific Reports*, *9*(1), 14291.

Bota, P. J., Wang, C., Fred, A. L. N., & Da Silva, H. P. (2019). A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. *IEEE Access*, *7*, 140990–141020.

Chashmi, A. J., & Amirani, M. C. (2019). An efficient and automatic ECG arrhythmia diagnosis system using DWT and HOS features and entropy-based feature selection procedure. *Journal of Electrical Bioimpedance*, *10*(1), 47–54.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Dar, M. N., Akram, M. U., Shaukat, A., & Khan, M. A. (2015). ECG based biometric identification for population with normal and cardiac anomalies using hybrid HRV and DWT features. *2015 5th International Conference on IT Convergence and Security (ICITCS)*, 1–5.

Devi, R., Tyagi, H. K., & Kumar, D. (2019). A novel multiclass approach for early-stage prediction of sudden cardiac death. *Biocybernetics and Biomedical Engineering*, *39*(3), 586–598.

Ebrahimzadeh, E., Foroutan, A., Shams, M., Baradaran, R., Rajabion, L., Joulani, M., & Fayaz, F. (2019). An optimal strategy for prediction of sudden cardiac death through a pioneering feature-selection approach from HRV signal. *Computer Methods and Programs in Biomedicine*, *169*, 19–36.

Henriques, T., Ribeiro, M., Teixeira, A., Castro, L., Antunes, L., & Costa-Santos, C. (2020). Nonlinear methods most applied to heart-rate time series: a review. *Entropy*, *22*(3), 309.

Kher, R., & others. (2019). Signal processing techniques for removing noise from ECG signals. *J. Biomed. Eng. Res*, *3*(101), 1–9.

Krstinić, D., Braović, M., Šerić, L., & Božić-Štulić, D. (2020). Multi-label classifier performance evaluation with confusion matrix. *Computer Science & Information Technology*, *1*, 1–14.

Lai, D., Zhang, Y., Zhang, X., Su, Y., & Heyat, M. B. Bin. (2019). An automated strategy for early risk identification of sudden cardiac death by using machine learning approach on measurable arrhythmic risk markers. *IEEE Access*, *7*, 94701–94716.

LaValley, M. P. (2008). Logistic regression. *Circulation*, *117*(18), 2395–2399.

Lopez-Caracheo, F., Camacho, A. B., Perez-Ramirez, C. A., Valtierra-Rodriguez, M., Dominguez-Gonzalez, A., & Amezquita-Sanchez, J. P. (2018). Fractal dimension-based methodology for sudden cardiac death prediction. *2018 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*, 1–6.

Murugappan, M., Murugappan, S., & Zheng, B. S. (2013). Frequency band analysis of electrocardiogram (ECG) signals for human emotional state classification using discrete wavelet transform (DWT). *Journal of Physical Therapy Science*, *25*(7), 753–759.

Panday, K. R., & Panday, D. P. (2018). Heart rate variability. *J Clin Exp Cardiol*, *9*, 1–12.

Rhif, M., Ben Abbes, A., Farah, I. R., Mart\'\inez, B., & Sang, Y. (2019). Wavelet transform application for/in non-stationary time-series analysis: A review. *Applied Sciences*, *9*(7), 1345.

Tavassoli, M., Ebadzadeh, M. M., & Malek, H. (2012). Classification of cardiac arrhythmia with respect to ECG and HRV signal by genetic programming. *Canadian Journal on Artificial Intelligence, Machine Learning and Pattern Recognition*, *3*(1), 1–8.

Wei, B., Zhang, W., Xia, X., Zhang, Y., Yu, F., & Zhu, Z. (2019). Efficient feature selection algorithm based on particle swarm optimization with learning memory. *IEEE Access*, *7*, 166066–166078.

Wong, C. X., Brown, A., Lau, D. H., Chugh, S. S., Albert, C. M., Kalman, J. M., & Sanders, P. (2019). Epidemiology of sudden cardiac death: global and regional perspectives. *Heart, Lung and Circulation*, *28*(1), 6–14.

World Health Organization. (2020). *The top 10 causes of death*. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death