

Particle Swarm Optimization For Improved Accuracy of Disease Diagnosis

Suamanda Ika Novichasari*¹, Iwan Setiawan Wibisono²

Informatics Engineering, Ngudi Waluyo University, Jl. Diponegoro no 186 Gedanganak - Ungaran Timur, Kab. Semarang Jawa Tengah, (024)-6925408.

*E-mail : vichareal0311@gmail.com*¹, loyal.wb99@gmail.com²*

**Corresponding author*

Abstract - The increasing number of patients suffering from various diseases and the impact and high cost of medical treatment for the community has made the government or health communities seek prevention early. This valuable information can be found using artificial intelligence and data mining. Most diseases are dangerous; if detected early and adequate diagnosis and treatment are available, there will be a chance for a cure. The main objective of this study was to use Particle Swarm Optimization (PSO) to improve the accuracy of several classification methods, namely Naive Bayes, C4.5, Support Vector Machine (SVM), and Neural Network (NN) to detect heart disease, hepatitis, kidney, and breast cancer. The method used in this research is the CRISP-DM model, with five stages. The data used were four disease data from UCI Machine Learning. This research shows that PSO can improve Naive Bayes, C4.5, SVM, and NN accuracy.

Keywords - NN, SVM, C4.5, Naïve Bayes, PSO

1. INTRODUCTION

The increasing number of patients suffering from various diseases, and the impact and high cost of medical treatment have made the government or health communities seek prevention early. This valuable information can be found using artificial intelligence and data mining.

Globally, cardiovascular disease and cancer are the leading causes of death worldwide. Cardiovascular disease is a disease caused by the heart and blood vessels' impaired function, such as coronary heart disease, heart failure, hypertension, and stroke [1]. One type of cancer that is very scary for women around the world is breast cancer. When breast cancer is detected early, and adequate diagnosis and treatment are available, there is a chance that it can be cured [2].

Chronic kidney disease (CKD) is a global public health problem with an increasing prevalence and incidence of kidney failure, poor prognosis, and high costs. Chronic kidney disease initially shows no signs and symptoms but can progress to kidney failure. Kidney disease can be prevented and managed, and the chances of getting effective therapy are greater if caught early [3].

Hepatitis is an infectious disease that is a public health problem, which affects the morbidity, mortality rate, public health status, life expectancy, and other socio-economic impacts [4]

Data mining is a process that aims to find patterns automatically or semi-automatically from existing data in a database that is used to solve a problem [5]. Data mining has several techniques, including classification and clustering. Classification technique is a learning

technique used to predict the value of the target category attributes [6]. Classification aims to divide objects assigned only to one number of categories called class [7].

From a dataset consisting of several variables that affect disease, a pattern will be sought to detect or diagnose whether a patient is a patient or not. The detection of disease diagnoses will use data mining classification techniques.

Several researchers have conducted studies on the same topic using the association rule [8] SVM and BPNN [9] for breast cancer detection [8], SVM, and Naïve Bayes for detection of liver disease [10], SVM and ANN for detection of Kidney disease [11].

One of the challenges in analyzing disease is the large data size and multiple attributes. In order to improve the model's performance, it requires more big feature selection than other features. Some of the methods used for this are Particle Swarm Optimization (PSO) and Genetic Algorithm (GA). PSO gives better results than GA in terms of each criterion's fitness function value [12] [13]. Chung-Jui Tu et al. In 2007 used PSO and GA for attribute weighting. The results prove that PSO-SVM is superior with a faster computation time than GA-SVM [14]. So that PSO can maximize attribute weighting or attribute selection to improve the classification.

Several other researchers have also proven that PSO has improved the classification method's accuracy in several different cases. Prisma Handayanna implemented SVM-PSO for diabetes prediction [15], and Junta Zeniarja in 2012 implemented SVM-PSO for opinion mining [16]. Farid Melgani and Yakoub Bazi in 2008 proved that PSO-SVM was superior to SVM, KNN, and RBF classifications for electrocardiogram signal classification [17]. Research on clove leaf classification conducted by Novichasari, SI (2015) proved that PSO is an attribute that enhances the relationship of SVM [18].

This study's main objective is to use PSO to improve several classification methods, namely C4.5, Naïve Bayes, SVM, and NN, to detect heart disease, hepatitis, Parkinson's, thyroid, breast cancer, and lung cancer. The experimental results will be compared so that the method with the best accuracy is known.

This study's results can be used as the basis for making a decision support system in diagnosing disease, making it easier for doctors to determine the right treatment and care for their patients. Assist public health personnel in determining disease prevention methods in the short and long term. It also contributes to data mining methods in comparing or comparing several data mining classification algorithms for different cases or in the same case and by using different algorithms.

2. RESEARCH METHOD

The method used in this research is the CRISP-DM model.

2.1 Business Understanding

The increasing number of patients suffering from various diseases and the impact and high cost of medical treatment have made the government or health communities seek prevention early. This valuable information can be found using artificial intelligence and data mining.

2.2 Data Understanding

The data used were taken from the University of California Irvine (UCI) Machine Learning with the title Kidney Disease [19], Heart Disease [20], Breast Cancer [21], and Hepatitis [22]. The data obtained from the UCI is in the form of text with type txt. To be used in Rapid Miner, the data must be converted into a sheet of type CSV or xls.

Table 1. list of raw data

No	Data Set Name	Number of Attributes	The amount of data
1.	Kidney Disease	25	400
2.	Heart Disease	14	303
3.	Breast Cancer	9	286
4.	Hepatitis	19	155

2.3 Data Preparation

The classification model using Naïve Bayes and C4.5 cannot handle dependent labels/attributes in the form of numerics, so changes are made manually by changing the values in the dependent attribute, as shown in the following table:

Table 2. Dependent attribute conversion

No	Data Set Name	Atribut Dependent	Initial Value	Conversion
1	Kidney Disease	Class	1-2	y-n
2	Heart Disease	Target	0-1	y-n
3	Breast Cancer	Class	recurrence events - no recurrence events	y-n
4	Hepatitis	Class	CKD-nonskid	y-n

The classification model using SVM and NN cannot handle type polynomial attributes. They cannot handle missing values so that the nominal attribute in all data sets is converted to numeric and eliminates all data containing missing values. The data processing uses Ms. Excel. So that the number of attributes and data in each data set has changed as follows:

Table 3. List of datasets ready to be modeled

No	Data Set Name	Number of Attributes	The amount of data
1.	Kidney Disease	24	189
2.	Heart Disease	14	303
3.	Breast Cancer	10	286
4.	Hepatitis	17	129

2.4 Modeling

Then the data is processed in such a way that it is ready for modeling as below.

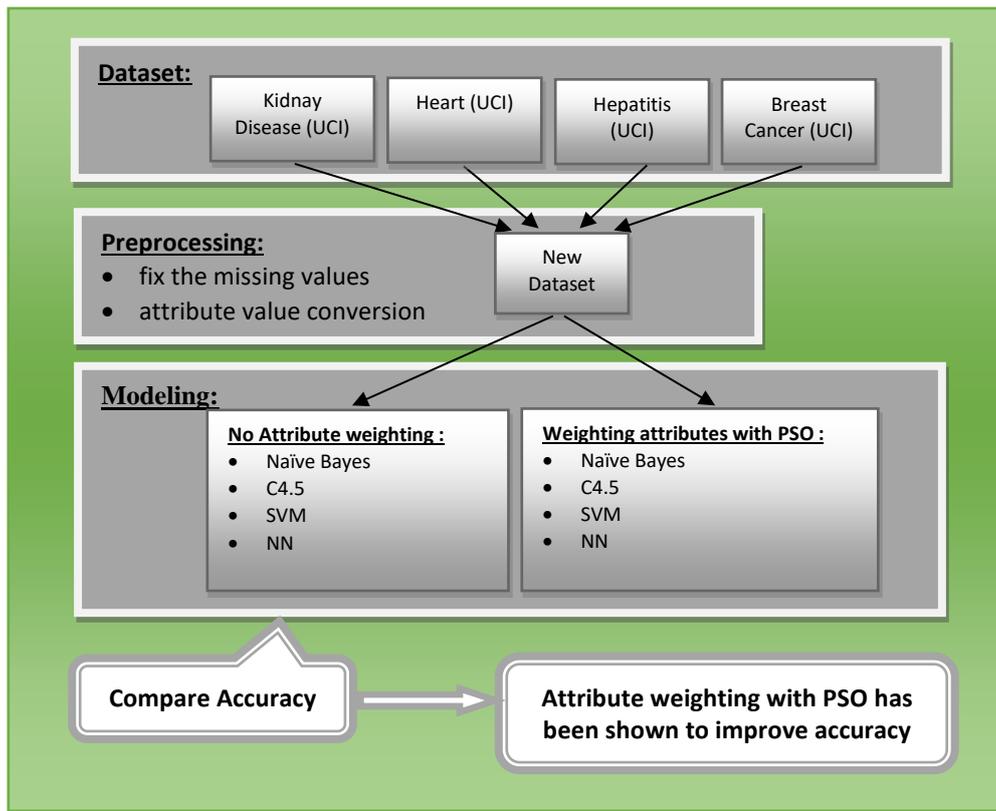


Figure 1. The proposed model

2.4.1 Particle Swarm Optimization (PSO)

Doctors Kennedy and Eberhart in 1995 introduced PSO as a global heuristic optimization method based on research on the behavior of groups of birds and fish [23]. Each particle is associated with the particles' speed moving through space at the speed of a dynamic search tailored to their historical behavior. Therefore, the particles tend to move towards a better search area during the search process [24]. The formula for calculating the velocity and velocity of a particle is :

$$V_i(t) = V_i(t - 1) + c_1 r_1 [X_{pbest_i} - X_i(t)] + c_2 r_2 [X_{Gbest} - X_i(t)] \quad (1)$$

$$X_i(t) = X_i(t - 1) + V_i(t) \quad (2)$$

$V_i(t)$ is the velocity of particle i during iteration t , $X_i(t)$ is the position of particle i when iterations t , c_1 , and c_2 are learning rates for an individual (cognitive) ability and social influence (group), r_1 and r_2 are random numbers uniformly distributed in the intervals of 0 and 1, X_{pbest_i} is the best position for particle i , X_{Gbest} is the best position globally.

In this study, PSO is used for attribute weighting. The flowchart of the proposed PSO implementation can be seen in Figure 2.

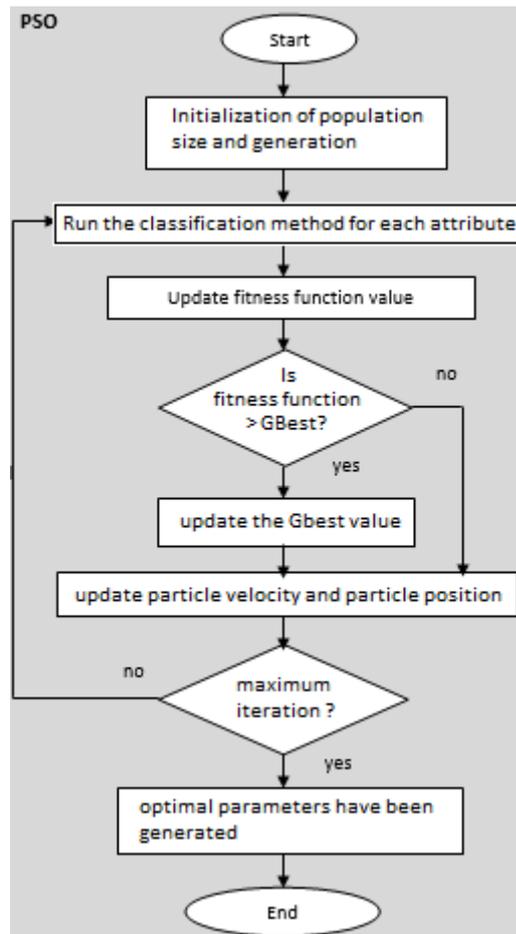


Figure 2. The proposed PSO attribute weighting model

2.4.2 Naïve Bayes

Naive Bayes is a statistical classification based on the Bayes theorem, which predicts the probability of class membership. Naive Bayes has been proven to have high accuracy and speed when applied to large databases [25].

The general form of the Bayes theorem is as follows:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (3)$$

X is data with an unknown class, H is data hypothesis X is a specific class, P (H | X) is the probability of hypothesis H based on condition X (posterior probability), P (H) is the probability of hypothesis H (prior probability).

2.4.3 C4.5

C4.5 is a development of the ID3 algorithm. Attributes of type numeric or continuous, which ID3 cannot handle, can be handled by C4.5. The root node selection is based on the highest Gain Ratio value, not on the highest gain value, such as ID3 [26]. For continuous attributes in C4.5, the data is divided by sorting the examples based on a continuous attribute A, then forming a minimum threshold M from existing examples from the majority class on each contiguous partition, then combining the adjacent partitions with the same majority class [27].

2.4.4 Support Vector Machine (SVM)

Problems in the real world are more non-linear. SVM is modified to solve non-linear problems by including kernel functions. The \vec{x} data is mapped by the function $\Phi(\vec{x})$ to a higher-dimensional vector space. The hyperplane is used to separate the two classes [26].

Usually, the transformation of the dot product Φ is unknown and very difficult to understand. Therefore the dot product calculation is replaced by the kernel function $K(\vec{x}_i, \vec{x}_j)$, which implicitly defines the transformation Φ . The Kernel Trick is formulated as below [22]:

$$K(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i) \cdot \Phi(\vec{x}_j) \quad (4)$$

$$f(\Phi(\vec{x})) = \vec{w} \cdot \Phi(\vec{x}) + b \quad (5)$$

2.4.5 Neural Network (NN)

NN adapts the human brain's neural network, consisting of a set of basic computational units called neurons. These neurons are connected through a weighted network and arranged in several layers. All neurons in the layer are connected exclusively from the previous layer and the next layer [28].

The similarities in principle between the workings of NN and the human brain are as follows [29]:

- There is a learning process to acquire knowledge;
- The knowledge gained is stored in the connections between neurons.

2.5 Evaluation

The validation process uses 10fold cross-validation; automatically, the data will be divided into two, namely 1/10 used as testing data and 9/10 for training data.

2.6 Deployment

This study's result is an analysis that leads to the Decision Support System (DSS) in diagnosing diseases, making it easier for doctors to determine the right treatment and care for their patients.

3. RESULTS AND DISCUSSION

There are four classification methods used: the Naïve Bayes algorithm, C4.5, SVM, and NN. The four methods are then combined with the PSO algorithm in attribute weighting using the Rapid Miner version 9.8.000 framework so that the attribute weights generated by the PSO will be seen.

The NN model's highest accuracy results were achieved from the first modeling experiment that was carried out without using PSO. Each classification method used uses the default parameters. The results show that SVM and NN have competitively superior accuracy than others on all datasets.

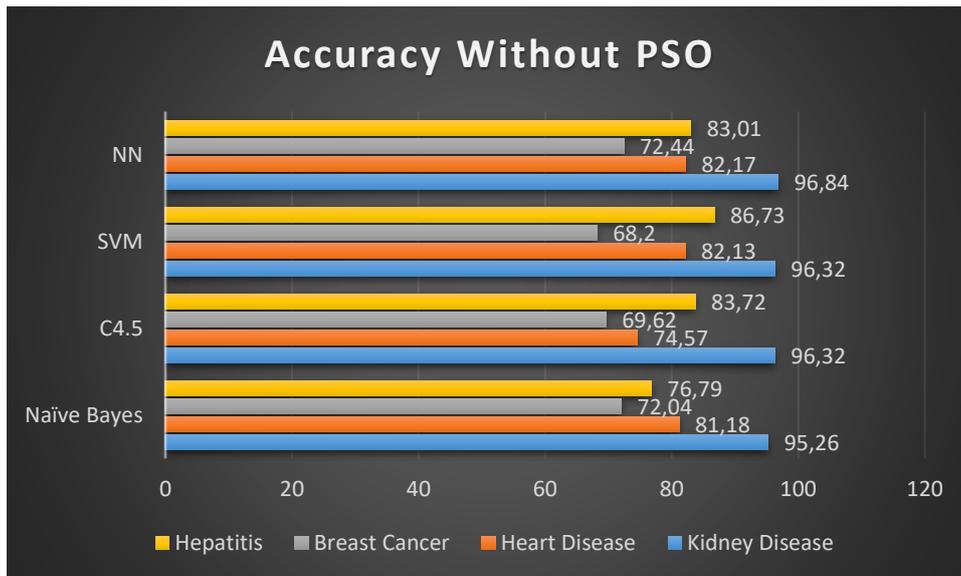


Figure 3. Accuracy Without PSO

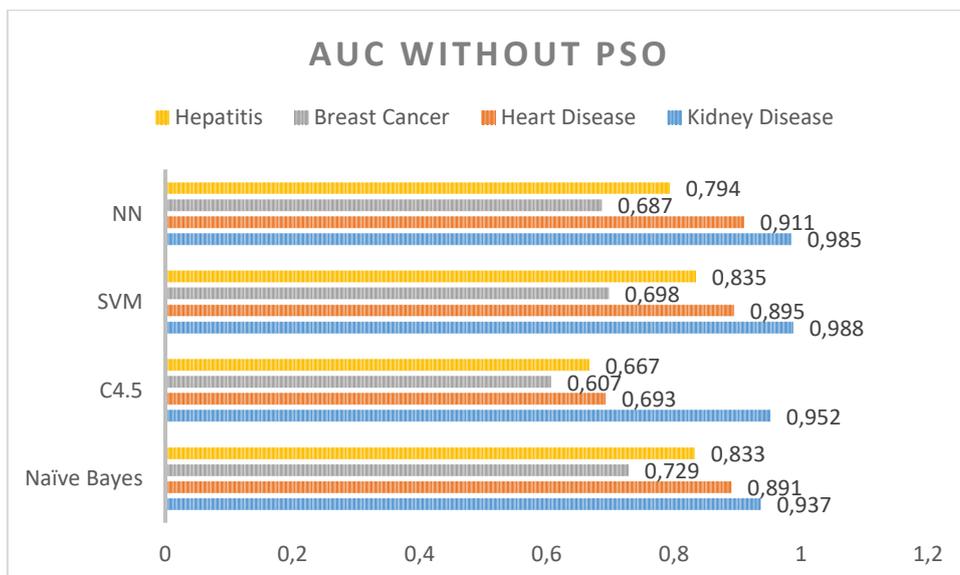


Figure 4. AUC Without PSO

From the graph above, it can be seen that the kidney disease data set is the ideal data because all the classification methods applied have an AUC value of more than 0.9, so that it can be categorized as having excellent performance.

Next experiment with the classification model using the PSO attribute weighting, the results are in line with expectations, namely PSO is proven to increase the accuracy of all classification models on all datasets as shown in the figure below:

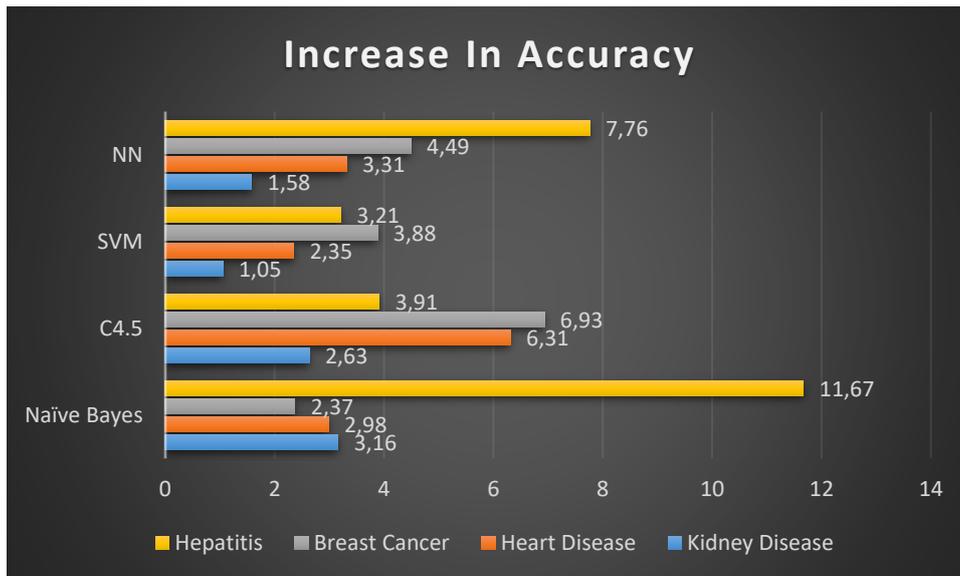


Figure 5. Increase In Accuracy

From the information in figure 5, it can be seen that all classification models have increased accuracy in all datasets. The highest increase in accuracy occurred in the Hepatitis dataset, which has 17 attributes.

Table 4. Results of Modeling Accuracy with PSO

Dataset	Without PSO				With PSO			
	Naïve Bayes	C4.5	SVM	NN	Naïve Bayes	C4.5	SVM	NN
Kidney Disease	95,26	96,3	96,3	96,8	98,42	99	97,4	98,4
Heart Disease	81,18	74,6	82,1	82,2	84,16	80,9	84,5	85,5
Breast Cancer	72,04	69,6	68,2	72,4	74,41	76,6	72,1	76,9
Hepatitis	76,79	83,7	86,7	83	88,46	87,6	89,9	90,8

Table 5. AUC results from modeling with PSO

Dataset	Without PSO				With PSO			
	Naïve Bayes	C4.5	SVM	NN	Naïve Bayes	C4.5	SVM	NN
Kidney Disease	0,937	0,95	0,99	0,99	0,99	0,99	0,98	0,99
Heart Disease	0,891	0,69	0,9	0,91	0,77	0,9	0,9	0,77
Breast Cancer	0,729	0,61	0,7	0,69	0,66	0,68	0,68	0,66
Hepatitis	0,833	0,67	0,84	0,79	0,69	0,88	0,87	0,69

The results of weighting attributes with PSO vary based on the classification method used. This is because the formula used to find the fitness function value is following the classification method used. If the attribute weight is "0," then the attribute does not affect and can be removed. The closer to the value "1", the attribute influences the classification. The data from the attribute weighting using PSO can be seen in Figure 6-9.

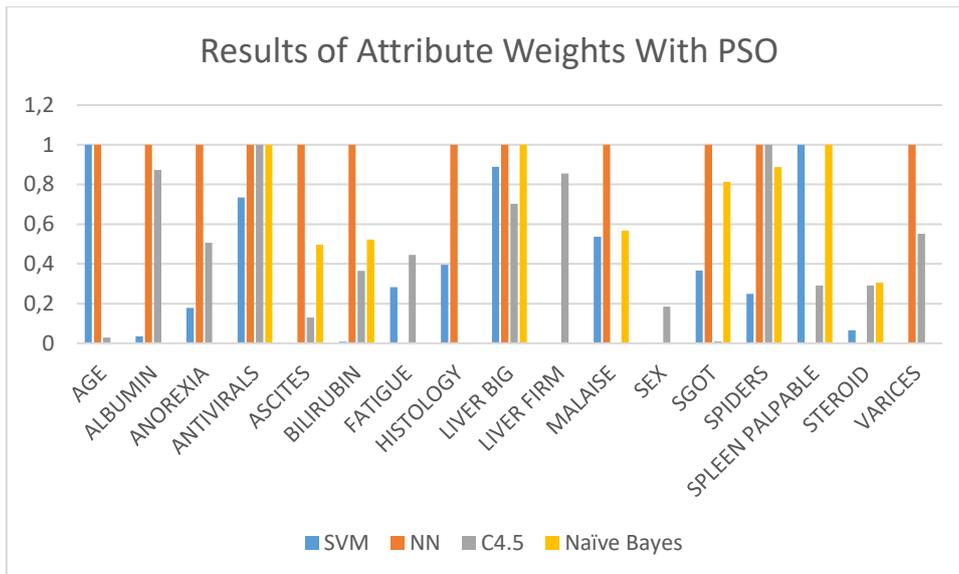


Figure 6. the results of the attribute weights with PSO from the hepatitis dataset

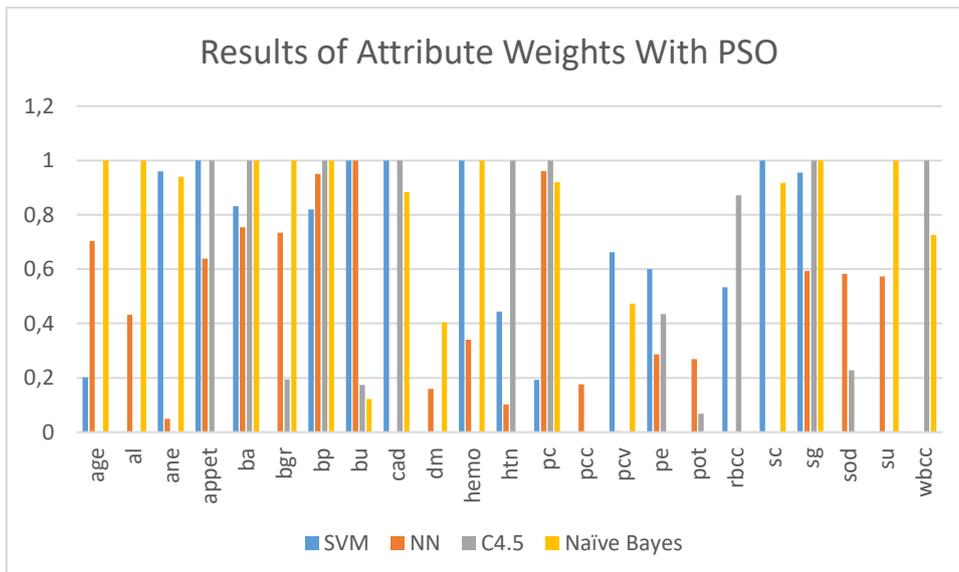


Figure 7. the results of the attribute weights with PSO from the Kidney Disease dataset

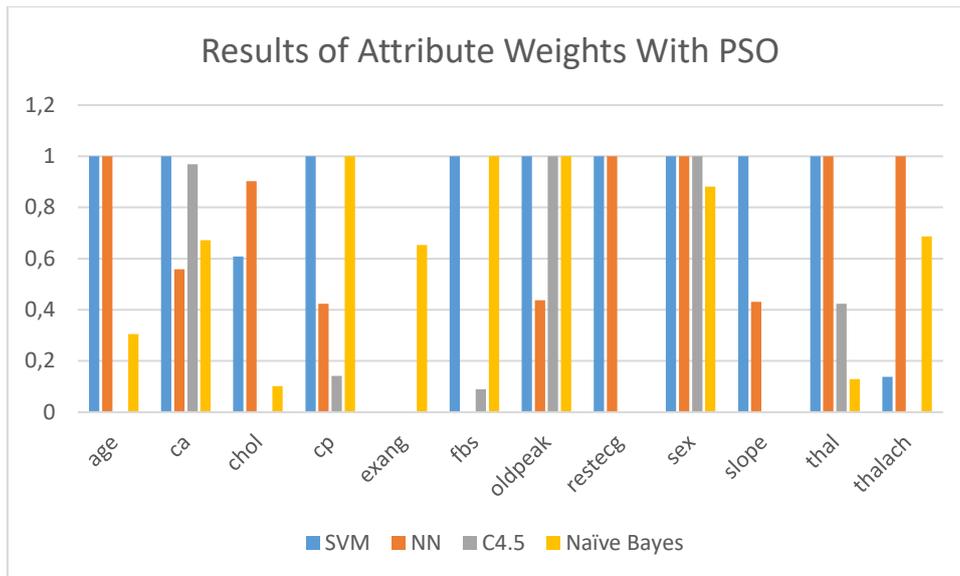


Figure 8. the results of the attribute weights with PSO from the Heart dataset

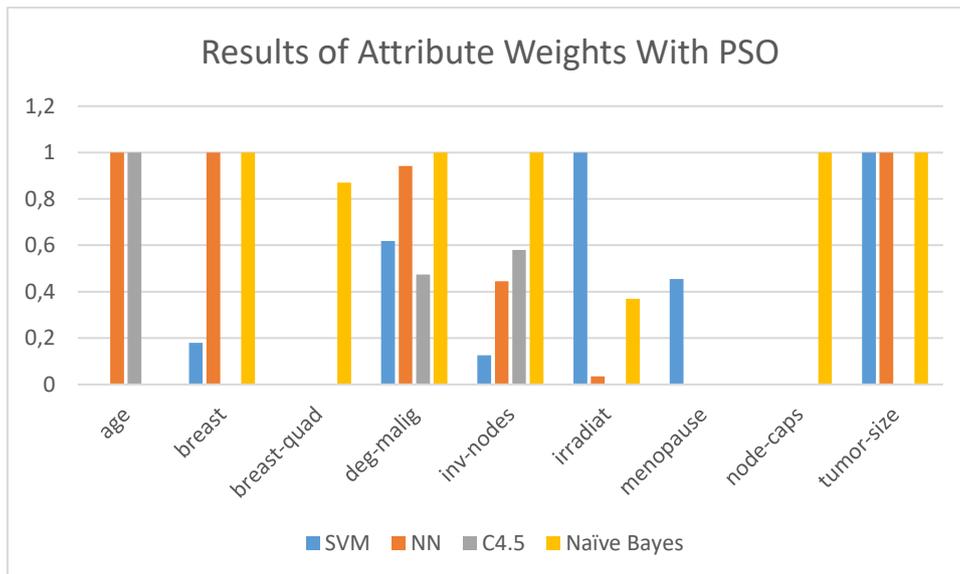


Figure 9. the results of the attribute weights with PSO from the Breast Cancer dataset

4. CONCLUSION

From the results and discussion above, it can be concluded that PSO has been shown to improve the accuracy of all Naive Bayes, C4.5, SVM, and NN classification methods used to detect disease in the four datasets used. This study's result is an analysis that leads to a Decision Support System (DSS) in diagnosing diseases, making it easier for doctors to determine the right treatment and care for their patients. Assist public health workers to determine disease prevention methods in the short and long term.

For further research, you can try to use other optimization methods or apply them to other cases.

REFERENCES

- [1] Kementerian Kesehatan Republik Indonesia, 2014, Infodatin Hipertensi, Jakarta, Indonesia.
- [2] Kementerian Kesehatan Republik Indonesia, 2016, Infodatin Bulan Peduli Kanker Payudara, Jakarta, Indonesia.
- [3] Kementerian Kesehatan Republik Indonesia, 2017, Infodatin Situasi Penyakit Ginjal Kronis, Jakarta, Indonesia.
- [4] Kementerian Kesehatan Republik Indonesia, 2014, Infodatin Situasi Kesehatan Jantung, Jakarta, Indonesia.
- [5] Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools And Techniques*. Burlington, USA: Morgan Kaufmann Publishers.
- [6] Vercellis, Carlo (2009). *Business Intelligent: Data Mining and Optimization for Decision Making*. Southern Gate, Chichester, West Sussex: John Willey & Sons, Ltd.
- [7] Bramer, Max. (, 2007). *Principles of Data Mining*. London: Springer.
- [8] Havard - Alwidian, J., Hammo, B.H. and Obeid, N., 2018. WCBA: Weighted classification based on association rules algorithm for breast cancer disease. *Applied Soft Computing*, 62, pp.536-549.
- [9] Abdel-Zaher, A.M. and Eldeib, A.M., 2016. Breast cancer classification using deep belief networks. *Expert Systems with Applications*, 46, pp.139-144.
- [10] Vijayarani, S., and Dhayanand, S., 2015. Liver disease prediction using SVM and Naïve Bayes algorithms. *International Journal of Science, Engineering and Technology Research (IJSETR)*, 4(4), pp.816-820.
- [11] Vijayarani, S., Dhayanand, S., and Phil, M., 2015. Kidney disease prediction using SVM and ANN algorithms. *International Journal of Computing and Business Research (IJCBR)*, 6(2).
- [12] Ramdania, D.R., Irfan, M., Alfarisi, F., and Nuraiman, D., 2019, December. Comparison of genetic algorithms and Particle Swarm Optimization (PSO) algorithms in course scheduling. In *Journal of Physics: Conference Series (Vol. 1402, No. 2, p. 022079)*. IOP Publishing.
- [13] Yazgan, H.R., Yener, F., Soysal, S. and Gür, A.E., 2019. Comparison Performances of PSO and GA to Tuning PID Controller for the DC Motor. *Sakarya University Journal of Science*, 23(2), pp.162-174.
- [14] Tu, C. J., Chuang, L. Y., Chang, J. Y., & Yang, C. H. (2007). Feature selection using PSO-SVM. *IAENG International Journal of computer science*, 33(1), 111-116.
- [15] Handayanna, F. "Penerapan Particle Swarm Optimization Untuk Seleksi Atribut Pada Metode Support Vector Machine Untuk Prediksi Penyakit Diabetes," Tesis Magister Ilmu Komputer. Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, 2012.
- [16] Zeniarja, J. "Opinion Mining of Movie Review On Twitter Using Support Vector Machine With Particle Swarm Optimization," Tesis Master of Computer Science. Universiti Teknikal Malaysia Melaka. 2012
- [17] Melgani, F and Bazi, Y., (2008). Classification of Electrocardiogram Signals With Support Vector Machines and Particle Swarm Optimization. *IEEE Transactions On Information Technology In Biomedicine*, Vol. 12, No. 5, September 2008
- [18] Novichasari, S.I., and Sipayung, Y.R., 2018. PSO-SVM Untuk Klasifikasi Daun Cengkeh Berdasarkan Morfologi Bentuk Ciri, Warna dan Tekstur GLCM Permukaan Daun. *Multimatrix*, 1(1).
- [19] https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease
- [20] <https://archive.ics.uci.edu/ml/datasets/heart+disease>

- [21] <https://archive.ics.uci.edu/ml/datasets/breast+cancer>
- [22] <https://archive.ics.uci.edu/ml/datasets/Hepatitis>
- [23] J. Kennedy and R. C. Eberhart. Particle swarm optimization. In Proceedings of the 1995 IEEE International Conference on Neural Networks. IEEE Service Center, Piscataway, 1995.
- [24] Abraham, A., Grosan, C., & Ramos, V. (2006). Swarm Intelligence In Data Mining. Verlag Berlin Heidelberg: Springer.
- [25] Lin, J dan Yu, J (2009). Weighted Naïve Bayes classification algorithm based on particle swarm optimization. The Yunnan University of Finance and Economics Yunnan Kunming, China.
- [26] Wu, Xindong, and Kumar, Vipin. (, 2009). The Top Ten Algorithms in Data Mining. Boca Raton: CRC Press
- [27] Larose, D. T. (2005). Discovering Knowledge in Data. New Jersey: John Willey & Sons, Inc.
- [28] Guidici, P., and Figini, S (2009). Applied Data Mining for Business and Industry. 2nd ed. United Kingdom: A John Wiley And Sons, Ltd., Publication.
- [29] Gorunescu, F. (2011). Data Mining Concepts, Models, And Techniques. Verlag Berlin Heidelberg: Springer.