



Improving Performance of KNN and C4.5 using Particle Swarm Optimization in Classification of Heart Diseases

Pareza Alam Jusia¹, Abdul Rahim², Herti Yani³, Jasmir Jasmir^{4*}

^{1,2}Informatic Engineering Universitas Dinamika Bangsa, Jambi, Indonesia

³Information System, Universitas Dinamika Bangsa Jambi Indonesia

⁴Computer Engineering, Universitas Dinamika Bangsa, Jambi, Indonesia

¹parezaalam@gmail.com, ²abdulrahim@unama.ac.id, ³adeherti@unama.ac.id, ⁴ijay_jasmir@yahoo.com

Abstract

Heart disease is a major problem that must be overcome for human life. In recent years, the volume of medical data related to heart disease has increased rapidly, and various heart disease data have collaborated with information technology such as machine learning in detecting, predicting, and classifying diseases. This research aims to improve the performance of machine learning classification methods, namely K-Nearest Neighbor (KNN) and Decision Tree (C4.5) with the particle swarm optimization (PSO) feature in cases of heart disease. In this research, a comparison was made of the performance of the PSO-based K-NN and C4.5 algorithms. Following experiments employing PSO optimization to improve the K-NN and C4.5 algorithms, the findings indicated that the K-NN algorithm performed exceptionally well with PSO, achieving an accuracy of 89.09%, precision of 89.61%, recall of 90.79%, and an AUC value of 0.935.

Keywords: machine learning; classification; improving; performance; heart disease

How to Cite: P. A. Jusia, A. Rahim, H. Yani, and J. Jasmir, "Improving Performance of KNN and C4.5 using Particle Swarm Optimization in Classification of Heart Diseases", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 8, no. 3, pp. 333 - 339, Jun. 2024. DOI: <https://doi.org/10.29207/resti.v8i3.5710>

1. Introduction

In recent years, the volume of computerized medical data has increased rapidly [1], including medical data on heart disease [2], [3]. Manipulating large amounts of data to extract knowledge from them is a complex task [4], [5]. However, machine learning techniques, which are one of the fields of artificial intelligence, can be used to explore meaningful information from medical data storage for heart disease [6]. Moreover, machine learning techniques are widely applicable for versatile applications such as classification [7], [8], clustering [9], prediction [10], [11], and others.

On the other hand, heart disease is an important problem that must be overcome for human life [12]. Although real-life consultants can predict and classify diseases with huge amounts of test data and require a long processing time, they can sometimes be wrong due to a lack of proper knowledge and experience in this regard.

Therefore, computer-based classification of heart disease can be a more effective and time-saving way for the betterment of human life [6]. Machine learning is

applied to find the initial stages of heart disease classification. Since the development of efficient classification techniques is growing rapidly for various types of classification tasks [13], [14], it is important to choose the right classification model approach for effective heart disease cases.

In the current era of technological and information development, various heart disease data have collaborated with the IT world in terms of detecting, predicting and classifying disease. Various studies have also used the latest methods according to their capacity. Samosir Amril, MS Hasibuan [15] use several machine learning methods such as Random Forest (RF), Naïve Bayes (NB) and K-Nearest Neighbor (KNN) for heart disease classification from the Cleveland Clinic Foundation dataset of 304 records with an accuracy evaluation value for each method of RF = 0.84; NB = 0.84, and KNN = 0.839. Sahar [16] also conducted research on heart disease using KNN, with an accuracy value of 67%, the dataset used was taken from UCI machine learning. Alham [17] used another machine learning, namely C4.5, for the classification of heart

disease and produced very good accuracy, namely 94.4% of the heart disease dataset at RSUD Dr. Soedarso Pontianak. Arni Sepharni[18], researched the classification of heart disease which also used the C4.5 algorithm, and produced an accuracy of 79%, using a public dataset from Kaggle.

KNN and C45 methods have been widely used as models for classifying heart disease. However, most of the performance evaluation values for the classification of heart disease above are not optimal, meaning they can still be improved with optimization methods. One method that can increase the classification performance evaluation value is Particle Swarm Optimization (PSO)[19], [20]

Challenges in classification frequently arise when encountering an extensive set of features in the dataset, with not all being utilized. Factors that can diminish classification performance include irrelevant and redundant features [21]. To enhance classification accuracy, the adoption of feature selection becomes crucial in determining the features to incorporate. Following the preprocessing phase, feature selection is employed to diminish attributes that do not contribute significantly to the improvement of classification accuracy [22]

Several previous studies have used PSO to increase the evaluation value of classification performance. As researched by Warid Yunus [23] using the PSO-based KNN method in the classification of chronic kidney disease. Uma N Dulhare [24] uses the PSO-based NB method in detecting heart disease. Tya Septiani Nurfauzia[25] conducted research by applying PSO to increase the accuracy of predictions of hepatitis diagnosis using the Naive Bayes method. Lis Saumi Ramdhani [26] also conducted research by applying PSO to increase the accuracy of predicting hepatitis diagnosis using the C4.5 method. From several studies above, especially the classification of diseases discussed by Arni Sepharini, and the classification of diseases discussed by Sahar, the author tries to combine the two studies into one study and at the same time contributes to this research, namely by utilizing the Particle Swarm Optimization optimization method to increase the value evaluation of the performance of KNN and C4.5 classification on heart disease datasets taken from the same source, namely the Kaggle public dataset. Therefore, the aim of this research is to increase the performance evaluation value of KNN and C4.5 classification with PSO features and compare the performance evaluation value of KNN and C4.5 without PSO and with PSO. , especially on the heart disease dataset taken from the Kaggle public dataset.

2. Research Methods

The flow of the classification process carried out in this research can be seen in Figure 1. In Figure 1 it is explained that the first step is dataset selection. The dataset used in this research is the public heart failure

dataset obtained on the page <https://www.kaggle.com/datasets/fedoriano/heart-failure-prediction>.

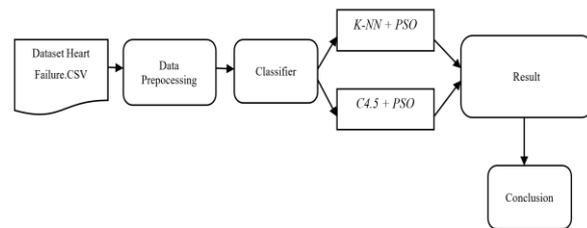


Figure 1. Research Framework

The number of data records in the dataset is 918 data records with 11 attributes and 1 class attribute, with 2 class labels, namely normal and heart failure, where this dataset consists of 508 heart failure patient data and 410 normal patient data. A snapshot of the dataset can be seen in Figure 2.

Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAn	Oldpeak	ST_Slope	Label
40	M	ATA	140	289	0	Normal	172	N	0	Up	Normal
49	F	NAP	160	180	0	Normal	156	N	1	Flat	Heart Failure
37	M	ATA	130	283	0	ST	98	N	0	Up	Normal
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	Heart Failure
54	M	NAP	150	195	0	Normal	122	N	0	Up	Normal
39	M	NAP	120	339	0	Normal	170	N	0	Up	Normal
45	F	ATA	130	237	0	Normal	170	N	0	Up	Normal
54	M	ATA	110	208	0	Normal	142	N	0	Up	Normal
37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	Heart Failure
48	F	ATA	120	284	0	Normal	120	N	0	Up	Normal
37	F	NAP	130	211	0	Normal	142	N	0	Up	Normal
58	M	ATA	136	164	0	ST	99	Y	2	Flat	Heart Failure
39	M	ATA	120	204	0	Normal	145	N	0	Up	Normal
49	M	ASY	140	234	0	Normal	140	Y	1	Flat	Heart Failure
42	F	NAP	115	211	0	ST	137	N	0	Up	Normal
54	F	ATA	120	273	0	Normal	150	N	1.5	Flat	Normal
38	M	ASY	110	196	0	Normal	166	N	0	Flat	Heart Failure
43	F	ATA	120	201	0	Normal	165	N	0	Up	Normal
60	M	ASY	100	248	0	Normal	125	N	1	Flat	Heart Failure
36	M	ATA	120	267	0	Normal	160	N	3	Flat	Heart Failure
43	F	TA	100	223	0	Normal	142	N	0	Up	Normal
44	M	ATA	120	184	0	Normal	142	N	1	Flat	Normal
49	F	ATA	124	201	0	Normal	164	N	0	Up	Normal
44	M	ATA	150	288	0	Normal	150	Y	3	Flat	Heart Failure
40	M	NAP	130	215	0	Normal	138	N	0	Up	Normal
36	M	NAP	130	209	0	Normal	178	N	0	Up	Normal

Figure 2. Snapshot of Dataset

Next, preprocessing is carried out, namely cleaning the data and preparing the data before processing it to the classification stage, including checking inconsistent data and correcting data errors. The next stage is the RapidMiner tools classification stage, data that is ready to be processed continues to the testing stage. The method used in this research is K-NN and C4.5 based on PSO (Particle Swarm Optimization), or the tests that will be carried out are classification with the K-NN and C4.5 algorithms without PSO and classification with the K-NN and C45 algorithms using PSO. For experimental purposes, the data is shared. In this study, the dataset was divided using a ratio of 70:30 or 70% of the data as training data and 30% of the data as test data.

The K-NN algorithm falls within the category of instance-based learning and is classified as one of the lazy learning techniques. In K-NN, the process involves identifying a set of k data points in the training dataset that closely resemble the objects in the new or testing data[27][28]. The effectiveness of the K-NN algorithm is significantly impacted by the existence of irrelevant features, whether they are included or excluded, and the proper alignment of their weights based on their importance in the classification process.[29]

The K-NN approach strives to categorize new objects by considering their attributes and training samples. In the process of selecting an attribute comprising n neighbors (commonly referred to as k), the testing phase determines the k parameter based on the optimal k value identified during training. This optimal k value is achieved through a process of trial and error [30].

The C4.5 algorithm is an algorithm that is commonly used, especially in the machine learning area. The C4.5 algorithm is an algorithm used to classify data that has numeric attributes. The results obtained from implementing C4.5 are the result of developing the ID3 algorithm[31]. C4.5 has advantages in terms of handling missing values in data, as well as dealing with continuous data. Saruni Dwiasnati et. al. [32] states that the C4.5 Algorithm is a model produced by forming a decision tree.

According to Seruni Dwiasnati et. al [32], Particle Swarm Optimization (PSO) is an optimization algorithm that can be used to increase the performance evaluation value of classification algorithms so that more optimal results are known. Warid Yunus [23] Particle Swarm Optimization (PSO) is a straightforward optimization method designed for the application and adjustment of multiple parameters. PSO is widely used to solve weight optimization and feature selection problems. Ujang Juhardi [20] stated that In 1995, Kennedy and Eberhart introduced the Particle Swarm Optimization (PSO) technique, inspired by the collective behaviour observed in animal groups, like the coordinated movement of a flock of birds.

The PSO algorithm mimics the social behaviour observed in organisms, where individual actions and the impact of others in a group play a crucial role. The term "particle" can be likened to a bird within a flock. Each individual or particle acts independently, relying on its own intelligence while being influenced by the overall collective behaviour of the group. When a particle or bird discovers an optimal or shorter path to a food source, the entire group can swiftly follow that path, even if they are situated at a distance from the original finding[33].

Thus it is known that Particle Swarm Optimization (PSO) was developed based on the following model[34]: When a bird approaches its target or destination, it quickly sends information to certain birds; Other birds will follow the direction of the food but not directly; There is a component that depends on each bird's mind, namely its memory of what has passed in the previous time.

3. Results and Discussions

To observe the accuracy comparison results in this study, four experiments were conducted. These included experiments using the KNN method without the PSO feature, followed by the KNN method with the PSO feature, then the C4.5 method without PSO, and

finally, the C4.5 method with PSO. The tool employed for these experiments was RapidMiner.

3.1 KNN without PSO

The following is an overview of the K-NN (K-Nearest Neighbor) algorithm testing carried out using RapidMiner:

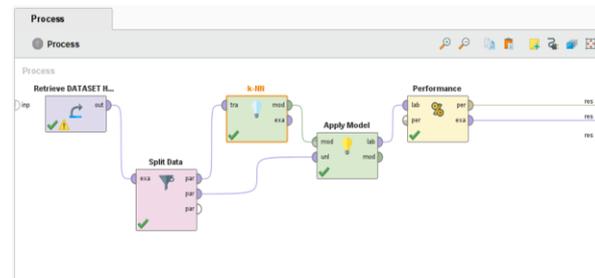


Figure 3. Classification process of KNN without PSO

Figure 3 explains the process chart for obtaining classification performance values without KNN without using the PSO feature, starting from calling the dataset, namely the heart disease dataset, then splitting the data, namely dividing the training data and test data with a ratio of 70:30, after the data is split, the process continues to the next stage. The method used is KNN. The next step is to apply the model and after that get the evaluation results. The results obtained can be seen in the confusion matrix as shown in Figure 4.

accuracy: 69.09%			
	true Normal	true Heart Failure	class precision
pred Normal	86	48	64.18%
pred Heart Failure	37	104	73.76%
class recall	69.92%	68.42%	

Figure 4. Confusion Matrix K-NN without PSO

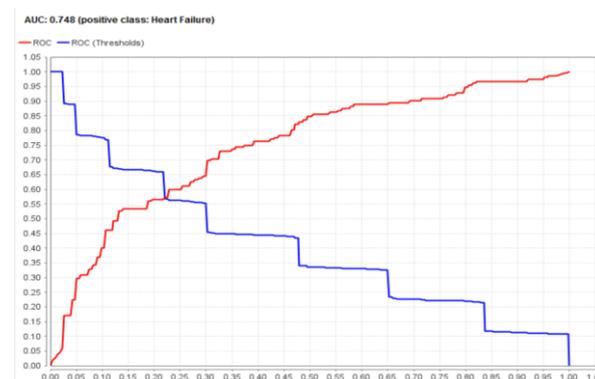


Figure 5. Result in AUC K-NN without PSO in PSO graph

Based on the confusion matrix in Figure 4 with a value of $K=9$, it can be seen that in the "Normal" class there are 86 data which are correctly predicted to be in the Normal category, and 48 data which are incorrectly categorized as Normal should be Heart Failure. In the "Heart Failure" class, 37 data were incorrectly categorized as Heart Failure which should be Normal, and 104 data were correctly predicted as Heart Failure.

So we obtained an accuracy value of 69.09%, precision of 73.76%, recall of 68.42% and AUC value = 0.748

which is included in the Fair Classification category. These values can be seen as shown in Figure 5

3.2 KNN with PSO

The following is an overview of the PSO (Particle Swarm Optimization) based K-NN algorithm testing carried out using RapidMiner:

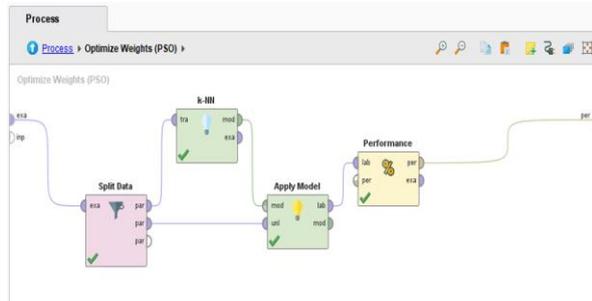


Figure 6. Classification Process of KNN using PSO

Figure 6 depicts the procedural flowchart for acquiring classification performance metrics utilizing the KNN method, commencing with the division of data into training and testing subsets with a 70:30 ratio. Subsequent to the data partitioning, the process proceeds with the utilization of the KNN method. Following this, the model is executed, culminating in the retrieval of evaluation results.

At the testing stage using PSO (Particle Swarm Optimization) the attribute values in the dataset will be weighted to see which attributes have good and relevant weights which can increase the accuracy value. The results of weighting with PSO (Particle Swarm Optimization) can be seen in Table 1.

Table 1. Results of PSO Weighting in K-NN

Attribute	Weight
ExerciseAngina	1
Oldpeak	1
ST_Slope	1
FastingBS	0,520
ChestPainType	0,301
Age	0,235
RestingECG	0,032
Sex	0
RestingBP	0
Cholesterol	0
MaxHR	0

Table 1 is the result of attribute selection carried out using PSO in the K-NN method, obtaining the 4 lowest attributes, namely with a weight of 0 for the Sex, RestingBP, Cholesterol and MaxHR attributes, which means that these attributes do not have any influence on accuracy. obtained. So from the attribute selection results, there are 7 attributes that have a weight of more than 0, which means these attributes have an influence on the accuracy value in the dataset being tested. Following the preprocessing of the dataset, RapidMiner is employed to perform calculations for assessing the accuracy of the applied method. The outcome includes the derivation of a confusion matrix, as depicted in Figure 7.

accuracy: 89.09%

	true Normal	true Heart Failure	class precision
pred. Normal	107	14	88.43%
pred. Heart Failure	16	138	89.61%
class recall	88.99%	90.79%	

Figure 7. Confusion Matrix K-NN using PSO

Using the confusion matrix depicted in Figure 7, when employing the K-NN method with a k-value of 9 and integrating PSO, it is evident that within the "Normal" class, 107 data points are correctly classified as Normal, but 14 data points are mistakenly assigned to the Normal category instead of Heart Failure. In the "Heart Failure" class, 16 data were incorrectly categorized as Heart Failure which should be Normal, and 138 data were correctly predicted as Heart Failure.

So the accuracy value was obtained at 89.09%, precision at 89.61%, recall at 90.79% and AUC value = 0,935 which was included in the Excellent Classification category. These values can be seen as shown in Figure 8

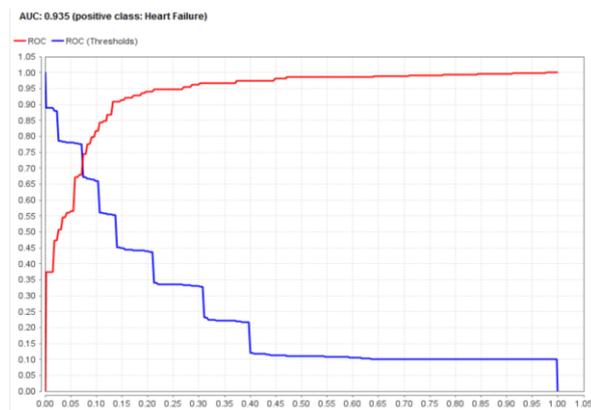


Figure 8. Result in AUC K-NN using PSO in ROC graph

3.3 C4.5 without PSO

The following is an overview of the C4.5 algorithm testing carried out using RapidMiner:

The diagram depicted in Figure 9 outlines the procedural flow for deriving classification performance metrics without employing C45 or utilizing the PSO functionality. Initially, it involves accessing the dataset, specifically the heart disease dataset. Subsequently, the data is partitioned, segregating it into training and testing subsets with a 70:30 ratio. Once the data is divided, the process advances to the subsequent phase, employing the Decision Tree (C4.5) methodology. Following this, the model is executed, and thereafter, the evaluation outcomes are obtained. The results obtained can be seen in the confusion matrix as shown in Figure 10:

Based on the confusion matrix in Figure 10 using the C4.5 method, it can be seen that in the "Normal" class there are 92 data correctly predicted for the Normal category, and 23 data incorrectly categorized as Normal should be Heart Failure. In the "Heart Failure" class, 31

data were incorrectly categorized as Heart Failure which should be Normal, and 129 data were correctly predicted as Heart Failure.

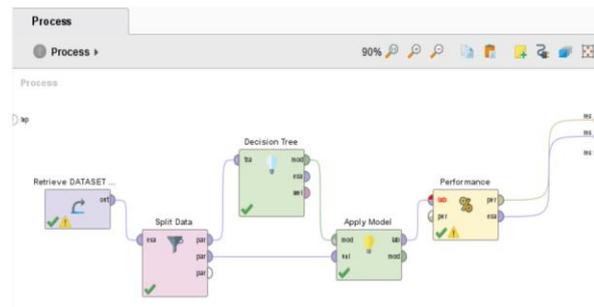


Figure 9. Classification Process C4.5 without PSO

accuracy: 80.36%

	true Normal	true Heart Failure	class precision
pred. Normal	92	23	80.00%
pred. Heart Failure	31	129	80.62%
class recall	74.80%	84.87%	

Figure 10. Confusion Matrix C4.5 without PSO

So we obtained an accuracy value of 80.36%, precision of 80.62%, recall of 84.87% and AUC value = 0.788 which is included in the Fair Classification category. These values can be seen as shown in Figure 11.



Figure 11. Result in AUC C4.5 without PSO in ROC graph

3.4 C4.5 with PSO

The following is an overview of the PSO (Particle Swarm Optimization) based C4.5 algorithm testing carried out using RapidMiner:

Figure 12 illustrates the process flowchart for acquiring classification performance metrics, beginning with the division of data into training and testing sets at a ratio of 70:30. Following the data split, the procedure advances to employing the Decision Tree (C4.5) method. Subsequently, the model is implemented, leading to the retrieval of evaluation results.

At the testing stage using PSO (Particle Swarm Optimization) the attribute values in the dataset will be weighted to see which attributes have good and relevant weights which can increase the accuracy value. The results of weighting with PSO (Particle Swarm Optimization) can be seen in Table 2.

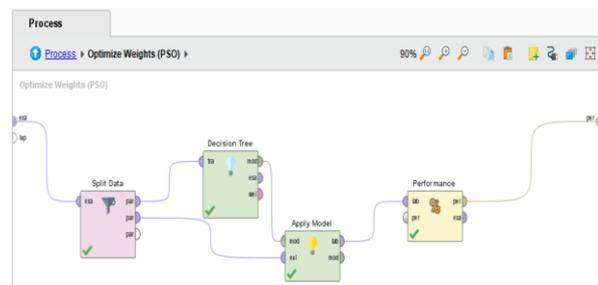


Figure 12. Classification Process C4.5 using PSO

Table 2. PSO Weighting Results in C4.5

Attribute	Weight
Age	1
ChestPainType	1
RestingBP	1
MaxHR	1
Oldpeak	1
Sex	0,487
ST_Slope	0,300
Cholesterol	0,103
ExerciseAngina	0,076
FastingBS	0
RestingECG	0

The results of the attribute selection carried out using PSO in the C4.5 method obtained the 2 lowest attributes, namely with a weight of 0 for the FastingBS and RestingECG attributes, which means that these attributes do not have any influence on the accuracy obtained. So from the attribute selection results, there are 9 attributes that have a weight of more than 0, which means these attributes have an influence on the accuracy value in the dataset being tested.

From the dataset that has been processed, calculations are then carried out using RapidMiner to determine the accuracy of the method used, a confusion matrix is obtained as shown in Figure 13.

accuracy: 86.91%

	true Normal	true Heart Failure	class precision
pred. Normal	101	14	87.83%
pred. Heart Failure	22	138	86.25%
class recall	82.11%	90.79%	

Figure 13. Confusion Matrix C4.5 using PSO

Based on the confusion matrix in Figure 13 which uses the C4.5 method based on PSO, it can be seen that in the “Normal” class there are 101 data correctly predicted for the Normal category, and 14 data incorrectly categorized as Normal should be Heart Failure. In the “Heart Failure” class, 22 data were incorrectly categorized as Heart Failure which should be Normal, and 138 data were correctly predicted as Heart Failure.

So we get an accuracy value of 86.91%, precision of 86.25%, recall of 90.79% and AUC = 0.855 which is included in the Good Classification category. These values can be seen as shown in Figure 14.

To increase the classification performance evaluation value, we set special PSO parameters in the Population Size, Max Generation, Inertia Weight, Local Best

Weight and Global Best Weight sections. The aim of setting parameters is to be able to reduce false positive values and produce good evaluation values if the false positives are reduced. Apart from that, we added the Optimize Weights (PSO) operator, an optimization method that helps strengthen the evaluation value. The parameter values are Population Size = 5, Max of Generation = 30, Inertia Weight = 1.0, Local Best Weight = 1.0, and Global Best Weight = 1.0. We give this value the same for KNN and C4.5.

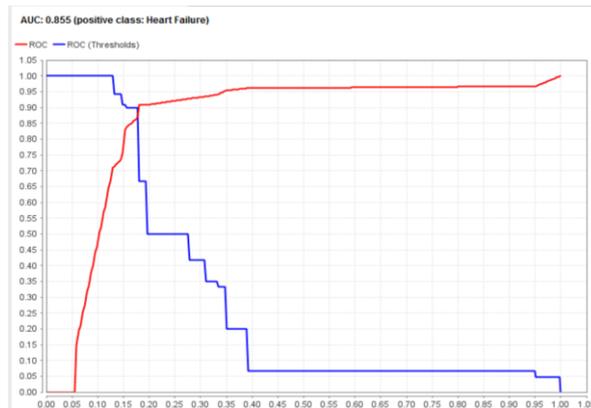


Figure 14. Result in AUC C4.5 using PSO in ROC graph

Testing using a rapid miner to test the K-NN and C4.5 methods has been carried out in the previous process to produce accuracy, precision, recall and AUC values which can be shown in Table 3.

Table 3. Classification Results Using Rapid Miner

Algorithm	Accuracy	Precision	Recall	AUC
K-NN	69,09	73,76	68,42	0,748
K-NN + PSO	89,09	89,61	90,79	0,935
C4.5	80,36	80,62	84,67	0,788
C4.5 + PSO	86,91	86,25	90,79	0,855

Table 3 shows a comparative comparison of the accuracy values obtained from the two classification algorithms used, namely K-NN and C4.5 using PSO and without using PSO. From the results obtained, it can be seen that both classification methods have increased the accuracy value using PSO. Compared to the results without using PSO in calculations using RapidMiner, especially the K-NN algorithm which experienced an increase in accuracy of 20% from 69.09% to 89.09% in testing with RapidMiner, for the C4.5 algorithm there was also an increase in accuracy of 6.55% from 80.36% to 86.91% in testing with RapidMiner. The highest accuracy value was obtained in the PSO-based K-NN algorithm carried out on RapidMiner, namely 89.09%.

Even though it uses a simple tool such as a rapid miner, the model obtained produces good scores in the classification of heart disease. Several PSO parameters that are set are able to increase the evaluation value of classification performance, especially the value of accuracy, precision-recall and f1-score.

4. Conclusions

The conclusion of this research states that the PSO

feature is able to increase the performance evaluation value of KNN and C4.5 classification of heart disease datasets using the RapidMiner tool. The K-NN algorithm when used with PSO in RapidMiner achieved the highest accuracy of 89.09%, precision of 89.61%, recall of 90.79%, and AUC value of 0.935. This performance is considered very good in terms of classification accuracy. The application of Particle Swarm can increase the accuracy resulting from the K-NN and C4.5 algorithms in classifying heart disease data where the results obtained are higher compared to the values obtained before implementing PSO. Even though it uses a simple tool such as a rapid miner, the model obtained produces a good score in the classification of heart disease. Several PSO parameters determined are able to increase the evaluation value of classification performance, especially the value of accuracy, precision, recall and f1-score.

References

- [1] A. Razzaque and D. A. Badholia, "PCA based feature extraction and MPSO based feature selection for gene expression microarray medical data classification," *Meas. Sensors*, vol. 31, no. November 2023, p. 100945, 2024, doi: 10.1016/j.measen.2023.100945.
- [2] B. Domenech *et al.*, "Valvular Heart Disease Epidemics," *J. Heart Valve Dis.*, vol. 25, no. 1, pp. 1–7, 2016.
- [3] A. Darmawahyuni and S. Nurmaini, "Coronary Heart Disease Interpretation Based on Deep Neural Network," *Comput. Eng. Appl.*, vol. 8, no. 1, 2019.
- [4] K. Vassakis, E. Petrakis, and I. Kopanakis, "Big data analytics: Applications, prospects and challenges," *Lect. Notes Data Eng. Commun. Technol.*, vol. 10, pp. 3–20, 2018, doi: 10.1007/978-3-319-67925-9_1.
- [5] S. Athmaja, M. Hanumanthappa, and V. Kavitha, "A survey of machine learning algorithms for big data analytics," *Proc. 2017 Int. Conf. Innov. Information, Embed. Commun. Syst. ICIIECS 2017*, vol. 2018-Janua, no. June, pp. 1–4, 2018, doi: 10.1109/ICIIECS.2017.8276028.
- [6] C. W. Tsao *et al.*, *Heart Disease and Stroke Statistics-2022 Update: A Report from the American Heart Association*, vol. 145, no. 8, 2022. doi: 10.1161/CIR.0000000000001052.
- [7] Rusydi Umar, Sunardi, and M. N. A. Nuriyah, "Comparing the Performance of Data Mining Algorithms in Predicting Sentiments on Twitter," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 4, pp. 817–823, 2023, doi: 10.29207/resti.v7i4.4931.
- [8] J. Jasmir, W. Riyadi, S. R. Agustini, Y. Arvita, D. Meisak, and L. Aryani, "Bidirectional Long Short-Term Memory and Word Embedding Feature for," *J. RESTI (Rekayasa Sist. Dan Teknol. Informasi)*, vol. 6, no. 4, pp. 505–510, 2022, [Online]. Available: <https://jurnal.iaii.or.id/index.php/RESTI/article/view/4005/606>
- [9] T. Sutanto and R. Nayak, "Fine-grained document clustering via ranking and its application to social media analytics," *Soc. Netw. Anal. Min.*, 2018, doi: 10.1007/s13278-018-0508-z.
- [10] J. Chambua and Z. Niu, "Review text based rating prediction approaches: preference knowledge learning, representation and utilization," *Artif. Intell. Rev.*, vol. 54, no. 2, pp. 1171–1200, 2021, doi: 10.1007/s10462-020-09873-y.
- [11] J. Jasmir, D. Z. Abidin, E. Rasywir, and P. A. Jusia, "Prediksi Mahasiswa Drop Out dengan menggunakan Algoritma Klasifikasi Data Mining," *Pros. Annu. Res. Semin. 2018*, vol. 4, no. 1, pp. 978–979, 2018.
- [12] D. Prihatiningsih and T. Sudyasih, "Perawatan Diri Pada Pasien Gagal Jantung," *J. Pendidik. Keperawatan Indones.*, vol. 4, no. 2, 2018, doi: 10.17509/jpki.v4i2.13443.
- [13] J. Jasmir, W. Riyadi, and P. A. Jusia, "Feature Extraction for Improvement Text Classification of Spam YouTube," *J. RESTI*

- (*Rekayasa Sist. dan Teknol. Informasi*), vol. 5, no. 158, pp. 1371–1377, 2023.
- [14] Y. N. Nanik Rahmawati, Jasmir Jasmir, “Klasifikasi Kondisi Gizi Balita Menggunakan Metode Naive Bayes (Studi Kasus Posyandu Melati IV),” *J. Ilm. Mhs. Tek. Inform.*, vol. 2, no. 3, 2020.
- [15] A. Samosir, M. S. Hasibuan, W. E. Justino, and T. Hariyono, “Komparasi Algoritma Random Forest, Naive Bayes dan K-Nearest Neighbor Dalam klasifikasi Data Penyakit Jantung,” *Pros. Semin. Nas. Darmajaya*, vol. 1, no. 0, pp. 214–222, 2021, [Online]. Available: <https://jurnal.darmajaya.ac.id/index.php/PSND/article/view/2955>
- [16] S. S. Tabrizi and N. Cavus, “A hybrid KNN-SVM model for Iranian license plate recognition,” *Procedia - Procedia Comput. Sci.*, vol. 102, no. August, pp. 588–594, 2016, doi: 10.1016/j.procs.2016.09.447.
- [17] S. R. J. I. Alham, “Sistem Diagnosis Penyakit Jantung Koroner Dengan Menggunakan Algoritma C4.5 Berbasis Website (Studi Kasus: RSUD Dr. Soedarso Pontianak),” *Petir*, vol. 14, no. 2, pp. 214–222, 2021, doi: 10.33322/petir.v14i2.1338.
- [18] A. Sefhami, I. E. Hendrawan, and C. Rozikin, “Klasifikasi Penyakit Jantung dengan Menggunakan Algoritma C4.5,” *STRING (Satuan Tulisan Ris. dan Inov. Teknol.)*, vol. 7, no. 2, p. 117, 2022, doi: 10.30998/string.v7i2.12012.
- [19] K. Kurniabudi, A. Harris, and A. E. Mintaria, “Komparasi Information Gain, Gain Ratio, CFs-Bestfirst dan CFs-PSO Search Terhadap Performa Deteksi Anomali,” *J. Media Inform. Budidarma*, vol. 5, no. 1, p. 332, 2021, doi: 10.30865/mib.v5i1.2258.
- [20] U. Juhardi, “Optimalisasi Penjualan Motor Menggunakan Algoritma Particle Swarm Optimization (PSO),” vol. 15, no. 2, 2019.
- [21] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, and F. E. Alsaadi, “Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods,” no. xxxx, 2019.
- [22] Kurniabudi, D. Stiawan, Darmawijoyo, M. Y. Bin Bin Idris, A. M. Bamhdi, and R. Budiarto, “CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection,” *IEEE Access*, vol. 8, pp. 132911–132921, 2020, doi: 10.1109/ACCESS.2020.3009843.
- [23] W. Yunus, “Algoritma K-Nearest Neighbor Berbasis Particle Swarm Optimization Untuk Prediksi Penyakit Ginjal Kronik,” *J. Tek. Elektro CosPhi*, vol. 2, no. 2, pp. 51–55, 2018, [Online]. Available: <https://cosphijournal.unisan.ac.id/index.php/cosphihome/article/view/43%0Ahttps://cosphijournal.unisan.ac.id/index.php/co>
- sphihome/article/download/43/20
- [24] U. N. Dulhare, “Prediction system for heart disease using Naive Bayes and particle swarm optimization,” *Biomed. Res.*, vol. 29, no. 12, pp. 2646–2649, 2018, doi: 10.4066/biomedicalresearch.29-18-620.
- [25] T. Septiani Nurfauzia Koeswara, M. Sukrisno Mardiyanto, and M. Abdul Ghani, “Penerapan Particle Swarm Optimization (PSO) Dalam Pemilihan Atribut Untuk Meningkatkan Akurasi Prediksi Diagnosis penyakit Hepatitis Dengan Metode Naive Bayes,” *J. Speed – Sentra Penelit. Eng. dan Edukasi*, vol. 12, no. 1, pp. 1–10, 2020.
- [26] L. S. Ramdhani, “Penerapan Particle Swarm Optimization (PSO) untuk Seleksi Atribut dalam Meningkatkan Akurasi Prediksi Diagnosis Penyakit Hepatitis dengan Metode Algoritma C4.5,” *Swabumi*, vol. 4, no. 1, pp. 1–15, 2016.
- [27] J. Jasmir, S. Nurmaini, and B. Tutuko, “Fine-grained algorithm for improving knn computational performance on clinical trials text classification,” *Big Data Cogn. Comput.*, vol. 5, no. 4, 2021, doi: 10.3390/bdcc5040060.
- [28] M. Azam, T. Ahmed, F. Sabah, and M. I. Hussain, “Feature Extraction based Text Classification using K-Nearest Neighbor Algorithm,” *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 18, no. 12, pp. 95–101, 2018, [Online]. Available: http://paper.ijcsns.org/07_book/201812/20181213.pdf
- [29] J. Jasmir, D. Z. Abidin, S. Nurmaini, and R. F. Malik, “Penerapan Metode K-Nearest Neighbor dalam Memprediksi Masa Studi Mahasiswa (Studi Kasus : Mahasiswa STIKOM Dinamika Bangsa),” *Pros. Annu. Res. Semin. 2017*, vol. 3, no. 1, pp. 133–138, 2017.
- [30] M. Fayaz, A. Khan, J. U. Rahman, A. Alharbi, M. I. Uddin, and B. Alouffi, “Ensemble machine learning model for classification of spam product reviews,” *Complexity*, vol. 2020, 2020, doi: 10.1155/2020/8857570.
- [31] V. Junita and F. A. Bachtiar, “Klasifikasi Aktivitas Manusia menggunakan Algoritme Decision Tree C4.5 dan Information Gain untuk Seleksi Fitur,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 10, pp. 9426–9433, 2020, [Online]. Available: <http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/6446>
- [32] S. Dwiasnati and Y. Devianto, “Optimasi Prediksi Keputusan Calon Nasabah Potensial Berbasis Particle Swarm Optimization,” vol. 6, no. 2, pp. 286–292, 2019.
- [33] D. Wang, D. Tan, and L. Liu, “Particle swarm optimization algorithm: an overview,” *Soft Comput.*, vol. 22, no. 2, pp. 387–408, 2018, doi: 10.1007/s00500-016-2474-6.
- [34] J. C. Bansal, *Particle Swarm Optimization*. Springer International Publishing, 2019. doi: 10.1007/978-3-319-91341-4.