

# Performance Comparison of K-Nearest Neighbor and Decision Tree C4.5 by Utilizing Particle Swarm Optimization for Prediction of Liver Disease

Zainatul Fadilah, Murnawan

**Abstract**— As time goes by, the data owned by a sector will accumulate if not used properly. To process large-scale data, we need a technique called data mining. Data mining can process large-scale data quickly so that many sectors use this technique for classification, clustering, etc. based on the cases they have. The sector that often applies data mining is the health sector. Usually, activities often found in the health sector are activities to diagnose a disease in patients. This study will discuss performance comparison between two data mining algorithms, namely K-Nearest Neighbor (KNN) and Decision Tree C4.5. Then, the algorithms combine with an optimization, namely Particle Swarm Optimization (PSO). The purpose of this study is to compare the performance between the two algorithms to obtain the best performance, which can later be used as a decision to predict disease.

The results obtained indicate that Decision Tree C4.5 with PSO has a better level of performance than KNN with PSO, so Decision Tree C4.5 with PSO can be used in predicting disease. The results obtained are the accuracy value of Decision Tree C4.5 with PSO is 91.26%, and the AUC value is 0.935. Then, Decision Tree C4.5 with PSO in processing data only takes 25 seconds of execution. In this study, to ensure that the two algorithms have significant differences, a trial use t-Test conduct. The results obtained  $\alpha = 0.007$ , meaning that the two algorithms have a significant difference, which the parameter requirements, namely  $\alpha < 0.050$ .

**Keywords**— Data Mining, K-Nearest Neighbor, Decision Tree C4.5, Particle Swarm Optimization, Predict Disease.

## I. INTRODUCTION

The liver (in Greek: *Hepar*) is one of the most important organs in the human body, but it is susceptible to damage if not treated properly. Liver damage by inflammation is mostly the result of viral infections, alcohol exposure, drug or chemical poisoning [1]. If this disease attacks and damages the liver, the body's ability will slowly decrease, especially the ability to neutralize toxins that enter the body, it will harm the body if not treated immediately [2]. Therefore, this disease cannot underestimate because the

liver plays a role in maintaining human life. Cases that are often encountered regarding this liver disease are almost all people experience delays in treatment because they only check their health when liver function is not functioning properly, this is due to early symptoms that are not realized by many people. Based on this statement, it is necessary to treat it as early as possible before the liver disease gets worse. The way is to do regular health checks and pay attention to unusual symptoms in the body such as changes in body color and yellow eyes, enlarged stomach, and bad breath.

Currently, the development of technology and information in the health sector has widely applied, one of which the application of data mining to help predict a disease. With data mining, health workers can consider decisions regarding disease diagnoses given to patients. The basis used in predicting disease is the past data of patients undergoing routine health checks. The application of data mining is expected to assist health workers in increasing the efficiency of diagnoses given to patients.

The following is the application of data mining in the health sector from previous researchers: Taghifirul Azhima Yoga Siswa, et.al (2018) explained that the application of Particle Swarm Optimization (PSO) produced significant results in providing improved performance on the Naïve Bayes and K-Nearest Neighbor algorithms. The purpose of this study is to apply and compare the best performance of the data mining classification method of the Naïve Bayes and K-Nearest Neighbor algorithms with PSO to detect breast cancer. Based on testing, the Naïve Bayes with PSO has a great value compared to the K-Nearest Neighbor with PSO an acquisition value of 0.978 [3]. Tya Septiani Nurfauzia Koeswara, et.al (2020) explained the application of Particle Swarm Optimization (PSO). This study states that the application of PSO can select attributes on Naive Bayes and the quality of diagnostic accuracy of hepatitis is better than using the individual Naive Bayes method. Results Based on the research, the accuracy value of Naïve Bayes is 84.85%, while the accuracy value of Naive Bayes optimization using PSO is 92.50%. While the evaluation using the ROC curve obtained for the Naïve Bay algorithm is feasible 0.894 with a Good Classification diagnosis level and the Naïve Bayes optimization using PSO is feasible 0.941 with an Excellent Class diagnosis level, the difference in UC values is 0.047 [4].

Manuscript received June 16, 2021.

Zainatul Fadilah is with Department of Information System, Widyatama University, West Java, Indonesia, (email: zainatul.fadilah@widyatama.ac.id)

Murnawan is with Department of Information System, Widyatama University, West Java, Indonesia, (email: murnawan@widyatama.ac.id)

Various data mining methods can use as needed. However, this study only discusses the classification method by comparing two algorithms, namely K-Nearest Neighbor (KNN) and Decision Tree C4.5. The choice of the KNN algorithm and the Decision Tree C4.5 algorithm was due to the absence of previous researchers who compared the performance results of the two algorithms. In addition, this comparison can facilitate decision-makers in predicting liver disease. To improve accuracy in predicting liver disease, these two algorithms are combined with Particle Swarm Optimization (PSO) to help optimize accuracy performance and determine which algorithm is more suitable to be applied.

## II. LITERATURE REVIEW

### A. Data Mining

Research on data mining is in great demand by technology and information science experts. The research was carried out by utilizing a set of data to find new patterns and knowledge. The data that initially accumulates in a database reaches hundreds, thousands, and even billions of data it will refer to as database garbage. This course will make the computer collect a lot of database garbage. The development of knowledge about data mining will make these data useful [5].

The definition of Data mining is processing that has meaning to discover new knowledge by cleaning the large-scale data, then storing it in storage, and utilizing a technique based on statistical and mathematical techniques [4]. Then, Data Mining can also refer to as the process of getting useful information from a large database warehouse. The data mining technique is how to trace existing data to build a model. The model is used to recognize other data patterns that are not in the stored database [6].

### B. Supervised Learning and Unsupervised Learning

In general, data mining has two approaches to data training, namely Unsupervised Learning and Supervised Learning.

Unsupervised Learning has an approach by grouping or categorizing data. Unsupervised Learning is descriptive so that it does not get a training dataset, and requires learning from existing data. While Supervised Learning, the system gives a training data set in the form of desired input and output information so that the system will learn based on existing data. The system will look for patterns from the dataset then the pattern will be used as a reference next dataset. Supervised Learning is predictive so that from the results of the many approaches taken, a prediction will obtain. Supervised Learning can solve data problems linearly, multilinearly, or polynomial [7].

### C. Classification Method

Classification is a method of finding patterns that can distinguish and describe labels on the data held. Existing patterns will be made based on training data analysis [8]. The classification method includes supervised learning. This method will find characteristic patterns of data sets in a database, then classify them into different groups according to the resulting classification model. In the classification process, training data from available attributes are used to

produce a particular model. The resulting model will then be used to classify different data for which the previous pattern has not been known [9].

The following is an example of a categorical technique for predicting the target class for any given information: For example, patients are often classified as "high risk" or "low risk" based on their disease exploit information. It's a supervised learning approach having best known category classes. Binary and construction area unit the 2 ways of classification. In binary classification, solely 2 potential categories, "high" or "low" risk patient is also thought-about whereas the multiclass approach has over 2 targets for instance, "high", "medium" and "low" risk patient. information set is divided as coaching and testing dataset [10].

### D. K-Nearest Neighbor (KNN)

KNN is one approach that is simple to implement and is an old method used in classification [9]. KNN has a concept looking at the distance of the nearest neighbor with new data or training data. The training data placed in a class, where this room represents the features of the data. This space divide into many parts according to the classification of the training data. KNN will look for the closest distance between the test data and the nearest neighbor "k" in the training data [2].

K-Nearest Neighbor (KNN) classifier is one in all the only classifier that discovers the unidentified information exploitation the antecedently best-known information points (nearest neighbor) and classified information points in keeping with the legal system. KNN contains a range of applications in numerous areas like health datasets, image fields, online selling, etc. [10].

### E. Decision Tree C4.5

Decision Tree C4.5 is one of the supervised learning algorithms in data mining. This algorithm contains various factors that can use as problem solvers. The use of the decision tree algorithm as a classification problem solver is classified as very good because it can find out only based on the pattern of the tree shape [2]. The main purpose of using the C4.5 decision tree algorithm is to be able to produce a specific prediction model in the form of rules that are easy to implement [11].

Decision Tree associate degree analogous to the flow chart within which each non-leaf node denotes a check on a specific attribute and each branch denotes an outcome of that check and each leaf node has a category label. The node at the highest most labels within the tree is named the root node. For instance, we have a financial organization call tree that is employed to choose that someone should grant the loan or not. Building a call for any drawback does not like any variety of domain data. Call Trees could be a classifier that uses a tree-like graph. The foremost common use of call Tree is in research analysis for shrewd conditional chances [10].

### F. Particle Swarm Optimization (PSO)

PSO is population-based on random initialization tracing of particles and interactions between particles in the population [4]. The population in the PSO call the swarm, while the individual is called the particle. Each particle will move at a speed that is adapted from the search area and

save as the best position ever reached. In addition, the particle in PSO always associates with the speed of the particle flying through the search space, which has a dynamic speed based on historical behavior. So, the particles in the PSO algorithm will always tend to fly towards a better search during the search process [5].

**G. RapidMiner**

In simple terms, RapidMiner is an application used for data processing with various open-source methods [12]. Using data mining principles and algorithms, RapidMiner extracts patterns from large data sets by combining statistical methods, artificial intelligence, and databases. This operator serves to modify the data [13]. RapidMiner contains tools for data preprocessing, classification, clustering, regression, association, and visualization [14].

**H. Confusion Matrix**

Confusion Matrix is an evaluation method used in calculating data mining accuracy. This method is in the form of a table consisting of many rows of test data that are positive and negative predictive with a classification model. The benefit of the confusion matrix is to determine the performance of the classification model [15].

		Actual Class	
		Positive	Negative
Predicted Class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Figure 1. Confusion Matrix

Evaluation using this confusion matrix will produce accuracy, precision, recall, and f-measure:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$F - \text{measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (4)$$

These values are described now [16]:

TP = True Positive: the predicted data is in a positive category, and the actual data is indeed positive.

TN = True Negative: the predicted data is in a negative category, and the actual data is indeed negative.

FP = False Positive: the predicted is in a negative category, and the actual data is positive.

FN = False Negative: the predicted is in a positive category, and the actual data is negative.

**I. Area Under Curve (AUC)**

AUC showing the measurement of the performance of the classification method based on the Receiver Operating Characteristic (ROC) curve. The scale in qualifying the AUC performance is 0 to 1. The following is a performance table from the AUC [9].

Performance	Scale
Excellent classification	0.90 – 1.00
Good classification	0.80 – 0.90
Fair classification	0.70 – 0.80
Poor classification	0.60 – 0.70
Failure classification	0.50 – 0.60

**III. RESEARCH METHODOLOGY**

This study using the research methodology by Ranjit Kumar. Research Methodology is useful in providing an overview of the research journey. This methodology consists of 8 steps:

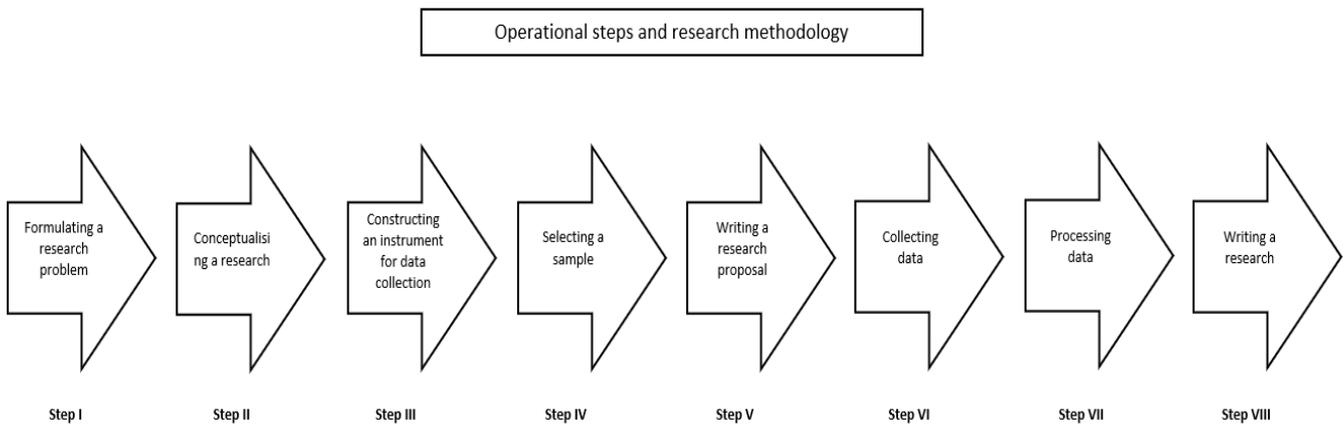


Figure 2. Operational Steps and Research Methodology [17]

Based on Figure 2., the following is an explanation of this research:

#### A. Step I – Formulating a Research Problem

This stage is very important in the research process because it decides what you want to find out. This study has two problems, namely health workers need a technique that can predict a disease quickly and accurately. Then, previous researchers have never discussed the performance comparison between KNN with PSO and Decision Tree C4.5 with PSO. The purpose of this study is to determine the best performance between the two algorithms so that later it can be used to predict a disease by health workers.

#### B. Step II – Conceptualising a Research Design

The concept of research design is needed because it explains how researchers find answers to the problems encountered. The design concept of this research is to compare algorithm modeling the KNN with PSO and Decision Tree C4.5 with PSO using an application, namely RapidMiner. From the application, the researchers can find out the performance difference between the two algorithms. The best performance results can be used to predict a disease.

#### C. Step III - Constructing an Instrument for Data Collection

This study uses secondary data in the form of documents based on the UCI Machine Learning Repository website.

#### D. Step IV - Selecting a Sample

The sample used in this study is one type of disease that has a high mortality rate in Indonesia, namely liver disease. Samples of liver disease used in this study were 2,093 records.

#### E. Step V - Writing a Research Proposal

The following is the proposal for this research:

1. The purpose of this study is to determine the performance comparison between the two algorithms so that the best performance results can be used to predict a disease by health workers.
2. The design concept in this research is to use algorithm modeling from KNN with PSO and Decision Tree C4.5 with PSO. This modeling is made using an application, namely RapidMiner.
3. This study only discusses the comparison of the performance of KNN with PSO and Decision Tree C4.5 with PSO. Also, the application used is only the RapidMiner application.
4. The data used is secondary data in the form of documents that are guided, by the UCI Machine Learning Repository.
5. In this study, 2093 records of liver disease dataset will be used.
6. Data processing is carried out using the RapidMiner application. The results of the data processing are in the form of accuracy values that are represented through a matrix table. In addition to the accuracy value, the AUC value will be obtained which is represented by the ROC curve.
7. The research report that will be proposed consists of several chapters, namely introduction, literature review, research methods, results and discussion, conclusions and suggestions, references.

8. The limitations of this study are that it only discusses the performance comparison between the two algorithms, the data used is the liver disease dataset and the application that processes the data, namely RapidMiner.

#### F. Step VI - Collecting Data

The collection of data used in this study is guided by the UCI Machine Learning Repository. The data used is in the form of a dataset, which consists of 11 attributes. One of the attributes is a class. The following is a table of the attribute in the liver disease dataset:

No	Attributes	Data Types
1	Age	Numeric
2	Gender	Nominal
3	Total Bilirubin	Numeric
4	Direct Bilirubin	Numeric
5	Alkaline Phosphatase	Numeric
6	Alamine Aminotransferase	Numeric
7	Aspartate Aminotransferasi	Numeric
8	Total Proteins	Numeric
9	Albumin	Numeric
10	Albumin dan Globulin Ratio	Numeric
11	Class	Nominal

#### G. Step VII - Processing and Displaying Data

To process the data that is owned then in this study using an application, namely RapidMiner. In this application, the process of cleaning data that has no meaning will be carried out using the "missing value" operator. Furthermore, the data that has been cleaned can then be processed to display visualizations in the form of matrix tables and curves.

#### H. Step VIII - Writing a Research Report

The last stage is to make a research report consisting of several chapters, namely introduction, literature review, research methods, results and discussion, conclusions and suggestions, references.

## IV. RESULT AND DISCUSSION

The performance of K-Nearest Neighbor (KNN) and Decision Tree C4.5 combined with Particle Swarm Optimization (PSO) can be seen from the values generated after processing and testing data using the RapidMiner application. The performance consists of accuracy, precision, recall, and f-measure. This study found that Decision Tree C4.5 with PSO has a better level of performance than KNN with PSO. The following are the performance results of each algorithm:

**A. Performance of K-Nearest Neighbor (KNN) with Particle Swarm Optimization (PSO)**

The results obtained from the confusion matrix for the KNN algorithm with PSO can be seen in the following table:

Table 3. Confusion Matrix of KNN with PSO

	True Positive	True Negative	Class Precision
Predicted Positive	1591	138	92.02%
Predicted Negative	90	274	75.27%
Class Recall	94.65%	66.50%	

Based on Table 3., there are details that True Positive (TP) is 1591, True Negative (TN) is 274, False Positive (FP) is 138, False Negative (FN) is 90. The measurement results obtained are as follows:

$$\text{Accuracy} = \frac{1591 + 274}{1591 + 274 + 138 + 90} = 89.11\%$$

$$\text{Precision} = \frac{1591}{1591 + 138} = 92.02\%$$

$$\text{Recall} = \frac{1591}{1591 + 90} = 94.65\%$$

$$\text{F - measure} = \frac{2 \times 0.9711 \times 0.9185}{0.9711 + 0.9185} = 93.01\%$$

In addition to the measurement performance of the confusion matrix, performance can also be seen from the AUC. AUC shows the measurement of performance base on the ROC curve:

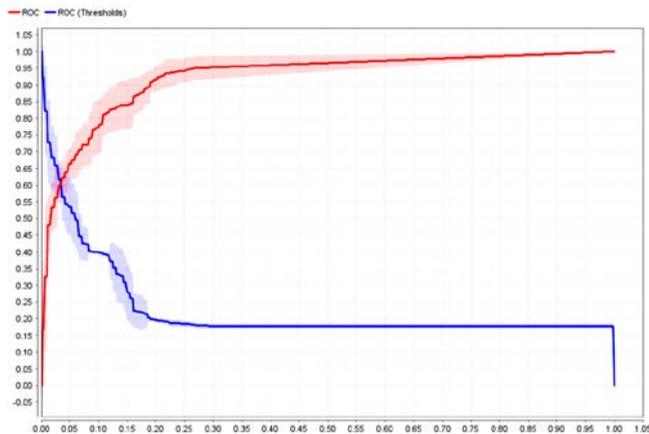


Figure 3. ROC Curve of KNN with PSO

Based on Figure 3., the Y-axis is almost close to 1.00, so that the AUC on KNN with PSO has an Excellent Classification parameter. The AUC value for KNN with PSO is 0.923.

**B. Performance of Decision Tree C4.5 with Particle Swarm Optimization (PSO)**

The results obtained from the confusion matrix for the KNN algorithm with PSO can be seen in the following table:

Table 4. Confusion Matrix of Decision Tree C4.5 with PSO

	True Positive	True Negative	Class Precision
Predicted Positive	1544	46	97.11%
Predicted Negative	137	366	72.76%
Class Recall	91.85%	88.83%	

Based on Table 4., there are details that True Positive (TP) is 1544, True Negative (TN) is 366, False Positive (FP) is 46, False Negative (FN) is 137. The measurement results obtained are as follows:

$$\text{Accuracy} = \frac{1544 + 366}{1544 + 366 + 46 + 137} = 91.26\%$$

$$\text{Precision} = \frac{1544}{1544 + 46} = 97.11\%$$

$$\text{Recall} = \frac{1544}{1544 + 137} = 91.85\%$$

$$\text{F - measure} = \frac{2 \times 0.9711 \times 0.9185}{0.9711 + 0.9185} = 94.25\%$$

AUC shows performance measurement based on the ROC curve, as shown below:

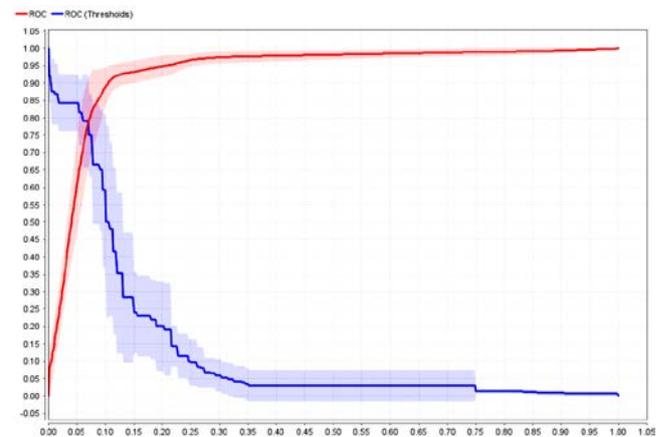


Figure 4. ROC Curve of Decision Tree C4.5 with PSO

Based on Figure 4., the AUC result from Decision Tree C4.5 with PSO is 0.935. Seen from the Y-axis, which is almost close to 1.00, the AUC value on KNN with PSO has an Excellent Classification parameter because it is in the range of 0.90 – 1.00.

**C. Performance Comparison of K-Nearest Neighbor with PSO and Decision Tree C4.5 with PSO**

Based on Figure 5, the comparison of the performance of the two algorithms is represented through the measurement results of accuracy, precision, recall, and f-measure. When viewed from the accuracy indicator, Decision Tree C4.5 with PSO obtained a value of 91.26%, so it is better than KNN with PSO only has a value of 89.11%. Likewise, with indicator precision, Decision Tree C4.5 with PSO has a better score of 97.11%, but KNN with PSO is only 92.02%.

In contrast to the recall indicator, KNN with PSO is better because it has a value of 94.65%, while Decision Tree C4.5 with is only 91.85%. When viewed from the f-measure indicator which is a representation of a combination of precision and recall, Decision Tree C4.5 with PSO is better of 94.25%, while KNN with PSO only gets a value of 93.01%. Based on the performance comparison, Decision Tree C4.5 with PSO has better performance than KNN with PSO.

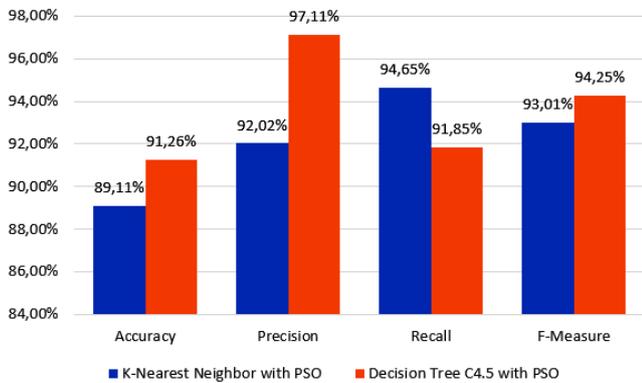


Figure 5. Measurement Comparison of KNN with PSO and Decision Tree C4.5 with PSO

For the AUC performance generated through the ROC curve, these two algorithms have the Excellent Classification parameter, because the AUC value is almost close to 1.00. However, Decision Tree C4.5 with PSO still has a better AUC performance of 0.935. Meanwhile, KNN with PSO only has an AUC of 0.923. The following is a comparison chart of AUC values between KNN with PSO and Decision Tree C4.5 with PSO:

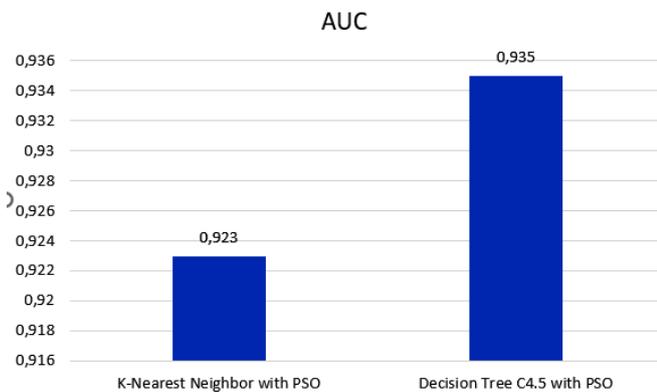


Figure 6. AUC Comparison of KNN with PSO and Decision Tree C4.5 with PSO

For execution time in processing and testing data, the KNN algorithm with PSO takes 94 seconds, whereas Decision Tree C4.5 with PSO only takes 25 seconds, meaning that the execution time of Decision Tree C4.5 with PSO is shorter than the KNN algorithm with PSO. PSO. Here is a comparison of the execution times between the two algorithms:

To determine whether the two algorithms have a significant difference in accuracy, then a test is carried out to compare the performance of an algorithm. This test is carried out using the t-Test operator on RapidMiner. The results obtained indicate that the value  $\alpha = 0.007$ .



Figure 7. Execution Time Comparison of KNN with PSO and Decision Tree C4.5 with PSO

This means that the two algorithms have significant differences, and have met the parameter requirements, namely  $\alpha < 0.050$ . So, Decision Tree C4.5 with PSO proved to have the best performance compared to KNN with PSO. The following are the t-test results for the KNN algorithm with PSO and Decision Tree C4.5 with PSO:

A	B	C
	0.891 +/- 0.013	0.913 +/- 0.019
0.891 +/- 0.013		0.007
0.913 +/- 0.019		

V. CONCLUSION

The conclusion of this research is the Decision Tree C4.5 with PSO has a better performance than KNN with PSO. The accuracy of Decision Tree C4.5 with PSO is 91.26%, while KNN with PSO is only 89.11%. Then, the AUC result between the two algorithms has a value that is almost close to 1.00, so that it is included in the Excellent Classification parameter. But, Decision Tree C4.5 with PSO still has a better AUC performance of 0.935. Meanwhile, KNN with PSO only has an AUC of 0.923.

However, other comparisons such as execution time in processing and testing data using RapidMiner can be seen that Decision Tree C4.5 with PSO has a very short time, which is only 25 seconds. Unlike the case with KNN with PSO, it takes longer, which is 94 seconds. Thus, proving that Decision Tree C4.5 with PSO is very appropriate to be applied because it does not need to spend time to execute an algorithm and provides effectiveness for other researchers who want to try to implement Decision Tree C4.5 with PSO.

This research also conducted a t-test, where the results obtained were  $\alpha = 0.007$ . So, the two algorithms have a significant effect. Therefore, this research proves that Decision Tree C4.5 with PSO has better performance, compared to KNN with PSO.

For this research to obtain varied results, further researchers can develop this research by adding other algorithms and replacing optimizations such as Genetic Algorithm or Ant Colony Optimization.

REFERENCES

[1] M. Maulina, "Zat-Zat yang Mempengaruhi Histopatologi Hepar,"

- Unimal Press*, vol. 49, p. 13, 2018, [Online]. Available: [http://repository.unimal.ac.id/4189/1/%5BMeutia Maulina%5D Zat Zat Yang Mempengaruhi Histopatologi Hepar.pdf](http://repository.unimal.ac.id/4189/1/%5BMeutia%20Maulina%5D%20Zat%20Zat%20Yang%20Mempengaruhi%20Histopatologi%20Hepar.pdf).
- [2] A. P. Ayudhitama and U. Pujiyanto, "Analisa 4 Algoritma Dalam Klasifikasi Penyakit Liver Menggunakan," *J. Inform. Polinema*, vol. 6, pp. 1–9, 2020.
- [3] T. A. Y. Siswa, "Analisis Penerapan Optimasi Perbandingan Kinerja Algoritma C4.5 Dan Naive Bayes Berbasis Particle Swarm Optimization (Pso) Untuk," *J. Bangkit Indones.*, vol. 7, no. 2, p. 1, 2018, doi: 10.52771/bangkitindonesia.v7i2.48.
- [4] T. S. N. Koeswara, M. S. Mardiyanto, and M. A. Ghani, "Penerapan Particle Swarm Optimization (Pso) Dalam Pemilihan Atribut Untuk Meningkatkan Akurasi Prediksi Diagnosispenyakit Hepatitis Dengan Metode Naive Bayes," *J. Speed – Sentra Penelit. Eng. dan Edukasi*, vol. 12, no. 1, pp. 1–10, 2020.
- [5] Nurahman, "Evaluasi Performa Algoritma C4.5 dan C4.5 Berbasis PSO untuk Memprediksi Penyakit Diabetes," *J. E-Kompetek*, vol. 4, no. 1, pp. 30–47, 2020.
- [6] A. Noviriandini, P. Handayani, and Syahriani, "Prediksi Penyakit Liver Dengan Menggunakan Metode Naive Bayes dan K-Nearest Neighbor (KNN)," *Pros. TAU SNAR-TEK Semin. Nas. Rekayasa dan Teknol.*, no. November, 2019.
- [7] H. Abijono, P. Santoso, and N. L. Anggreini, "Algoritma Supervised Learning Dan Unsupervised Learning Dalam Pengolahan Data," *J. Teknol. Terap. G-Tech*, vol. 4, no. 2, pp. 315–318, 2021, doi: 10.33379/gtech.v4i2.635.
- [8] A. Setiawati, Intan; Wibowo, Adityo Permana; Hermawan, "Implementasi Decision Tree untuk Mendiagnosis Penyakit Liver," *J. Inf. Syst. Manag.*, vol. 1, no. 1, pp. 13–17, 2019.
- [9] A. R. Kadafi, "Perbandingan Algoritma Klasifikasi Untuk Penjurusan Siswa SMA," *J. ELTIKOM*, vol. 2, no. 2, pp. 67–77, 2018, doi: 10.31961/eltikom.v2i2.86.
- [10] M. A. Salih, A. H. Alobaidi, and A. M. Alsamarai, "Obesity as a Risk Factor for Disease Development:Part-I Cardiovascular Diseases and Renal Failure," *Indian J. Public Heal. Res. Dev.*, vol. 11, no. 1, p. 1926, 2020, doi: 10.37506/v11/i1/2020/ijphrd/194136.
- [11] Noviandi, "Implementasi Algoritma Decision Tree C4.5 Untuk Prediksi Penyakit Diabetes," *Inohim*, vol. 6, no. 1, pp. 1–5, 2018.
- [12] N. T. Rahman, F. I. Komputer, U. Darwan, and A. Sampit, "Analisa Algoritma Decision Tree dan Naive Bayes pada Pasien Penyakit Liver," *J. Fasilkom*, vol. 10, no. 2, pp. 144–151, 2020.
- [13] D. Novianti, "Implementasi Algoritma Naive Bayes Pada Data Set Hepatitis Menggunakan Rapid Miner," *Paradig. - J. Komput. dan Inform.*, vol. 21, no. 1, pp. 49–54, 2019, doi: 10.31294/p.v21i1.4979.
- [14] A. Ridwan, "Penerapan Algoritma Naive Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 15–21, 2020, doi: 10.47970/siskom-kb.v4i1.169.
- [15] F. S. Nugraha, M. J. Shidiq, and S. Rahayu, "Analisis Algoritma Klasifikasi Neural Network Untuk Diagnosis Penyakit Kanker Payudara," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 149–156, 2019, doi: 10.33480/pilar.v15i2.601.
- [16] A. A. K. Qodrat, "Perbandingan Algoritma Naive Bayes Dan K-Nearest Neighbor Untuk Sistem Kelayakan Kredit Pada Nasabah ( Studi Kasus : PT . Armada Finance Cabang Makassar )," 2017.
- [17] R. Kumar, *Research Methodology - A step by step guide for beginners - 3rd edition*.