Contents lists available at https://citedness.com/index.php/jdsi

Data Science Insights

Journal Page is available to https://citedness.com/index.php/jdsi

Research article

Prediction of Heart Disease Attack Risk using Deep Learning Algorithm

Michelle Virya Effendy
Department of Teknik Informatika, Institut Bisnis dan Teknologi Pelita Indonesia, Pekanbaru, Riau, Indonesia
email: michelle.virya@student.pelitaindonesia.ac.id

ARTICLE INFO

Article history:
Received: 14 December 2024
Revised 01 January 2025
Accepted 01 February 2025
Available online 15 February 2025

Keywords:
Algorithm
Deep Learning
Disease Attack
Heart
Prediction

Please cite this article in IEEE style as:
M. V. E. Mishel, "A Prediction of Heart Disease Attack Risk using Deep Learning Algorithm: Journal Data Science", Data Science Insights, pp. 10–18, Feb. 2025.

ABSTRACT

The heart is a muscular organ that acts as the main pump in the human circulatory system, pumping oxygen-rich blood throughout the body and returning blood containing carbon dioxide to be purified. Coronary heart disease, caused by arterial blockages due to plaque buildup (fat, cholesterol, and other substances), is often the leading cause of heart attacks as blood flow to the heart muscle is reduced. This condition is one of the leading causes of death worldwide, making it necessary to have an accurate method to detect this disease early. This study aims to help predict the risk of heart disease based on gender using data mining. Data mining facilitates heart disease diagnosis, particularly in helping doctors determine whether a patient suffers from heart disease based on early symptoms that appear. The author uses five data mining algorithms: Naïve Bayes, K-Nearest Neighbor (KNN), Decision Tree, Random Forest, and Deep Learning. The research results show that the Deep Learning model is the best algorithm for predicting heart disease symptoms. Additionally, using the right predictive model can help reduce the risk of delayed diagnosis. Therefore, the predictive model with this algorithm is recommended for implementation in hospitals to help detect heart disease symptoms in patients more accurately and efficiently. This way, early diagnosis can be made to improve patient recovery chances and reduce mortality rates due to heart disease.

Correspondence:
Michelle Virya Effendy
Department of Teknik Informatika,
Institut Bisnis dan Teknologi Pelita Indonesia, Pekanbaru, Riau, Indonesia

1. Introduction

Heart disease is one of the leading causes of death worldwide. Risk factors such as hypertension, diabetes, high cholesterol, and unhealthy lifestyles can increase the likelihood of developing this condition. Early detection and proper management are crucial to reducing mortality rates from heart disease, but there are still many challenges in diagnosing it accurately and quickly.

Research using deep learning algorithms offers significant benefits in diagnosing heart disease with higher accuracy. This algorithm is capable of analyzing large volumes of medical data, recognizing patterns that are difficult for humans to detect, and providing faster and more efficient prediction results. In addition, the author uses four other algorithms, namely Naïve Bayes, K-Nearest Neighbor (KNN), Decision Tree, and Random Forest.

The purpose of this study is to explore the application of deep learning algorithms in diagnosing heart disease, with the expectation of providing a more accurate and efficient solution for identifying this condition. By using this technology, it is hoped that diagnosis errors can be reduced and early awareness and intervention for heart disease can be improved.

2. Literature Review

- **Prof. Michael Lee (2022)**
  Prof. Michael Lee from the University of Auckland developed a deep learning model to analyze medical images, such as CT scans, in predicting heart disease. His research aimed to improve

the speed and accuracy of diagnoses, as well as assist doctors in identifying signs of heart disease that are difficult to detect with traditional methods.

- **Dr. Jane Smith (2023)**
  Dr. Jane Smith from the University of California led research that used deep learning to analyze ECG data, enhancing early detection of heart disease with high accuracy. She developed a neural network-based model to recognize patterns of arrhythmia and other symptoms, speeding up diagnosis and enabling heart disease detection before clinical symptoms appear.
- **Dr. Maria Gonzalez (2024)**
  Dr. Maria Gonzalez developed a deep learning algorithm to analyze biomarker data and patient medical records in detecting heart disease. Her research aimed to improve diagnostic accuracy by utilizing more comprehensive data, while also aiding early detection and more efficient management of heart disease.

3. Research Methodology

There are several stages in the data model development cycle in Data Science or Machine Learning, which consist of multiple phases.

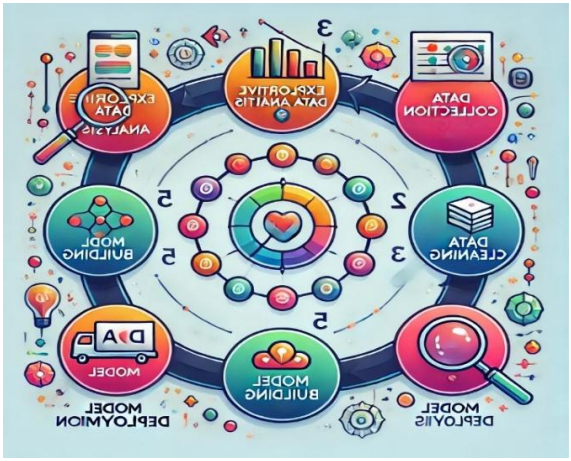


Figure 1. Data Science Process

In Data Science, procedures are needed to ensure data accuracy to support research credibility and be useful in various aspects in the future. Therefore, minimizing errors in data processing is very important. The procedures carried out by the author include:

3.1 Data Collection

The first step taken is the process of collecting data from various sources. This data can be in the form of numbers, text, images, and so on. Data can also be obtained by conducting direct surveys. It should be noted that the data must be accurate, complete and reliable. The data used in this study were obtained from the Kaggle website (www.kaggle.com). Kaggle is an online community and platform that focuses on data science and machine learning. Founded in 2010, Kaggle has become a leading destination for data scientists, data analysts, and learners who want to develop their skills, collaborate with others, and stay up to date with the latest developments in the field.

3.2 Data Cleaning

Data that has been collected usually needs to be selected so that unimportant data can be removed from the dataset. At this stage, cleaning is carried out from errors, redundancies and data inconsistencies. This process is important to ensure that the data to be studied is accurate and reliable. This data cleaning process is carried out using Microsoft Excel software. At this stage, problematic data is selected and data format standardization is also carried out so that the data to be processed gets accurate results.

3.3 Exploratory Data Analysis

After the data has gone through the cleaning process, data scientists begin using various statistical and visualization techniques to look for hidden patterns, trends, and relationships among the data. At this stage, it is important to remain critical and not jump to conclusions from what is seen. Data scientists must consider various possibilities and look for relevant evidence according to the research objectives.

At this stage, visualization is carried out on the data using Tableau to help facilitate the visualization process. Using Tableau as a tool to visualize data has benefits such as offering great flexibility in processing and visualizing data from various sources. This allows for a more holistic

and comprehensive analysis. Tableau also provides a variety of powerful visualization tools, ranging from simple graphs to complex shapes.

3.4 Model Development

With the knowledge gained from the data analysis process, the author can build a model that can predict, group, or classify new data accurately. In the process of building this model, critical thinking and a deep understanding of the data and algorithms that will be used on the data are required. At this stage, data model development is carried out using rapidminer software. RapidMiner is a data analysis platform that helps understand patterns and trends in large data sets. In rapidminer, various algorithms can also be applied and analyzed which algorithms are suitable for use in research purposes.

3.5 Model Application

The model that has been successfully built will be applied in various aspects that are in accordance with the research objectives, such as sales strategies, strategies for retaining customers, and so on. By carrying out this entire process carefully and critically, data scientists can help solve various problems in the world, especially completing the research that has been done.

4. Results and Discussion

In this section, the results will be provided according to the stages carried out:

4.1 Data Collection

The dataset taken consists of heart disease data stored to predict symptoms that can be concluded from the heart disease.

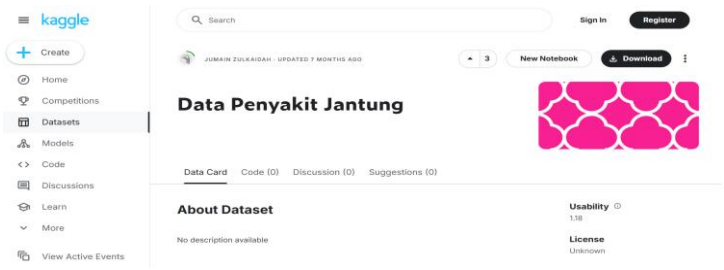


Figure 2. Heart Disease Dataset

The purpose of this data collection is to analyze early symptoms of heart disease, thus helping doctors determine whether a patient has heart disease based on symptoms that are close to the condition. (https://www.kaggle.com/datasets/jumainzulkaidah/data-penyakit-jantung)

4.2 Data Cleaning

Data cleaning on this dataset is important to ensure the quality of the data used in analysis or model creation.

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
58	0	0	100	248	0	0	122	0	1	1	0	2	1
58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
43	0	0	132	341	1	0	136	1	3	1	0	3	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
52	1	0	128	204	1	1	156	1	1	1	0	0	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
54	1	0	124	266	0	0	109	1	2.2	1	1	3	0
50	0	1	120	244	0	1	162	0	1.1	2	0	2	1
58	1	2	140	211	1	0	165	0	0	2	0	2	1

Figure 3. Data Cleaning

There are several main reasons for performing data cleansing, namely: dealing with missing values, correcting invalid data, eliminating data duplication, and ensuring data consistency.

4.3 Data Cleaning Results

In the data cleaning process, the author uses Google Colab. First, we will enter the data set "heart.csv" into the Google Colab file, then import the data and include the required libraries and run this code (Figure 4) to display the top 4 rows of the data set.

```
import pandas as pd
```

```
[ ] heart = pd.read_csv('/content/heart.csv')
```

```
heart.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

Figure 4. Read Dataset

Then, insert the heart.info() function into the code line to find out each data type for each variable and df.shape to find out each number of rows and columns from the dataset.

```
heart.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1025 non-null   int64  
 1   sex       1025 non-null   int64  
 2   cp        1025 non-null   int64  
 3   trestbps  1025 non-null   int64  
 4   chol      1025 non-null   int64  
 5   fbs       1025 non-null   int64  
 6   restecg   1025 non-null   int64  
 7   thalach   1025 non-null   int64  
 8   exang     1025 non-null   int64  
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64  
 11  ca        1025 non-null   int64  
 12  thal      1025 non-null   int64  
 13  target    1025 non-null   int64  
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
```

Figure 5. Information about the Dataset

Enter the heart.describe() function. The heart.describe() command is used to provide a descriptive statistical summary of a numeric column in a DataFrame.

```
heart.describe()
```

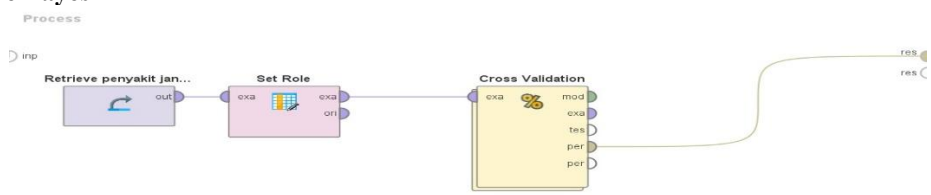
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak
count	1025.000000	1025.000000	1025.000000	1025.000000	1025.00000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000
mean	54.434146	0.695610	0.942439	131.611707	246.00000	0.149268	0.529756	149.114146	0.336585	1.071512
std	9.072290	0.460373	1.029641	17.516718	51.59251	0.356527	0.527878	23.005724	0.472772	1.175053
min	29.000000	0.000000	0.000000	94.000000	126.00000	0.000000	0.000000	71.000000	0.000000	0.000000
25%	48.000000	0.000000	0.000000	120.000000	211.00000	0.000000	0.000000	132.000000	0.000000	0.000000
50%	56.000000	1.000000	1.000000	130.000000	240.00000	0.000000	1.000000	152.000000	0.000000	0.800000
75%	61.000000	1.000000	2.000000	140.000000	275.00000	0.000000	1.000000	166.000000	1.000000	1.800000
max	77.000000	1.000000	3.000000	200.000000	564.00000	1.000000	2.000000	202.000000	1.000000	6.200000

Figure 6. Dataset Description

4.4 Model Building Results

Model building is done by testing five different algorithms and comparing the algorithms to obtain the appropriate algorithm.

1. Naïve Bayes



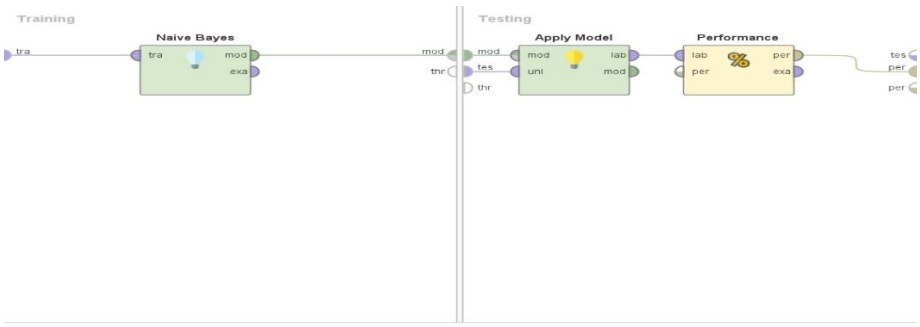


Figure 7. Naïve Bayes algorithm using Rapid Miner

The first algorithm tested using rapidminer is the Naïve Bayes algorithm. The rapidminer structure as in Figure 7 begins with reading the dataset then dividing the data by 75% as test data and 25% as training data. After that, validation is carried out in which classification has been carried out using the first stage naïve bayes algorithm. Then the second stage classification is carried out to obtain the results of accuracy, precision and recall on the data.

accuracy: 75.51% +/- 3.20% (micro average: 75.51%)

	true 1	true 0	class precision
pred. 1	569	107	84.17%
pred. 0	144	205	58.74%
class recall	79.80%	65.71%	

Figure 8. Naïve Bayes Algorithm Results

The results of the naïve Bayes algorithm can be seen in Figure 8. The algorithm has an accuracy rate of 75.51%. This accuracy rate is lower in testing on data.

2. KNN (K-Nearest Neighbors)

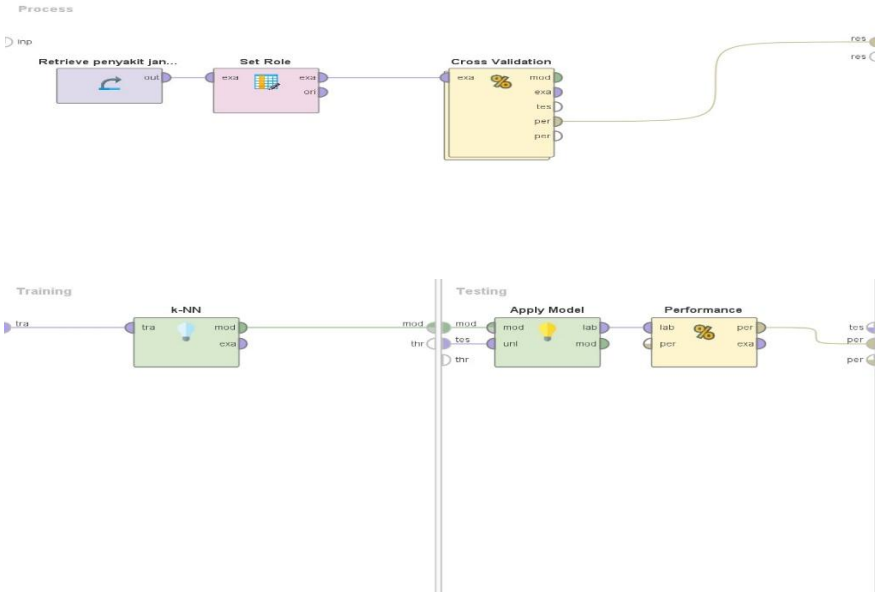


Figure 9. K-Nearest Neighbors Algorithm Using Rapidminer

The second algorithm tested using rapidminer is the KNN (K-Nearest Neighbors) algorithm. The rapidminer structure as in Figure 9 begins with reading the dataset then dividing the data by 75% as test data and 25% as training data. After that, cross-validation is carried out in which classification has been carried out using the first stage KNN algorithm. Then the second stage classification is carried out to obtain the results of accuracy, precision and recall on the data.

accuracy: 86.83% +/- 3.41% (micro average: 86.83%)

	true 1	true 0	class precision
pred. 1	693	115	85.77%
pred. 0	20	197	90.78%
class recall	97.19%	63.14%	

Figure 10. Results of the K-Nearest Neighbors Algorithm

The results of the KNN algorithm can be seen in Figure 10. The algorithm has an accuracy rate of 86.83%. This level of accuracy is higher than testing on data using the naïve bayes algorithm.

3. Decision Tree

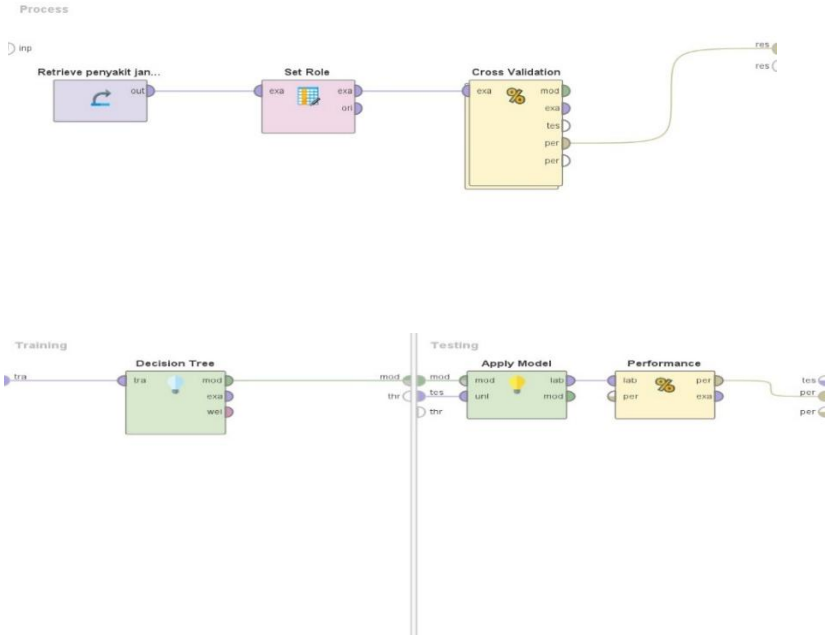


Figure 11. Decision Tree Algorithm Using Rapidminer

The third algorithm tested using rapidminer is the Decision Tree algorithm. The rapidminer structure as in Figure 11 begins with reading the dataset then dividing the data by 75% as test data and 25% as training data. After that, cross validation is carried out in which classification has been carried out using the first stage decision tree algorithm. Then the second stage classification is carried out to obtain the results of accuracy, precision and recall on the data.

accuracy: 76.79% +/- 2.28% (micro average: 76.78%)

	true 1	true 0	class precision
pred. 1	710	235	75.13%
pred. 0	3	77	96.25%
class recall	99.58%	24.68%	

Figure 12. Decision Tree Algorithm Results

The results of the decision tree algorithm can be seen in Figure 12. The algorithm has an accuracy rate of 76.79%. This accuracy rate is lower than testing on data with the KNN algorithm.

4. Random Forest

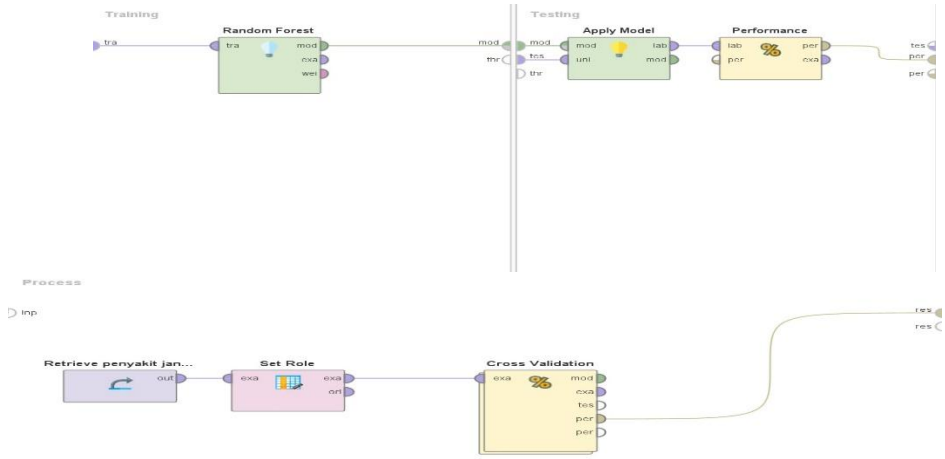


Figure 13. Random Forest Algorithm Using Rapidminer

The algorithm tested using RapidMiner is the Random Forest algorithm. The rapidminer structure as in Figure 13 begins with reading the dataset then dividing the data by 75% as test data and 25% as training data. After that, cross validation is carried out in which classification has been carried out using the first stage random forest algorithm. Then the second stage classification is carried out to obtain the results of accuracy, precision and recall on the data.

accuracy: 86.03%

	true No	true Yes	class precision
pred. No	3608	465	88.58%
pred. Yes	273	937	77.44%
class recall	92.97%	66.83%	

Figure 14. Random Forest Algorithm Results

The results of the random forest algorithm can be seen in Figure 14. The algorithm has an accuracy rate of 86.03%. This level of accuracy is higher than testing on data using the naïve bayes and decision tree algorithms but higher than the knn algorithm.

5. Deep Learning

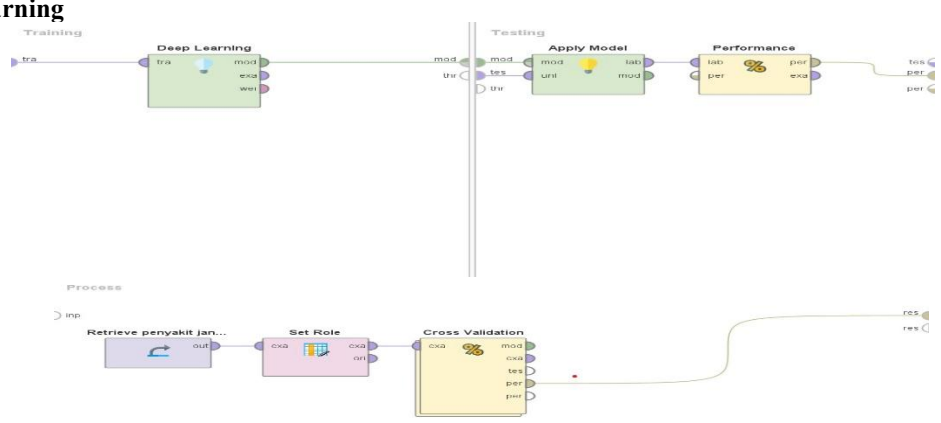


Figure 15. Deep Learning Algorithm Using Rapidminer

The last algorithm tested using Rapidiner is the Random Forest algorithm. The rapidmine structure as shown in Figure 15 begins with reading the dataset and then dividing the data by 75% as test data and 25% as training data. After that, cross-validation is carried out in which classification has been carried out using the first stage random forest algorithm. Then the second stage classification is carried out to obtain the results of accuracy, precision and recall on the data.

accuracy: 89.27% +/- 4.03% (micro average: 89.27%)

	true 1	true 0	class precision
pred. 1	655	52	92.64%
pred. 0	58	260	81.76%
class recall	91.87%	83.33%	

Figure 16. Deep Learning Algorithm Results

The results of the deep learning algorithm can be seen in Figure 16. The algorithm has an accuracy rate of 89.27%. This accuracy rate is the highest compared to the previous four algorithms.

After testing with 5 different algorithms, namely Naïve Bayes, K-Nearest Neighbor (KNN), Decision Tree, Random Forest and Deep Learning. The author builds a graphical model using the Tableau application. The results can be seen in Figure 17.

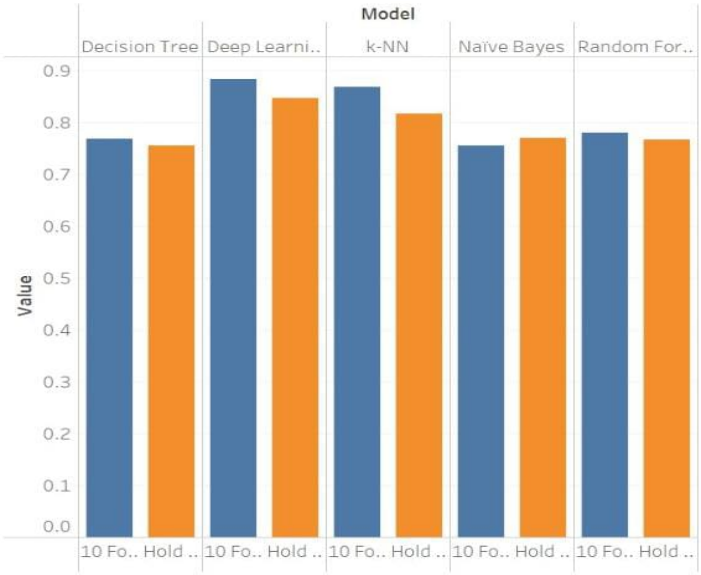


Figure 17. Accuracy Based on Algorithm Comparison

Model	Hold Out	10 Fold Cross Validation
Naïve Bayes	**76,95%**	**75,51%**
KNN (K- Nearest Neighbors)	**81,64%**	**86,83%**
Decision Tree	**75,59%**	**76,79%**
Random Forest	**76,76%**	**86,03%**
Deep Learning	**84,77%**	**89,27%**

Table 1. Dataset Accuracy Results

After testing the five algorithms, it was found that the Deep Learning Algorithm is the most suitable algorithm for the purpose of this study. As seen in Figure 16, the Deep Learning Algorithm has the highest accuracy compared to the other four algorithms tested in this study with an accuracy rate of 88.27%.

5. Conclusion and Suggestions

Based on the test results of the k-Nearest Neighbor, Random Forest, Naïve Bayes, Decision Tree, and Deep Learning Algorithms to solve the problem of classifying patients with heart disease or not using the RapidMiner application, Data taken from Kaggle, namely Heart Disease Data. Showing the results that of the five algorithms, the best and most suitable algorithm for classifying liver patients is Deep Learning with an accuracy of 89.27%. In addition to its highest accuracy, Deep Learning is also able to classify patients with heart disease in greater numbers so that it is considered accurate. So far, many algorithms have high accuracy values but are unable to classify correctly, even many detect patients who do not have heart disease even though the original data is heart disease. Based on the results of the algorithm, it shows that Deep Learning has the best results.

Reference

[1] & H. Cardemil, Parashar and Kemenkes RI, "Your Browser is out of date.," Pusdatin.Kemenkes.Go.Id, vol. 0, p. Ministry of Health of the Republic of Indonesia, 2017, [Online].
[2] M. Vaduganathan, G. A. Mensah, J. V. Turco, V. Fuster, and G. A. Roth, “The Global Burden of Cardiovascular Diseases and Risk: A Compass for Future Health,” J. Am. Coll. Cardiol., vol. 80, no. 25, pp. 2361–2371, 2022, doi: 10.1016/j.jacc.2022.11.005.
[3] A. BenTaieb and G. Hamarneh, “Deep Learning Models for Digital Pathology,” 2019, doi: 10.1109/access.2020.3001149.R.
[4] R. H. Laftah and K. H. K. Al-Saedi, “Explainable Ensemble Learning Models for Early Detection of Heart Disease,” J. Robot. Control, vol. 5, no. 5, pp. 1412–1421, 2024, doi: 10.18196/jrc.v5i5.22448.
[5] I. Vaccari, V. Orani, A. Paglialonga, E. Cambiaso, and M. Mongelli, “A generative adversarial network (GAN) technique for internet of medical things data,” Sensors, vol. 21, no. 11, p. 1273253, 2021, doi: 10.3390/s21113726.
[6] M. Priya, A. Anitha, M. K. Nallakaruppan, J. Raju, R. Raghavan, and D. Srivastava, “Heart Disease Prediction Using Machine Learning Algorithms,” 2023 Innov. Power Adv. Comput. Technol. i-Pact 2023, vol. 13, no. 4, 2023, doi: 10.1109/I-PACT58649.2023.10434931.
[7] A. Bustamam, "Arrhythmia Classification Using the Deep Learning Visual Geometry Group (VGG) Model," Nusant. Sci. Technol. Proc., vol. 2023, no. Dl, pp. 7–18, 2023.
[8] L. F. Tampubolon, A. Ginting, and F. E. Saragi Turnip, “A Description of Factors Affecting the Incidence of Coronary Heart Disease (CHD) at the Integrated Heart Center (PJT),” J. Ilm. Permas J. Ilm . Stikes Kendal, vol. 13, no. 3, pp. 1043–1052, 2023, doi: 10.32583/pskm.v13i3.1077.
[9] M. H. Ramadhan, “Risk Factors for Coronary Heart Disease (CHD),” J. Kedokt . Syariah Kuala, pp. 1–15, 2022.
[10] T. Arini and F. N. Umam, “Measurement of Body Fat and Body Mass Index as an Effort to Prevent the Risk of Cardiovascular Disease,” J. Bhakti Civ. Akad., vol. IV, no. 1, pp. 25–30, 2021, [Online].
[11] S. Sidaria, E. Huriani, and S. D. Nasution, “Self Care and Quality of Life of Coronary Heart Disease Patients,” Jik J. Health Sciences., vol. 7, no. 1, p. 41, 2023, doi: 10.33757/jik.v7i1.631.
[12] H. Salman, J. Grover, and T. Shankar, “Hierarchical Reinforcement Learning for Sequencing Behaviors,” vol. 2733, no. March, pp. 2709–2733, 2018, doi: 10.1162/Neco.
[13] M. S. Al Reshan, S. Amin, M. A. Zeb, A. Sulaiman, H. Alshahrani, and A. Shaikh, “A Robust Heart Disease Prediction System Using Hybrid Deep Neural Networks,”
[14] N. A. Usri, Wisudawan, Nurhikmawati , Nesyana Nurmadilla, and Irmayanti, “Characteristics of Risk Factors for Coronary Heart Disease Patients at Ibnu Sina Hospital Makassar in 2020,” Fakumi Med. J. J. Mhs. Kedokt., vol. 2, no. 9, pp. 619–629, 2022, doi: 10.33096/fmj.v2i9.117. [15] R. Sumara, N. Ari, and I. Indarti, “ Identification of Factors of Coronary Heart Disease Incidence in Women Aged ≤ 50 Years at Rsu Haji Surabaya, ” J. Manaj. Nursing Care, vol. 6, no. 2, pp. 53 – 59, 2022, doi: 10.33655/mak.v6i2.134.