 ISSN:2302-8149
e-ISSN:2540-9719

Comparison of Decision Trees, Naïve Bayes and Random Forest in Detecting Heart Disease

[1]Erni*, [2]Rabiatus Sa'adah
[1,2]Sistem Informasi Kampus Kota Pontianak, Fakultas Teknik dan Informatika,
Universitas Bina Sarana Informatika
Jl. Abdul Rahman Saleh No.18, Bangka Belitung Laut, Kec. Pontianak Tenggara, Kota Pontianak, Kalimantan Barat
*e-mail: erni.erx@bsi.ac.id

(***received*:** 17 May 2024, ***revised*:** 10 June 2024, ***accepted*:** 20 June 2024)

Abstract

Heart disease (HD) is the main cause of human death throughout the world, which generally occurs when the heart is unable to distribute enough fresh and oxidized blood throughout the body. Doctors often use an electrocardiogram (EKG) to detect abnormal heartbeats or heart rhythm disturbances, which provides important data for assessing a patient's heart condition. In determining whether the heart is functioning normally or not, machine learning methods can be applied for classification. This research compares three classification methods, namely Random Forest, Extra Trees Classifier, and Naïve Bayes, using split validation and train-test techniques. The test results show that the Extra Trees Classifier method provides the highest accuracy of 86.93%, compared to the Naïve Bayes and Random Forest Classifier methods, which each have an accuracy of 84.21%.

Keywords: heart disease, random forest, extra trees classifier and naïve bayes

1 Introduction

One of the main causes of human death is heart disease (HD) worldwide which generally occurs when the heart is unable to push enough fresh, oxidized blood to the rest of the body [1][2]. Starting with the definition of Heart Disease, that heart disease is a disturbance in the heart's normal electrical system and pumping function. Where this disease makes it more difficult for the heart muscle to pump blood efficiently and causes chest pain, chest pressure, shortness of breath, pain in the neck and jaw [3][4].

Having high blood pressure is also one of the main causes of heart disease. A survey stated that from 2011 to 2014, the incidence of hypertension in the world was around 35%, which is also a cause of heart disease. Likewise, there are many more reasons for heart disease such as obesity, not consuming proper nutrition, increased cholesterol and lack of physical activity. So, prevention is very necessary. For prevention, awareness of heart disease is important. About 47% of people die outside of hospital and it shows that they do not act on early warning signs [5].

The risk factors that can cause a person to develop heart disease are factors that can be controlled and factors that cannot be controlled. Factors that cannot be controlled, namely age, gender and hereditary factors, are factors that trigger heart disease. Meanwhile, factors that can be controlled include lifestyle such as smoking habits, unhealthy diet, lack of physical activity, and obesity as well as a history of diseases including hypertension [6].

Detection of heart disease is a major challenge in the medical field, because predicting whether someone will develop heart disease is very difficult. Usually, doctors use an electrocardiogram (ECG) to detect heartbeat abnormalities or abnormalities in the heart. However, Machine Learning offers innovative solutions in predicting heart disease [6][7][8]. Classification methods such as Random Forest, Extra Trees Classifier, and Naïve Bayes have shown great potential in improving the accuracy of heart disease predictions. So a comparison is made between the three methods to find out the best results.

Random Forest is an ensemble learning method that applies the bootstrap method to the CART algorithm, allowing the calculation of non-linear variable functions that show interactions between

 ISSN:2302-8149
e-ISSN:2540-9719

variables [9][10]. Extra Trees Classifier, on the other hand, uses an attribute randomization approach and random selection of cut points to separate the nodes of a tree, utilizing all training samples to grow the tree [11][12]. Naïve Bayes, although based on the assumption of class independence, still provides good results in many cases even though it has weaknesses in accuracy [13][14].

This study aims to explore the best prediction accuracy of the three methods using the cardiovascular disease detection dataset from Kaggle. Through this research, it is hoped that it can support efforts for early detection and more effective prevention of heart disease, thereby reducing death rates and improving the quality of life for heart disease sufferers.

2 Literature Review

Previous research on Estimating Predictions for the Use of Heart Disease Machine Learning Logistic Regression Models by Montu Saw in 2020 with 0.87 accuracy obtained. The overall model can be improved with more data and by using more Machine Learning models [15].

Another similar research on Diagnosing heart disease uses machine learning and also data mining. The research results showed that through a literature survey, they concluded that a combinational and more complex model was needed to increase the accuracy of heart disease prediction [15]. Research related to heart disease prediction was carried out by [16]. The research carried out is a quantitative approach to predicting heart disease with the aim of predicting the possibility of heart-related disease more accurately using the Hoeffding tree algorithm, LTM. The results obtained from the Hoeffding tree algorithm were 81.24%, and LTM were 80.69% but with different data and attributes.

The same research on heart disease used the Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Random Forest, and Decision Tree methods. The results of the study revealed that Naïve Bayes was better using cross validation and train test techniques with an accuracy of 82.17% each. , 84.28%[17]. Based on the literature review above, there has been no research related to heart disease prediction datasets that uses a comparison of the three algorithms (Random Forest, Decision Tree and Naïve Bayes) and uses cross-validation and train-test split techniques. This research was conducted using Python tools because in previous research related to heart disease prediction datasets, only a few studies used Python tools. And related research using the same dataset, produces accuracy that still needs to be improved.

3 Research Method

The method used in this research is Cross Industry Standard for Data Mining (CRISP-DM). In this study, only five stages of CRISP-DM were used. The description of the CRISP-DM model phases in this research is as follows.

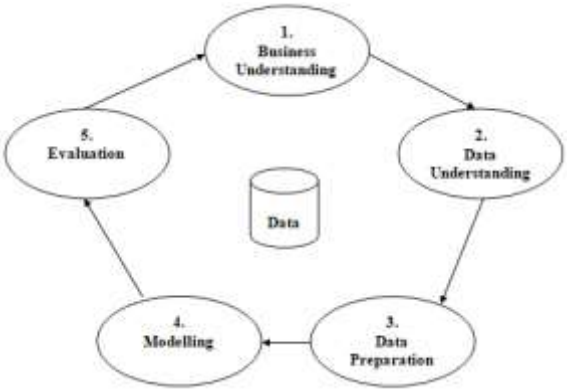

Figure 1. Research stages

3.1.1 Business Undersanding

By utilizing existing data sources, it can be analyzed using data mining techniques which aim to predict heart disease datasets using the Random Forest, Extra Trees Classifier and Naïve Bayes algorithms.

 ISSN:2302-8149
e-ISSN:2540-9719

3.1.2 Data Understanding

It is a heart disease dataset obtained from Kaggle with the URL address (https://www.kaggle.com/dileep070/heart-disease-predictionusing-logistic-regression). The stage of understanding the data is by describing the attributes contained in the dataset and describes the type of attribute value in the dataset used.

3.1.3 Data Preparation

The data obtained in this research amounted to 918 data with 12 attributes which will be processed to produce predictions of heart disease. The class used in this research is TenYearCHD (10 year risk of coronary heart disease CHD) which is used to divide the data into two sets (Label) with values 1 and 0. Where, 1 is Yes and 0 means no.

3.1.4 Modeling

At this stage, the step taken is to apply the dataset used, namely the Heart Disease dataset using Python 3 tools with *.ipynb format. After applying the dataset, the modeling stage in this research is carrying out a classification process using the Random Forest, Extra Trees Classifier and Naïve Bayes algorithms. To evaluate the performance of each algorithm used, this research divides testing data and training data. A K-Fold value of 10 was used. The data was separated using the split validation method with training data of 80% and testing data of 20%.

3.1.5 Evaluation

To compare the overall performance of the proposed research scheme, an evaluation was carried out using: accuracy, recall, precision, F1-Score and AUC. The following is an explanation of each evaluation matrix:

1. Accuracy
   Accuracy is the most commonly used evaluation metric for classification. However, for imbalanced data classification problems, accuracy may not be a good choice because accuracy often has a bias towards the majority class [18].
   $Accuracy = \frac{(TP+TN)}{Totalsample}\ x\ 100\%$ ..... (1)
2. Precision
   Precision is the part of the data that is taken according to the required information [19].
   $Precision = \frac{(TP)}{(TP+FP)}\ x\ 100\%$ ..... (2)
3. Recall
   Recall is the collection of data that is successfully retrieved from the part of the data that is relevant to the query [19].
   $Recall = \frac{(TP)}{(TP+FN)}\ x\ 100\%$ ..... (3)
4. AUC/ROC
   The ROC (Receiver Operating Characteristics) curve is recognized as the most rational choice for unbalanced data, depicting the relative trade-off between benefits and costs (Fawcett). The ROC curve is generated based on a basic matrix in machine learning called the confusion matrix which is the confusion matrix of binary classification problems [18].
   The formula for calculating AUC/ROC is:
   $AUC = 1/2 \sum_{ki=1}^{n} (\text{xi} + 1 - \text{xi})(\text{yi} + 1 - \text{yi})$ …. (4)
5. F-1 Score
   Combines precision and recall into one measure. Mathematically, the F-1 score is the harmonic average of precision and recall. To calculate the F-1 score binary classification [19].
   The formula for calculating the F-1 score is:
   $F - 1 = 2\frac{(Precision.Recall)}{(Precision+Recall)}$ ……..(5)

A confusion matrix was also presented to describe the performance of each classification model and an assessment of important features to determine the attributes that most influence the results of the classification model..

 ISSN:2302-8149
e-ISSN:2540-9719

4 Results and Analysis

4.1 Dataset

At this stage we select data to be used as research material. The data used in this research was taken through the Kaggle repository with the name Cardio Vascular Disease Detection Dataset. The amount of data held is 69302 data with 12 columns.

Table 1. Dataset attributes

No.	Field Name	Contents	Description
1	*Age*	Age	Patient age
2	*Height*	Height	Patient height
3	*Weight*	Weight	Patient weight
4	*Gender*	1 or 2	Gender, female or male
5	*Systolic blood pressure*	Systolic blood pressure	Systolic blood pressure in the patient's body
6	*Diastolic blood pressure*	*Diastolic blood pressure*	Diastolic blood pressure in the patient's body
7	*Cholesterol*	*Cholesterol*	Cholesterol in the patient's body
8	*Glucose*	*Glucose*	The patient's blood glucose
9	*Smoking*	0 or 1	0 patients did not smoke or 1 patient smoked
10	*Alcohol intake*	0 or 1	0 of these patients drank alcohol or 1 patient did not drink alcohol
11	*Physical activity*	0 or 1	0 The patient does not exercise often or 1 patient exercises frequently
12	*Presence or absence of cardiovascular disease*	0 or 1	0 absence of cardiovascular disease in the patient's body or 1 presence of cardiovascular disease in the patient's body

4.2 *Pre-Processing* Data

This research carried out split validation which aims to divide training and testing data, training data is used as training data for model learning, testing data is used as testing data for model validation or evaluation. In this research, a split was carried out using a split percentage of 80:20, which is a division of 80% of testing data and 20% of training data.

4.3 Model Testing

At this stage, what the researcher does is test the dataset using the method used. The testing process was carried out using the Random Forest, Extra Trees Classifier and Naïve Bayes methods. The tool used for this testing uses Python. Before testing the method you want to test, the data is separated using the split validation method with 80% testing data and 20% training data that was not used in previous research.

Table 2. Method test results

Method	Accuracy	AUC	Recall	Prec	F1-score
Extra Trees Classifier	86.93%	93.81%	89.18%	87.57%	88.16%
Random Forest Classifier	84.21%	92.82%	87.18%	85.25%	85.77%
Naïve Bayes	84.21%	90.22%	84.09%	87.00%	85.09%

 ISSN:2302-8149
e-ISSN:2540-9719

Test results with three different methods resulted in the Extra Trees Classifier obtaining an accuracy of 86.93%, AUC 93.81%, recall 89.18%, precision 87.57% and f1-score 88.16%, Random Forest Classifier obtained an accuracy of 84.21%, AUC 92.82%, recall 87.18%, precision 85.25% and f1-score 85.77%, and Naïve Bayes obtained accuracy 84.21%, AUC 90.22%, recall 84.09%, precision 87.00% and f1-score 85.09%.

4.4 Analysis Results

The results displayed from this test are Accuracy, AUC, Recall, Precision and f1-score. The following are the results obtained from several methods.

Table 3. Accuracy comparison

Method	Accuracy
Extra Trees Classifier	86.93%
Random Forest Classifier	84.21%
Naïve Bayes	84.21%

In tests carried out using the Extra Trees Classifier method, accuracy results were 86.93%, there was an increase of 2.72% compared to the Random Forest Classifier method and the Naïve Bayes method, whose accuracy was 84.21%.

Table 4. AUC comparison

Method	AUC
Extra Trees Classifier	93.81%
Random Forest Classifier	92.82%
Naïve Bayes	90.22%

In testing carried out using the Extra Trees Classifier method, the AUC results were 93.81%, there was an increase of 0.99% compared to the Random Forest Classifier method, whose AUC was 92.82%, and the Naïve Bayes method, whose AUC was 90.22%, experienced a decrease of 2.6% compared to the method Random Forest Classifier.

Table 5. Recall comparison

Method	***Recall***
Extra Trees Classifier	89.18 %
Random Forest Classifier	87.18%
Naïve Bayes	84.09%

In testing carried out using the Extra Trees Classifier method, the Recall result was 89.18%, there was an increase of 1% compared to the Random Forest Classifier method, whose Recall was 87.18% and the Naïve Bayes method, whose Recall was 84.09%, experienced a decrease of 3.09% compared to the method Random Forest Classifier.

Table 6. Precision comparison

Method	***Precision***
Extra Trees Classifier	87.57%
Random Forest Classifier	85.25%
Naïve Bayes	87.00%

In testing carried out using the Extra Trees Classifier method, the Precision result was 87.57%, there was an increase of 0.57% compared to the Naïve Bayes method, whose Precision was 87.00%, and the Random Forest Classifier method, whose Precision was 85.25%, experienced a decrease of 1.75% compared to the method Naïve Bayes.

Table 7. F1-score comparison

Method	***F1-score***
Extra Trees Classifier	88.16%
Random Forest Classifier	85.77%
Naïve Bayes	85.09%

In testing carried out using the Extra Trees Classifier method, the F1-score was 88.16%, there was an increase of 2.39% compared to the Random Forest Classifier method, whose F1-score was 85.77% and the Naïve Bayes method, whose F1-score was 85.09%. a decrease of 0.68% compared to the Random Forest Classifier method of 85.77%.

 ISSN:2302-8149
e-ISSN:2540-9719

4.5 Evaluation

To see the visualization of the test, the ROC curve, confusion matrix, precision-recall curve and feature importance are displayed to find out the attributes that most influence the results of the classification model.

4.5.1 *Extra Trees Classifier*

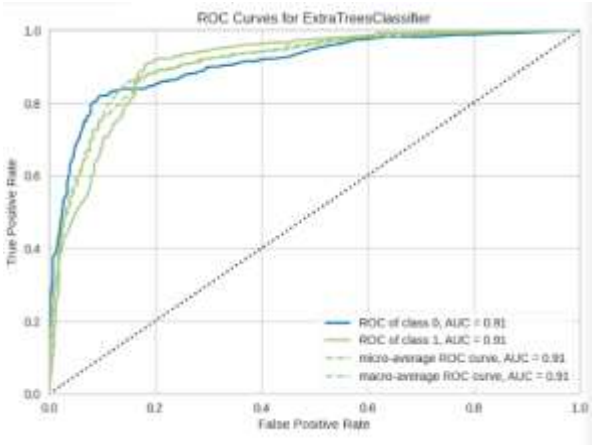

Figure 2. Extra trees urve

The Extra Trees ROC curve displays the ROC results for the class without cardiovascular disease of 0.91, the class with cardiovascular disease of 0.91. Meanwhile, the average ROC micro for the three classes is 0.91 and the average ROC macro is 0.91.

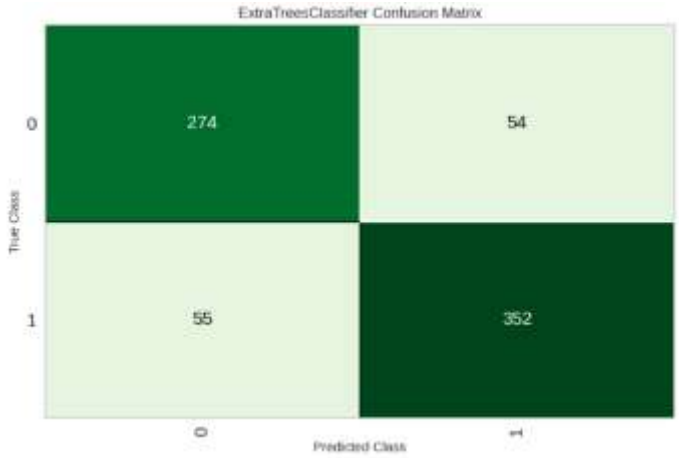

Figure 3. *Confusion Matrix Extra Trees*

Based on Figure 3, we can conclude that the use of the confusion matrix in predicting heart disease shows interesting results. There were 274 individuals who were predicted to be "Not/Healthy" and were truly healthy, and 54 individuals who were predicted to be "Not/Healthy" but turned out to be suffering from heart disease. In contrast, there were 55 individuals who were predicted "Yes/Infected" but were actually healthy, and 352 individuals who were predicted "Yes/Infected" actually had heart disease. Evaluation of this classification shows a fairly good level of accuracy between predictions and actual conditions.

 ISSN:2302-8149
e-ISSN:2540-9719

4.5.2 *Random Forest Classifier*

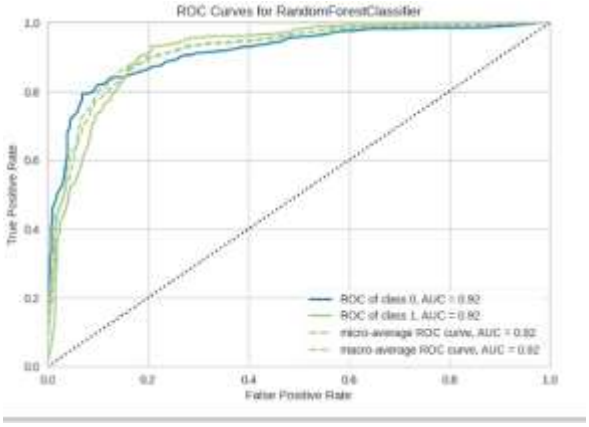

Figure 4. Random forest curve

The Random Forest ROC curve displays the ROC results for the class without cardiovascular disease of 0.92, the class with cardiovascular disease of 0.92. Meanwhile, the average ROC micro for the three classes is 0.92 and the average ROC macro is 0.92.

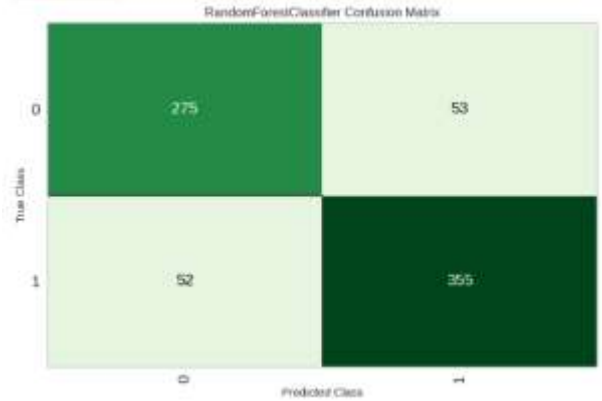

Figure 5. Confusion matrix random forest

Based on Figure 5, the confusion matrix shows that there were 275 individuals who were predicted to be "Not/Healthy" and were truly healthy, and 53 individuals who were predicted to be "Not/Healthy" but turned out to be suffering from heart disease. In contrast, there were 52 individuals who were predicted "Yes/Infected" but were actually healthy, and 355 individuals who were predicted "Yes/Infected" actually had heart disease. Evaluation of this classification shows a fairly good level of accuracy between predictions and actual conditions.

 ISSN:2302-8149
e-ISSN:2540-9719

4.5.3 *Naïve Bayes*

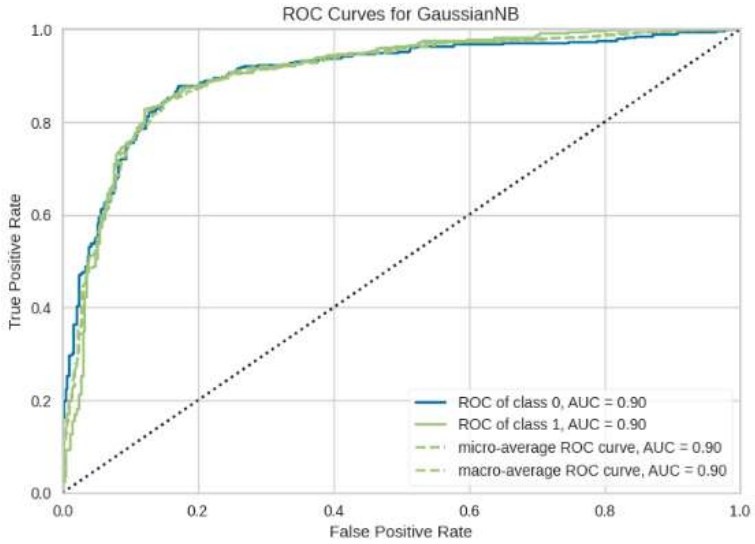

Figure 6. Naive bayes curve

The Extra Trees ROC curve displays the ROC results for the class without cardiovascular disease of 0.90, the class with cardiovascular disease of 0.90. Meanwhile, the average ROC micro for the three classes is 0.90 and the average ROC macro is 0.90.

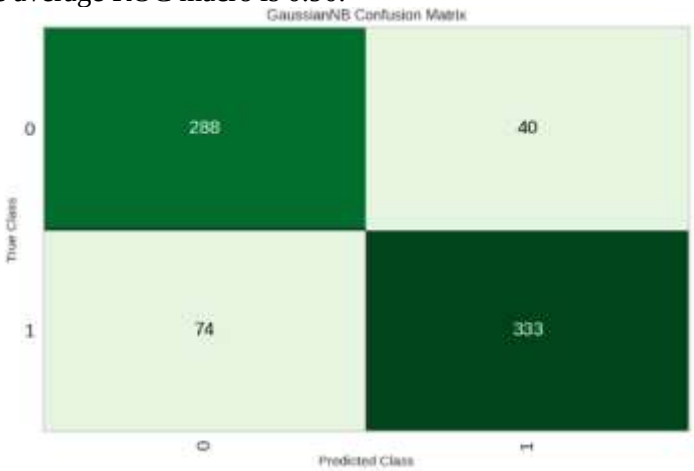

Figure 7. Confusion matrix naive bayes

Based on Figure 7, the confusion matrix shows that there were 288 individuals who were predicted to be "Not/Healthy" and were truly healthy, and 40 individuals who were predicted to be "Not/Healthy" but turned out to be suffering from heart disease. In contrast, there were 74 individuals who were predicted "Yes/Infected" but were actually healthy, and 333 individuals who were predicted "Yes/Infected" actually had heart disease. Evaluation of this classification shows a fairly good level of accuracy between predictions and actual conditions.

5 Conclusion

The conclusion of this research is that the Extra Trees Classifier, Naïve Bayes, and Random Forest Classifier methods have been compared in classifying heart disease status using a cardiovascular disease detection dataset. The dataset used consists of classes indicating Yes/Infected and No/Healthy conditions. This research uses split validation and train-test techniques to test these three methods, which has never been done in previous research. The results show that the Extra Trees Classifier method provides the highest accuracy of 86.93%, compared to the Naïve Bayes and Random Forest Classifier methods, which each have an accuracy of 84.21%. This research succeeded

 ISSN:2302-8149
e-ISSN:2540-9719

in showing that split validation and train-test techniques can improve the performance of heart disease classification models. With these results, the Extra Trees Classifier method can be considered a more effective option in detecting heart disease compared to other methods tested. This makes an important contribution to efforts to increase the accuracy of heart disease predictions, which in turn can help in earlier diagnosis and more effective disease prevention.

Reference

[1] A. U. L. H. et Al, "Identifying the Predictive Capability of Machine Learning Classifiers for Designing Heart Disease Detection System," *Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process*, pp. 130–138, 2019, doi: 10.1109/ICCWAMTIP47768.2019.9067519.

[2] C. Raju, E. Philipsy, S. Chacko, L. P. Suresh, and D. R. S, "A Survey on Predicting Heart Disease using Data Mining Techniques," *2018 Conf. Emerg. Devices Smart Syst.*, no. March, pp. 253–255, 2018.

[3] R. Jane *et al.*, "Prediction of Cardiac Disease using Supervised Machine Learning Algorithms Arun Raj Lakshminarayanan," no. Icicccs, pp. 570–575, 2020.

[4] S. Vineet, R. Akhtar, and H. Gauray, "Prediction of Heart Disease using DNN," *Int. Conf. Inven. Res. Comput. Appl.*, vol. 8, no. 2 Special Issue 6, pp. 486–489, 2020, doi: 10.35940/ijrte.B1092.0782S619.

[5] M. Saw, T. Saxena, S. Kaithwas, R. Yadav, and N. Lal, "Estimation of prediction for getting heart disease using logistic regression model of machine learning," *2020 Int. Conf. Comput. Commun. Informatics, ICCCI 2020*, pp. 20–25, 2020, doi: 10.1109/ICCCI48352.2020.9104210.

[6] J. R. Arkam, Wisudawan, A. S. F. Arsal, Nurhikmawati, and F. Sommeng, "Hubungan Faktor Resiko Penyakit Jantung terhadap Hasil Elektrokardiografi (EKG) pada Perawat UGD RS. Ibnu Sina," *Fakumi Med. J. J. Mhs. Kedokt.*, vol. 3, no. 1, pp. 36–44, 2023, doi: 10.33096/fmj.v3i1.177.

[7] M. Gandhi and S. N. Singh, "Predictions In Heart Disease Using Techniques of Data Mining," *2015 1st Int. Conf. Futur. Trends Comput. Anal. Knowl. Manag. ABLAZE 2015*, pp. 520–525, 2015, doi: 10.1109/ABLAZE.2015.7154917.

[8] E. K. Ampomah, Z. Qin, and G. Nyame, "Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement," *Inf.*, vol. 11, no. 6, 2020, doi: 10.3390/info11060332.

[9] Y. Guo, Y. Zhou, X. Hu, and W. Cheng, "Research on recommendation of insurance products based on random forest," *Proc. - 2019 Int. Conf. Mach. Learn. Big Data Bus. Intell. MLBDBI 2019*, pp. 308–311, 2019, doi: 10.1109/MLBDBI48998.2019.00069.

[10] A. Fattah, M. Mashat., P. S., and T. F., "A Decision Tree Classification Model for University Admission System," *Int. J. Adv. Comput. Sci. Appl.*, vol. 3, no. 10, pp. 17–21, 2012, doi: 10.14569/ijacsa.2012.031003.

[11] L. Abhishek, "Optical character recognition using ensemble of SVM, MLP and extra trees classifier," *2020 Int. Conf. Emerg. Technol. INCET 2020*, pp. 7–10, 2020, doi: 10.1109/INCET49848.2020.9154050.

[12] S. Liu, H. Li, Y. Zhang, B. Zou, and J. Zhao, "Random Forest-Based Track Initiation Method," *J. Eng.*, vol. 2019, no. 19, pp. 6175–6179, 2019, doi: 10.1049/joe.2019.0180.

[13] Heliyanti Susana, "Penerapan Model Klasifikasi Metode Naive Bayes Terhadap Penggunaan Akses Internet," *J. Ris. Sist. Inf. dan Teknol. Inf.*, vol. 4, no. 1, pp. 1–8, 2022, doi: 10.52005/jursistekni.v4i1.96.

[14] Naomi Chatrina Siregar, Riki Ruli A. Siregar, and M. Yoga Distra Sudirman, "Implementasi Metode Naive Bayes Classifier (NBC) Pada Komentar Warga Sekolah Mengenai Pelaksanaan Pembelajaran Jarak Jauh (PJJ)," *J. Teknol.*, vol. 3, no. 1, pp. 102–110, 2020.

[15] M. Marimuthu, M. Abinaya, K. S., K. Madhankumar, and V. Pavithra, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach," *Int. J. Comput. Appl.*, vol. 181, no. 18, pp. 20–25, 2018, doi: 10.5120/ijca2018917863.

[16] P. Motarwar, A. Duraphe, G. Suganya, and M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning," *Int. Conf. Emerg. Trends Inf. Technol. Eng. ic-*

ETITE 2020, 2020, doi: 10.1109/ic-ETITE47903.2020.242.

[17] H. El Hamdaoui, S. Boujraf, N. E. H. Chaoui, and M. Maaroufi, "A Clinical Support System for Prediction of Heart Disease using Machine Learning Techniques," *2020 Int. Conf. Adv. Technol. Signal Image Process. ATSIP 2020*, 2020, doi: 10.1109/ATSIP49331.2020.9231760.

[18] G. Haixiang, L. Yijing, L. Yanan, L. Xiao, and L. Jinling, "BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification," *Eng. Appl. Artif. Intell.*, vol. 49, pp. 176–193, 2016, doi: 10.1016/j.engappai.2015.09.011.

[19] H. M. Nawawi, S. Rahayu, J. J. Purnama, and S. I. Komputer, "Algoritma c4.5 untuk memprediksi pengambilan keputusan memilih deposito berjangka," vol. 16, no. 1, pp. 65–72, 2019.