Eksakta : Berkala Ilmiah Bidang MIPA
VOLUME 24 NO 04 2023, pp 513-522
ISSN : Print 1411-3724 — Online 2549-7464
DOI : https://doi.org/10.24036/eksakta/vol24-iss04/461

Eksakta
Berkala Ilmiah Bidang MIPA
http://www.eksakta.ppj.unp.ac.id/index.php/eksakta

Article

Machine Learning Classification for Detecting Heart Disease with K-NN Algorithm, Decision Tree and Random Forest

Ambran Hartono[1*], Lizky Aska Dewi[1], Elvan Yuniarti[1], Salsabila Tahta Hirani Putri[2], Try Surya Harahap[3]

[1]Department of Physics, Faculty of Science and Technology, UIN Syarif Hidayatullah Jakarta, Tangerang Selatan, Indonesia
[2]Department of Informatics System, Faculty of Science and Technology, UIN Syarif Hidayatullah Jakarta, Tangerang Selatan, Indonesia
[3]Southeast Asia Biodiversity Research Institute, Chinese Academy of Sciences & Center for Integrative Conservation, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, Yunnan 666303, China

Article Info

Article history :

Received August 25, 2023
Revised September 27, 2023
Accepted October 16,2023
Published December 30, 2023

Keywords :
Machine learning, K-NN algorithm, decision tree, random forest

Abstract. The heart is one of the most important organs for human; therefore, it needs to always be looked after and maintained properly. If it is not looked after and maintained properly, it will be at risk of disease. Currently, heart disease of various types still ranks first in deaths both in Indonesia and abroad. Various efforts continue to be developed by relevant scientists to detect it. Considering the importance of development efforts, in this research a machine-learning program was designing to classify heart disease as a detection system effort. In this article we will describe the analysis of the characteristics of the K-NN classifier, decision tree, random forest (accuracy, precision and recall), as well as determining the best classifier for detecting heart disease. To support the analysis of test results, Python and Google Colab programming has been implementation here. The best results obtained from the analysis of the application of these three models are the Decision Tree Classifier with accuracy, precision and recall values of 90%, 87% and 88% respectively. These results indicate that this model has been successfully developing

Corresponding Author :
Amdran Hartono
Department of Physics, Faculty of Science and Technology,
UIN Syarif Hidayatullah Jakarta, Tangerang Selatan, Indonesia
Email : ambran.hartono@uinjkt.ac.id

1. Introduction

Health is a condition that must be care and continuously good maintained, especially for the important organs of the human body, one of which is the heart. The heart itself is the most important organ for humans, where the heart is an organ the size of a human fist, which has the function of pumping blood throughout the body [1-5].

In Indonesia, cardiovascular disease ranks first according to the Sample Registration System survey with a coronary heart death rate of 12.9% of all deaths. Based on a doctor's diagnosis through Basic Health Research (Riskesdas) in 2013, this can indicate that heart disease ranks highest. For this reason, early detection is very necessary as a preventive measure [6-10]. Various studies related to this heart disease detection system continue to develop until now, such as Deep Learning, Genetic Algorithms and others with the accuracy value about 78 % [11-14]. However, it is still in dire need Innovation and development of models or systems to detect heart disease. Given the current detection, system is still complicated and sometimes expensive [15-17].

One possible development in the design of this detection system is machine learning. This very possible considering that this model has developed a lot but for other detection systems. To facilitate the process of identifying people who have heart disease, machine learning is used. Machine learning is a learning machine with a computational process and uses input data. Machine learning is a method to simplify the classification identification process in this study. This happens because heart disease is the most important cause of death in the world [12-13][18].

Machine learning is a technique for processing and analyzing previous data. The technique can predict new data and exploit hidden data. There are also several machine learning models such as classification, regression, and clustering. By training data processing into one pf the models, then the model predicts data is not visible from previous data. So that this machine learning technique can help to overcome data failures [7-8][12][19]. Machine learning can also explain the principles of algorithms, where with this ability machine learning can adapt for the purpose of efficiency and effectiveness of its function.

Machine learning is also concerned with how computers can build program performance with multiple task through experience. Learning is learning to predict supervised targets based on a training model using labeled data. The results of this model was using to predict new data. Supervised learning knows the relationship between input and target, and then it can be classification into main categories, i.e. Classification and regression [12][20-21].

Several previous studies has been carries out, including: implementation of classification using the k-NN algorithm, Decision Tree and Random Forest to determine Clean Water Quality. Between Classifiers Using Machine learning. The result is that machine learning able to read the cases in this study well [22]. Subsequent research by Haris and friends in 2021, implements machine learning to determine the value of radioactivity. As a result, machine learning can work well [23]. Then the third study (Nanik, Sarfiah, 2021), entitled "Random Forest Classifier for detection of Covid-19 Patients with CT Scan Images". The results of the random forest algorithm have a highest accuracy value compared to other method [24]. Based on these studies it is very possible to apply this Machine Learning application for classification and identification of heart disease. This factor is the focus of ttis research.

2. Experimental Section

2.1. Materials

In this research required some equipment and materials. The equipment used includesLenovo Idea pad Slim 3 brand laptop with Windows 11 operating system specifications, AMD 3020e CPU, 1.2GHz, 4GB RAM, 256GB SSD and Google Collabotary Python. While the material used in this research are kaggle.com dataset.

2.2. Experiment Stages

This research generally has several stages, namely problem identification, literature study, programming, retrieving data sets, then testing the program, and analyzing the results of the program that has been made. Schematically shown in Figure 1.

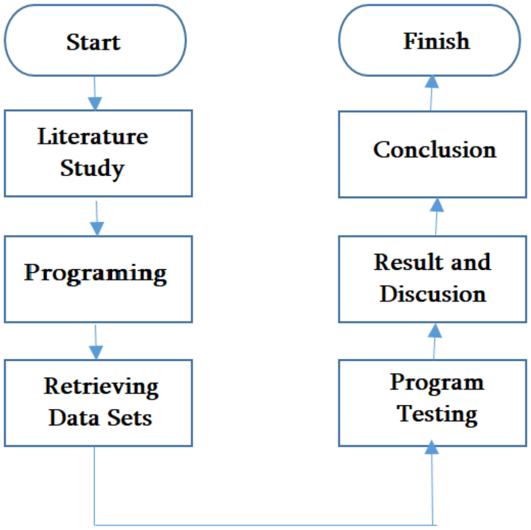


Figure 1. Schematically of research

1.2.1. Programming

Making this program was doing using Python with Google Collaboratory, which is an executable document, used to share programs that has been written via Google Drive. Through Google Collab, programs created using machine learning k-NN classification, decision tree and random forests [25-28]. The program is designing to study data patterns from existing data, so that the results of the program produce expenses as expected. Where the expected values are accuracy, precision and recall.

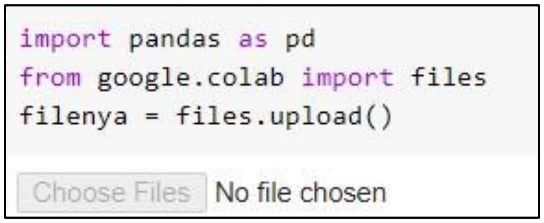
1.2.2. Retrieving Datasets

Retrieval of data sets is doing by taking data that is already available on the Kaggle website, which is a common data science learning resource platform, so there are many data sets available there. The data set used for this study is the heart data set [29-31]. By Ratndeep Chavan (2021) with a total of 917 data and parameters per person are 12 parameters with 21 parameters used.

1.2.3. Program Testing

Program testing is carries out with data sets that has obtained and then processed in machine learning. The data sets obtained from Kaggle is a Comma Separated Values or SCV file and will be retrieved using the pandas library and import files because it uses Google Colab, which will be read later [32-

34]. For datasets, processing doing with the program as well as the list of programs in the table as shown in Figure 2.



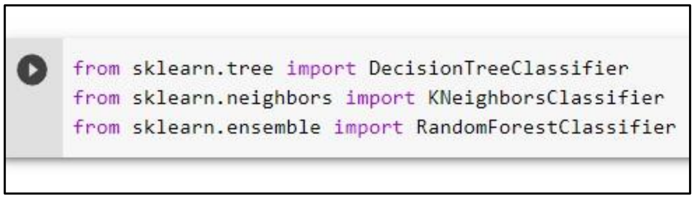
```
import pandas as pd
from google.colab import files
filenya = files.upload()
```

Figure 2. Program to include files in machine learning

After the program enters machine learning then the data set will be divided by df.isna.sum () which has a function to delete columns that have nan data, but data because nan data in this data set is 0 then it goes straight to the next process. In the df = pf.get dummies process, one hot encoding is carried out, which is a process for creating new columns with categorical variables, with each category being a new column, Then select feature data and target data, select heart disease which, is the target data selected to find out disease data.

Therefore, we get X for data features and Y for data target. After the process of dividing the data features and target data is done, from the sklearn.model library to be able to share test data and train data, by dividing 90% training data and 10% test data, after obtaining train data and data sets, scaling is done for better data obtained using MinMaxScaler [35-37].

Furthermore, the classifier needs to be taken from sklearn for the algorithm that will be used machine learning, for machine learning for the Decision Tree it will be taken "Decision Tree Classifier", for k-Nearest Neighbor taken from "K Neighbors Classifier", and Random Tree Forest taken from "Random Forest Classifier". Then all algorithm are tested using validation data and test data to be able to study machine learning with the library shown in Figure 3.

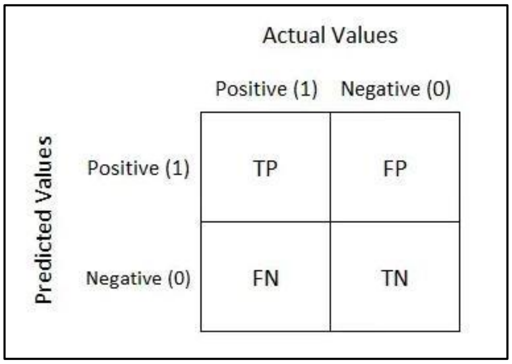


```
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
```

Figure 3. Library classifier for machine learning algorithm

At the programming stage, the data set has processed by separating the data set parameters. The separated parameter values are integers and objects, such as gender, age, type op pain, resting ECG, exercise Angina, and ST slope. Then the parameter values that has been separating are using as X and Y variables to determine the train set and test set. By dividing the 75% train and 25% test set, machine learning works randomly so that random results are obtain.

Then the training data obtained is 826 and 92 for the test set, in which the data is scaling to produce better data. Furthermore, data validation was carried out, this data was obtain with a total of 743 train sets and 83 test set, and then the data was processed using each classifier. The results that are issue from the algorithm are in the form of confusion matrix with several provisions for accuracy, precision, and recall values. As shown in Figure 4. The Confusion matrix is a representation of predictions and actual conditions from data generated by machine learning, in addition to the provisions of the values mentioned above, there are predicted values in the form of True Positive, True Negative, False Positive and False Negative.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 4. Confusion matrix [32]

True positive is the case where the predicted result actually occurs (true), then, for True Negative is the result where the predicted does not occurs. Then for False Positive is a case with a prediction that should have happened but did not happen or the prediction was wrong, and the last one is False Negative, which is a case where the predicted result did not happen but actually happened [34][36-37].

3. Results and Discussion

In this section, the results obtained from the characteristic testing process should be explaining, namely the accuracy and precision test results from machine learning. Here we will describe the accuracy and precision of machine learning based on the k-NN classifier, decision tree and random forest. Accuracy in machine learning is the ratio of correct predictions (positive and negative) of the entire data. If the accuracy of the algorithm is good then the data set has the amount of data for False Negative and False Positive are nearly symmetrical.

Accuracy of the data validation of the k-NN algorithm are highest at k = 0 with an accuracy of 84%, then for accuracy at k = 1 it is 77% and this the lowest point in accuracy, then at k = 2 it is 82% and stable at k = 3, then at k = 4 the accuracy decreased by 81 %. For test data, the highest value was obtained at k = 0 by 64%, decreased by k = 1 and k = 2 by 57% and 48% and at k = 3, the value was 54% and k = 4 by 69%, as shown in Figure 5.

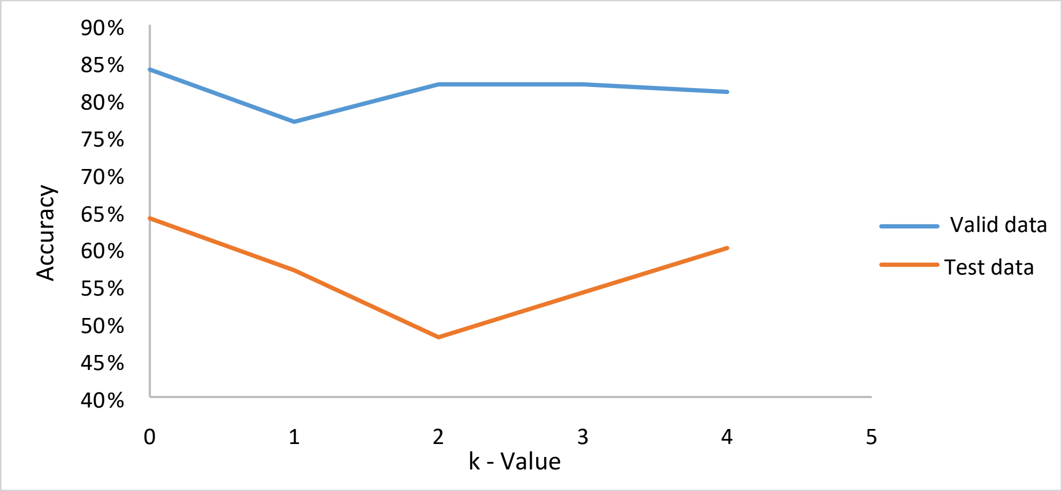


Figure 5. Plot graph of accuracy at k – NN algorithm

The accuracy value on the graph for the validation data k = 0 is 84% and the test data is 64%, this happens because k = 0 does not use the scikit knn.predict and knn.predict_probe libraries where these libraries can predict class labels for data that has been provided and can be returned to the estimated probability for the test data. Whereas for values k = 1 to k = 4 use the scikit library [33].

The value of the confusion matrix taken is the value k = 0 for test data of 64% which is the highest, the test data is the data used after finishing testing the training process, which means this test data cannot be seen before [34]. Then the Confusion Matrix test data as shown in Figure 6, the value k = 0 gets 12 True positive (heart disease) and 47 True Negative (no heart disease), then for False Positive it gets 4 (predicted heart disease but not heart disease) and False Negative obtained 29 (predicted heart disease but no heart disease). With an accuracy value of k = 0 for test data is 64%.

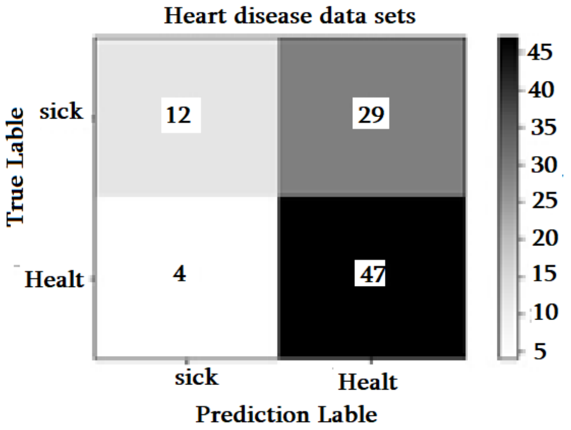


Figure 6. K-NN confusion matrix

In the decision tree calculation, two data has obtained namely validation data and test data. Test data has obtained with a true positive ratio of 90% and validation data of 84%. Just like the explanation in the k-NN algorithm that only test data has been using as the result in this study. The results of the accuracy of the test data for the Decision tree algorithm are obtain True Positive 41 (heart disease), the for False Positive of 3 (patients who are predicted to have heart disease but actually do not have heart disease) and False Negative of 6 (patients who are predicted not have heart disease but actually heart disease). In addition, the accuracy obtained from the test data is 90% with the Confusion Matrix as shown in Figure 7 detail.

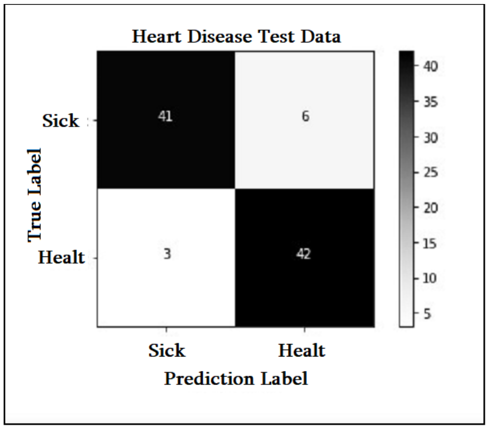


Figure 7. Decision tree confusion matrix

For Random Forest calculations, two data are obtain, namely validation data and test data. Data test was obtain with a true positive ratio 87% and validation data of 88%. The results of the accuracy of the data test on the Random Forest algorithm has obtain True Positive 34 (heart disease) and True Negative 46 (no heart disease), then for False Negative obtain 5 (predicted heart disease but not heart disease) and False Positive 7 (predicted no heart disease but heart disease), with accuracy value of 87%. As shown in Figure 8.

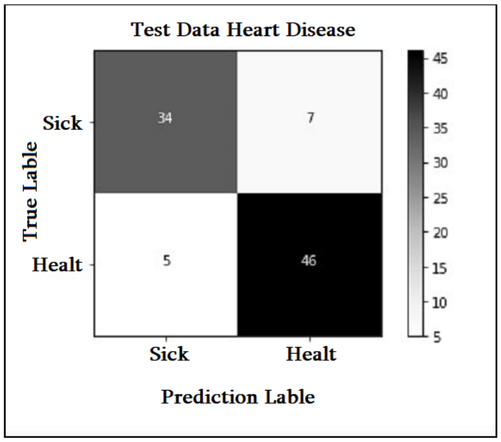


Figure 8. Random forest confusion matrix

Furthermore, the results of the Precision of true positive sick data on the k – NN Algorithm as shown in Figure 9. The results on the precision of k-NN pain for the highest validation data are at k = 0 with a result of 83% and the smallest is obtained at 74% at k = 4. For Values k =1 to 3, respectively obtained 75%, 75% and 78%. The results on precision pain k-NN for test data at value of k = 0 to 4 are obtained respectively 83%, 75%, 75% and 54%. Precision results for validation data on the Random Forest algorithm were obtain at 86% for sick and 89% for healthy, while for test data it was obtained for 87% for both healthy and sick.

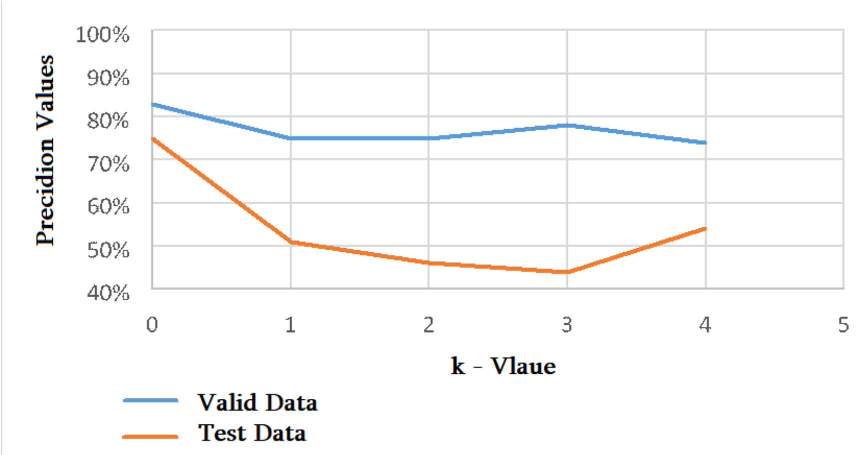


Figure 9. Precision of true positive sick data on the k – NN algorithm

Overall, the test results of the developed model as shown in Table 1. Table 1 shows the level accuracy, precision and recall of all classifiers.

Table 1. Table comparison of test data on all classifier algorithms

Algorithms	Accuracy	Precision	Recall	Validation Data of Accuracy Values (%)
k-NN	0.62	0.64	0.29	84
Decision Tree	0.90	0.88	0.93	84
Random Forest	0.87	0.87	0.90	88

From the overall results of the test conducted, it shows that the built model work well, this indicates that the developed model can be an alternative in detected heart disease, besides that it is also a breakthrough in heart disease detection systems, and the best model from this research is Decision Tree Classifier. The level of accuracy obtain from this research is better than that carried out by several previous researchers with an accuracy of around 78% [11 – 14]. These results indicate that this model has been successfully developing.

4. Conclusion

Based on the results of testing and analysis that has been carries out by researches, it can be conducted that accuracy, precision and recall values in the test data for the k-NN algorithm, decision tree and random forest obtained results for k-NN obtained accuracy of 62%, precision of 64% and recall of 29%. Decision tree with accuracy of 90%, precision of 88%, and recall of 93%. Then for the random forest, results obtained an accuracy of 87%, precision of 87%, and a recall of 90%. Based on the accuracy value, it shows that the best classifier is a Decision Tree.

5. Acknowledgement

This study was supported by Physics Department of Science and Technology Faculty at UIN Syarif Hidayatullah Jakarta.

References

[1] Aazmi, A., Zhou, H., Li, Y., Yu, M., Xu, X., Wu, Y., ... & Yang, H. (2022). Engineered vasculature for organ-on-a-chip systems. *Engineering*, *9*, 131-147.

[2] Roberts, W., Salandy, S., Mandal, G., Holda, M. K., Tomaszewksi, K. A., Gielecki, J., ... & Loukas, M. (2019). Across the centuries: Piecing together the anatomy of the heart. *Translational Research in Anatomy*, *17*, 100051.

[3] Alnour, H., Sharma, A., Halawa, A., & Alalawi, F. (2021). Global practices and policies of organ transplantation and organ trafficking. *Experimental and clinical transplantation*.

[4] Aubert, O., Yoo, D., Zielinski, D., Cozzi, E., Cardillo, M., Dürr, M., ... & Loupy, A. (2021). COVID-19 pandemic and worldwide organ transplantation: a population-based study. *The Lancet Public Health*, *6*(10), e709-e719.

[5] Gun, S. Y., Lee, S. W. L., Sieow, J. L., & Wong, S. C. (2019). Targeting immune cells for cancer therapy. *Redox biology*, *25*, 101174.

[6] Ghani, L., Susilawati, M. D., & Novriani, H. (2016). Faktor risiko dominan penyakit jantung koroner di Indonesia. *Buletin Penelitian Kesehatan*, *44*(3), 153-164.

[7] Brunese, L., Martinelli, F., Mercaldo, F., & Santone, A. (2020). Deep learning for heart disease detection through cardiac sounds. *Procedia Computer Science*, *176*, 2202-2211.

[8] Duchateau, N., King, A. P., & De Craene, M. (2020). Machine learning approaches for myocardial motion and deformation analysis. *Frontiers in cardiovascular medicine*, *6*, 190.

[9] Gu, X., Jiang, Y., & Ni, T. (2020). Discriminative neural network for coronary heart disease detection. *Journal of Medical Imaging and Health Informatics*, *10*(2), 463-468.

[10] Mercaldo, F., & Santone, A. (2020). Deep learning for image-based mobile malware detection. *Journal of Computer Virology and Hacking Techniques*, *16*(2), 157-171.

[11] Pathak, A. K., & Arul Valan, J. (2019). A predictive model for heart disease diagnosis using fuzzy logic and decision tree. In *Smart Computing Paradigms: New Progresses and Challenges: Proceedings of ICACNI 2018, Volume 2* (pp. 131-140). Singapore: Springer Singapore.

[12] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms*, *16*(2), 88.

[13] Brunese, L., Martinelli, F., Mercaldo, F., & Santone, A. (2020). Deep learning for heart disease detection through cardiac sounds. *Procedia Computer Science*, *176*, 2202-2211.

[14] Uyar, K., & İlhan, A. (2017). Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. *Procedia computer science*, *120*, 588-593.

[15] Halim, W., & Mudjihartono, P. (2022). Kecerdasan Buatan dalam Teknologi Kedokteran: Survey Paper. *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi*, *2*(1).

[16] Sapra, V., Sapra, L., Bhardwaj, A., Bharany, S., Saxena, A., Karim, F. K., ... & Mohamed, A. W. (2023). Integrated approach using deep neural network and CBR for detecting severity of coronary artery disease. *Alexandria Engineering Journal*, *68*, 709-720.

[17] Ueda, D., Matsumoto, T., Ehara, S., Yamamoto, A., Walston, S. L., Ito, A., ... & Miki, Y. (2023). Artificial intelligence-based model to classify cardiac functions from chest radiographs: a multi-institutional, retrospective model development and validation study. *The Lancet Digital Health*, *5*(8), e525-e533.

[18] Tajudin, T., & Nugroho, I. D. W. (2020). Analisis Kombinasi Penggunaan Obat pada Pasien Jantung Koroner (Coronary Heart Disease) dengan Penyakit Penyerta di Rumah Sakit X Cilacap tahun 2019. *Pharmaqueous: Jurnal Ilmiah Kefarmasian*, *1*(2), 6-13.

[19] Kumar, N., & Makkar, A. (2020). *Machine learning in cognitive IoT*. CRC Press.

[20] Heryadi, Y., & Wahyono, T. (2020). Machine learning konsep dan implementasi. *Yogyakarta: Gava Media*.

[21] Soofi, A. A., & Awan, A. (2017). Classification techniques in machine learning: applications and issues. *Journal of Basic & Applied Sciences*, *13*(1), 459-465.

[22] Sutisna, S., & Yuniar, M. N. (2023). Klasifikasi Kualitas Air Bersih Menggunakan Metode Naïve baiyes. *Jurnal Sains dan Teknologi*, *5*(1), 243-246.

[23] Xu, D., Shi, Y., Tsang, I. W., Ong, Y. S., Gong, C., & Shen, X. (2019). Survey on multi-output learning. *IEEE transactions on neural networks and learning systems*, *31*(7), 2409-2429.

[24] Normah, N., Rifai, B., Vambudi, S., & Maulana, R. (2022). Analisa Sentimen Perkembangan Vtuber Dengan Metode Support Vector Machine Berbasis SMOTE. *Jurnal Teknik Komputer AMIK BSI*, *8*(2), 174-180.

[25] Uddin, K. M. M., Ripa, R., Yeasmin, N., Biswas, N., & Dey, S. K. (2023). Machine learning-based approach to the diagnosis of cardiovascular vascular disease using a combined dataset. *Intelligence-Based Medicine*, *7*, 100100.

[26] D'Souza, A. (2015). Heart disease prediction using data mining techniques. *International Journal of Research in Engineering and Science (IJRES) ISSN (Online)*, 2320-9364.

[27] Loesche, W. J. (1994). Periodontal disease as a risk factor for heart disease. *Compendium (Newtown, Pa.)*, *15*(8), 976-978.

[28] Learning, M. (2017). Heart disease diagnosis and prediction using machine learning and data mining techniques: a review. *Advances in Computational Sciences and Technology*, *10*(7), 2137-2159.

[29] Chandrasekhar, N., & Peddakrishna, S. (2023). Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization. *Processes*, *11*(4), 1210.

[30] Pathan, M. S., Nag, A., Pathan, M. M., & Dev, S. (2022). Analyzing the impact of feature selection on the accuracy of heart disease prediction. *Healthcare Analytics*, *2*, 100060.

[31] Sivapalan, G., Nundy, K. K., Dev, S., Cardiff, B., & John, D. (2022). ANNet: a lightweight neural network for ECG anomaly detection in IoT edge sensors. *IEEE Transactions on Biomedical Circuits and Systems*, *16*(1), 24-35.

[32] Sivapalan, G., Nundy, K. K., James, A., Cardiff, B., & John, D. (2023). Interpretable rule mining for real-time ECG anomaly detection in IoT Edge Sensors. *IEEE Internet of Things Journal*.

[33] Gavhane, A., Kokkula, G., Pandya, I., & Devadkar, K. (2018, March). Prediction of heart disease using machine learning. In *2018 second international conference on electronics, communication and aerospace technology (ICECA)* (pp. 1275-1278). IEEE.

[34] Kumar, N. K., Sindhu, G. S., Prashanthi, D. K., & Sulthana, A. S. (2020, March). Analysis and prediction of cardio vascular disease using machine learning classifiers. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 15-21). IEEE.

[35] Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2018, July). Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)* (pp. 268-282). IEEE.

[36] Sharma, A., Kumar, N., Kumar, A., Dikshit, K., Tharani, K., & Singh, B. (2021). Comparative investigation of machine learning algorithms for detection of epileptic seizures. *Intelligent Decision Technologies*, *15*(2), 269-279.

[37] Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. *Computers in biology and medicine*, *112*, 103375.