Jurnal Informatika Universitas Pamulang
Penerbit: Program Studi Teknik Informatika Universitas Pamulang
Vol. 8, No. 3, September 2023 (420-426)
ISSN: 2541-1004
e-ISSN: 2622-4615
10.32493/informatika.v8i3.34103

Performance Evaluation of Popular Supervised Learning Algorithms Towards Cardiovascular Disease

Anis Fitri Nur Masruriyah[1], Hilda Yulia Novita[2], and Cici Emilia Sukmawati[3]

Department Information, Universitas Buana Perjuangan Karawang, Jl. HS.Ronggo Waluyo
Karawang, Indonesia, 41361
e-mail: [1]anis.masruriyah@ubpkarawang.ac.id, [2]hilda.yulia@ubpkarawang.ac.id,
[3]cici.emilia@ubpkarawang.ac.id

Submitted Date: September 11th, 2023 Reviewed Date: September 27th, 2023
Revised Date: September 29th, 2023 Accepted Date: September 29th, 2023

Abstract

Heart disease is a global health concern, being the leading cause of death worldwide. Annually, over 17 million people lose their lives to various types of heart disease, including coronary heart disease, heart failure, and stroke. Unfortunately, access to adequate and affordable healthcare remains a challenge in certain areas, which can hinder early detection and management of heart disease. Thus, addressing this problem necessitates a collaborative effort from various entities, including governments, health organizations, the medical community, and individuals. Prevention, early detection, and management of heart disease is a top global health priority. Advances in technology and computer science are expected to play a more significant role in the future in the prevention, diagnosis and management of heart disease. Several studies have discussed the benefits of supervised learning for dealing with extensive heart disease data. However, only a few have evaluated the performance of these algorithms. This study built a classification model utilizing supervised learning algorithms, including C4.5, Random Forest, Logistic Regression, and Support Vector Machine. The data processed are in the form of category data with character data types. The results of accuracy, precision, and performance evaluation indicate that Logistic Regression Algorithm has the most superior value compared to the other algorithms. However, it was found that the C4.5 and SVM algorithms had anomalous events. Although C4.5 had higher accuracy and precision than SVM, SVM had better performance.

Keywords: Cardiovascular Disease; Performance Evaluation; Supervised Learning

1. Introduction

Heart disease is a significant global health problem, including in Indonesia, and is the leading cause of death worldwide (Braunwald, 2019) (Ministry of Health Republic of Indonesia, 2014) (World Health Organization, 2019). According to the World Health Organization (WHO), more than 17 million cardiovascular disease deaths occur yearly. Heart disease also imposes a significant economic burden on healthcare systems worldwide, including medical care costs and lost economic productivity (Setyaji et al., 2018) (Utami, 2017). Despite improvements in the health system in Indonesia, access to quality health care is still limited in some areas, especially in rural areas (Maharani et al., 2019) (Setyaji et al., 2018) (Sofiana et al., 2021). In addition, public awareness about the importance of preventing and managing heart disease still needs to be increased. Health education and the promotion of healthy lifestyles are essential. Then, delays in treating heart disease are risky and can have serious medical consequences. Therefore, it is crucial to identify risk factors for heart disease, undergo regular health check-ups, and seek timely medical treatment if symptoms or signs of heart disease are present (Çakmak & Demir, 2020) (Ciumărnean et al., 2022) (Komalasari et al., 2019). Prevention and good management of heart disease is essential to avoid potentially fatal complications.

Previously, research in the field of computer science has been conducted, proving that data mining in heart disease has excellent potential to improve early recognition, more accurate diagnosis, more personalized treatment, and more effective management of heart disease (Balakrishnan et al., 2021) (Krittanawong et al., 2020) (Maiga et al., 2019) (Mathur et al., 2020) (Mezzatesta et al., 2019) (Padmanabhan et al., 2019). So, it can help reduce the death rate due to heart disease and improve the patient's quality of life.

Research conducted by Maiga(2019) employed Random Forest, Naïve Bayes, K-Nearest Neighbor (KNN), and Logistic Regression in modeling heart disease diagnoses. A total of 70,000 medical records of patient data were utilized in the study, and the comparison results revealed that Random Forest had a high classification accuracy of 73%, specificity of 65%, and sensitivity of 80%. Furthermore, Padmanabhan (2019) conducted a study to debunk the outdated belief that biomedical researchers must possess extensive knowledge of underlying algorithms in order to construct effective machine-learning classification models. The research utilized heart disease data to compare the efficacy of machine learning automation techniques with that of a graduate student, analyzing key metrics such as total time spent building the model and final classification accuracy on a brand new test dataset. The results revealed that with machine learning automation, a classification model can be assembled in just one hour and perform better than a graduate student's model built over the course of a month.

Moreover, extensive research has been conducted in the field of cardiovascular medicine and it has been found that artificial intelligence has been effectively used in cardiovascular imaging, predicting cardiovascular risk, and identifying potential drug targets. The study by Mathur et al. (2020) highlighted the significant advancements made in comprehending different heart failure and congenital heart disease phenotypes, which has led to the development of innovative treatment approaches for various cardiovascular conditions. These include drug therapy and post-marketing surveys of prescription drugs.

Krittanawong’s (2020) study revealed that machine learning (ML) algorithms, such as support vector machine (SVM) and boosting, are highly promising in predicting cardiovascular disease. However, their performance may vary based on different parameters, which is a vital factor for clinicians to consider when selecting the most optimal algorithm for their datasets. Specifically, the research showed that the boosting algorithm had an area under the curve (AUC) of 0.88 when predicting coronary artery disease, while the custom algorithm exhibited an even higher AUC of 0.93. These findings provide important insights into the efficacy of different ML algorithms in predicting cardiovascular disease and can aid clinicians in making informed decisions when selecting the most appropriate algorithm for their datasets.

Furthermore, Balakrishnan (2021) utilized pre-processing techniques such as eliminating noisy and missing data, filling in default values when necessary, and categorizing attributes at different levels for prediction and decision-making purposes. Additionally, the performance of the diagnosis model was assessed using classification, accuracy, sensitivity analysis, and specificity. To achieve this, the accuracy of implementing rules to individual outcomes from various algorithms, including Support Vector Machine, Gradient Boosting, Random Forest, Naive Bayes classifier, and logistic regression, was compared on datasets obtained from a specific region to develop a precise cardiovascular disease prediction model. The Random Forest algorithm yielded the highest accuracy rate of 92.4%.

The purpose of this study is to thoroughly examine the effectiveness of various algorithms that have been utilized in prior research studies involving heart disease data. This data has been sourced from the Heart Disease Maps and Data Sources(National Center for Chronic Disease Prevention and Health Promotion Division for Heart Disease and Stroke Prevention, 2022), which is a division of the National Center for Chronic Disease Prevention and Health Promotion that specifically focuses on heart disease and stroke prevention. Through this evaluation, we hope to gain a better understanding of how these algorithms operate and their potential implications for improving heart disease treatment and prevention.

2. Methods

For this study, we made use of a dataset sourced from an annual survey conducted by the Centers for Disease Control (CDC) (National Center for Chronic Disease Prevention and Health Promotion Division for Heart Disease and Stroke Prevention, 2022). The survey collected information from a significant number of adults, specifically 300000, with regard to their health status. In order to ensure the quality of the data, we conducted a thorough examination to ensure that it was complete. We then standardized and normalized the data, paying close attention to missing entries, which we addressed by either completing them or eliminating affected rows or columns. Additionally, we meticulously identified and resolved any outliers that could potentially affect our analysis or model accuracy. Finally, we corrected any inaccuracies or disparities present in the data, ensuring that our analysis was based on reliable and accurate information.

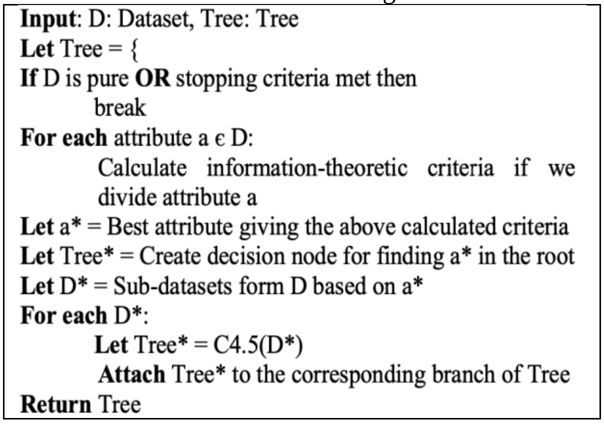
In order to create a model, we utilize several different algorithms including the C45 algorithm, random forest, Support Vector Machine (SVM), and Logistic Regression. The C4.5 algorithm is a crucial decision tree learning algorithm that constructs decision trees based on dataset attributes (Cherfi et al., 2018) (Hartshorn, 2020) (Jijo & Abdulazeez, 202) (Liu et al., 2017). The algorithm begins by selecting the best attribute using information measurements like entropy (equation 1) and information gain or gain ratio (equation 2) to divide the data into two branches or child nodes. This process is repeated recursively on each branch until a stopping criterion is met, such as when all data in a branch is of the same class or the maximum depth is reached. After the tree is formed, pruning is done to avoid overfitting. Certain branches or nodes that are not significant or cause overfitting can be removed. Finally, the model is tested by passing data through the tree, following the appropriate branch based on data attributes, until it reaches a leaf with a class label. Pseudocode 1 provides a clear overview of how the algorithm works.

$$E(S) = \sum_{i=1}^{c} -p_i log_2 \, p_i \tag{1}$$

$$Information\ Gain = Entropy(before) - \sum_{j=1}^{k} Entropy(j, after) \tag{2}$$

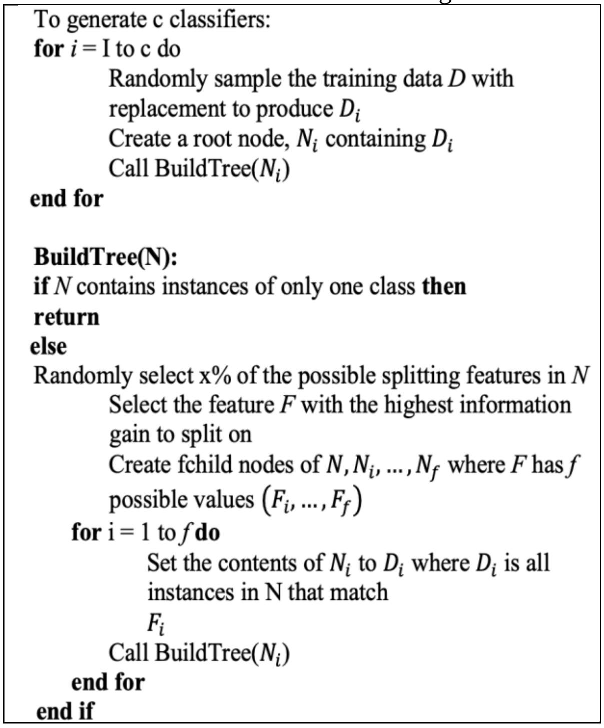
The Random Forest algorithm is a method of ensemble learning that employs multiple decision trees for classification purposes (Hartshorn, 2020) (Primajaya & Sari, 2018). In order to achieve this, Random Forest employs Bootstrap Aggregating (Bagging), where a random sample with replacement (bootstrap) is selected from the training dataset for each tree to be created. As a result, each tree in the forest sees a slightly different dataset. Random Feature Selection is then performed by randomly selecting attributes from all available attributes, which helps prevent overfitting and introduces variation between the trees in the forest. Finally, each tree provides a class prediction during the prediction stage, and the most common class is taken as the final result in classification. The Random Forest algorithm follows the process outlined in Pseudocode 2.

Pseudocode 1. C4.5 Algorithm



```
Input: D: Dataset, Tree: Tree
Let Tree = {
If D is pure OR stopping criteria met then
        break
For each attribute a ϵ D:
        Calculate information-theoretic criteria if we
        divide attribute a
Let a* = Best attribute giving the above calculated criteria
Let Tree* = Create decision node for finding a* in the root
Let D* = Sub-datasets form D based on a*
For each D*:
        Let Tree* = C4.5(D*)
        Attach Tree* to the corresponding branch of Tree
Return Tree
```

Pseudocode 2. Random Forest Algorithm

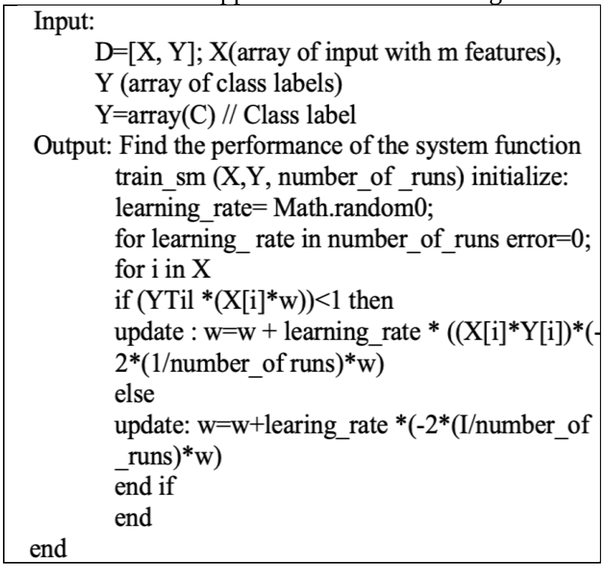


```
To generate c classifiers:
for $i$ = I to c do
        Randomly sample the training data $D$ with
        replacement to produce $D_i$
        Create a root node, $N_i$ containing $D_i$
        Call BuildTree($N_i$)
end for

BuildTree(N):
if $N$ contains instances of only one class then
return
else
Randomly select x% of the possible splitting features in $N$
        Select the feature $F$ with the highest information
        gain to split on
        Create fchild nodes of $N, N_i, \dots, N_f$ where $F$ has $f$
        possible values $(F_i, \dots, F_f)$
    for i = 1 to $f$ do
            Set the contents of $N_i$ to $D_i$ where $D_i$ is all
            instances in N that match
            $F_i$
        Call BuildTree($N_i$)
    end for
end if
```

Furthermore, the Support Vector Machine (SVM) algorithm is able to determine the hyperplane that maximizes the distance between two distinct groups(Koda et al., 2018) (Mahmoud & Ren, 2019) (Noor et al., 2019). This distance, referred to as the margin, represents the gap between the hyperplane and the closest data points of each group, known as support vectors. Utilizing a kernel can further optimize SVM's efficacy. A kernel is a function that transforms the initial data into a higher dimension, allowing for easier separation of the data. This approach empowers SVM to tackle non-linear classification problems in the original dimension. After identifying the hyperplane, SVM can predict the class of test data based on its position relative to the hyperplane. The SVM process is illustrated in Pseudocode 3.

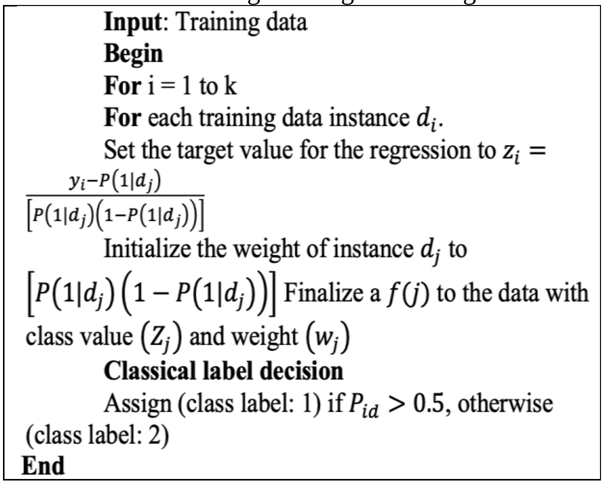
Pseudocode 3. Support Vector Machine Algorithm



```
Input:
        D=[X, Y]; X(array of input with m features),
        Y (array of class labels)
        Y=array(C) // Class label
Output: Find the performance of the system function
          train_sm (X,Y, number_of _runs) initialize:
          learning_rate= Math.random0;
          for learning_ rate in number_of_runs error=0;
          for i in X
          if (YTil *(X[i]*w))<1 then
          update : w=w + learning_rate * ((X[i]*Y[i])*(-
          2*(1/number_of runs)*w)
          else
          update: w=w+learing_rate *(-2*(I/number_of
          _runs)*w)
          end if
          end
end
```

In the realm of classification problems, the logistic regression algorithm is a commonly employed statistical method (Ath et al., 2022) (Handayani, 2021) (Ho et al., 2020). Its technique involves utilizing the logistic function (sigmoid) to model the probability that the data belongs to a specific class. The sigmoid function transforms linear values into probability values, ranging from 0 to 1. By selecting a threshold based on these probabilities, we make a classification decision. For instance, if the probability exceeds 0.5, we predict it as class 1; if it is less than or equal to 0.5, we predict it as class 0. Logistic regression seeks the best parameter through maximum entropy classification, where probability is represented by p and a, b are model parameters, with a as a factor. You can find a demonstration of the logistic regression workflow in pseudocode 4.

Pseudocode 4. Logistic Regression Algorithm



```
        Input: Training data
        Begin
        For i = 1 to k
        For each training data instance $d_i$.
        Set the target value for the regression to $z_i =$
$\frac{y_i - P(1|d_j)}{[P(1|d_j)(1-P(1|d_j))]}$
        Initialize the weight of instance $d_j$ to
$\left[P(1|d_j)\left(1-P(1|d_j)\right)\right]$ Finalize a $f(j)$ to the data with
class value $(Z_j)$ and weight $(w_j)$
        Classical label decision
        Assign (class label: 1) if $P_{id} > 0.5$, otherwise
(class label: 2)
End
```

In order to achieve the primary goal of this study, the data is subsequently partitioned into training and test data through the utilized of K-Fold Cross Validation, a method that impartially distributes data (Djatna et al., 2018) (Masruriyah et al., 2019) (Mia et al., 2022).

3. Result and Dicussion

After undergoing pre-processing, the data was categorized based on character types and then split into training and test sets using K-Fold with K-10. Table 1 showcases the accuracy and precision outcomes of the model. The Logistic Regression algorithm delivered the highest accuracy and precision results, at 91.5% and 88.9% respectively. In contrast, the SVM algorithm had the lowest values, with 80.4% accuracy and 84.2% precision. Random Forest and C.45 algorithms showed identical accuracy values but slightly different precision values.

Tabel 1 Accuracy and Precision Report

Model	Accuray (%)	Precision (%)
Logistic Regression	91.5	88.9
Random Forest	90.9	87.8
C4.5	90.9	87.4
SVM	80.4	84.2

When determining how dependable a built model is, simply relying on accuracy and precision may not be sufficient. For a more comprehensive understanding of the model's performance, an evaluation was carried out using Receiver Operating Characteristics (ROC) and Area Under the Curve (AUC). This particular evaluation technique allows for the measurement of True Positive Rate (sensitivity) versus False Positive Rate (1-specificity) trade-offs, particularly when comparing different models. The AUC metric is a singular value that gauges how effectively the model distinguishes between classes in classification tasks. It ranges from 0 to 1, with values under 0.5 requiring further evaluation due to random performance. A value closer to 1, ideally above 0.7, indicates a higher level of reliability and more resilient performance. Based on the evaluation results, as shown in Figure 1, the Logistic Regression model is represented by the red line, Random Forest by the orange line, C4.5 by the green line, and SVM by the purple line. Each model's corresponding AUC values are 0.91, 0.89, 0.43, and 0.66, respectively.

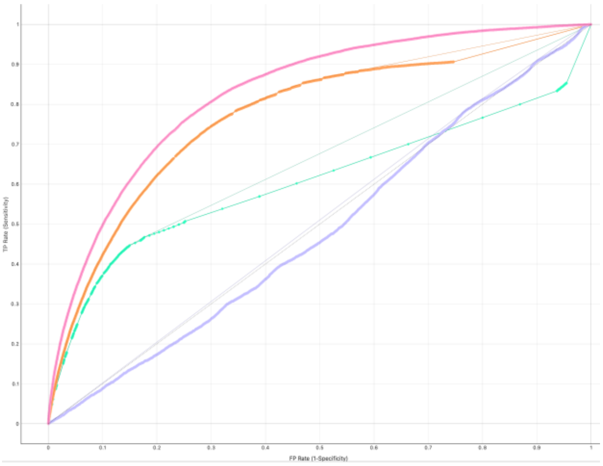


Figure 1 Evaluation Result

4. Conclusion

The findings demonstrate that logistic regression applied to categorical data with character types outperforms other algorithms. Nevertheless, it is crucial to also consider the C4.5 and SVM algorithms. Despite C4.5's superior accuracy and precision, performance assessment through ROC analysis reveals that the SVM algorithm is more effective.

5. Future Work

In future research, an in-depth evaluation needs to be carried out regarding unbalanced data, normalization and data transformation also need to be analyzed. The performance of the algorithm is truly reliable when all analysis schemes are carried out, but the evaluation results and accuracy only experience insignificant changes.

References

Ath, S., Al, T., Darmawan, D., Fahmi, N., Hakim, A., Qibtiya, M. Al, & Syafei, N. S. (2022). *Jurnal Teknologi Terpadu Hybrid Machine Learning Model untuk Memprediksi Penyakit Jantung dengan Metode Logistic Regression dan Random*. *8*(1), 40–46.

Balakrishnan, M., Arockia Christopher, A. B., Ramprakash, P., & Logeswari, A. (2021). Prediction of Cardiovascular Disease using Machine Learning. *Journal of Physics: Conference Series*, *1767*(1), 1–7. https://doi.org/10.1088/1742-6596/1767/1/012013

Braunwald, E. (2019). Braunwald’s Heart Disease: A Textbook of Cardiovascular Medicine. In *Elsivier* (Vol. 7, Issue 2).

Liu, H., Cocea, M., & Ding, W. (2017). Decision Tree Learning Based Feature Evaluation and Selection for Image Classification. *Proceedings of 2017 International Conference on Machine Learning and Cybernetics, ICMLC 2017*.

Çakmak, H. A., & Demir, M. (2020). Microrna and cardiovascular diseases. *Balkan Medical Journal*, *37*(2). https://doi.org/10.4274/balkanmedj.galenos.2020.2020.1.94

Cherfi, A., Nouira, K., & Ferchichi, A. (2018). Very Fast C4.5 Decision Tree Algorithm. *Applied Artificial Intelligence*, *32*(2), 119–137. https://doi.org/10.1080/08839514.2018.1447479

Ciumărnean, L., Milaciu, M. V., Negrean, V., Orășan, O. H., Vesa, S. C., Sălăgean, O., Iluţ, S., & Vlaicu, S. I. (2022). Cardiovascular risk factors and physical activity for the prevention of cardiovascular diseases in the elderly. In *International Journal of Environmental Research and Public Health* (Vol. 19, Issue 1). https://doi.org/10.3390/ijerph19010207

Djatna, T., Hardhienata, M. K. D., & Masruriyah, A. F. N. (2018). An intuitionistic fuzzy diagnosis analytics for stroke disease. *Journal of Big Data*, *5*(1). https://doi.org/10.1186/s40537-018-0142-7

Handayani, F. (2021). Komparasi Support Vector Machine, Logistic Regression Dan Artificial Neural Network Dalam Prediksi Penyakit Jantung. *Jurnal Edukasi Dan Penelitian Informatika*

(JEPIN), *7*(3). https://doi.org/10.26418/jp.v7i3.48053

Hartshorn, S. (2020). *Machine Learning with Random Forest and Decision Tree*.

Ho, C. C., Su, E., Li, P. C., Bolger, M. J., & Pan, H. N. (2020). Machine vision and deep learning based rubber gasket defect detection. *Advances in Technology Innovation*, *5*(2). https://doi.org/10.46604/aiti.2020.4278

Jijo, B. T., & Abdulazeez, A. M. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

Koda, S., Zeggada, A., Melgani, F., & Nishii, R. (2018). Spatial and Structured SVM for Multilabel. *Ieee Transactions on Geoscience and Remote Sensing*, 1–13.

Komalasari, R., Nurjanah, & Yoche, M. M. (2019). Quality of life of people with cardiovascular disease: A descriptive study. *Asian Pacific Island Nursing Journal*, *4*(2). https://doi.org/10.31372/20190402.1045

Krittanawong, C., Virk, H. U. H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., Zhang, H. J., Kaplin, S., Narasimhan, B., Kitai, T., Baber, U., Halperin, J. L., & Tang, W. H. W. (2020). Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-72685-1

Maharani, A., Sujarwoto, Praveen, D., Oceandy, D., Tampubolon, G., & Patel, A. (2019). Cardiovascular disease risk factor prevalence and estimated 10-year cardiovascular risk scores in Indonesia: The SMARThealth Extend study. *PLoS ONE*, *14*(4). https://doi.org/10.1371/journal.pone.0215219

Mahmoud, M. A. I., & Ren, H. (2019). Forest fire detection and identification using image processing and SVM. *Journal of Information Processing Systems*, *15*(1), 159–168. https://doi.org/10.3745/JIPS.01.0038

Maiga, J., Hungilo, G. G., & Pranowo. (2019). Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data. *Proceedings - 1st International Conference on Informatics, Multimedia, Cyber and Information System, ICIMCIS 2019*, 45–48. https://doi.org/10.1109/ICIMCIS48181.2019.8985205

Masruriyah, A. F. N., Djatna, T., Dewi Hardhienata, M. K., Handayani, H. H., & Wahiddin, D. (2019). Predictive Analytics For Stroke Disease. *Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019*. https://doi.org/10.1109/ICIC47613.2019.8985716

Mathur, P., Srivastava, S., Xu, X., & Mehta, J. L. (2020). Artificial Intelligence, Machine Learning, and Cardiovascular Disease. In *Clinical Medicine Insights: Cardiology* (Vol. 14). https://doi.org/10.1177/1179546820927404

Mezzatesta, S., Torino, C., De Meo, P., Fiumara, G., & Vilasi, A. (2019). A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. *Computer Methods and Programs in Biomedicine*, *177*, 9–15. https://doi.org/10.1016/j.cmpb.2019.05.005

Mia, M., Masruriyah, A. F. N., & Pratama, A. R. (2022). The Utilization of Decision Tree Algorithm In Order to Predict Heart Disease. *JURNAL SISFOTEK GLOBAL*, *12*(2), 138. https://doi.org/10.38101/sisfotek.v12i2.551

Ministry of Health Republic of Indonesia. (2014). *Penyakit Tidak Menular*. http://www.depkes.go.id/folder/view/01/structure-publikasi-pusdatin-buletin.html

National Center for Chronic Disease Prevention and Health Promotion Division for Heart Disease and Stroke Prevention. (2022, June). *Heart Disease Maps and Data Sources*. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/statistical_reports.htm

Noor, N. Bin, Anwar, M. S., & Dey, M. (2019). Comparative Study between Decision Tree, SVM and KNN to Predict Anaemic Condition. *BECITHCON 2019 - 2019 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health, December*, 24–28. https://doi.org/10.1109/BECITHCON48839.2019.9063188

Padmanabhan, M., Yuan, P., Chada, G., & Van Nguyen, H. (2019). Physician-friendly machine learning: A case study with cardiovascular disease risk prediction. *Journal of Clinical Medicine*, *8*(7). https://doi.org/10.3390/jcm8071050

Primajaya, A., & Sari, B. N. (2018). Random Forest Algorithm for Prediction of Precipitation. *Indonesian Journal of Artificial Intelligence and Data Mining*, *1*(1), 27. https://doi.org/10.24014/ijaidm.v1i1.4903

Setyaji, D. Y., Prabandari, Y. S., & Gunawan, I. M. A. (2018). Aktivitas fisik dengan penyakit jantung koroner di Indonesia The relationships of physical activity with coronary heart disease in Indonesia. *Jurnal Gizi Klinik Indonesia*, *14*(3), 115–121. https://jurnal.ugm.ac.id/jgki

Sofiana, L., Rokhmayanti, R., Sulistyawati, Nurfita, D., Astuti, F. D., & Sholekhati, P. A. (2021).

Evaluation of cardiovascular disease program in Sleman district, Indonesia. *International Journal of Public Health Science, 10*(2). https://doi.org/10.11591/ijphs.v10i2.20492

Utami, R. Y. (2017). *Pengembangan Media Ajar Anatomi Jantung dengan Low Cost Material Development of Anatomy of Heart Learning Resource with Low Cost Material*. *2*(1).

World Health Organization. (2019). *Cardiovascular diseases (CVDs)*. who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)