Journal of Informatics and Data Science (J-IDS)

Vol. 3, No. 2, November 2024
e-ISSN: 2964-0415

Application of Random Forest for Heart Disease Classification with SMOTE Approach to Balance Data

Fachriz Effendy[1]*, GiaColin Alfaro Samuel Sianturi[2], Dewi Fortuna Silaban[3] , M Fachri Aqil[4],Enjelita Simangunsong[5], Yolanda Angel Lina Sitorus[6], Arnita[7]
[1,2,3,4,5,6,7]Statistics Study Program, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Medan, Indonesia

In order to increase the accuracy and efficiency in heart disease detection, this work intends to develop a Random Forest algorithm based on machine learning into a heart disease prediction model. There are 255 samples in the dataset including 17 independent variables covering lifestyle and health elements. This work uses the SMote (Synthetic Minority Over-sampling Technique) technique to balance the class distribution by including synthetic data to the minority class given the data imbalance between the "Yes" (heart disease) and "No" (no heart disease) classes. With an accuracy of 94.7% and an AUC of 0.983, the Random Forest model built showed quite good results indicating that this model can effectively separate persons with and without heart disease. This work shows that the application of SMOTE considerably enhances model performance in handling data imbalance issues and helps to build machine learning-based predictive systems for heart disease classification. This work is novel in the use of the SMOTE technique to overcome data imbalance in heart disease prediction, so providing an efficient solution for data-driven medical decision making.

Keywords: Random Forest Algorithm, SMOTE, Imbalance Data, Heart Disease, Science Data
Received Month 20xx / **Revised** Month 20xx / **Accepted** Month 20xx

INTRODUCTION

Heart disease is one of the major health problems that causes the most deaths in the world. Risk factors such as high blood pressure, diabetes, obesity, an unhealthy lifestyle, smoking, and genetics can trigger heart disease. [1]. There are several ways that can be done to help medical personnel in finding out whether someone is indicated as having heart disease or not so that when a patient experiences this disease, they can find out quickly, one of which is by using machine learning. With this use, it has been proven to be able to solve classification topics and optimization in creating a health service provider system [2] [3].

Machine learning has three types, namely supervised learning, unsupervised learning, and reinforcement learning. Each type has a different type of data set. In the supervised learning algorithm, the system is given a training data set in the form of desired input and output information so that the system will learn based on existing data. The system will look for patterns from the data set, and then the pattern will be used as a reference for the next data set. [4]. According to [5], one part of supervised learning is random forest. So based on this information, the data used for this study is in accordance with the criteria for supervised learning and is suitable for processing using the Random Forest algorithm.

Numerous previous studies have explored the prediction of heart disease using various methods. In a study conducted by [6], they compared the decision tree, Naïve Bayes, and random forest models for heart disease classification prediction and found the accuracy value of the decision tree method was 0.71%, Naïve Bayes 0.72%, and random forest 0.75%. This implies that Random Forest outperforms the other two tested algorithms in terms of accuracy.
In addition, a study conducted by [7] tried to predict heart failure disease using Random Forest and found an accuracy value of 82.6087%. Additionally, a study from [8] concluded that the prediction results from Random Forest demonstrated better performance compared to other algorithms. Therefore, this study will

[1]*Fachriz Effendy.
Email addresses: fachrizeffedyk@gmail.com (Effendy)
DOI: 10.24114/j-ids.xxxxx

compare the performance of the Random Forest model against previous studies on heart disease prediction. In this study, SMOTE (Synthetic Minority Over-Sampling Technique) will also be used, which is tasked with adding synthetic data to minor data so that there will be a balance in the data to be tested. According to [9], the application of the SMOTEENN technique to the Random Forest algorithm has succeeded in significantly increasing the accuracy of heart disease prediction.

In addition to quantitative results showing the high performance of the Random Forest model in detecting heart disease, there are several important aspects that can be discussed further. Random Forest is a good algorithm for the medical field because it can handle complex data with many variables. Breiman (2001) stated that Random Forest can provide high accuracy and is resistant to overfitting because it combines the results of many decision trees through a majority voting mechanism [10]. This conclusion supports the results of this study, which show very high accuracy and AUC values.

One common challenge in medical classification is the imbalance of data between positive and negative classes. In this study, the model may be biased because the number of people with heart disease is very small compared to those without. Therefore, the application of the SMOTE technique is crucial because it is able to add synthetic data to the minority class and make the distribution balanced. Chawla et al. (2002) pointed out that SMOTE can greatly improve how well the model detects heart disease and performs overall when the data is unbalanced [11]. This benefit can be seen in the increase in sensitivity (recall) and balanced accuracy values after SMOTE was applied.

Furthermore, the results of the variable importance in the model show that features such as DiffWalking, BMI, GenHealth, and PhysicalHealth have a significant contribution to the classification. This finding is consistent with a study by Weng et al. (2017), which indicated that lifestyle factors and general health conditions such as physical activity and body mass index greatly affect the risk of cardiovascular disease [12]. This evidence shows that the results obtained are not only statistically significant but also clinically relevant.

Model evaluation using AUC also provides a strong picture of the model's performance. With an AUC value of 0.983, this model can be categorized as excellent, considering that an AUC above 0.9 is usually considered almost perfect in the context of medical classification. According to Mandrekar (2010), an AUC value approaching 1 indicates the model's ability to effectively distinguish positive and negative classes [13].

Finally, in terms of methodology, the selection of Random Forest and SMOTE is very appropriate for the dataset used. Han, Kamber, and Pei (2012) explained that Random Forest is effective for data with mixed variables between numeric and categorical, such as this heart disease dataset, and is very commonly used in binary classification cases [14]. Therefore, the application of the SMOTE technique is crucial because it is able to add synthetic data to the minority class and make the distribution balanced [15].

METHODS

With secondary data from the Heart Disease Dataset, which is openly accessible on Kaggle, this study employs a quantitative methodology. In addition to the variables associated with heart disease risk, such as age, gender, blood pressure, blood cholesterol, blood sugar, and other factors like smoking, alcohol use, stroke history, physical activity, sleep quality, physical and mental health, difficulty walking, race, diabetes history, kidney disease, and skin cancer, the dataset includes patient medical records.

Data was chosen based on a few important factors: it needed to have all the necessary information about heart disease risk, no repeated or incorrect entries, and very few missing values. The data used also underwent preprocessing to prepare it for analysis using the Random Forest algorithm. We used additional credible data sources like the UCI Machine Learning Repository as supplementary data when necessary. We collected the data by directly downloading the dataset file in CSV format from the Kaggle platform. We then performed feature selection to identify the variables most influential on heart disease classification, drawing from medical literature and initial analysis. We conducted data exploration using descriptive statistical analysis to understand the distribution patterns and correlations between variables.

For data processing and analysis, this study uses RStudio software due to its ability to manage large-scale data and support machine learning algorithms. The analysis process includes modeling with the Random Forest algorithm to classify the risk of heart disease based on the variables that were selected earlier. From an ethical standpoint, the data used is anonymous and does not contain personal identifying information, making it safe for research purposes without violating individual privacy. The references explicitly cite the dataset sources as a form of academic responsibility.

Overall, this study uses the SMOTE technique to balance classes in an imbalanced dataset and evaluates the effectiveness of the Random Forest model in predicting heart disease. This procedure provides a clear picture of how the method used can improve the performance of classification models in a medical context.

RESULT AND DISCUSSION

We obtained the primary indicator data for heart disease in 2022 through Kaggle. The data consists of 17 independent variables about various lifestyle and health factors that affect 1 dependent variable, namely heart disease itself. Before entering the random forest model, we need to convert most of the categorical data into numeric form.

This dataset is very suitable for binary classification using Random Forest because the target is categorical and the variables cover various lifestyles that can help with predictions. In addition, Random Forest is chosen here because its nature can handle mixed variables, such as numeric (BMI, physical health, etc.) and categorical (smoking, sex, race, etc.). We carry out several stages of data analysis using the random forest method:

1. Exploratory Data Analysis (EDA)

This EDA aims to analyze the data's structure and distribution and find the initial pattern in the data to be tested. So the first thing to do is to input data into R and make sure it's clean and readable. Next, we conduct checks to identify missing values, test the distribution of the target variables, and ensure class balance. Once all the results are secure, you can proceed to the next stage of data testing.

2. Data Pre-Processing

Based on the commonly used guidelines in data science, there are seven main stages in data pre-processing, namely:

a) Collecting Data

The data has been obtained previously and has been checked in the previous stage, and it has been ensured that all data can be read properly in R with a total of 255 data and 18 variables.

b) Data integration.

Data integration has been carried out by ensuring that the data has been transformed into a single, consistent data unit that is ready for analysis.

c) Data Transformation

We transform all categorical data into numeric data to facilitate analysis and obtain valid decisions later.

d) Data Cleaning

The data has been checked for missing values, outliers, data that is too similar (duplicate), and data with a high level of noise, but R detects that there is no data that meets these criteria and shows that the data is clean and ready to use.

e) Data Reduction

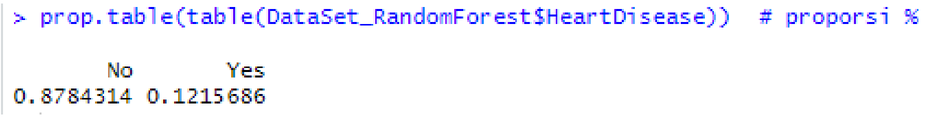
Data reduction is usually used in the pre-processing stage of data when the data is too large and there is a lot of data or features that are not relevant to the data. However, this data does not require reduction, as all the data holds importance and none is irrelevant.

f) Data Discretization & Binning

Discretization is done by changing categorical data into numeric, and binning is also done, for example, for the variable "Sex," where Female is replaced with 0 and Male is replaced with 1. Additionally, the HeartDisease variable assigns a label of 1 to Yes and 0 to No.

g) Handling Imbalanced Data

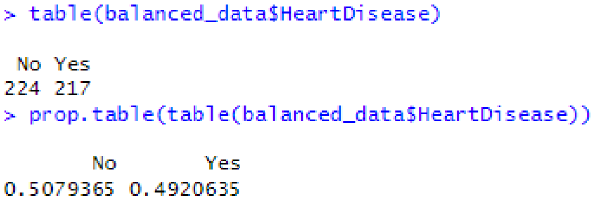
The data analysis revealed unbalanced data proportions because R produced the following output:



```
> prop.table(table(DataSet_RandomForest$HeartDisease))  # proporsi %

       No       Yes 
0.8784314 0.1215686 
```

Figure 1. Variable proportion of initial data

The results above show imbalanced data. The percentage of Yes is relatively low at 12.16%, whereas the percentage of No is significantly higher at 87.84%. This will later be able to create confusion in predictions because the model that is formed will continue to choose NO and cause high accuracy in choosing NO without really recognizing YES because of the unbalanced proportion. Therefore, the SMOTE method handles this unbalanced data effectively. The SMOTE method balances the data by assigning 50% of the proportion to YES and 50% to NO.



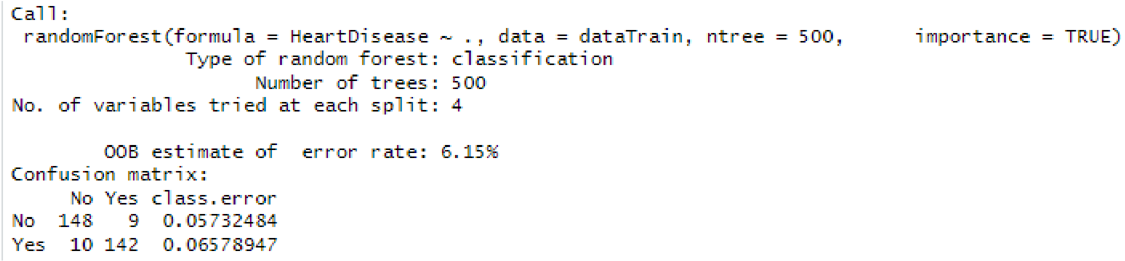
```
> table(balanced_data$HeartDisease)

 No Yes 
224 217 
> prop.table(table(balanced_data$HeartDisease))

       No       Yes 
0.5079365 0.4920635 
```

Figure 2. Proportion of data after using the SMOTE method

After using the SMOTE method, the data proportion has been changed to almost 50/50. The model's predictions for heart disease are unbiased when the data proportion is balanced, allowing it to classify "YES" or "NO" predictions.

3. Building a Random Forest Model

Based on the training results of the random forest model, the following model results are obtained:



```
Call:
 randomForest(formula = HeartDisease ~ ., data = dataTrain, ntree = 500,      importance = TRUE) 
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 6.15%
Confusion matrix:
     No Yes class.error
No  148   9  0.05732484
Yes  10 142  0.06578947
```

Figure 3. Creating a random forest model

We divide this model into 70% training data and 30% test data (commonly used comparison values). Training data is used to study patterns and relationships between variables, while test data is later used to measure the model's performance on the data. In addition to training data, 500 random forests (commonly used values) are also used to form this random forest model. More trees yield more consistent results, as they are more accurate, stable, and resistant to overfitting.

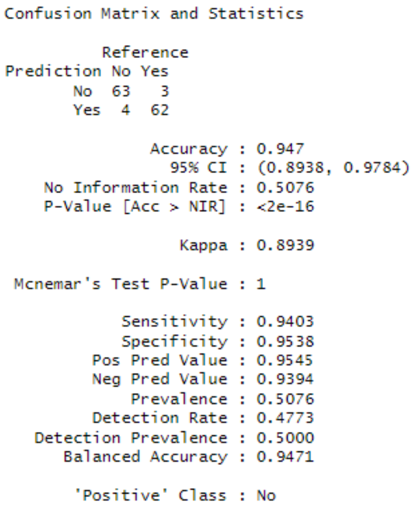
In addition to the model, checks were also carried out on the most influential variables, and the results are as follows:

Table 1. Variable importance scores from the random forest model

Variable	No	Yes	MeanDecreaseAccuracy	MeanDecreaseGini
BMI	11.875293	28.444121	28.856774	19.1682636
Smoking	4.5151722	13.547082	11.885057	6.4465038
AlcoholDrinking	0.2737687	2.937867	2.436109	0.4705426
Stroke	2.8188356	5.856832	5.973015	1.6093761
PhysicalHealth	1.7165086	18.812524	17.468756	10.2116378
MentalHealth	3.9791819	14.493976	13.664607	5.138081
DiffWalking	21.4109942	29.541796	31.698759	26.7821629
Sex	6.0813502	11.935872	11.779483	4.769359
AgeCategory	7.7236491	25.701336	24.879226	16.1089138
Race	2.4369315	12.479701	11.021865	5.1077099
Diabetic	11.1290025	14.443298	16.041891	9.6445823
PhysicalActivity	11.4181505	14.735706	15.417552	9.9932383
GenHealth	13.840704	21.566876	22.145223	15.7894234
SleepTime	3.2486138	17.901869	15.368013	9.1454792
Asthma	0.8226926	10.165364	7.99271	3.2971345
KidneyDisease	6.4618016	8.554629	9.554196	3.6919403
SkinCancer	3.7207522	12.861403	11.606701	5.0024551

Figure 4 above shows the important values of the variables that have the most influence on a person's decision whether they have heart disease or not. We can identify the most significant variables in the data by examining the MeanDecreaseAccuracy and MeanDecreaseGini values. So MeanDecreaseGini measures how well a feature separates the data in a decision tree, while MeanDecreaseAccuracy measures the feature's contribution to the overall model performance in terms of accuracy.

Thus, the created model identifies several variables that significantly influence heart disease conditions. Some of the most influential variables include DiffWalking, BMI, AgeCategory, GenHealth, PhysicalHealth, Diabetic, PhysicalActivity, and SleepTime. We also checked the accuracy of the most influential variables, yielding the following results:



```
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  63   3
       Yes  4  62
                                          
               Accuracy : 0.947           
                 95% CI : (0.8938, 0.9784)
    No Information Rate : 0.5076          
    P-Value [Acc > NIR] : <2e-16          
                                          
                  Kappa : 0.8939          
                                          
 Mcnemar's Test P-Value : 1               
                                          
            Sensitivity : 0.9403          
            Specificity : 0.9538          
         Pos Pred Value : 0.9545          
         Neg Pred Value : 0.9394          
             Prevalence : 0.5076          
         Detection Rate : 0.4773          
   Detection Prevalence : 0.5000          
      Balanced Accuracy : 0.9471          
                                          
       'Positive' Class : No              
```

Figure 5. Accuracy model

Based on the output above, some information was obtained:

a) Accuracy: 0.947

The model managed to correctly predict about 91% of the total test data. This is excellent accuracy.

b) Kappa: 0.8939

Kappa value measures the agreement between predictions and actual labels, corrected for chance. Values between 0.8 and 0.8–1.0 fall into the category of "almost perfect agreement." Thus, the model predictions are highly reliable.

c) Sensitivity (Recall for class "NO"): 0.9403

Of all the data that were truly "no," about 94.03% were successfully recognized. The model is excellent at recognizing patients who do not have heart disease.

d) Specifity (Recall for class "YES"): 0.9538

The model successfully identified 95.38% of the truly "yes" data (i.e., diseases present). This is important because heart disease detection is the main goal.

e) Positive Predictive Value (PPV) for "No": 0.9545

Of all the "No" predictions, about 95.45% actually did not hurt.

f) Negative Predictive Value (NPV) for "Yes": 0.9394

Of all the "Yes" predictions, about 93.94% actually got sick.

g) Balanced Accuracy: 0.9471

This value represents the average of both sensitivity and specificity. The high value of 94.71%, suitable for balanced data, signifies consistent performance in both classes.

4. Random Forest Model Evaluation

Before the main conclusion is drawn from the prior random forest model, it is crucial to evaluate the model. The main goal of this review is to see how well the model that was already made works as a whole. not just its accuracy but also its F1-Score, ROC-AUC, recall (sensitivity), and precision. Particularly because it is responsible for binary classification (HeartDisease: Yes/No). Accuracy, precision, and sensitivity were also addressed in stage 3, and the ROC curve and AUC (Area Under Curve) can be verified using R.

ROC Curve (Receiver Operating Characteristic Curve) is a graph that illustrates the correlation between Sensitivity (True Positive Rate) and 1 - Specificity (False Positive Rate). While AUC (Area Under the Curve) is a comprehensive measure of the model's capacity to differentiate between the "Yes" (positive) and "No" (negative) classes. In order to get the AUC and ROC output:

With an AUC of 0.9831228, it gauges the capacity of the model to differentiate between positive and negative classes at the output. AUC spans 0 to 1:

- ≤ 0.5 → no classification ability (just like coin toss),
- > 0.7 → good
- > 0.9 → very good,
- = 1 → perfect.

AUC = 0.9831228 means that the model formed has a very high accuracy of separation between classes, almost perfect.

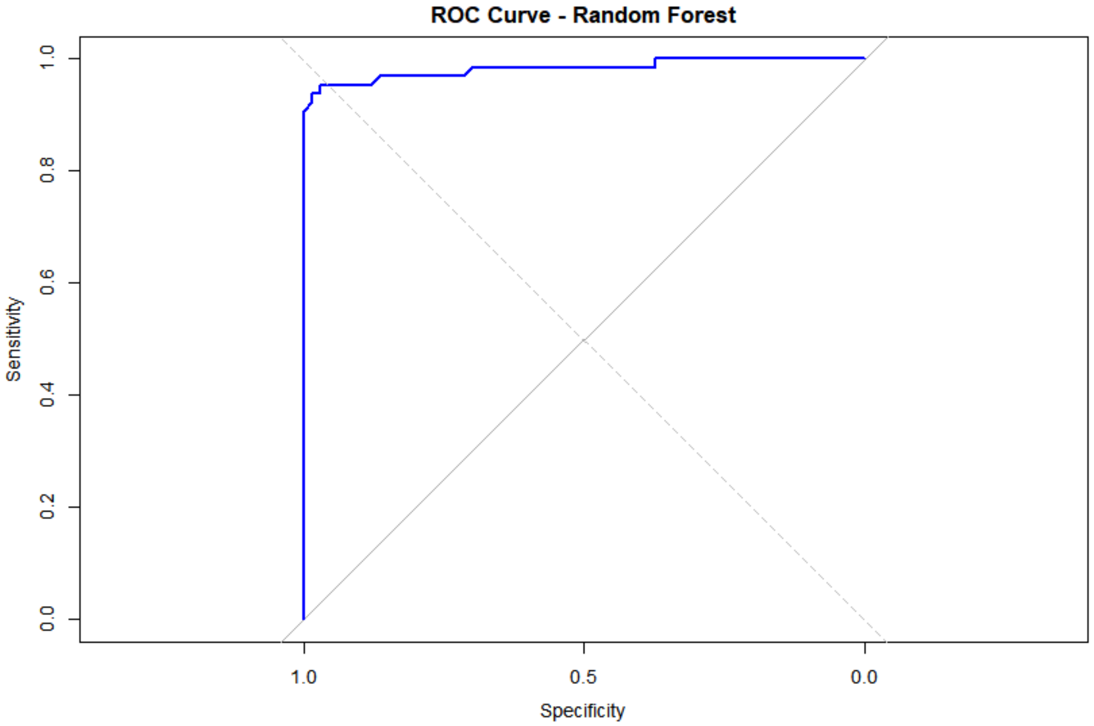


Figure 6. ROC curve

In the ROC (Receiver Operating Characteristic) output, the ROC curve illustrates the relationship between True Positive Rate (Sensitivity) and False Positive Rate (1 - Specificity). The blue line that sticks close to the upper left corner means the model has a very high classification ability, because that point means high sensitivity (few positive cases fail to be detected) and low false positives (rarely wrongly detecting negatives as positive).

CONCLUSION

SMOTE has been successfully used to stabilize unbalanced data, where SMOTE here is the core of the strategy to ensure the reliability of predictions. The random forest algorithm is also the best model for predicting various medical decisions with a high level of accuracy. The random forest model that has been created is able to detect heart disease (yes) and not heart disease (no) with high precision. It is feasible and valid to be used for medical decision-making or early screening for possible heart disease in the context of this dataset. This performance is also supported by high accuracy (94.7%) and balanced accuracy (94.7%) from the previous confusion matrix. Based on the model formed, it is known that the variables that have the most influence on patient heart disease include DiffWalking, BMI, AgeCategory, GenHealth, PhysicalHealth, Diabetic, PhysicalActivity, and SleepTime. So in the future, perhaps there can be more socialization of several main symptoms of heart disease sufferers so that it can become a concern for many people to maintain their health.

REFERENCES

[1] N. H. Alfajr and S. Defiyanti, “PREDIKSI PENYAKIT JANTUNG MENGGUNAKAN METODE RANDOM FOREST DAN PENERAPAN PRINCIPAL COMPONENT ANALYSIS (PCA),” *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3S1, Okt 2024, doi: 10.23960/jitet.v12i3S1.5055.

[2] Hidayat, A. Sunyoto, and H. Al-Fatta, “Klasifikasi Penyakit Jantung Menggunakan Random Forest Clasifier,” *Jurnal Sistem Komputer dan Kecerdasan Buatan*, vol. 7, no. 1, hlm. 31–40, Sep 2023.

[3] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4).

[4] P. Santoso, H. Abijono, and N. L. Anggreini, "ALGORITMA SUPERVISED LEARNING DAN UNSUPERVISED LEARNING DALAM PENGOLAHAN DATA," *Unira Malang* |, vol. 4, no. 2, hlm. 315–318, Apr 2021.

[5] R. Sharma, "Study of Supervised Learning and Unsupervised Learning," *Int J Res Appl Sci Eng Technol*, vol. 8, no. 6, hlm. 588–593, Jun 2020, doi: 10.22214/ijraset.2020.6095.

[6] D. Haganta Depari, Y. Widiastiwi, and M. Mega Santoni, "Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung," *JURNAL INFORMATIK Edisi ke*, vol. 18, no. 3, hlm. 239–248, Des 2022.

[7] Edric and Saut Parsaoran Tamba, "PREDIKSI PENYAKIT GAGAL JANTUNG DENGAN MENGGUNAKAN RANDOM FOREST," *Jurnal Sistem Informasi dan Ilmu Komputer Prima)*, vol. 5, no. 2, hlm. 176–181, Feb 2022.

[8] N. Utami, K. A. Baihaqi, E. E. Awal, and D. Waiddin, "Analisis Kinerja Algoritma Decision Tree Dan Random Forest Dalam Klasifikasi Penyakit Kardiovaskular," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 2, hlm. 970–981, Sep 2024, doi: 10.47065/bits.v6i2.5722.

[9] A. Rahmada and E. R. Susanto, "Peningkatan Akurasi Prediksi Penyakit Jantung dengan Teknik SMOTEENN pada Algoritma Random Forest," *Jurnal Pendidikan dan Teknologi Indonesia*, vol. 4, no. 12, hlm. 795–803, Jan 2025, doi: 10.52436/1.jpti.524.

[10] Breiman, L. (2001). *Random forests*. Machine learning, 45(1), 5-32.

[11] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research, 16, 321–357.

[12] Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). *Can machine-learning improve cardiovascular risk prediction using routine clinical data?*. PLoS one, 12(4), e0174944.

[13] Mandrekar, J. N. (2010). *Receiver operating characteristic curve in diagnostic test assessment*. Journal of Thoracic Oncology, 5(9), 1315-1316.

[14] Han, J., Kamber, M., & Pei, J. (2012). *Data mining: concepts and techniques* (3rd ed.). Morgan Kaufmann.

[15] Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences*, 465, 1-20.