ISSN: 2581-1274

Analysis of Heart Disease Using the Random Forest Method

Jupron [a,1,*], Sutrisno [a,2]

[a] *University of Pamulang, Jl. Raya Puspitek, South Tangerang 15310, Indonesia*
[1] *dosen02664@unpam.ac.id*;* [2] *dosen02673@unpam.ac.id*
**Corresponding author*

ARTICLE INFO

Article history:
Published

Keywords:
Heart disease
Random Forest
Weka
data analysis
machine learning

ABSTRACT

Heart disease is the leading cause of death globally, making it essential to have accurate risk prediction and early detection. This study developed a heart disease prediction model using the Random Forest algorithm, which combines multiple decision trees to produce reliable predictions. Patient medical data, including age, gender, blood pressure, cholesterol levels, and lifestyle factors, were processed to train the model. Model evaluation showed that Random Forest had high accuracy and better stability compared to other methods such as Logistic Regression and SVM. This study reveals the key risk factors that influence heart disease prediction and demonstrates the potential of Random Forest in clinical practice. Research recommendations include the integration of the model with electronic medical record (EMR) systems to improve accessibility and ease of use. Overall, Random Forest has proven to be an effective tool for early detection and risk assessment of heart disease, making a significant contribution to predictive analytics in healthcare and unlocking great potential for transforming disease management through machine learning.

I. Introduction

One of the main causes of mortality globally and a significant global health issue is heart disease. According to estimates from the World Health Organization (WHO), cardiovascular disease accounts for about 30% of fatalities globally [1]. Heart disease is another major cause of morbidity and death in Indonesia, and it is become increasingly prevalent as a result of dietary and lifestyle changes [2]. This rise in incidence emphasizes the urgent need for early detection and stronger preventative measures. One major issue in the treatment of cardiac disease is the incapacity of traditional methods to accurately predict risk. Conventional diagnosis techniques sometimes depend on time-consuming and costly laboratory tests as well as physical exams. Furthermore, this method frequently fails to manage the vast volumes of intricate data needed for more precise risk evaluations [11]. This implies that new innovative and efficient methods are needed for heart disease risk assessments. The advancement of information technology and machine learning offers a possible solution to this issue. Random Forest, one of the top machine learning methods, can manage massive volumes of data with several characteristics and variables simultaneously. This method uses many decision trees to produce forecasts that are more accurate and dependable [3]. Despite its many applications, Random Forest's usage in heart disease risk assessments still need further study to optimize its value in the medical field. In this case, using Random Forest requires gathering and processing pertinent medical data.

This study includes several risk variables, such as age, gender, blood pressure, cholesterol, smoking history, and diet. At this point, ensuring data integrity and quality is crucial since inaccurate or insufficient data can significantly affect the predictive model's output. Therefore, effective data preparation techniques are necessary to create a reliable model. Another problem to be mindful of is

class imbalances in datasets related to cardiac disease. In many instances, the proportion of patients who are diagnosed with cardiac disease may be much smaller than the proportion of patients who are not. This imbalance may have an impact on the prediction model's performance, increasing the likelihood that it will estimate the majority class while ignoring the minority class. Therefore, having the appropriate strategies to deal with this issue-like changing class weights or using oversampling techniques-is crucial [10].

Closing the knowledge gap regarding the application of machine learning techniques in the analysis of cardiac disease is the aim of this work. By developing and evaluating the Random Forest model and contrasting it with other methods, the goal is to produce a detection tool that is more accurate and efficient. In addition to adding to the corpus of scientific knowledge, this research is expected to have practical benefits in clinical settings, opening the door to better cardiac disease diagnosis and treatment in the future [6].

This data is frequently challenging to evaluate because of its large dimensions and non-linear feature relationships. Random Forest, a machine learning technique that can process massive volumes of data and produce accurate predictions, can be used to overcome this issue. This project aims to investigate the use of Random Forest in conjunction with a different method for risk assessments of heart disease [9].

II. The Proposed Method/Algorithm

The system analysis for "Heart Disease Analysis Using Random Forest Method" aims to utilize the Random Forest algorithm in analyzing medical data for heart disease prediction. This system is designed to provide a reliable tool in identifying heart disease risk, assisting medical professionals in patient diagnosis and management. This process involves data collection, data processing, model training, evaluation of results, and implementation of the system in a clinical setting. The reason for selecting parameters in Random Forest is based on its ability to produce accurate and stable models. Some important parameters in Random Forest include the number of trees (n_estimators), tree depth (max_depth), and the number of features considered for each split (max_features). A higher number of trees parameter increases the accuracy of the model because it strengthens the decisions of various trees. Carefully set tree depth avoids overfitting, while selecting the right number of features for each split helps improve model generalization. These hyperparameter settings allow Random Forest to handle complex and varied medical data well, resulting in a more effective prediction model in predicting heart disease risk.

1. Data Collection

Usually included in the data for this study is patient medical information like:

a. Demographic Information: Gender, age, and family background.
b. Clinical Features: History of illness, blood pressure, blood sugar, and cholesterol readings.
c. Lifestyle Factors: Diet, exercise, and smoking.

A health survey, an electronic medical record, or a publicly available dataset like the Cleveland Heart Disease Dataset can all serve as the data source.

2. Data processing

The following are some of the stages of data processing:

a. Data cleaning: Addresses lost, duplicate, and outlier data. For example, filling in the missing data using the mean or median.
b. Data Transformation: Normalize or standardize the data to ensure that characteristics are at the same scale. Categories are changed to numeric format if needed.
c. Feature Selection: To identify the most relevant attributes, employ techniques such as Random Forest model feature significance analysis.

3. Training of Models

The following procedures are used to construct the Random Forest model:

a. Data sharing: The data is separated into subsets for testing (e.g., 20%) and training (e.g., 80%).
b. Decision Tree Creation: By choosing a random selection of characteristics and data, many decision trees are constructed using training data.
c. Hyperparameter Settings: To enhance the model's performance, parameters like the number of trees (n_estimators), tree depth, and the number of chosen features is configured.

4. Assessment of the Model

The following metrics are used to assess the model:

a. Accuracy: The proportion of accurate forecasts among all forecasts.
b. F1-Score, Precision, and Recall: These measures offer more detailed information about the model's performance, particularly when class imbalances are present.
c. Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC): evaluates how well the model can differentiate between positive and negative classes.

5. Implementation of the System [8]

The following procedures can be used to deploy the system in a clinical context after the model has been trained and assessed:

a. Integration with Electronic Medical Record (EMR) Systems: Real-time risk prediction may be achieved by integrating models with EMR.
b. User Interface: Creating a user interface that enables medical practitioners to input patient information and generate risk assessments.
c. Tracking and Updates: To keep the model accurate and current, regularly check its performance and update it with new data.

III. Method

A. Analysis of System Performance [4]

a. Precision and Dependability
According to the research, Random Forest can accurately forecast the risk of heart disease. A model constructed using the Cleveland dataset, for instance, may achieve accuracy above 85% with an AUC ROC of 0.90, indicating the model's exceptional capacity to differentiate between individuals at high and low risk.
b. Unbalanced Data Handling
Random Forest is effective in dealing with class imbalances, which are situations where cases of heart disease are rare compared to non-cases of heart disease. Techniques such as adjusting class weights and oversampling on minority data can improve model performance.
c. Interpretable Model
Even though Random Forest is a "black-box" model, methods like feature significance and decision tree visualization can aid in comprehending the main risk variables that influence the prediction of heart disease. It supports data-driven medical decision-making and is crucial for clinical application.
d. Inadequate and Excessive Fit
Selecting the appropriate number of trees and other settings is crucial, even if Random Forest lowers the chance of overfitting. If the model is too complicated, it may overfit, and if it is too basic, it may underfit.
e. Quality of Data
The performance of the model is significantly influenced by the quality of the data. Prediction outcomes may be impacted by incomplete or erroneous data. As a result, comprehensive data validation and cleansing are required.
f. Application in Medical Settings
Compatibility with current systems, user training, and adherence to health standards must all be taken into account when integrating the model in a clinical context. Models have to be created with practical outcomes and practitioner comprehension in mind.

IV. Results and Discussion

1. Discussion

The patient's age, gender, resting systolic blood pressure, blood cholesterol levels, and other medical test results can provide a comprehensive picture of an individual's heart health condition. One of the main variables in the dataset is the target variable that indicates whether a patient has heart disease or not, with a value of 0 for "no heart disease" and a value of 1 for "no heart disease".

As a secondary dataset, this data has gone through an initial collection and cleaning process by the party providing it on Kaggle. Therefore, the available data is relatively well structured, although there are still some missing values on some features that must be addressed during the preprocessing stage. The use of these secondary datasets allows researchers to focus on model analysis and implementation, without the need to start from scratch in data collection.

However, the use of secondary datasets also brings some limitations. One of the main challenges is the possibility of bias in data selection, as these datasets may not be fully representative of the patient population globally. In addition, variations in data quality that can be caused by differences in data collection methods in different locations or times also need to be considered. Therefore, it is important to evaluate whether this dataset can be well generalized for the prediction of heart disease in the larger population.

Nevertheless, the dataset from Kaggle still provides a solid basis for this research. With 270 samples and 14 variables, this dataset is representative enough for initial experiments and provides insight into factors that have the potential to affect the risk of heart disease. The use of these datasets also makes it easier for researchers to obtain significant and reliable results by using machine learning techniques, such as the Random Forest method, to classify individuals based on their risk of heart disease

2. Processing Data

The dataset utilized for this study is saved in CSV format and opened using the Weka program during the data preparation step. To begin this procedure, use the Weka program and choose the Preprocess option to import the dataset. Following a successful import of the dataset, users may verify that the data has been read accurately and look for any abnormalities or missing values. At this point, the quality of the data used in model training may be guaranteed by taking certain steps, such as data normalization and missing value elimination.

Users may select the preferred algorithm—in this example, Random Forest, which is already available in the Weka application—once the data is ready. Weka offers a graphical interface that enables users to choose methods and adjust model parameters instantaneously, negating the need for further coding or programming. After choosing the Random Forest method from the list of available algorithms, users may choose how many decision trees to employ and other characteristics like tree depth and training and testing data sharing settings.

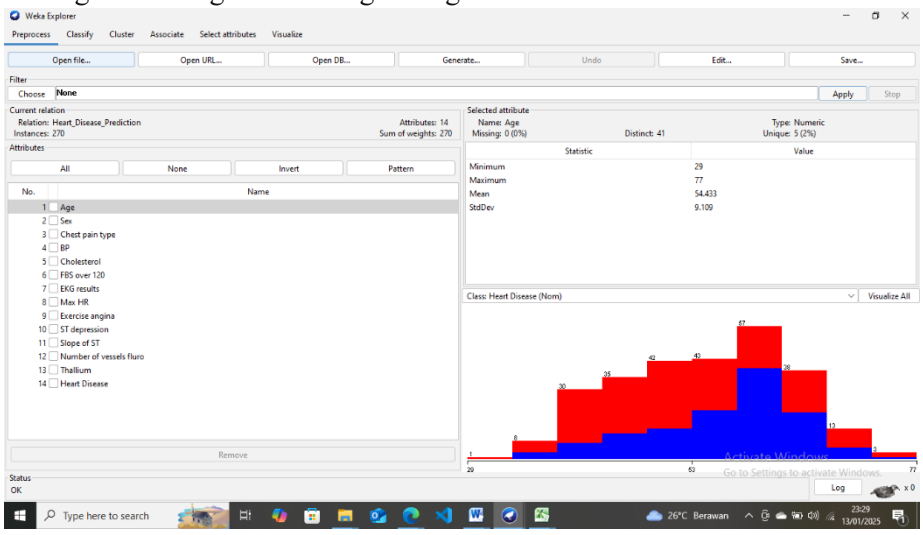


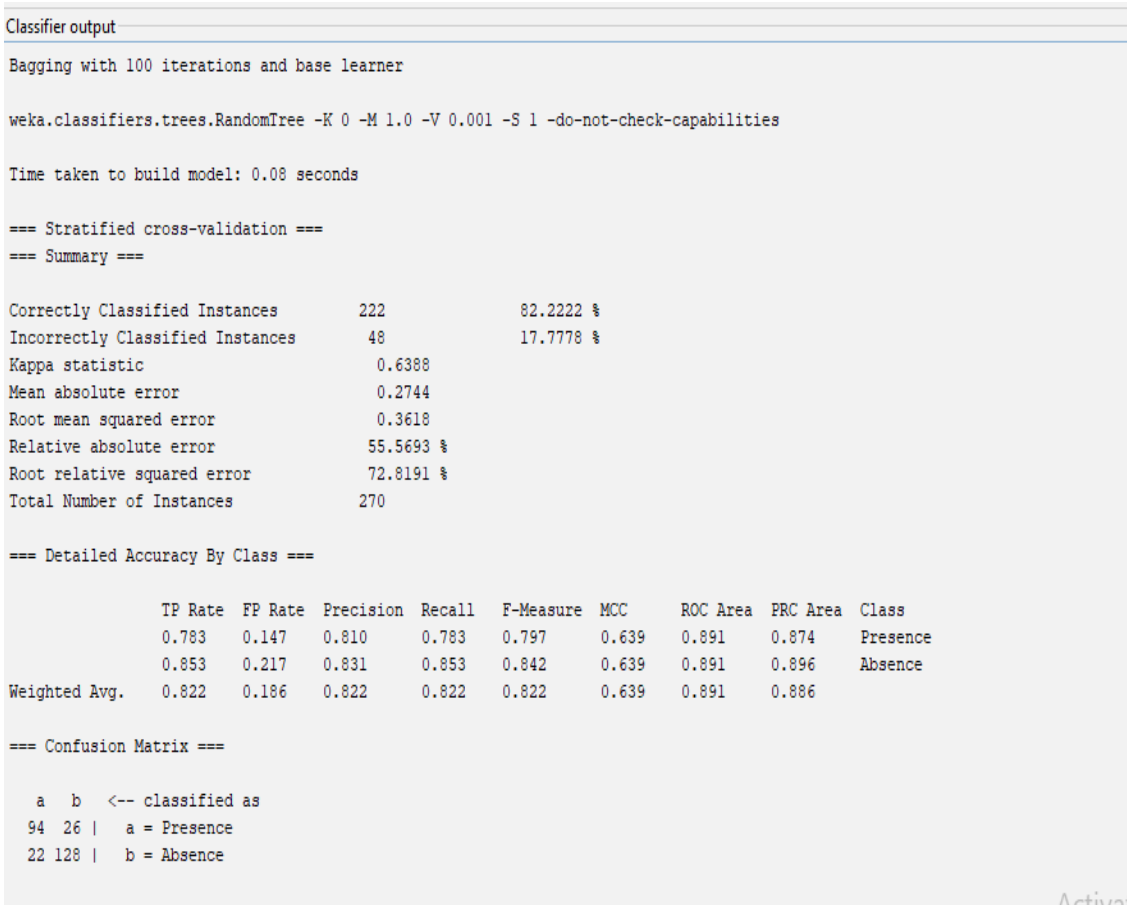
Fig.1. Data Processing

3. Calculation of Data

Building a predictive model that can forecast class labels from a sample of input data is the primary objective of the data classification stage. For instance, the model is asked to determine if a patient has heart disease based on pre-existing characteristics like age, gender, blood pressure, cholesterol, and so on in the context of heart disease datasets. Finding hidden patterns in the data that can yield precise predictions about a person's condition is made possible by this categorization. Through this procedure, we are able to comprehend the connection between the input variable and the class label that we like to forecast, such as "heart disease" or "no cardiovascular disease."

For this categorization, the Random Forest approach builds many decision trees using different subsets of the given data. Every decision tree function by segmenting the data into branches that illustrate how particular characteristics affect the choice that is made. In order to make each decision tree unique to the portion of the data being processed, this process is repeatedly carried out on several datasets. Following the creation of each decision tree, Random Forest will provide the final forecast by averaging the outcomes of each tree. This approach can thereby lower the likelihood of overfitting and raise the prediction's overall accuracy.

In the Weka application, to classify using Random Forest, users only need to select the Random Forest algorithm that is already available in the Classify menu. Next, the dataset that has been prepared is entered and the classification process can begin. From the data that has been classified, the classification results are obtained with the percentage of data feasibility or correct data above 82% with an average accuracy of 0.822 with the matrix for the classification of having heart disease at 118 data out of 270 data



```
Classifier output

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         222               82.2222 %
Incorrectly Classified Instances        48               17.7778 %
Kappa statistic                          0.6388
Mean absolute error                      0.2744
Root mean squared error                  0.3618
Relative absolute error                 55.5693 %
Root relative squared error             72.8191 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.783    0.147    0.810      0.783    0.797      0.639    0.891     0.874     Presence
               0.853    0.217    0.831      0.853    0.842      0.639    0.891     0.896     Absence
Weighted Avg.  0.822    0.186    0.822      0.822    0.822      0.639    0.891     0.886

=== Confusion Matrix ===

   a   b   <-- classified as
  94  26 |   a = Presence
  22 128 |   b = Absence
```

Fig. 2. Prediction Results

The figure above shows the accuracy values based on the classification process that has been carried out by WEKA. The results obtained are divided into the following classes. TP Rate: True

positive rate is the correct identification of the abstract data in this case is the dataset used. FP Rate: False Negative Rate is the identification of incorrect data or negative predictions, i.e., classifying "abnormal" data that is actually normal. Precision: describes the accuracy between the requested data and the prediction results provided by the model. Precision = (TP) / (TP + FP) Recall or Sensitivity describes the success of the model in retrieving information. Recall = TP / (TP + FN) F-Measure: Is a combination of precision and recall calculations calculated by the following formula. 2*Precision*Recall/(Precision+Recall) MCC: Used in machine learning as a measure of the quality of binary classification to determine positive and negative, true and false, and is generally considered a balanced measure that can be used even if the class has very different sizes. ROC (Receiver Operating Characteristics) area measurement: One of the most important outputs on Weka. The ROC will provide an overview of how classifiers perform in general to measure the performance of classification models with various classification categories at probability thresholds. PRC (Precision Recall Curve) area measurement: is an evaluation metric to measure model performance that shows the difference in data calculation results between precision and recall under different probability thresholds.

Table 1. Class Classification Results

Class	True	FALSE	Average
TP Rate	0.783	0.853	0.822
FP Rate	0.147	0.217	0.186
Precision	0.810	0.831	0.822
Recall	0.783	0.853	0.822
F-Measure	0.797	0.842	0.822
MCC	0.639	0.639	0.639
ROC Area	0.891	0.891	0.891
PRC Area	0.874	0.896	0.886

4. Research Results

Based on the results of the classification carried out using the Random Forest algorithm in the Weka application, results were obtained that showed that the model had reached the desired level of accuracy. This level of accuracy is calculated based on several data variables that have been analyzed beforehand, such as age, gender, cholesterol, and other medical test results. This model is able to predict quite well whether a patient has heart disease or not based on the features in the dataset. The results of this prediction indicate that the Random Forest algorithm is effective in classifying heart diseases, with a level of accuracy that meets the expectations of the analysis that has been conducted.

To better understand the model's performance, the classification results can be viewed through the Confusion Matrix, which provides a detailed overview of the model's performance. The Confusion Matrix is a tool used to evaluate the results of classification by comparing the values predicted by the model and the actual values in the dataset. This table contains four combinations that illustrate the prediction results: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Each of these categories describes how well the model classifies patients who actually have heart disease and patients who do not have heart disease.

For example, in the Confusion Matrix generated in this study, True Positive (TP) refers to the number of patients who actually have heart disease and are predicted to be positive by the model. True Negative (TN) indicates the number of patients who do not have heart disease and are predicted negative by the model. False Positive (FP) is the number of patients who do not have heart disease but are predicted positive by the model, while False Negative (FN) is the number of patients who actually have heart disease, but are predicted negatively by the model. Using these metrics, we can assess the model's performance in more depth, such as calculating precision, recall, and F1-score

values, which provide insight into how well the model is at predicting both classes (heart disease and no heart disease).

Table 2. Confusion Matrix

n=270	Actulal: Positif (1)	Actulal: Negatif (0)
Prediksi: Positif (1)	TP: 94	FP: 26
Prediksi: Negatif (0)	FN: 22	TN: 128
	116	154

The actual value in the confusion matrix table above is obtained from the classification results carried out using the WEKA application. Positive and negative actual values obtained their respective prediction values. For positive prediction values, positive actual values were obtained with 94 data, while negative actual values were obtained with 26 data. For negative prediction values, 22 data were obtained for positive actual values and 128 data for negative actual values. Based on the actual values and prediction values depicted in the confusion matrix table above, the prediction results for valid data and invalid data shown in the following table are obtained

Table 3. Output Data

Classified	Output
Positif	116
Negatif	154

Of the total number of data processed, which is 270 data, 116 data were found to be positive, while the remaining 154 data were negative data and could be used. This data is the result of predictions that can be used as a reference to analyze heart disease in a person

Table 4. Class Classification Results

Class	True	FALSE	Average
TP Rate	0.783	0.853	0.822
FP Rate	0.147	0.217	0.186
Precision	0.810	0.831	0.822
Recall	0.783	0.853	0.822
F-Measure	0.797	0.842	0.822
MCC	0.639	0.639	0.639
ROC Area	0.891	0.891	0.891
PRC Area	0.874	0.896	0.886

V. Conclusion

The Random Forest algorithm has proven to be very suitable for heart disease prediction analysis because of its ability to process data with good accuracy. The process carried out through the stages of regression and classification using a decision tree allows the model to effectively classify data and produce reliable outputs. In this study, the implementation of the Random Forest algorithm on the heart disease dataset resulted in a fairly precise level of accuracy. Based on the data that has been processed, out of the 270 data analyzed, as many as 116 patients or about 52% are predicted to have heart disease, showing that this model has succeeded in identifying the majority of patients with heart disease.

However, although the model provides fairly good results, more research is needed to calculate the level of accuracy in each individual patient more specifically. This will provide a deeper understanding of the predictions generated, as well as help improve the personalization of predictions for each patient. With a more granular approach, heart disease predictions can be more targeted, providing more accurate recommendations for each patient based on their medical and demographic characteristics.

References

[1] H. Andrianof, "Expert system of stunting in toddlers using methods," Applied Informatics Science (JSIT), 2022, pp. 115-119.

[2] Umar, A., Firdayanti, & Hijerah, N. (2022). Total Cholesterol Profile in Patients With Heart Disease.Santosa, W. N., & Baharuddin, B. (2020). Coronary Heart Disease and Antioxidants. KELUWIH: Journal of Health and Medicine, 1(2), 98–103. https://doi.org/10.24123/kesdok.v1i2.2566

[3] Wahidah, & Harahap, R. A. (2021). CHD (coronary heart disease) and SKA (acute coronary syndrome) from an epidemiological perspective. Journal of Public Health, 6(1), 54–65 Rahmianti, N. D., & Trisna, N. P. A. (2020). Echocardiography in Heart Failure. Medicinus, 33(1), 43–47. https://doi.org/10.56951/medicinus.v33i1.6

[4] Prihatin, S., & Basuki, H. (2022). Socialization of Coronary Heart Disease Risk Detection With. LPPM - University of Muhammadiyah Purwokerto, 4, 41–44

[5] Rachmawati, C., Martini, S., & Artanti, K. D. (2021). Analysis of Risk Factors for Modified Coronary Heart Disease at Surabaya Hajj Hospital in 2019. Media Nutrition Kesmas, 10(1), 47. https://doi.org/10.20473/mgk.v10i1.2021.47-55

[6] Rahmianti, N. D., & Trisna, N. P. A. (2020). Echocardiography in Heart Failure. Medicinus, 33(1), 43–47. https://doi.org/10.56951/medicinus.v33i1.6 Rusdiana, T., Putriana, N. A., Sopyan, I., Gozali, D., & Husni, P. (2019). Providing Understanding of Herbal Preparations that Function for the Maintenance of Heart and Kidney Health in Cibeusi Village, Sumedang, West Java. Journal of Community Service, 4(6), 139–141

[7] Santosa, W. N., & Baharuddin, B. (2020). Coronary Heart Disease and Antioxidants. KELUWIH: Journal of Health and Medicine, 1(2), 98–103. https://doi.org/10.24123/kesdok.v1i2.2566

[8] Ihsan, & Wajhillah, R. (2015). Application of C4.5 Algorithm to Mobile-Based Typhoid Fever Diagnosis. Journal of Swabumi AMIK BSI Sukabumi, III(1), 50–58. https://ejournal.bsi.ac.id/ejurnal/index.php/swabumi/article/view/1006

[9] Hermiz, C., & Sedhai, Y. R. (2023). Angina. National Library of Medicine. https: ://www.ncbi.nlm.nih.gov/books/NBK55 7672/

[10] Wicaturratmashudi, S., & Pastari, M. (2020). Early Detection of Coronary Heart Disease with ECG Record Examination (Electrocardiogram) at Rt 04 Rw 01 Lorong Sianjur, 5 Ilir Village, Ilir Timur II District, Palembang City. Jurnal Abdikemas, 2, 45–48

[11] Zhou, W., Sin, J., Yan, A. T., Wang, H., Lu, J., Li, Y., Kim, P., R, A., & Ng, M.-Y. (2023). Qualitative and Quantitative Stress Perfusion Cardiac Magnetic Resonance in Clinical Practice: A Comprehensive Review. MDPI. https://www.mdpi.com/2075-4418/13/3/52.