Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1015−1024
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10372

Improved Genetic Algorithm with Adaptive Operators and Elitism for Random Forest Feature Selection in Heart Disease Classification

Rahma Dhea Safitri[1,*], Solikhun[2], Timbo Faritcan P. Siallagan[3]

[1]Information System, STIKOM Tunas Bangsa, Pematangsiantar
Jl. Sudirman Blok No. 1, 2 dan 3, Banjar, Kec. Siantar Bar., Kota Pematang Siantar, Sumatera Utara, Indonesia
[2]Informatics Engineering, STIKOM Tunas Bangsa, Pematangsiantar
Jl. Sudirman Blok No. 1, 2 dan 3, Banjar, Kec. Siantar Bar., Kota Pematang Siantar, Sumatera Utara, Indonesia
[3] Faculty of Engineering, Computer and Network Engineering, Universitas Mandiri, Subang
Jl. Marsinu No.5, Dangdeur, Tegalkalapa, Kabupaten Subang, Jawa Barat, Indonesia
Email: [1,*]rahmasafitri164@gmail.com, [2]solikhun@amiktunasbangsa.ac.id, [3]timbofaritcansiallagan@gmail.com
Correspondence Author Email: rahmasafitri164@gmail.com
Submitted: **20/06/2026**; Accepted: **15/07/2026**; Published: **15/07/2026**

Abstract−Heart disease is one of the leading causes of mortality worldwide, and accurate prediction models are essential to support early diagnosis. However, conventional Random Forest classifiers generally utilize all available features, although not all features contribute equally to classification performance, resulting in unnecessary model complexity. This study proposes an Improved Genetic Algorithm (IGA) that extends the conventional Genetic Algorithm through elitism, adaptive crossover, and adaptive mutation operators to optimize feature selection for Random Forest-based heart disease classification. The proposed method was evaluated using the Cardiovascular Disease Dataset from Kaggle, which consisting of 1,000 records and 14 variables, where 12 predictor features were used for model development. The experimental procedure included data preprocessing, train-test splitting, class imbalance handling using SMOTE on the training set, feature normalization, Random Forest modeling, feature selection using the proposed IGA, and model evaluation. The proposed IGA selected six important features slope, chestpain, restingBP, restingelectro, oldpeak, and gender. The optimized Random Forest model achieved an accuracy of 99.50%, precision of 99.15%, recall of 100.00%, F1-score of 99.57%, and AUC-ROC of 99.90%. These findings indicate that feature selection can simplify the model without compromising classification performance, making the Random Forest + IGA approach a viable alternative for developing more efficient heart disease prediction models.

Keywords: Heart Disease; Feature Selection; Improved Genetic Algorithm; Random Forest; Classification

1. INTRODUCTION

Heart disease is a health disorder that has a high risk level because it is directly related to the function of the heart in circulating blood throughout the body[1]. Report World Health Organization The World Health Organization (WHO) in 2019 stated that cardiovascular disease, including heart disease, causes approximately 17.9 million deaths annually worldwide, equivalent to approximately 31% of all global deaths[2]. In Indonesia, this condition is not much different. Data from the 2018 National Riskesdas Report shows that the number of heart disease cases based on doctor diagnoses across all age groups in Indonesia is at 1.5%[3]. This high figure confirms that heart disease is not only an individual health issue, but has become a serious public health burden and requires more attention in prevention and early detection efforts.

Various studies have been conducted in an effort to build a machine learning-based heart disease classification model. Reference[4] tested three classification algorithms, namely Decision Tree, Naive Bayes, And Random Forest Classifier on the heart disease dataset. Test results and processing hyperparameter tuning, Random Forest Classifier became the best model, with an F1-score of 0.868 using grid search. Reference[5] also compared several models, namely Decision Tree, Naive Bayes, And Random Forest on the classification of heart disease, with the results showing that Random Forest obtained the highest accuracy compared to other models, which is 0.75. Reference[6] applies ensemble voting classifier algorithm based Naive Bayes and Random Forest with the SMOTE method to handle data imbalance in heart disease datasets. The results of this study indicate that ensemble voting classifier achieved the best performance with accuracy 98.28%. References[7] use various algorithms to choose the most optimal model in heart disease analysis, namely random forest, C45, logistic regression, and Support Vector Machine (SVM). Categorization performance compared to the adaptive synthetic (ADASYN) method and synthetic minority oversampling (SMOTE). Test result Random Forest became the best model, with an initial accuracy of 90.71% increasing to 94.54% with the use of the oversampling approach. References[8] using the Random Forest algorithm to classify heart disease by applying data normalization, handling class imbalance with SMOTE, and hyperparameter tunning. The results show that the Random Forest model can reach accuracy 92% with the optimal parameter configuration obtained through the tuning process. Overall, previous studies consistently demonstrate that Random Forest provides competitive performance for heart disease classification. Several studies improved predictive performance through hyperparameter tuning, oversampling techniques, or ensemble learning. Nevertheless, these approaches generally rely on the complete feature set and pay limited attention to feature selection optimization.

Based on previous research, the use of Random Forest has been shown to perform well in heart disease classification. However, most existing studies have focused on comparing classification algorithms, applying oversampling techniques, or optimizing hyperparameters to improve predictive performance[9]. These approaches

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1015−1024
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10372

generally utilize all available input features during model construction, although not every feature contributes equally to the classification results. The inclusion of redundant or less informative features may increase model complexity, reduce computational efficiency, and potentially affect model generalization. Therefore, an effective feature selection strategy is required to identify the most relevant feature subset while maintaining high predictive performance. To address this limitation, this study proposes an Improved Genetic Algorithm (IGA) that extends the conventional Genetic Algorithm by incorporating three optimization strategies, namely elitism, adaptive crossover, and adaptive mutation. These strategies are designed to preserve high-quality candidate solutions, balance exploration and exploitation during the search process, and maintain population diversity, thereby improving the effectiveness of feature selection before Random Forest classification.[10],[11],[12],[13].

This study aims to optimize feature selection for heart disease classification by proposing an Improved Genetic Algorithm (IGA) with adaptive operators and elitism for Random Forest. The data used comes from Cardiovascular Disease Dataset available on Kaggle, with 1,000 data points and 14 variables. The research framework includes data preprocessing, feature selection using the proposed IGA, Random Forest classification, and performance evaluation based on accuracy, precision, recall, F1-score, confusion matrix, and AUC-ROC [14]. By integrating adaptive crossover, adaptive mutation, and elitism into the Genetic Algorithm, this study aims to improve the effectiveness of feature selection, reduce redundant features, and maintain high classification performance while producing a more efficient prediction model. This contribution provides an alternative feature selection strategy for Random Forest that enhances model efficiency without compromising predictive performance.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This research was conducted through several systematic stages to produce an optimal heart disease classification model. The following is a diagram of the research stages, as seen in Figure 1.

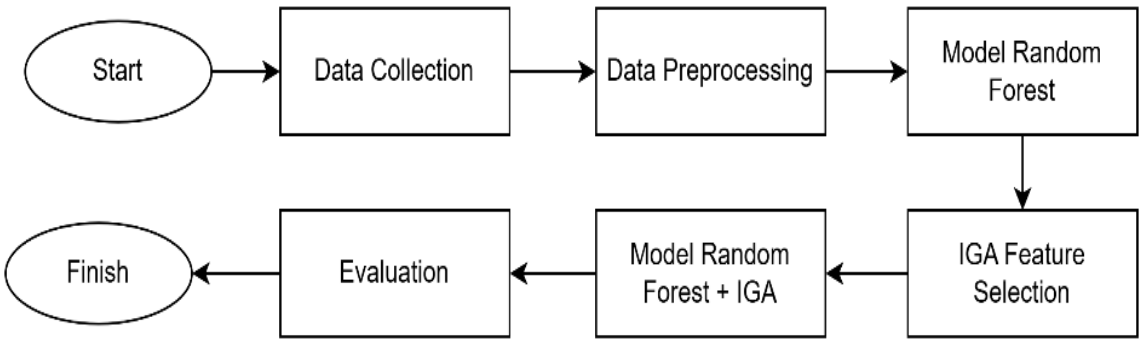

Figure 1. Research Stages

The research process began with the collection of heart disease datasets obtained from Kaggle. Data preprocessing was then performed by removing the patientid variable, as it only serves as patient identification and is not clinically relevant. While the target variable is used as a class label and is therefore not counted as a feature. Thus, the model Random Forest was built using 12 features. The dataset is then divided into training and testing sets using an 80:20 ratio before applying the (SMOTE), which is performed only on the training data to address class imbalance and prevent data leakage[15]. Next, numerical features are normalized using Min-Max Scaling before constructing the baseline Random Forest model. Subsequently, the proposed Improved Genetic Algorithm (IGA), which incorporates elitism, adaptive crossover, and adaptive mutation, is employed to select the most relevant feature subset for Random Forest classification[16],[17]. Finally, the performance of the baseline Random Forest model and the Random Forest model with IGA-based feature selection is evaluated and compared using accuracy, precision, recall, F1-score, confusion matrix, and AUC-ROC to determine the effectiveness of the proposed feature selection approach[18].

2.2 Data Collection

The Cardiovascular Disease Dataset, which can be accessed publically on the Kaggle platform at kaggle.com/datasets/jocelyndumlao/cardiovascular-disease-dataset, was utilized in this study. Twelve predictor factors, one patient ID variable, and one classification target variable make up the 14 variables in the dataset, which comprises 1,000 patient medical records. The patient ID variable is not utilized in the modeling process because it just functions as patient ID. In contrast, the target variable serves as a binary class label, with class 0 designating patients without heart disease and class 1 designating those who do. Table 1 lists all of the variables utilized in the dataset.

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1015−1024
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10372

Table 1. Variables in the Dataset

No	Variabel	Tipe Data
1	patientid	Numeric
2	age	Numeric
3	gender	Categorical
4	chestpain	Categorical
5	restingBP	Numeric
6	serumcholestrol	Numeric
7	fastingbloodsugar	Categorical
8	restingrelectro	Categorical
9	maxheartrate	Numeric
10	exerciseangia	Categorical
11	oldpeak	Numeric
12	slope	Categorical
13	noofmajorvessels	Categorical
14	target	Categorical

Initial analysis of the dataset showed no missing values or duplicate data across all 1,000 samples. However, the class distribution of the target variable indicated data imbalance[19],[20]. Class 0, or no heart disease, accounted for 420 samples (42%), while class 1, or heart disease, accounted for 580 samples (58%). The class distribution is shown in Figure 2.

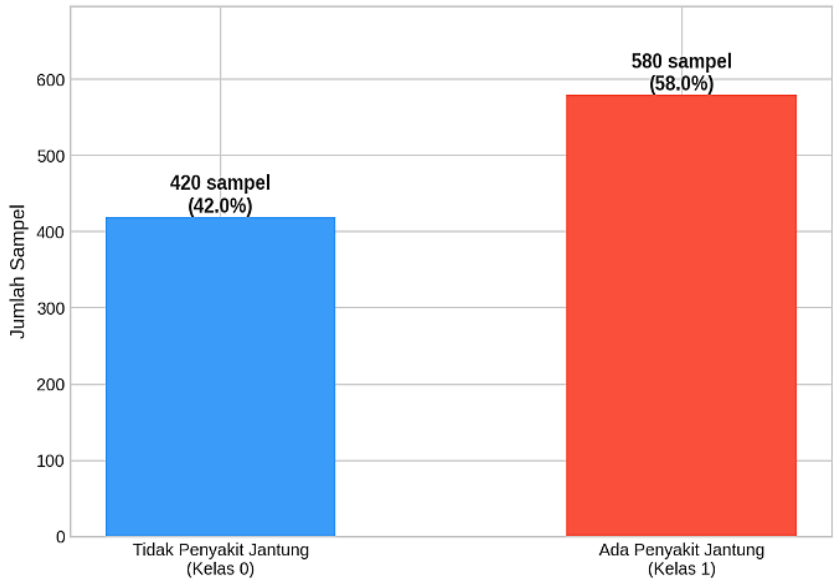

Figure 2. Distribution of dataset classes

2.3 Random Forest Algorithm

Random Forest is a classification method based on ensemble learning algorithm which benefits the group decision tree to produce a final decision. In the formation process, each tree is trained using a portion of the data sample selected randomly through a mechanism bootstrap. In addition, feature selection at each tree branch is also done randomly so that each tree has a different learning pattern. The classification results from all trees are then combined, and the class with the most votes is used as the final prediction result[21]. This mechanism makes Random Forest more stable than a single decision tree, because the model's decisions do not depend solely on a single tree structure. In addition, this approach can help reduce the risk overfitting and remains suitable for use on data that has many features[22]. In this study, the Gini index is used as a node separation criterion which is expressed in the following equation:

$$Gini = 1 - \sum_{i=1}^{c}(p_i)^2 \quad (1)$$

with p_i is the relative frequency of the class i whereas c shows the number of classes in the dataset. Nodes with values $Gini$ approaching 0 indicates a purer and more optimal separation[23].

2.4 Improved Genetic Algorithm

The proposed Improved Genetic Algorithm (IGA) extends the conventional Genetic Algorithm (GA) to address the premature convergence problem that often occurs during the evolutionary search process. GA has been widely

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1015−1024
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10372

applied for feature selection because it can identify informative feature subsets that improve classification performance. Previous studies have also demonstrated that optimization algorithms can be integrated with Random Forest to enhance classification performance through more effective feature selection[24]. Unlike the conventional Genetic Algorithm that employs fixed crossover and mutation probabilities, the proposed IGA combines elitism with adaptive crossover and adaptive mutation operators. The combination is intended to preserve high-quality solutions while dynamically balancing exploration and exploitation throughout the evolutionary search process, thereby improving feature subset optimization for Random Forest. In this study, proposed IGA integrates three main strategies. First, an elitism strategy that retains a number of the best individuals from each generation to prevent them from being lost during the evolutionary process. Second, the adaptive crossover rate is gradually decreased from 0.85 to 0.60 as the generations progress, allowing broader exploration in the early generations and stronger exploitation in the later generations. Third, the adaptive mutation rate is gradually increased from 0.05 to 0.20 to maintain population diversity and reduce the risk of premature convergence[25]. The fitness function used combines cross-validation accuracy with a feature count penalty as follows:

$$Fitness = \alpha \times \text{Accuracy}_{cv} + \beta \times \left(1 - \frac{\text{fitur terpilih}}{\text{total fitur}}\right) \quad (2)$$

where α and β are weights that regulate the balance between feature accuracy and efficiency. The values of α = 0.92 and β = 0.08 were empirically determined through preliminary experiments involving several α–β combinations. Among the evaluated combinations, α = 0.92 and β = 0.08 achieved the highest classification performance while selecting six informative features. Therefore, these values were adopted in the final experiments to balance prediction accuracy and feature reduction.

2.5 Model Evaluation

The model's capacity to categorize heart disease data in the test data was assessed. Several assessment measures, including accuracy, precision, recall, and F1-score, were used in this study to compare the model's performance using Random Forest before feature selection and Random Forest after feature selection using Improved Genetic Algorithm. These metrics are computed using the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values found in the confusion matrix. The following equation displays the evaluation formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1\text{-}Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (6)$$

In this study, these four metrics are used as a basis for comparing model performance before and after feature selection, so that it can be seen whether the use of Improved Genetic Algorithm able to improve classification performance and produce more efficient models[26].

3. RESULT AND DISCUSSION

3.1 Data Preprocessing

To make the dataset ready for use in the classification procedure, the preprocessing step was completed. The results of the inspection showed that all the data did not contain missing value and duplication, so that 1,000 data points could still be used in the study. Furthermore, the patientid variable was removed from the modeling process because it only serves as a patient identifier and has no clinical contribution to prediction. The target variable was separated as a class label, so the classification process was carried out using 12 predictor features. Some numeric features were then scaled using Min-Max Scaling to ensure a uniform range of values. After preprocessing, the dataset is divided into training and test data in an 80:20 ratio.

Based on the class distribution of the original dataset, there were 420 samples (42%) belonging to class 0 (no heart disease) and 580 samples (58%) belonging to class 1 (heart disease), indicating a moderate class imbalance, as illustrated in Figure 3. After the dataset was divided into training and testing sets, the training data contained 336 samples from class 0 and 464 samples from class 1. To address this imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied only to the training data, while the testing data remained unchanged to ensure an unbiased evaluation of model performance. As shown in Figure 3, the application of SMOTE successfully balanced both classes to 464 samples each (50%). This balanced training set helps reduce model bias toward the majority class and improves the learning process without introducing synthetic samples into the testing phase.

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1015−1024
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10372

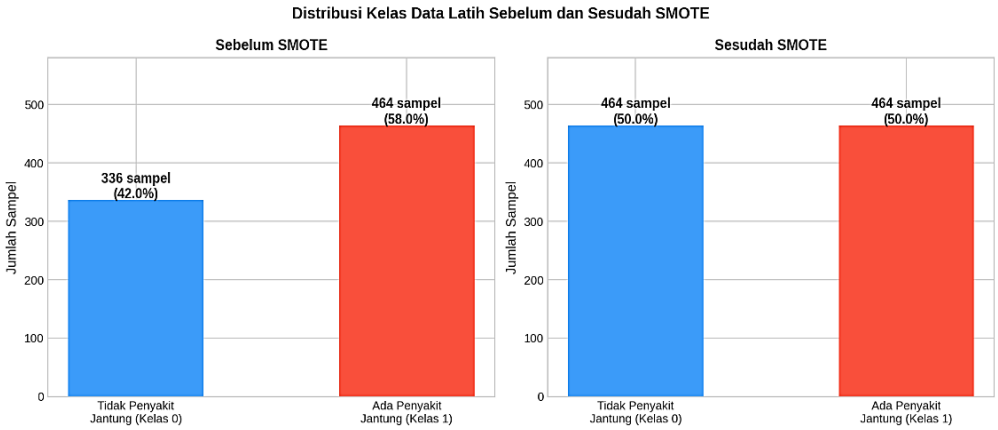

Figure 3. Distribution of training data classes before and after SMOTE

3.2 Model Random Forest

Model Random Forest built as an initial comparison model before feature selection optimization is carried out using Improved Genetic Algorithm At this stage, the Model Random Forest using 12 predictor features, namely age, gender, chestpain, restingBP, serumcholestrol, fastingbloodsugar, restingrelectro, maxheartrate, exerciseangia, oldpeak, slope, And noofmajorvessels. The patientid variable is not used because it only functions as a patient identifier, while the target variable is used as a class label.

Testing on test data shows that the model Random Forest produced an accuracy of 99.00%, a precision of 99.14%, a recall of 99.14%, and an F1-score of 99.14%, as shown in Figure 4. These results indicate that the Random Forest model is capable of classifying heart disease very well using 12 predictor features.

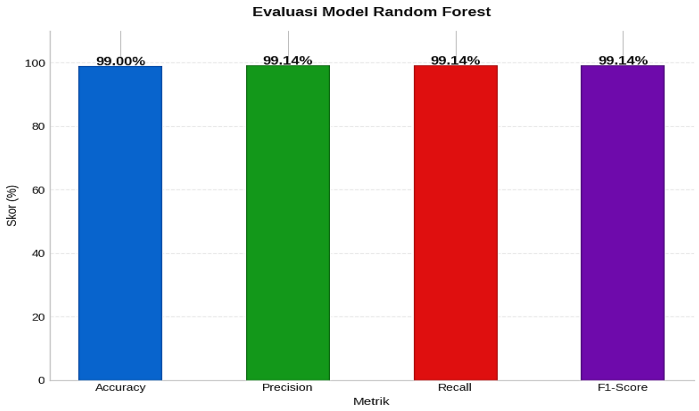

Figure 4. Random Forest Model Evaluation Results

Results confusion matrix Figure 5 shows that from the total test data, the model successfully classified 83 data points as having no heart disease and 115 data points as having heart disease. However, there was still 1 data point as having no heart disease that was predicted as having heart disease and 1 data point as having heart disease that was predicted as not having heart disease. This very small classification error indicates that the model Random Forest has high classification capabilities, but can still be optimized to improve performance.

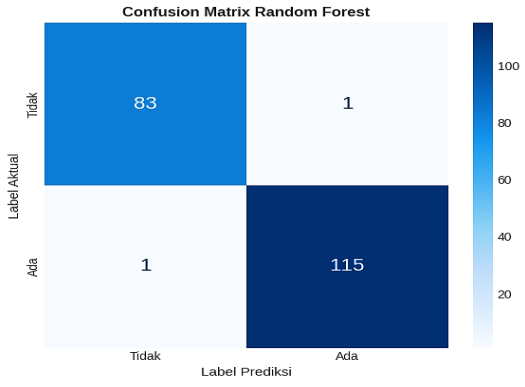

Figure 5. Confusion Matrix Random Forest

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1015−1024
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10372

3.3 Improved Genetic Algorithm Feature Selection Results

Feature selection is done using Improved Genetic Algorithm (IGA) to obtain the best combination of features from the 12 predictor features used in the model. Random Forest The IGA process is carried out by considering a fitness value that combines the model's accuracy performance and the number of features used. In this study, the fitness value uses an alpha weight of 0.92 for accuracy and a beta weight of 0.08 for the feature number penalty, so the selection process focuses not only on improving accuracy but also on reducing less relevant features.

Based on the convergence graph in Figure 6, the value best fitness experienced an increase since the first generation and tended to be stable until the 10th generation. This behavior indicates that the proposed IGA was able to efficiently identify a near-optimal feature subset while maintaining solution stability during the remaining evolutionary process. Meanwhile, the average fitness fluctuated across generations because different candidate solutions were continuously generated through the selection, crossover, and mutation operators, reflecting the balance between exploration and exploitation.

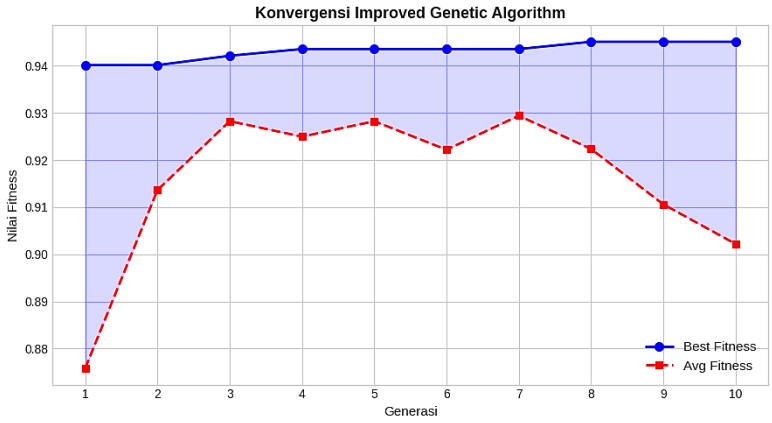

Figure 6. Convergence of Improved Genetic Algorithm

Feature selection results using Improved Genetic Algorithm produce 6 selected features from a total of 12 initial features, namely slope, chestpain, restingBP, restingrelectro, oldpeak, And gender. Based on value importance in Figure 7, the feature that has the greatest contribution isslope 48.08%, followed by chestpain 18,38%, restingBP 16,48%, restingrelectro 7,26%, old peak 6.44%, andgender 3.36%. These results indicate that IGA is capable of reducing features from 12 to 6, allowing the model to work with more concise and relevant features for heart disease classification.

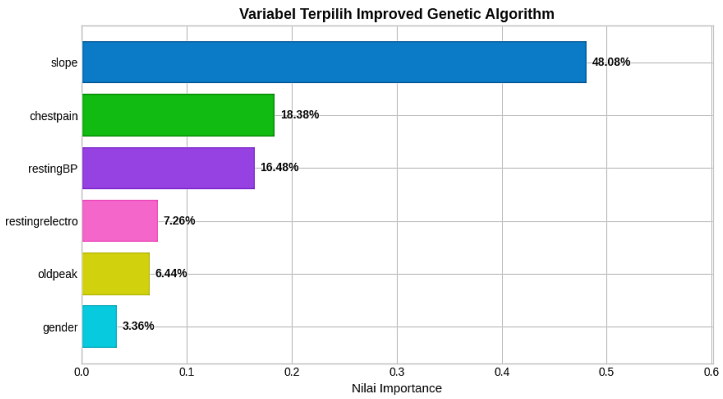

Figure 7. Selected Variables of Improved Genetic Algorithm

The dominance of the slope feature is clinically plausible because it represents the ST-segment slope observed during exercise electrocardiography, which is widely recognized as one of the important indicators for detecting myocardial ischemia and coronary artery disease. Therefore, its high importance is consistent with existing medical knowledge. Nevertheless, the feature importance values obtained in this study are specific to the employed dataset and should not be interpreted as universal clinical evidence.

3.4 Discussion

This research succeeded in creating a classification model for heart disease after obtaining the 6 best features through a feature selection process using Improved Genetic Algorithm, the next stage is to build the model Random Forest by using only selected features, namely slope, chestpain, restingBP, restingrelectro, oldpeak, And gender.

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1015−1024
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10372

As seen in Figure 8, the model Random Forest + IGA produced high classification performance with 99.50% accuracy, 99.15% precision, 100.00% recall, and 99.57% F1-score. The 100.00% recall value indicates that all patient data belonging to the heart disease class were correctly recognized by the model. This condition shows that the results of feature selection using IGA are able to maintain and even improve model performance. Random Forest even though the number of features used is less.

Visualization confusion matrix Figure 9 shows that the model was able to correctly predict 83 data points in the non-heart disease class and 116 data points in the present heart disease class. Misclassification only occurred in one data point in the non-heart disease class that was predicted as having heart disease, while there were no errors in the present heart disease class. These results indicate that the RF + IGA model has good classification capabilities, particularly in minimizing errors in the class of patients with heart disease.

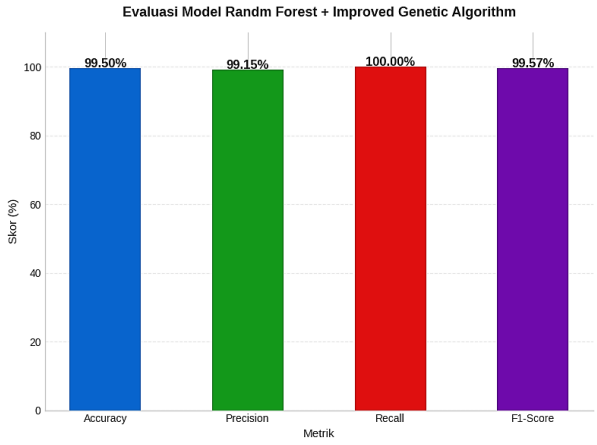

Figure 8. Evaluation Results of the Random Forest + Improved Genetic Algorithm Model

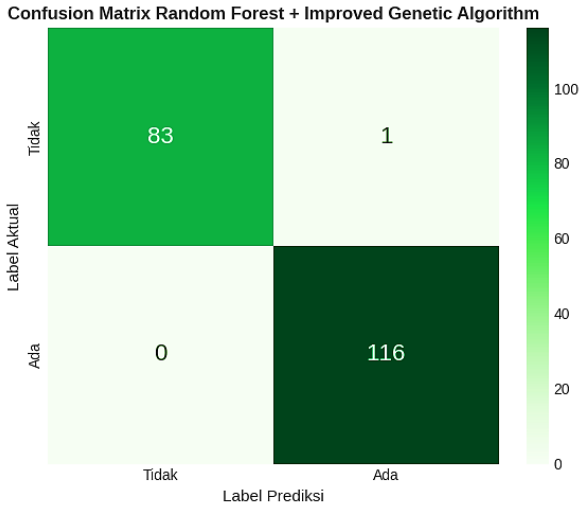

Figure 9. Confusion Matrix Random Forest + Improved Genetic Algorithm

Model comparison Random Forest and Random Forest + Improved Genetic Algorithm (RF + IGA) was conducted to determine the effect of feature selection on the performance of heart disease classification. The model Random Forest the baseline uses 12 predictor features, while the RF + IGA model only uses 6 selected features, namely slope, chestpain, restingBP, restingrelectro, oldpeak, And gender. The results of the performance comparison of the two models are presented in Table 3.

Table 3. Comparison of Random Forest and RF + IGA Performance

Classification Model	matrix %				
	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Random Forest	99,00	99,14	99,14	99,14	99,90
RF + IGA	99,50	99,15	100,00	99,57	99,90

Based on Table 3, the model Random Forest produces 99.00% accuracy, 99.14% precision, 99.14% recall, 99.14% F1-score, and 99.90% AUC-ROC. After feature selection using Improved Genetic Algorithm, the RF + IGA model achieved 99.50% accuracy, 99.15% precision, 100.00% recall, 99.57% F1-score, and 99.90% AUC-ROC. The difference in these values indicates that the use of IGA is able to provide improvements in several

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1015−1024
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10372

evaluation metrics, especially on recall. Mark recall which reached 100.00% indicating that all data in the heart disease class was successfully classified correctly by the model. Although the proposed model achieved an accuracy of 99.50%, this result should be interpreted carefully because the dataset consists of only 1,000 publicly available samples. High classification performance on relatively small datasets may be influenced by the characteristics of the dataset itself, including the separability of the classes and the distribution of predictor variables. To minimize the risk of data leakage, the dataset was first divided into training and testing subsets, and SMOTE was applied exclusively to the training data. Consequently, the test set remained untouched throughout the training process. Nevertheless, further validation using larger and more diverse clinical datasets is required to confirm the generalizability of the proposed approach.

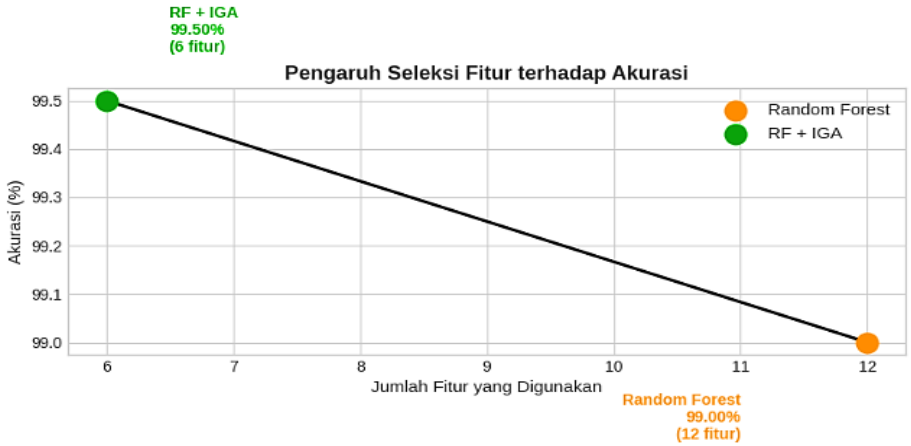

Figure 10. Effect of Feature Selection on Accuracy

Figure 10 shows the effect of feature selection on model accuracy. Random Forest The baseline uses 12 features with 99.00% accuracy, while RF + IGA only uses 6 features but produces a higher accuracy of 99.50%. This condition indicates that the application of Improved Genetic Algorithm able to filter out more relevant features, reducing the number of features by 50% without decreasing model performance. In fact, this feature reduction actually improves classification accuracy. Thus, IGA can help shape the model. Random Forest which is more concise, efficient, and still has high classification capabilities.

Table 4. Results of Research Accuracy Comparison

Penelitian	Metode	Akurasi
This research	Random Forest, Random Forest + IGA	Random Forest: 99,00 Random Forest + IGA: 99,50
[4]	Decision tree, naive Bayes, and random forest classifier	Decision tree: 84,40 Naïve Bayes: 85,00 Random forest classifier: 85,20
[5]	Decision tree, naive Bayes, and random forest	Decision tree: 72,00 Naïve Bayes: 71,00 Random forest: 75,00
[6]	Ensemble voting classifier, naive Bayes and random forest algorithm	Naïve Bayes: 67,24 Random forest: 97,41 Ensemble voting classifier: 98,28
[7]	Random forest classifier, C45 algorithm, logistic regression, and (SVM)	Random forest classifier: 94,54 C45 algorithm: 91,74 logistic regression: 76,27 SVM: 76,24
[8]	Random Forest classifier + SMOTE + Hyperparameter Tuning	Random Forest classifier + SMOTE + Hyperparameter Tuning: 92%

Based on Table 4, this study shows that the implementation of Improved Genetic Algorithm on Random Forest can contribute to the heart disease classification process. Unlike previous studies that focused more on algorithm comparisons or model combination, this study emphasizes optimizing feature selection to obtain more relevant feature combinations. Thus, the resulting model not only has competitive classification performance but is also more efficient because it does not use all the initial features. This demonstrates that IGA-based feature selection can be an effective approach in building simpler heart disease classification models.

The modeling of heart disease classification in this study shows that the application of Improved Genetic Algorithm on Random Forest can contribute to the development of more targeted medical prediction models. By

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1015−1024
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10372

selecting the most relevant features, the resulting model has the potential to help more efficiently identify heart disease risk, as not all patient attributes need to be used in the classification process. This approach can form the basis for developing simpler, faster, and more user-friendly clinical decision support systems. This study evaluates the proposed IGA against the Random Forest baseline to demonstrate the effectiveness of feature selection. However, a direct comparison with the conventional Genetic Algorithm and other feature selection methods such as Boruta or Recursive Feature Elimination (RFE) was not included. Such comparisons constitute an important direction for future work to further quantify the advantages of the proposed approach. Therefore, further validation using independent clinical datasets is required to confirm whether the same feature ranking can be consistently reproduced in different patient populations.

4. CONCLUSION

This study demonstrates that the proposed Improved Genetic Algorithm (IGA), which integrates elitism with adaptive crossover and adaptive mutation operators, can effectively optimize feature selection for Random Forest based heart disease classification. The proposed Random Forest + IGA model achieved an accuracy of 99.50%, recall of 100.00%, and an F1-score of 99.57%, while reducing the number of predictor features from 12 to 6, representing a 50% reduction in the feature set. These findings indicate that the proposed feature selection strategy can improve model efficiency without reducing predictive performance, suggesting that the eliminated features contributed relatively little to the classification process. From a methodological perspective, this study demonstrates that integrating adaptive evolutionary operators with Random Forest provides an effective approach for identifying relevant feature subsets in heart disease classification. Nevertheless, the results should be interpreted with caution because the experiments were conducted on a single public Kaggle dataset containing 1,000 records, and no external validation was performed. Therefore, the generalizability of the proposed model to broader clinical populations remains to be verified. Future work should evaluate the proposed IGA against conventional Genetic Algorithm and other feature selection methods, such as Recursive Feature Elimination (RFE) and Boruta, using larger and more diverse clinical datasets.

REFERENCES

[1] S. D. Sawu, A. A. Prayitno, and Y. I. Wibowo, “Analisis Faktor Risiko pada Kejadian Masuk Rumah Sakit Penyakit Jantung Koroner di Rumah Sakit Husada Utama Surabaya,” J. Sains dan Kesehat., vol. 4, no. 1, pp. 10–18, 2022, doi: 10.25026/jsk.v4i1.856.

[2] World Health Organization, “Cardiovascular diseases,” Drug Discovery Today: Disease Models. Accessed: May 28, 2026. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases?utm_source#tab=tab_1

[3] Tim Riskesdas 2018, “Laporan Nasional Riskesdas 2018,” 2019. [Online]. Available: https://repository.kemkes.go.id/book/1323?utm_source

[4] J. D. Muthohhar and A. Prihanto, “Analisis Perbandingan Algoritma Klasifikasi untuk Penyakit Jantung,” J. Informatics Comput. Sci., vol. 04, pp. 298–304, 2023, doi: 10.26740/jinacs.v4n03.p298-304.

[5] D. H. Depari, Y. Widiastiwi, and M. M. Santoni, “Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung,” Inform. J. Ilmu Komput., vol. 18, no. 3, p. 239, 2022, doi: 10.52958/iftk.v18i3.4694.

[6] Dede Kurniadi, Asri Indah Pertiwi, and Asri Mulyani, “Ensemble Voting Classifier Berbasis Multi-Algoritma dan Metode SMOTE untuk Klasifikasi Penyakit Jantung,” J. Nas. Tek. Elektro dan Teknol. Inf., vol. 14, no. 2, pp. 145–153, 2025, doi: 10.22146/jnteti.v14i2.17157.

[7] Anis Fitri Nur Masruriyah, Hilda Yulia Novita, Cici Emilia Sukmawati, Siti Novianti Nuraini Arif, Angga Ramda Ramadhan, and P. Studi Informatika, “Evaluasi Algoritma Pembelajaran Terbimbing terhadap Dataset Penyakit Jantung yang telah Dilakukan Oversampling,” J. MIND J. | ISSN, vol. 8, no. 2, pp. 242–253, 2023, [Online]. Available: https://doi.org/10.26760/mindjournal.v8i2.242-253

[8] A. M. A. Rahim, Inggrid Yanuar Risca Pratiwi, and Muhammad Ainul Fikri, “Klasifikasi Penyakit Jantung Menggunakan Metode Synthetic Minority Over-Sampling Technique Dan Random Forest Clasifier,” Indones. J. Comput. Sci., vol. 12, no. 5, 2023, doi: 10.33022/ijcs.v12i5.3413.

[9] A. S. Prabowo and F. I. Kurniadi, “Analisis Perbandingan Kinerja Algoritma Klasifikasi dalam Mendeteksi Penyakit Jantung,” J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan), vol. 7, no. 1, pp. 56–61, 2023, doi: 10.47970/siskom-kb.v7i1.468.

[10] T. Gori and A. Hestiningtyas, “Optimasi Pemilihan Fitur untuk Prediksi Penyakit Jantung Menggunakan Algoritma Genetika dan Random Forest,” Indones. J. Comput. Sci., vol. 13, no. 5, 2024, doi: 10.33022/ijcs.v13i5.4214.

[11] M. K. Suryadi, R. Herteno, S. W. Saputro, M. R. Faisal, and R. A. Nugroho, “A Comparative Study of Various Hyperparameter Tuning on Random Forest Classification with SMOTE and Feature Selection Using Genetic Algorithm in Software Defect Prediction,” J. Electron. Electromed. Eng. Med. Informatics, vol. 6, no. 2, pp. 137–147, 2024, doi: 10.35882/jeeemi.v6i2.375.

[12] D. N. Jawza, M. I. Mazdadi, A. Farmadi, T. H. Saragih, D. Kartini, and V. Abdullayev, “Enhancing Diabetes Prediction Accuracy Using Random Forest and XGBoost with PSO and GA-Based Feature Selection,” J. Electron. Electromed. Eng. Med. Informatics, vol. 7, no. 2, pp. 295–306, 2025, doi: 10.35882/jeeemi.v7i2.626.

[13] H. Ghinaya, R. Herteno, M. R. Faisal, A. Farmadi, and F. Indriani, “Analysis of Important Features in Software Defect Prediction using Synthetic Minority Oversampling Techniques (SMOTE), Recursive Feature Elimination (RFE) and Random Forest,” J. Electron. Electromed. Eng. Med. Informatics, vol. 6, no. 3, pp. 276–288, 2024, doi:

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1015−1024
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10372

10.35882/jeeemi.v6i3.453.
[14] S. P. R. Yulianto, A. Z. Fanani, A. Affandy, and M. I. Aziz, "Analisis Metode Smoote pada Klasifikasi Penyakit Jantung Berbasis Random Forest Tree," J. Media Inform. Budidarma, vol. 8, no. 3, p. 1460, 2024, doi: 10.30865/mib.v8i3.7712.
[15] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, "Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang," MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput., vol. 21, no. 3, pp. 677–690, 2022, doi: 10.30812/matrik.v21i3.1726.
[16] B. Satria, T. A. Y. Siswa, and W. J. Pranoto, "Optimasi Random Forest dengan Genetic Algorithm dan Recursive Feature Elimination pada High Dimensional Data Stunting Samarinda," J. Media Inform. Budidarma, vol. 8, no. 3, p. 1778, 2024, doi: 10.30865/mib.v8i3.7883.
[17] C. Cahyaningtyas, D. Manongga, and I. Sembiring, "Algorithm Comparison and Feature Selection for Classification of Broiler Chicken Harvest," J. Tek. Inform., vol. 3, no. 6, pp. 1717–1727, 2022, doi: 10.20884/1.jutif.2022.3.6.493.
[18] G. Guntoro, L. Lisnawita, and L. Costaner, "Optimizing Random Forest for IoT Cyberattack Detection using SMOTE: A Study on CIC-IoT2023 Dataset," MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput., vol. 25, no. 1, pp. 83–96, 2025, doi: 10.30812/matrik.v25i1.5382.
[19] C. Kaope and Y. Pristyanto, "The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance," MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput., vol. 22, no. 2, pp. 227–238, 2023, doi: 10.30812/matrik.v22i2.2515.
[20] M. A. Nugraha, M. I. Mazdadi, A. Farmadi, Muliadi, and T. H. Saragih, "Penyeimbangan Kelas SMOTE dan Seleksi Fitur Ensemble Filter pada Support Vector Machine untuk Klasifikasi Penyakit Liver," J. Teknol. Inf. dan Ilmu Komput., vol. 10, no. 6, pp. 1273–1284, 2023, doi: 10.25126/jtiik.2023107234.
[21] W. N. Rohman and I. M. A. Agastya, "Evaluation of SMOTE Technique in the Comparison of XGBoost and Random Forest Algorithms for Liver Disease Prediction," J. Appl. Informatics Comput., vol. 9, no. 6, pp. 3157–3167, 2025, doi: 10.30871/jaic.v9i6.10239.
[22] S. D. Amalia, M. A. Barata, and P. E. Yuwita, "Optimization of Random Forest Algorithm with Backward Elimination Method in Classification of Academic Stress Levels," J. Appl. Informatics Comput., vol. 9, no. 3, pp. 633–641, 2025, doi: 10.30871/jaic.v9i3.9280.
[23] M. H. Dzaki, A. Nugraha, A. Luthfiarta, A. A. R. Riyanto, and Y. D. Novandian, "Accelerating Classification For Iot Attack Detection Using Decision Tree Model With Gini Impurity Tree-Based Feature Selection Technique," J. Tek. Inform., vol. 6, no. 5, pp. 3405–3418, 2025, doi: 10.52436/1.jutif.2025.6.5.3817.
[24] M. D. Wardana et al., "Implementation of Ant Colony Optimization in Obesity Level Classification Using Random Forest," J. Tek. Inform., vol. 6, no. 5, pp. 3543–3557, 2025, doi: 10.52436/1.jutif.2025.6.5.4696.
[25] M. A. Al Ghifari, I. Budiman, T. H. Saragih, M. I. Mazdadi, R. Herteno, and H. A. A. Rozaq, "Implementation of Extra Trees Classifier and Chi-Square Feature Selection for Early Detection of Liver Disease," J. Tek. Inform., vol. 6, no. 5, pp. 3925–3937, 2025, doi: 10.52436/1.jutif.2025.6.5.4261.
[26] T. N. Annisa, J. Jasmir, and N. Nurhadi, "Comparison of ANOVA and Chi-Square Feature Selection Methods to Improve Machine Learning Performance in Anemia Classification," J. Tek. Inform., vol. 6, no. 4, pp. 1925–1940, 2025, doi: 10.52436/1.jutif.2025.6.4.5017.