International Journal of Reconfigurable and Embedded Systems (IJRES)
Vol. 12, No. 2, July 2023, pp. 205~214
ISSN: 2089-4864, DOI: 10.11591/ijres.v12.i2.pp205-214

Heart failure prediction based on random forest algorithm using genetic algorithm for feature selection

Yudi Ramdhani[1], Cakra Mahendra Putra[2], Doni Purnama Alamsyah[3]
[1]Department of Informatic Engineering, Adhirajasa Reswara Sanjaya University, Bandung, Indonesia
[2]Department of Information Systems, Adhirajasa Reswara Sanjaya University, Bandung, Indonesia
[3]Entrepreneurship Department, BINUS Business School Undergraduate Program, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Aug 7, 2022
Revised Oct 3, 2022
Accepted Dec 10, 2022

Keywords:

Cardiovascular
Cross validation
Feature selection
Genetic algorithm
Heart failure
Random forest
Split validation

ABSTRACT

A disorder or illness called heart failure results in the heart becoming weak or damaged. In order to avoid heart failure early on, it is crucial to understand the causes of heart failure. Based on validation, two experimental processing steps will be applied to the dataset of clinical records related to heart failure. Testing will be done in the first step utilizing six different classification algorithms, including K-nearest neighbor, neural network, random forest, decision tree, Naïve Bayes, and support vector machine (SVM). Cross-validation was employed to conduct the test. According to the results, the random forest algorithm performed better than the other five algorithms in tests employing the algorithm. Subsequent testing uses an algorithm with the best accuracy value, which will then be tested again using split validation with varying split ratios and genetic algorithms as a selection feature. The value generated from testing using the genetic algorithm selection feature is better than the random forest algorithm alone, which is recorded to produce an accuracy value of 93.36% in predicting the survival of heart failure patients.

Corresponding Author:

Yudi Ramdhani
Department of Informatic Engineering, Adhirajasa Reswara Sanjaya University
Bandung, Indonesia
Email: yudi@ars.ac.id

1. INTRODUCTION

The heart is a vital organ that is also known as the "centrality of the human body" since it is responsible for supplying blood to all of the other organs and when it is unable to fulfill these duties, a person may pass away instantly. According to study, the majority of adults who are susceptible to heart-related disorders have unhealthy eating habits, mental stress, despair, and excessive work hours. These factors are also present in those who have heart failure [1]. Heart failure is essentially a disorder where the heart is unable to adequately pump blood to the body's organs. This typically happens as a result of diseases like diabetes, hypertension, or other heart ailments like human immunodeficiency virus (HIV), thyroid disorders, alcoholism, or genetic diseases [2]. When the heart's muscles deteriorate, the heart's ability to pump blood is restricted [3]. It is significant to mention that one of the most prevalent ailments among middle-aged persons is heart disease [4]. As confirmed by the World Health Organization (WHO) [3], cardiovascular disease (CVD), a condition affecting the heart and blood vessels, is responsible for 31% of annual deaths. Therefore, in order to be able to give the proper treatment process and save the lives of many patients, it is essential to recognize the condition early and precisely [5]. Medical professionals have effectively used machine learning algorithms like principal component analysis (PCA) and support vector machine (SVM) to assist in the diagnosis of a variety of disorders including diabetes and heart failure. Researchers have also used artificial neural networks (ANN) in the medical field [6].

Additionally, a significant part in obtaining useful information from massive data is played by data mining. It is extensively employed in practically all spheres of life, including business, engineering, medical, and education. Data mining is employed to examine data by minimizing mistakes in forecasts and factual outcomes, several machine learning algorithms have been used to comprehend the intricacy and non-linear interactions between diverse components [7]. Additionally, this prospective strategy opens a significantly better resource window, improving the sensitivity and specificity of disease detection and diagnosis [8]. Machine learning algorithms are required as medical data keeps growing in order to help the medical team analyze data and make precise and accurate diagnostic conclusions [9]. To forecast cardiovascular disease in patients and cardiac death, many classification algorithms are utilized in medical data mining [10].

Machine learning algorithms typically handle research prediction and classification tasks that use ready-to-use data sets. Researchers have tried a variety of methods to increase the precision of data classification to locate possible patients [11]. In this work, a machine learning system was used to classify conditions of death from heart disease in accordance with measurement results and life information gathered from individuals. One application of labeled data in supervised learning is classification. A new class of predictions is created by category class, and label class after classification splits the dataset into training and test data. One method that is quite well-known in classification is random forest [12]. Ho originally put forth the random forest algorithm in 1995, and claims that the algorithm can achieve greater accuracy without overtraining if the decision tree can recognize skewed hyperplanes [13]. The underlying algorithm's "randomness" is a requirement for the random forest algorithm's improved accuracy. The random forest technique, which integrates genetic algorithms, will be used in this study to pick features for categorizing heart failure. The primary goal of this work is to overcome the issue of dataset imbalance and choose feature selection to obtain more accuracy. To do this, we will employ the random forest algorithm and genetic algorithm as machine feature selection to predict the survival of heart failure patients suitable in comparison to earlier studies.

2. METHOD

The data mining paradigm, which describes the process of looking for or mining knowledge, is the foundation of this research which optimization function based on the random forest algorithm. The genetic algorithm will search the "heart failure clinical record dataset" using this objective function to uncover significant features. Figure 1 shows the overall progression of this study.

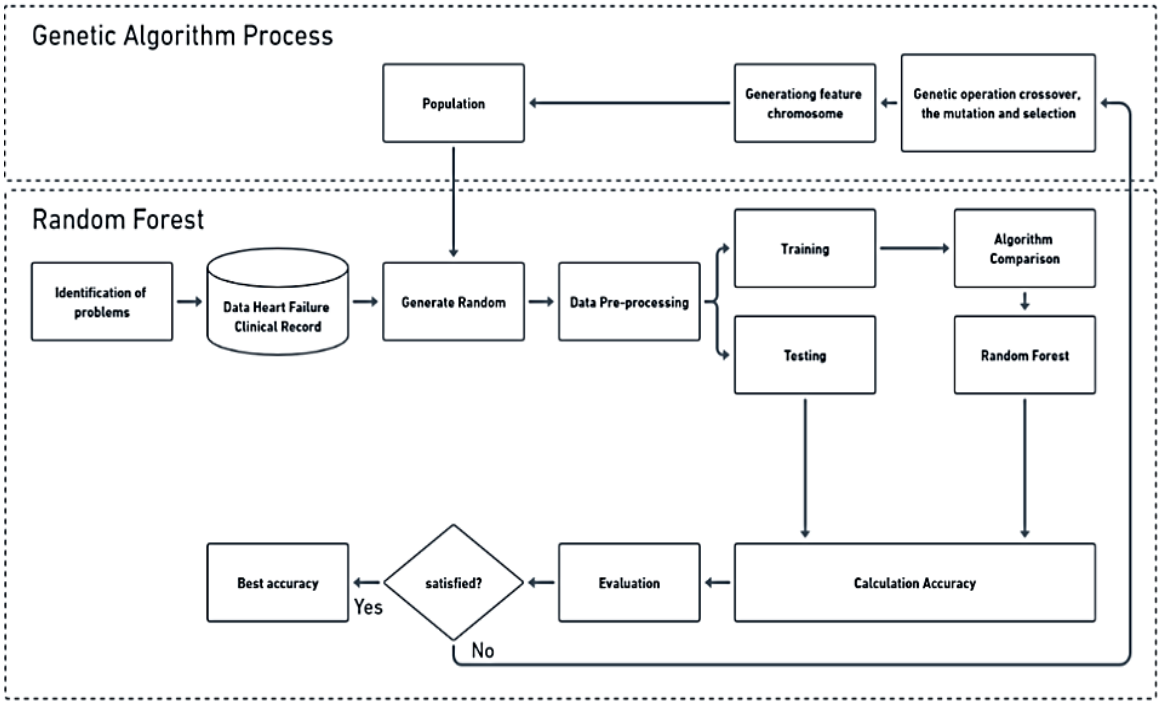

Figure 1. Workflow of the proposed framework

2.1. Problem identification

The aim of this study is to create a reliable predictive model for heart failure patients. The researchers will use a random forest algorithm to predict patient survival and a genetic algorithm as a feature selection technique to achieve this goal. The accuracy of the model is expected to increase significantly by selecting the most important features. Accurate heart failure classification is essential for early diagnosis of heart failure and appropriate treatment of heart failure, and this study has the potential to contribute to better clinical records and, as a result, better patient outcomes. Overall, this study has the potential to improve clinical decision making for heart failure patients and advance our understanding of prediction of survival in heart failure patients.

2.2. Dataset

The dataset is the primary component that will be analyzed using an algorithm in order to perform research. Heart failure clinical records data from the UCI repository website were the dataset used in this investigation. The collection includes the medical histories of 299 patients with heart conditions. 194 males and 105 females were found in the 299 records (older than 40). In the target class, there were 203 survivors (incidence of death=0), compared to 96 fatalities (incidence of death=1). In terms of statistics, there were 67.89% negative and 32.11% positive. The dataset was published in 2020 with 13 features with 299 events of class_death, the information shown in Table 1.

Table 1. The feature of heart failure record dataset

No	Features	Description	Data type
1.	Age	Patient age (years)	Numeric
2.	Anemia	Decrease in red blood cells or hemogoblin	Boolean
3.	Creatine	CPK enzyme level in the blood (mcg/L)	Numeric
4.	Diabetes	If the patient has diabetes	Boolean
5.	Ejection_fraction	The percentage of blood that leaves the heart with each contraction	Numeric
6.	High_blood_pressure	If people with hypertension	Boolean
7.	Platelets	Platelets in the blood (kiloplatelet/mL)	Numeric
8.	Serum_creatinine	Serum creatinine level in the blood(mg/dL)	Numeric
9.	Serum_sodium	Serum sodium level in the blood (mEq/L)	Numeric
10.	Sex	Gender: female of male	Boolean
11.	Smoking	If the patient smokes or not	Boolean
12.	Time	Follow-up period (days)	Numeric
13.	Death_event	If the patient dies during follow-up	Boolean

2.3. Pre-processing

Data preprocessing for data mining is the collection of approaches utilized prior to the use of data mining methods, and it is acknowledged as one of the most important difficulties in the renowned knowledge discovery of data processing [14]. Data cannot be applied directly to begin the data mining process since it is likely incomplete, inconsistent, and redundant. More complex analysis methods are required when data collection scales up. Preprocessing data enables the processing of data that would not otherwise be possible by adjusting the data to the constraints set by the different data mining algorithms [15].

2.4. The comparison algorithm

Classification and supervised machine learning models were applied to forecast results from the data. This study proposes a method for heart disease prediction employing a classification method and an ensemble of classifiers to increase classification accuracy. Individual classifiers were trained using a train set after the data were split into a training set and a test set. Using test data, classifier effectiveness is evaluated [9]. This study tested some of the most frequently used algorithms such as K-nearest neighbors, support vector machine, random forest, neural network, Naïve Bayes, and decision tree.

2.5. Random forest

Scientists are aware of any modifications to genes' DNA that occur in any particular disease [16]. As a result, it will be exceedingly challenging to identify the symptoms caused by these alterations, particularly in heart failure. Finding information is done through classification [17]. Random forest is one of the techniques used to categorize the "heart failure clinical record dataset" dataset. Every sample set builds a decision tree with attributes chosen at random [18]. The random forest is a combination of classifiers formed by combining the basic K from the decision tree [19]. Random forest is also a method that produces very accurate predictions and can handle many input variables without overfitting [20]. Random forest uses the classification and regression tree (CART) algorithm to build trees. Given the huge number of trees

constructed, the random forest approach is distinguished by excellent performance in classifiability and good noise resistance [18]. The decision tree is grown using the CART method, grows to its maximum size, and is then left unpruned. As a result, a forest is created, which is a collection of trees [21]. Additionally, the classification approach incorporates random forest, which consists of a structured collection of decision trees that each cast a unit vote for the most prevalent class on input x using independent random vectors that are distributed uniformly, as shown in Figure 2 [22].

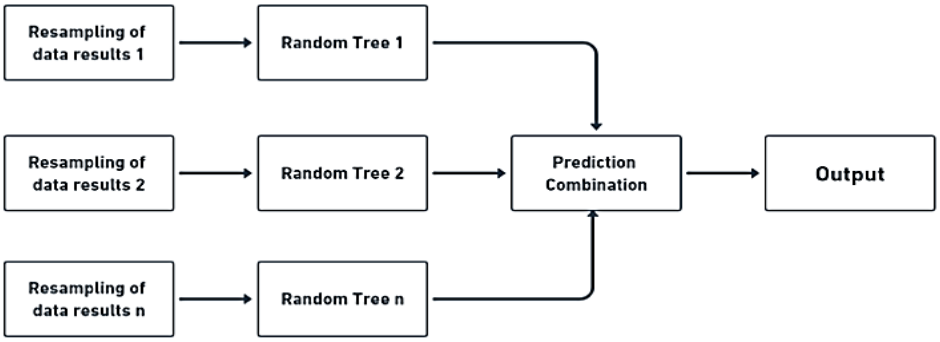

Figure 2. Random forest operation

Each decision tree will produce a result based on the input data, and the integration of several outputs will produce the ultimate result of a random forest. For the original training set, the bagging method was used to select random data by replacement and form the training set by difference. The features are also selected using a sampling approach. If it is assumed that a data set has N features, then M features will be sampled from N, where M<<N. for each extracted training set, only feature M selected randomly rather than all N features will be used for node splitting in constructing the tree. All decision trees built will grow freely without pruning [18].

2.6. Genetic algorithm

Genetic algorithm (GA) is an algorithm for search and optimization based on genetic principles and natural selection processes [23]. An optimization issue is constructed using the GA search algorithm to find the best solution from the population of candidate solutions. Figure 3 shows how GA functions.

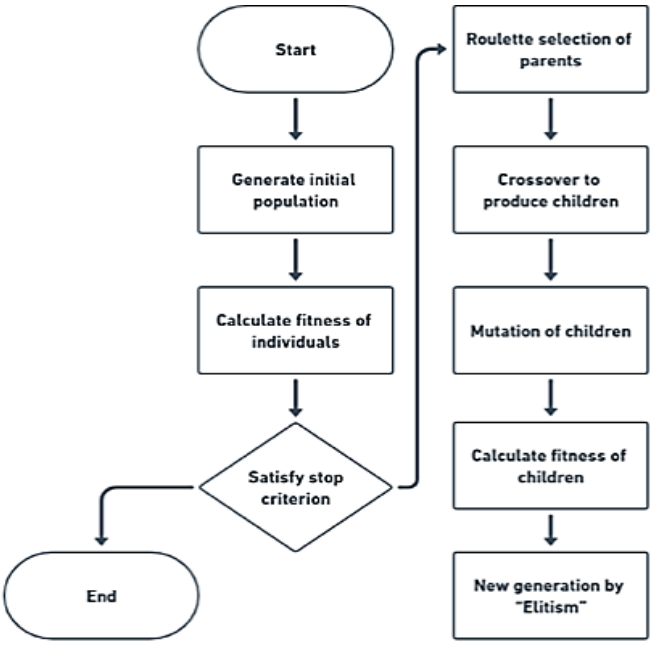

Figure 3. Working principle of GA

Candidate solutions are more broadly referred to as components of a potential answer to a specific problem. The population as a whole is made up of several potential solutions that GA is creating. Additionally, candidate solutions include a collection of parameters known as the genotype or chromosome. The genome is a representation of this set of variables [24]. A randomly generated initial population of genomes is also used as the starting point for the GA evolution mechanism. In each iteration, the fitness value is computed using the fitness function. The optimum overall fitness value was calculated based on a comparison of the present population. In this study, random forest was utilized to categorize the features; the fitness function was calculated using these results in the suggested framework. Genotypes represent the original dataset's feature vectors. Chromosomal phenotypes represent the mask of the feature vector. As a result, the phenotype labeled "0" stands for the trait that was removed, whereas the phenotype labeled "1" stands for the characteristic that was chosen. As a result, a phenotype labeled "0" is considered a less significant trait, whereas "1" is considered a highly significant feature. Each genotype produces a collection of subsets based on phenotype. The training set for the suggested framework is this subset and the operating basis of genetic processes is shown in Figure 4.

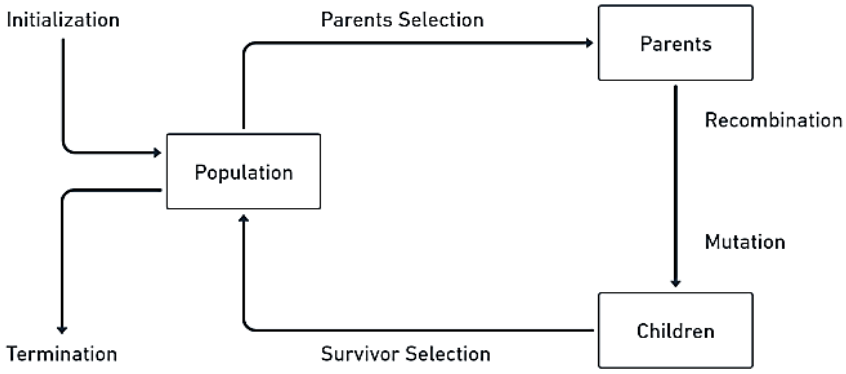

Figure 4. GA operation

2.7. Classification

The classification stage is the stage for classifying the quality of the "heart failure clinical record dataset" dataset. In this study, validation was carried out using cross validation, which previously shared training data and testing data to determine which model has the best level of accuracy. Then validate using split validation to test the model that has been taken using cross validation. This model will later be optimized using the genetic algorithm selection feature.

2.8. Matrix evaluation

To ascertain if a model is accurate or not during the classification process, the model's performance must be calculated. There are numerous ways to assess the performance of machine learning models. The progress of analytical research is anticipated to be supported by the combination of several evaluation instruments [9]. In this study, differences in machine learning-based algorithms will be investigated using four fundamental measures (accuracy, precision, recall, and F-score). The variables in the confusion matrix are employed in performance calculations for accuracy, precision, and recall. The confusion matrix elements are true positive (TP), true negative (TN), false positive (FP) and false negative (FN). If the most important forecast is based on information about medical false negatives [25]. Negative labels are classified as normal or not cancer, but positive labels are classified as cancer. Each variable in Table 2 is explained.

Table 2. Confusion matrix

Parameter	Predicted Positive	Predicted Negative
Actual positive	True positive (TP)	False positive (FP)
Actual negative	False positive (FP)	True negative (TN)

2.9. Area under cover

Area under curve (AUC) is one of the common methods used to calculate the value of under the receiver operating characteristic (ROC) curve. Area under the curve can be interpreted as a probability value,

if you choose one positive and negative example at random, the classification method will give a higher score on the positive example than the negative example. Therefore, a higher AUC value can indicate a better classification method [26]. The following is a formula for finding the AUC value seen from the results of the confusion matrix [27].

$$AUC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

The AUC value will always be in the range 0-1, because part of the unit square area with the x-axis and y-axis has a value from 0 to 1. Values above 0.5 are said to be interesting values because random predictions produce diagonal lines between (0,0) and (1,1) which has an area of 0.5. The quality of classification accuracy of diagnostic tests using AUC values is shown in Table 3 [26].

Table 3. Accuracy of classification results based on AUC

0	Category
0.90 – 1.00	Best
0.80 – 0.90	Good
0.70 – 0.80	Average
0.60 – 0.70	Poor
0.50 – 0.60	Bad

3. RESULTS AND DISCUSSION

In five classification algorithm models, the research was carried out across ten trials with cross validation, and the best accuracy value was chosen. According to the experimental findings shown in Table 1, the random forest algorithm generates results with a higher degree of accuracy than other algorithms. Then, based on variations in the split validation ratio for the categorization of heart failure, the best model will be optimized using the genetic selection algorithm feature. Table 4 shows that compared to other optimization features, the accuracy value produced by the genetic feature selection technique is better. Based on the comparison of these algorithms, it is known that the random forest algorithm produces the best accuracy and AUC values compared to other algorithms, namely 82.93% and 0.892, respectively. The following is a table of confusion matrix generated by the random forest classification model, which can be seen in Table 5. The random forest algorithm model's AUC score, according to the test findings, is 0.892. The test result demonstrates that the random forest algorithm achieves the good classification. ROC-AUC curve, which illustrates the algorithm's performance, can be seen in Figure 5. The ROC curve shows the correlation between the test data and the prediction data. The area under the curve, or AUC value, which represents the ROC curve is calculated from this curve. AUC score was 0.892, which is good. The categorization outcomes attained can be deemed to be extremely impressive. Next is the approach stage using the optimization feature.

Table 4. Comparison algorithm

Algoritma	Validasi	AUC	Accuracy
K-NN	Cross Validation	0.462	62.20%
Naïve Bayes	Cross Validation	0.855	76.59%
Decision tree	Cross Validation	0.699	76.59%
Neural network	Cross Validation	0.828	80.24%
SVM	Cross Validation	0.862	81.60%
Random forest	Cross Validation	0.892	82.93%

Table 5. Confusion matrix random forest

	True 1	True 0	Class Precision
Pred 1	**64**	22	74.42%
Pred 0	32	181	84.98%
Class Recall	66.67%	89.16%	

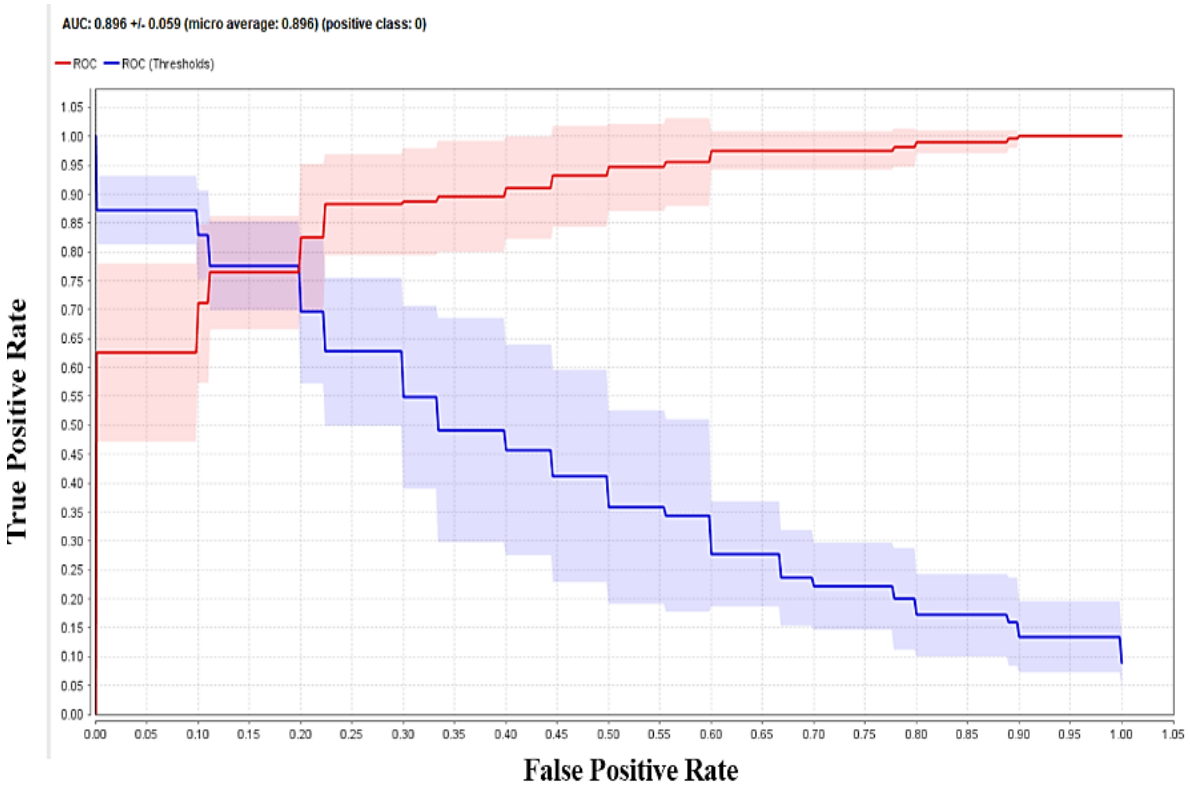

Figure 5. ROC average

After obtaining the best model, the next step is to implement feature optimization. In this study, five experiments were carried out based on variations in the split validation ratio. In each experiment, the best algorithm will be applied from the results of the comparison, namely random forest and optimization of genetic algorithm features. The experimental results can be seen in Table 6. The random forest algorithm yields an accuracy value of 82.55% on average, with the greatest results coming from a split validation ratio of 0.7 and 0.9, which yields an accuracy value of 83.33%. The best results were obtained with a split validation ratio of 0.9 at 100%, while the feature optimization approach can yield an average accuracy value of 93.36%. It is clear from the trial findings in Table 6 that the random forest classification is doing better than before. The average improvement in performance is 10.812%, and the biggest improvement, 16.670%, was at a split validation ratio of 0.9. Furthermore, a t-Test paired two samples for means test was performed to see whether the technique employed may significantly increase the performance of the random forest algorithm [28]. The results of the t-Test paired two sample for means test results in a P value of t-Test of 0.00695. These results indicate that the application of genetic algorithms for feature selection in the random forest algorithm can significantly improve accuracy performance, indicated by the P value of t-Test <0.05. The results of the t-Test test can be seen in Table 7.

According to the results of the t-Test significance test, the genetic algorithm has an excellent performance to improve the performance of the classification algorithm through the feature selection stage. The random forest algorithm will undergo another experiment in which selection criteria other than genetic algorithms, such as greedy forward selection and greedy backward selection, will be used. To ascertain whether the genetic algorithm is the optimal algorithm for feature selection, this experiment was carried out. Table 8 displays the outcomes of the experiment. The experimental results in Table 5 state that the genetic algorithm is a feature selection algorithm that can produce the best classification performance when compared to other feature selection algorithms. The average value of accuracy for each feature selection algorithm is 90.18% for forward selection, 89.56% for backward selection, 90.52% for greedy forward selection and 88.32% for greedy backward selection while the average gains the average value of the accuracy of the genetic algorithm is 93.36%. Based on all the experimental results that have been obtained, it can be concluded that the genetic algorithm has succeeded in improving the performance of the random forest algorithm for the classification of heart disease through feature selection. In previous studies, several

classifications of heart disease have been obtained using various algorithms and classification methods. A comparison of the results of this study with the results of other studies can be seen in Table 9.

This study's focus on data mining is not based on extremely big amounts of data, but it is an illustration of a field that will likely be developing in the future and have an impact on heart failure and many other aspects of health. Based on tests performed on the "heart failure clinical records dataset," it can be seen that the application of the genetic algorithm for feature selection in the random forest algorithm has a good accuracy of 93.36%, making it suitable for use by experts in the field of medical personnel. For programmers, it can also be a reference method that can be used to implement the method into a program related to heart failure. The use of this new science to the prevention of diseases, especially heart failure, and the promotion of health will have significant and potentially extremely favorable effects.

Table 6. Classification experiment results

Validation	Ratio	RF	RF+GA
Split Validation	0,9	83.33%	100.00%
Split Validation	0,8	81.67%	96.67%
Split Validation	0,7	83.33%	91.11%
Split Validation	0,6	83.19%	89.08%
Split Validation	0,5	81.21%	89.93%
Split Validation	0,9	83.33%	100.00%

Table 7. Result of t-Test

Parameter	RF	RF+GA
Mean	0.82546	6.545138889
Variance	0.000104908	0.00165459
Observations	5	5
Pearson Correlation	0.504165571	
Hypothesized Mean Difference	0	
df	4	
t Stat	-7.150.962.362	
P(T<=t) one-tail	0.001011728	
t Critical one-tail	2.131.846.786	
P(T<=t) two-tail	0.002023455	
t Critical two-tail	2.776.445.105	

Table 8. Comparison of genetic algorithms with other selection features

Rasio split validation	Forward selection	Backward selection	Greedy forward selection	Greedy backward selection	Genetic algorithm
0,9	96.67%	93.33%	93.33%	93.33%	100.00%
0,8	88.33%	88.33%	91.67%	90.00%	96.67%
0,7	88.89%	90.00%	91.11%	86.67%	91.11%
0,6	89.08%	88.24%	89.92%	85.71%	89.08%
0,5	87.92%	87.92%	86.58%	85.91%	89.93%
Average	90.18%	89.56%	90.52%	88.32%	93.36%

Table 9. Comparison of previous research results

No	Method	Accuracy
1	Classification algorithm [3]	83.00%
2	Naïve Bayes + PSO [29]	92.67%
3	Random forest+SMOTE-ENN [9]	90.00%
4	Multi-layer perceptron (MLP)	88.00%
5	Extreme gradient boosting (XGBoost)	83.00%
6	This research (random forest dan algoritma genetika)	93.36%

4. CONCLUSION

This study has succeeded in measuring the accuracy level of the random forest algorithm on the “heart failure clinical record dataset” by applying the genetic algorithm selection feature with an accuracy rate of 93.36% higher than other algorithms. Genetic algorithm applied to feature selection and random forest algorithm to improve the accuracy of the heart failure clinical record dataset. In five testing experiments using split validation with varying ratios, the genetic algorithm proved to be effective in increasing accuracy significantly. Comparison of genetic algorithms with other selection features such as forward selection, backward selection, greedy forward

selection, and greedy backward selection to compare which selection features can increase the accuracy value of the heart failure clinical record dataset. Genetic algorithm is proven to have better performance among other selection features. In this study, genetic algorithms are generally applied for feature selection along with the random forest algorithm which aims to improve the performance of heart failure classification. Several things that can be done to improve this research include using other parameter optimization algorithms or also using other optimization features such as feature weighting or feature generation against other classification algorithms such as neural networks, decision tree, or extreme gradient boosting (XGBoost).

REFERENCES

[1] O. N. Emuoyibofarhe, S. Adebayo, A. Ibitoye, M. O. Ayomide, and A. Taye, "Predictive system for heart disease using a machine learning trained model," *International Journal of Computer (IJC)*, vol. 34, no. 1, pp. 140–152, 2019.

[2] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, 2020, doi: 10.1186/s12911-020-1023-5.

[3] O. O. Oladimeji and O. Oladimeji, "Predicting survival of heart failure patients using classification algorithms," *JITCE (Journal of Information Technology and Computer Engineering)*, vol. 4, no. 02, pp. 90–94, 2020, doi: 10.25077/jitce.4.02.90-94.2020.

[4] J. H. Joloudari *et al*., "Coronary artery disease diagnosis; ranking the significant features using a random trees model," *International Journal of Environmental Research and Public Health*, vol. 17, no. 3, 2020, doi: 10.3390/ijerph17030731.

[5] M. F. Aslan, K. Sabanci, and A. Durdu, "A CNN-based novel solution for determining the survival status of heart failure patients with clinical record data: numeric to image," *Biomed Signal Process Control*, vol. 68, 2021, doi: 10.1016/j.bspc.2021.102716.

[6] M. T. Le, M. T. Vo, L. Mai, and S. V. T. Dao, "Predicting heart failure using deep neural network," in *2020 International Conference on Advanced Technologies for Communications (ATC)*, 2020, pp. 221–225, doi: 10.1109/ATC50776.2020.9255445.

[7] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," *PLoS One*, vol. 12, no. 4, 2017, doi: 10.1371/journal.pone.0174944.

[8] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance analysis of data mining classification techniques to predict diabetes," *Procedia Computer Science*, vol. 82, pp. 115–121, 2016, doi: 10.1016/j.procs.2016.04.016.

[9] A. Ishaq *et al*., "Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

[10] M. M. A. Mary and T. L. A. Beena, "Heart disease prediction using machine learning techniques: A survey," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, vol. 8, no. 10, pp. 441–447, 2020, doi: 10.22214/ijraset.2020.31917.

[11] R. T. Prasetio, A. A. Rismayadi, N. Suryana, and R. Setiady, "Features selection and k-NN parameters optimization based on genetic algorithm for medical datasets classification," *Heart Disease (SPECTF)*, pp. 3080–3086, 2020, doi: 10.5220/0009947130803086.

[12] M. Huljanah, Z. Rustam, S. Utama, and T. Siswantining, "Feature selection using random forest classifier for predicting prostate cancer," *In IOP Conference Series: Materials Science and Engineering*, vol. 546, no. 5, 2019, doi: 10.1088/1757-899X/546/5/052031.

[13] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[14] S. García, J. Luengo, and F. Herrera, "Data preprocessing in data mining," *Part of the book series: Intelligent Systems Reference Library*, Springer Cham, 2015, doi: 10.1007/978-3-319-10247-4.

[15] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," *Big Data Analytics*, vol. 1, no. 1, pp. 1–22, 2016, doi: 10.1186/s41044-016-0014-0.

[16] T. N. Nuklianggraita, A. Adiwijaya, and A. Aditsania, "On the feature selection of microarray data for cancer detection based on random forest classifier," *Jurnal Infotel*, vol. 12, no. 3, pp. 89–96, 2020, doi: 10.20895/infotel.v12i3.485.

[17] M. J. H. Mughal, "Data mining: Web data mining techniques, tools and algorithms: an overview," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 6, 2018, doi: 10.14569/IJACSA.2018.090630.

[18] L. Yingchun and Y. Liu, "Random forest algorithm in big data environment," *Computer Modelling and New Technologies*, vol. 18, no. 12A, pp. 147–151, 2014.

[19] B. Dai, R. C. Chen, S. Z. Zhu, and W. W. Zhang, "Using random forest algorithm for breast cancer diagnosis," *Proceedings - 2018 International Symposium on Computer, Consumer and Control, IS3C 2018*, pp. 449–452, 2019, doi: 10.1109/IS3C.2018.00119.

[20] A. D. Kulkarni and B. Lowe, "Random forest algorithm for land cover classification," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 4, no. 3, pp. 58–63, 2016.

[21] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, USA: Elsevier Science Ltd, 2011.

[22] E. Goel, E. Abhilasha, E. Goel, and E. Abhilasha, "Random forest: A review," *International Journal of Advanced Research in Computer Science*, vol. 7, no. 1, pp. 251–257, 2017, doi: 10.23956/ijarcsse/V7I1/01113.

[23] D. Riana, Y. Ramdhani, R. T. Prasetio, and A. N. Hidayanto, "Improving hierarchical decision approach for single image classification of pap smear," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 6, pp. 5415-5424, 2018, doi: 10.11591/ijece.v8i6.pp5415-5424.

[24] C. B. Gokulnath and S. P. Shantharajah, "An optimized feature selection based on genetic approach and support vector machine for heart disease," *Cluster Computing*, vol. 22, pp. 14777–14787, 2019, doi: 10.1007/s10586-018-2416-4.

[25] A. M. Hay, "The derivation of global estimates from a confusion matrix," *International Journal of Remote Sensing*, vol. 9, no. 8, pp. 1395–1398, 1988, doi: 10.1080/01431168808954945.

[26] F. Gorunescu, *Data Mining: Concepts, models and techniques*, Springer Science and Business Media, vol. 12, 2011.

[27] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management* vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[28] R. T. Prasetio and D. Riana, "A comparison of classification methods in vertebral column disorder with the application of genetic algorithm and bagging," in *2015 4th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME)*, 2015, pp. 163–168.

[29] F. Novaldy and A. Herliana, "Application of PSO to Naïve Bayes for prediction of life expectancy in heart failure patients," *Jurnal Responsif: Riset Sains dan Informatika*, vol. 3, no. 1, pp. 37–43, 2021, doi: 10.51977/jti.v3i1.396.

BIOGRAPHIES OF AUTHORS

Yudi Ramdhani is faculty member and researcher from Adhirajasa Reswara Sanjaya University in Department of Informatic Engineering. Have focused research on data mining, data science and decision support system. He can be contacted at email: yudi@ars.ac.id.

Cakra Mahendra Putra is a student of Department of Information Systems at Adhirajasa Reswara Sanjaya University. He has a high interest in data mining processing, has HTML, PHP, and graphic design skills. He can be contacted at email: cakramahendra70@gmail.com.

Doni Purnama Alamsyah is faculty member from Bina Nusantara University, teaching at Entrepreneurship Department in BINUS Business School Undergraduate Program. He received a Doctorate degree from Padjadjaran University in the field of Management Science. Currently has an interest and research focus on consumer behavior, and is very open to conducting research collaboration. He also can be contacted at email: doni.syah@binus.ac.id.