Implementation of Genetic Algorithm and Adaptive Neuro Fuzzy Inference System in Predicting Survival of Patients with Heart Failure

Dian Alya Korzhakin[1], Endang Sugiharti[*2]

[1,2]Computer Science Department, Faculty of Mathematics and Natural Science,
Universitas Negeri Semarang, Indonesia

Abstract.

Purpose: Heart failure is a disease that is still a global threat and plays a major role as the number one cause of death worldwide. Therefore, accurate predictions are needed to determine the survival of heart failure patients. One technique that can be used to predict a decision is classification. Adaptive Neuro-Fuzzy Inference System (ANFIS) is an algorithm that can be used in the classification process in making predictions. Genetic Algorithms can help improve the performance of classification algorithms through the feature selection process.

Methods/Study design/approach: In this study, predictions or diagnoses were made on the survival of heart failure patients based on the heart failure clinical record dataset obtained from the UCI Machine Learning Repository. The data used is 299 data with 12 attributes and 1 class. The result of this research is the comparison of the accuracy of the ANFIS algorithm before and after using the Genetic Algorithm.

Result/Findings: The ANFIS algorithm produces the highest accuracy of 94.444%. While the ANFIS algorithm after attribute selection using the Genetic Algorithm produces the highest accuracy of 96.667%. This shows that the Genetic Algorithm is able to improve the performance of the ANFIS classification algorithm through the attribute selection process.

Keywords: Prediction, Heart Failure, Adaptive Neuro-Fuzzy Inference System, Genetic Algorithm.

INTRODUCTION

Currently, advances in information technology have brought significant changes in human life. The development of information technology encourages the creation of several methods that can be used to help daily life, one of which is to predict disease. Health care needs to use such patterns for the diagnosis. However, forming a pattern requires large, heterogeneous, and row data, and at the same time, this data is widely distributed. Before extracting, the data must be fully organized and collected first. Then, the collected structured data must be combined to obtain a system that contains useful medical information through which extraction techniques are applied [1].

Machine learning (ML) is a learning machine that focuses on data analysis using various statistical tools to gain more knowledge from data. Data mining is an interdisciplinary field affected by other disciplines, including statistics, ML, and database management [2]. Data Mining or Knowledge Discovery in Databases (KDD) process is used to discover new patterns from large datasets and profoundly impact society by solving real-life problems. Data mining aims to extract useful knowledge and represent the new knowledge to make it understandable [3]. Data mining tools can be very useful to control limitations of people such as subjectivity or error due to fatigue and to provide indications for the decision-making processes. The

---

[*]Corresponding Author

Email addresses: dianalya15@students.unnes.ac.id (Korzakhin), endangsugiharti@mail.unnes.ac.id (Sugiharti)

essence of data mining is to identify relations of patterns and models that provide support for prediction, decision-making processes, and diagnosis [4].

Heart failure is a pathophysiological condition in which there are abnormalities in the heart, resulting in heart failure to pump blood to tissues throughout the body to meet the body's metabolic needs [5]. Heart disease is the first cause of death in different countries and accounts for approximately 80% of all deaths. Based on WHO report, about 12 million deaths per year occur in the world due to heart diseases [6]. Information technology applied in the health sector can effectively predict various diseases from patient medical record data, including survival in heart failure patients [7]. Presently, there is some heart disease prediction system that is based on soft computing paradigms. Most of these models comprise two portions, feature selection (FS) and classification. In FS, the most relevant features of heart disease are selected. Whereas the selected subset features are used as an input in the classification part [8].

The classifier method in this study uses Neuro Fuzzy. Neuro Fuzzy is a method that combines Artificial Neural Networks with Fuzzy Logic [9]. In 1992, J.S.R. Jang developed an adaptive neuro fuzzy inference system based on a fuzzy inference system called the adaptive neuro fuzzy inference system (ANFIS). ANFIS is a fuzzy system combined with a neural network to determine fuzzy sets and fuzzy rules. An adaptive network is a network structure consisting of many interconnected nodes [10]. The advantage of ANFIS is that it can convert knowledge from experts into the form of rules, but it usually takes a long time to define its membership function [11].

To get optimal results in making a prediction or diagnosis, you can use data preprocessing techniques. Preprocessing techniques can help improve data quality and provide more accurate results because data quality can determine the prediction method's performance and the usefulness of the extracted knowledge [12]. One of the processes in the preprocessing technique is feature selection. Feature selection is a process of selecting relevant and informative features that can help reduce the features to improve the prediction accuracy and reduce computation time [13]. Genetic Algorithm (GA) is one of the algorithms that can be used to perform attribute selection. The Genetic Algorithm was chosen because it can reduce attributes in the data without reducing the data information. The genetic algorithm process such as selection, crossover, and mutations to produce the best individuals [14].

Several studies have proved that AG can be used to get more optimal results in improving the performance of classification algorithms to predict disease. Siahaan et al. [15] conducted a study using Genetic Algorithm to select attributes and ANFIS to classify hepatitis. The accuracy value obtained from the ANFIS-AG method is 98.73%, while the accuracy value if only using the ANFIS method is 86.67%. This proves that AG can optimize a method in predicting the attribute selection process.

Based on this, this study uses the method of combining the ANFIS algorithm with GA. The combination of the ANFIS-GA method is expected to produce a maximum level of accuracy in predicting or diagnosing the survival rate of heart failure patients.

METHODS
The method used in this study is ANFIS-GA with ANFIS as a classification algorithm and GA (Genetic Algorithm) as a feature selector to predict the survival of heart failure patients. In this study, there are five main stages to increase the accuracy of ANFIS using genetic algorithms in predicting the survival of heart failure patients. The five stages are the data collection stage, the parameter testing stage, the feature selection stage, the classification stage, and the accuracy calculation stage. This study will determine the comparison of accuracy before and after the application of the Genetic Algorithm to the ANFIS Algorithm. Flow chart The ANFIS Algorithm with the Genetic Algorithm is shown in Figure 1.

Dataset Collection
This study used a public dataset, namely heart failure clinical record from UCI Machine Learning Repository. This data has 12 attributes, including one attribute as class and 299 instances, where x attributes are numeric, and x attributes are nominal. We divide the data into 70% training data and 30% test data. The description of the dataset can be seen in Table 1.
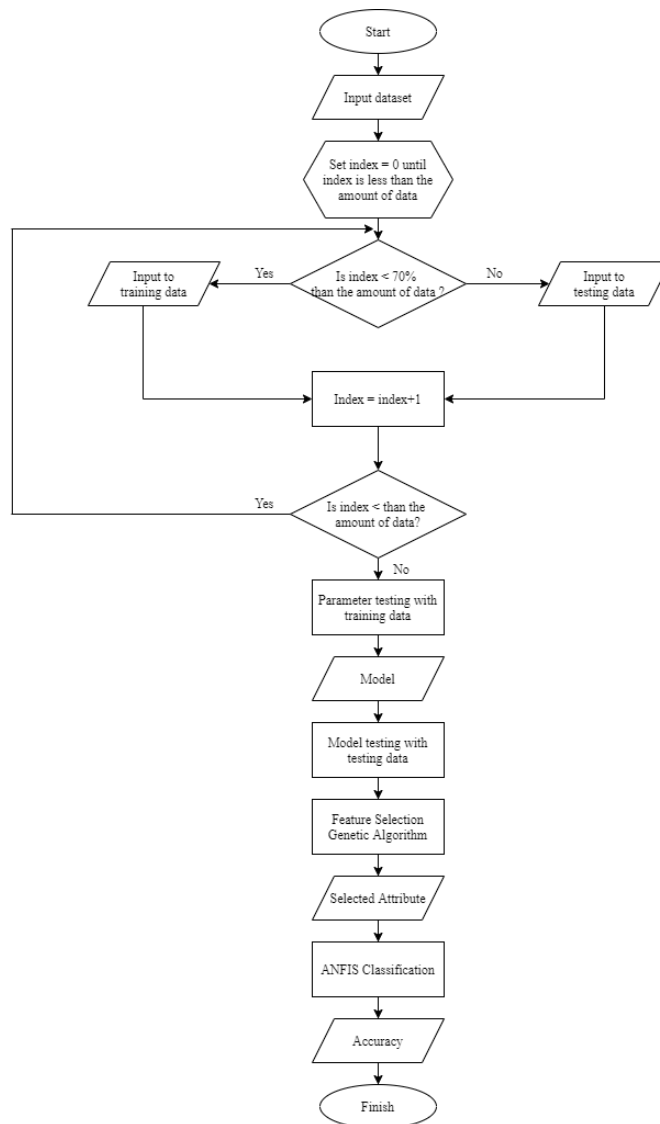
Figure 1. Flowchart ANFIS with Genetic Algorithm

Table 1. Description of Heart Failure Clinical Record Dataset

No	Attributes	Types	Description
1	Age	Numeric	Age of patient
2	Anemia	Boolean	Decrease of red blood cell/hemoglobin
3	High Blood Pressure	Boolean	If patient has hypertension
4	Creatinine phosphokinase	Numeric	Level of CPK enzyme in blood
5	Diabetes	Boolean	If patient has diabetes
6	Ejection fraction	Numeric	Percentage of blood leaving the heart at each contraction
7	Platelets	Numeric	Platelets in blood
8	Serum creatinine	Numeric	Level of serum creatinine in blood
9	Serum sodium	Numeric	Level of serum sodium in blood
10	Sex	Binary	Woman or man
11	Smoking	Boolean	Smokers or not
12	Time	Numeric	Follow-up period
13	Death Time	Boolean	If patient deceased during floow-up period

Parameter Testing

At this stage, the parameters of the Genetic Algorithm are tested. The parameters tested include the pop size value, crossover probability, mutation probability, and the maximum number of generations. Parameter testing is done by using training data as much as 70% of the total data. After testing the parameter values, the best parameter values are selected based on the highest accuracy values. The results of the best-selected parameters can be seen in Table 2.

Table 2. Genetic Algorithm Best Parameters Result

No	Parameter	Optimum Value
1	Popsize	50
2	Crossover Probability	0.5
3	Mutation Probability	0.2
4	Maximum Generation	70

Feature Selection

The stages of the Genetic Algorithm in attribute selection are as follows.
1. Initialize individuals in order to know what type of data will be used as a calculation which will be represented on each chromosome.
2. Calculate the fitness value of each particle in the population.
3. Selection to choose the chromosome with the best fitness value that will be used as prospective parents using the roulette wheel method.
4. Recombination to produce a new chromosome with a better fitness value than the previous chromosome using the one point crossover method.
5. Change the value of genes in a chromosome with a mutation process.
6. Update the value of the old chromosome with the fitness value of the new chromosome.
7. Stop the iteration if the best fitness value or maximum generation is met. If not, return to step 2.

Classification

At this stage, data classification is carried out using the ANFIS algorithm. The ANFIS architecture representation for this system is as shown in Figure 2.
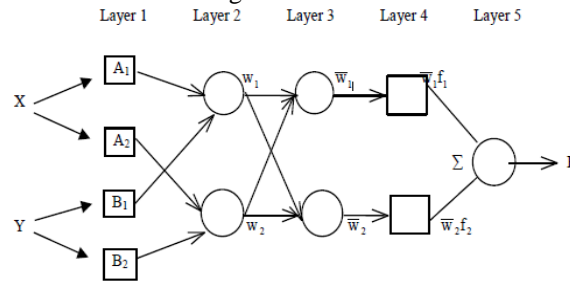

Figure 2. ANFIS Architecture

1) Layer 1

This layer is fuzzification layer. Every node in this layer is an adaptive node with a node function. The output of layer 1 given in equation (1).

$$\mu(A) = \frac{1}{1+\left|\frac{A-c}{a}\right|^{2b}} \tag{1}$$

2) Layer 2

This layer includes the nodes which represents the antecedent part of association rule. The output of each node is given in equation (2) and (3).

$$w_1 = \mu_{A1} \times \mu_{B1} \tag{2}$$
$$w_2 = \mu_{A2} \times \mu_{B2} \tag{3}$$

3) Layer 3

Each neuron in this layer is a fixed neuron which is the result of the calculation of the ratio of the i-th firing power (wi) to the total number of combustion forces in the second layer. The output of the this layer is given in equation (4).

$$\overline{w}_t = \frac{w_i}{w_1 + w_2} , i = 1,2 \tag{4}$$

4) Layer 4

This layer is in the form of neurons which are adaptive neurons to an output. The output of the this layer is given in equation (5).

$$\overline{w}_i f_i = \overline{w}_i(p_i x + q_i y + r_i) \tag{5}$$

pi, qi, ri are the consequent parameters of the neuron. Least Square Estimation (LSE) is a method used to obtain consequent parameter in equation (6).
.

$$\theta_k = (A^T A)^{-1} A^T Z_t \tag{6}$$

5) Layer 5

This layer is a single neuron which is the result of the sum of all the outputs of the fourth layer. The output of the this layer is given in equation (7).

$$\sum_i \overline{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \tag{7}$$

Accuracy Calculation

The last stage is calculating accuracy using a formula like the following.

$$\text{Accuracy} = \frac{correct\ amount\ of\ data}{total\ data\ overall} \ x\ 100\%$$

RESULTS AND DISCUSSION

This study uses a heart failure clinical record dataset which is a public dataset from the UCI Machine Learning Repository, where this dataset has 12 attributes with 1 class attribute, and 299 data samples. This research was conducted using the PHP programming language with the CodeIgniter framework. The result of this study is the accuracy of classification in diagnosing the survival rate of patients with heart failure.

Result

This research is divided into 2 test data. The first is testing the ANFIS classification algorithm. The second is testing the ANFIS classification algorithm, which has gone through the feature selection preprocessing process by the Genetic Algorithm. This testing process uses test data as much as 30% of the total data. The results of the accuracy value of the ANFIS classification algorithm test can be seen in Table 3.

Table 3. Accuracy Value of ANFIS

No	Data	Accuracy Value
1	Training	83.732%
2	Testing	94.444%

The results of testing the test data using the ANFIS algorithm get an accuracy value of 94.444%. The second test is testing the ANFIS classification algorithm with Genetic Algorithm as feature selection. Genetic Algorithm parameters used are parameters that have been obtained from the parameter testing process. The results of the accuracy value of ten executions of the ANFIS-GA algorithm can be seen in Table 4.

Tabel 4. Accuracy Value of ANFIS-GA

Number of Execution	Feature Selected	Fitness Value	Accuracy
1	*Age, creatinine phosphokinase, ejection fraction, platelets, serum sodium, sex, time*	52	96.667%
2	*Age, anemia, high blood pressure, platelets, serum sodium, smocking, time*	53	93.333%
3	*Age, creatinine phosphokinase, diabetes, high blood pressure, sex, smocking, time*	61	92.222%
4	*Age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, smocking, time*	59	96.667%
5	*Age, creatinine phosphokinase, diabetes, ejection fraction, platelets, serum creatinine, serum sodium, smocking, time*	61	95.556%
6	*Age, anaemia, creatinine phosphokinase, diabetes, high blood pressure, platelets, serum creatinine, sex, smocking*	60	93.333%
7	*Age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, sex, time*	63	95.556%
8	*Anaemia, diabetes, ejection fraction, high blood pressure, serum sodium, sex, smocking, time*	58	94.444%
9	*Creatinine phosphokinase, diabetes, high blood pressure, platelets, serum creatinine, serum sodium, sex, smocking, time*	63	93.333%
10	*Age, anaemia, creatinine phosphokinase, ejection fraction, platelets, serum sodium, sex, smocking , time*	66	96.667%

The highest accuracy value from the implementation of ANFIS-GA can reach 96.667%, while the average accuracy value of the ten executions of the ANFIS-GA algorithm is 94.778%.

Discussion

This study applies the Genetic Algorithm in the attribute selection process to obtain optimal accuracy values in the data classification process of patients with heart failure using the ANFIS method. The dataset used in this study is the heart failure clinical records dataset obtained from the UCI Machine Learning Repository. This study aims to determine the workings and accuracy results obtained from the Genetic Algorithm and ANFIS methods in diagnosing survival in heart failure patients. The results of the accuracy of the ANFIS classification process with and without using the Genetic Algorithm are shown in Table 5.

Tabel 5. Accuracy Value

Method	Accuracy
ANFIS	94.444%
ANFIS-GA	94.778%

Based on the accuracy results from Table 5, we can see that there is an increase in accuracy of 0.334% in the average value of ten ANFIS classification executions after using the Genetic Algorithm. The percentage result produced by the ANFIS-AG method is greater than the previous study [16] which using 10 methods to predict the survival of heart failure patients using only 2 attributes, namely Serum Creatinine and Ejection Fraction. The study resulted in an accuracy value of 74% for the Random Forest method, 73.7% for Decision Tree, 73.8% for Gradient Boosting, 73% for Linear Regression, 72.9% for One Rule, 68% for ANN, 69, 6% for Naïve Bayes, 69% for SVM Radial, 68.4% for SVM Linear, and 62.4% for KNN. Another research [17] also conducted with the same dataset but using all attributes in the Random Forest method. This study resulted in an accuracy rate of 85.82% using the SMOTE technique and 94.31% using the resample sampling technique.

The application of the Genetic Algorithm uses random numbers so that the selected attribute is different every time it is executed. However, the attributes that have an effect on increasing the accuracy value are serum sodium, smoking, time, ejection fraction, and high blood pressure.

CONCLUSION

In this study, the application of the ANFIS algorithm was combined with the Genetic algorithm as a selection feature for the diagnosis of survival of heart failure patients in the heart failure clinical record dataset obtained from the UCI Machine Learning Repository. This study resulted in an accuracy value of 94.444% on the application of the ANFIS algorithm without preprocessing. While the results of the application of the ANFIS algorithm with the feature selection preprocessing process using the Genetic Algorithm resulted in an average accuracy value of ten executions of 94.778%. In addition, the results of the application of ANFIS-GA were able to achieve the highest accuracy value of 96.667%. This proves that the Genetic Algorithm can improve the performance of the classification algorithm to produce better data quality and higher accuracy values.

REFERENCES

[1] D. Kinge and S. K. Gaikwad, "Survey on Data Mining Techniques For Disease Prediction," Int. Res. J. Eng. Technol., vol. 9001, no. 01, pp. 630–636, 2008.

[2] A. Faruk and E. S. Cahyono, "Prediction and classification of low birth weight data using machine learning techniques," *Indones. J. Sci. Technol.*, vol. 3, no. 1, p. 18, 2018.

[3] S. Zainudin, D. S. Jasim, and A. Abu Bakar, "Comparative analysis of data mining techniques for Malaysian rainfall prediction," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 6, no. 6, p. 1148, 2016.

[4] B. Milovic and M. Milovic, "Prediction and decision making in health care using data mining," *Kuwait Chapter of the Arabian Journal of Business and Management Review*, vol. 1, no. 12, p. 126, 2012.

[5] A. M. Usman, U. K. Yusof, and S. Naim, "Cuckoo inspired algorithms for feature selection in heart disease prediction," *Int. J. Adv. Intell. Inform.*, vol. 4, no. 2, p. 95, Jul. 2018.

[6] M. W. Bloom *et al.*, "Heart failure with reduced ejection fraction," *Nat. Rev. Dis. Primers*, vol. 3, no. 1, p. 17058, 2017.

[7] H. Karim, S. R. Niakan, and R. Safdari, "Comparison of neural network training algorithms for classification of heart diseases," *IAES Int. J. Artif. Intell. (IJ-AI)*, vol. 7, no. 4, p. 185, 2018.

[8] S. J. Al'Aref *et al.*, "Determinants of In-Hospital Mortality After Percutaneous Coronary Intervention: A Machine Learning Approach," *J. Am. Heart Assoc.*, vol. 8, no. 5, p. e011160, Mar. 2019.

[9] M. N. Fata, R. Arifudin, and B. Prasetiyo, "Optimization Neuro Fuzzy Using Genetic Algorithm For Diagnose Typhoid Fever," *Scientific Journal of Informatics*, vol. 6, no. 1, pp. 1–11, May 2019.

[10] H. Turabieh, "GA-based feature selection with ANFIS approach to breast cancer recurrence," *Int. J. Comput. Sci. Issu.*, vol. 13, no. 1, pp. 36–41, 2016.

[11] Y. Sinatrya and L. A. Wulandhari, "Deteksi diabetes melitus untuk wanita dan penyusunan menu sehat dengan pendekatan adaptive neuro fuzzy inference system (ANFIS) dan algoritma genetika (GA)," *J. tek. inform.*, vol. 12, no. 1, pp. 39–58, 2019.

[12] R. Asgarnezhad, M. Shekofteh, F. Z. Boroujeni, "Improving diagnosis of diabetes melitus using combination of preprocessing techniques," *J. Theor. Appl. Inf. Technol.,* vol. 95, no. 13, pp. 2889-1895, 2017

[13] M. Sharma and S. Mukherjee, "Fuzzy C-Means, ANFIS and Genetic Algorithm for segmenting Astrocytoma –A type of Brain Tumor," *IAES Int. J. Artif. Intell. (IJ-AI)*, vol. 3, no. 1, p. 16, 2014.

[14] K. T. Hidayat, R. Arifudin, and A. Alamsyah, "Genetic Algorithm for relational database optimization in reducing query execution time," *Sci. J. Inform.*, vol. 5, no. 1, p. 27, 2018.

[15] W. F. A. Siahaan, O. S. Sitompul, and Z. Situmorang, "The accuracies of ANFIS and genetic algorithm with tournament selection on classifying hepatitis data," *J. Phys. Conf. Ser.*, vol. 1566, no. 1, p. 012121, Jun. 2020.

[16] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, p. 16, 2020.

[17] A. Newaz, F. S. Haq, and N. Ahmed, "A case study on risk prediction in heart failure patients using random survival forest," in *2021 5th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, 2021.