An Intelligent System for Heart Disease Prediction using Adaptive Neuro-Fuzzy Inference Systems and Genetic Algorithm

Jindong Feng1*, 2, Qian Wang1 and Na Li1
1Northwest Minzu University, Lanzhou, China
2Key Laboratory of Streaming Data Computing and Application, Northwest Minzu University, Lanzhou
*Corresponding author’s e-mail: jindong81@xbmu.edu.cn

Abstract. Cardiovascular disease remains the leading cause of death worldwide over the past two decades. Because of a large number of clinical data and the complexity of the disease, it is often challenging to diagnose and make the proper treatment. Over the past decade, as a soft computing method, fuzzy expert systems have been applied in disease diagnosis by many researches because of its superiority in dealing with uncertain and ambiguous problems. This study proposes an Adaptive Neuro-Fuzzy Inference Systems (ANFIS) to diagnose heart disease. Parameters related to membership functions in ANFIS are optimized by applying the genetic algorithm. The experiment was conducted on the public UCI heart disease datasets. The experimental result shows that 91.25% accuracy was obtained on the testing set, which was found to be satisfying based on comparison.

1. Introduction
According to the World Health Organization, in the past two decades, heart disease is still the leading cause of death worldwide [1]. Nowadays, unhealthy lifestyle, overwork and dietary changes cause an increase in cardiovascular diseases. Since 2000, the number of deaths from heart disease has increased by more than 2 million, to nearly 9 million in 2019 [2]. In most cases of heart disease, ECG cannot detect abnormalities when there is no heart attack, which makes early identification of the heart disease very difficult. In order to obtain more reliable medical diagnosis, various artificial intelligence methods, such as machine learning, fuzzy logic and three-way decisions are applied to the investigation of heart disease. Among these techniques, fuzzy logic theory simulates the reasoning and thinking mode of human brain under uncertain environment. Numerous experiments have shown that fuzzy inference system (FIS) has strong fault tolerance and robustness when dealing with problems in nonlinear or uncertain environments, such as medical diagnosis, financial analysis and industrial control.

In the past decade, numerous scientific researches focus on the application of fuzzy inference system in medical diagnosis [3, 4, 5]. In [3], a decision support system for automatic diagnosis of heart disease was developed based on Mamdani fuzzy logic model, which obtained 68.75% specificity when determines the potential risk of heart disease. Luo and Zhao [4] proposed a novel model to solve medical diagnosis problems by using intuitionistic fuzzy sets theory. The basic idea of this model is to realize pattern recognition by calculating the difference degree between the information carried by intuitive fuzzy sets. Alzheimer's disease (AD) occurs in central nervous degenerative diseases in early
and old age. The etiology is not yet quite clear. Krashenyi et al. [5] developed a FIS to diagnose AD, which processed patients’ MRI images and used means and standard deviations in intensities from most descriptive brain regions as the system inputs.

Once proposed by Jyh-Shing and Roger Jang in 1993 [6], ANFIS has shown excellent advantages in dealing with uncertain problems. ANFIS uses the learning mechanism of neural network to adjust premise parameters and consequence parameters, and can automatically generate If-THEN rules. In [7], Sungging et al. presented medical prognosis using ANFIS to predict lung cancer. Their prediction system used suspected patients’ medical history such as characteristic data and pulmonary x-ray image as input of membership functions. In [8], Ubeyli and Guler invented a novel ANFIS model to diagnose internal carotid artery stenosis and occlusion. Total classification accuracy achieved in their approach is 92.65%, which is better than conventional neural network model. Unfortunately, in the medical domain, many terrible diseases (such as cancer and heart disease) may not be detected until the terminal stage, because there are no signs. For the purpose of assisting clinicians to diagnose esophageal cancer and predict the survival probability of patients at an early stage, Wang et al. [9] invented a new medical prediction system based on ANFIS method.

In this study, an ANFIS model trained by genetic algorithm (GA) is proposed to assist clinicians in diagnosing heart disease. The rest of this paper is organized as follows: The mathematical framework of ANFIS and the implementation method of GA is demonstrated in section 2. Section 3 details the information of dataset and Generation of fuzzy rules. The evaluation metrics and experimental results is given in section 4 and finally the concluding remarks and the future research is presented.

2. Overview of ANFIS and GA

2.1. Mathematical Framework of ANFIS

ANFIS is a suitable intelligent computing technique based on Takagi-Sugeno fuzzy model. Combining the advantages of artificial neural networks and fuzzy logic, ANFIS not only acquires learning and adaptive ability, but also ensures that the weight values obtained can be explainable [10]. The architecture of ANFIS consist of five layers. To better understand the system, a basic ANFIS structure consisting of two inputs and one output is shown in Figure 1. By analysing Figure 1, it is shown that this structure can generate up to four rules that illustrated as follows:

**Rule n:** IF $x$ is $A_i$ and $y$ is $B_j$, THEN $f_n = p_n + q_n + r_n$  
$i, j \in \{1, 2\}$ and $n \in \{1, 2, 3, 4\}$

In the following part, the function of every layer in ANFIS will be explained.

Layer 1: This layer is the input layer, also named as the fuzzification layer. The function of this layer is to obtain fuzzy clustering from input features by using membership function. Each node in this layer represents the values of language variables related to membership functions (such as big and small, high and low). As given in Eq. (1), the output of this layer represents the membership degree of the fuzzy set of input variables transformed into each language variable value.

$$O_i^1 = \mu_{A_i}(x), \quad i = 1, 2$$  
(1)

where $x$ and $y$ are the inputs to node $i$. Here, fuzzification was done through Gaussian membership function as it is comprehensible and appropriate to the problem:

$$O_i^1 = \mu_{A_i}(x) = e^{-\frac{(x-c_i)^2}{2\sigma_i^2}}$$  
(2)

where $\{c_i, \sigma_i\}$ is a parameter set. $C_i$ denote the centres of the curve and $\sigma_i$ determine the Gaussian functions’ width. These parameters are called premise parameters.
Layer 2: The function of this layer is to match the premises of fuzzy rules and calculate the firing strength ($w_i$) of each rule. The output of each node is the product of the input signals:

$$O^2_i = w_i = \mu_{A_i}(x) \cdot \mu_{B_i}(y), \quad i = 1, 2$$

Layer 3: The function of this layer is to implement the normalized calculation. This layer calculates the trigger proportion of rule $i$ in the whole rule base, representing the degree to which this rule is used throughout the reasoning process:

$$O^3_i = \frac{w_i}{w_1 + w_2 + w_3 + w_4}, \quad i \in \{1, 2, 3, 4\}$$

Layer 4: This layer is the defuzzification layer. In this layer, weighted values of rules are calculated in each node:

$$O^4_i = \overline{w_i f_i} = \overline{w_i (p_i x + q_i y + r_i)}$$

where $\overline{w_i}$ is the output of layer 3 and $\{p_i, q_i, r_i\}$ is the consequence parameters set of fuzzy rules.

Layer 5: This layer is the output layer. The overall output of ANFIS is computed by summing all incoming signals:

$$O^5_i = \sum_{i=1}^{4} \overline{w_i f_i} = \frac{\sum_{i=1}^{4} w_i f_i}{\sum_{i=1}^{4} w_i} = \text{overall output}$$

2.2. Training ANFIS based on Genetic algorithm

After the fuzzy logic model of ANFIS is established, the next main task is to optimize the parameters via training algorithms. The adjustable parameters in ANFIS are mainly concentrated in layer 1 and layer 4. The parameters of layer 1 are membership function parameters, also known as premise parameters, and the parameters of layer 4 are consequence parameters of fuzzy rules. Inspired by Jang’s paper in 1993 [6], conventional ANFIS models often adopt LSE-GD hybrid learning algorithm. To be specific, premise parameters are trained with gradient descent (GD), and consequence parameters are trained with least squares estimation (LSE) method. Nevertheless, after a large number of experiments, it is found that these derivative-based optimization algorithms have a risk of local minimum [10]. Recently, there is a trend that heuristic algorithms are widely used in training ANFIS to achieve better performance, such as genetic algorithm (GA), simulated annealing algorithm (SAA), particle swarm optimization (PSO), and Artificial immune system (AIS). In this study, GA is applied to optimize the parameters of ANFIS. GA was proposed by Holland in 1975 [11], which is often applied to search the optimal solution by simulating the process of natural evolution. Firstly, the
algorithm expresses the problem to be solved as chromosome or individual in genetic space by coding process, which was the algorithm's input. And then the algorithm performs genetic processes, such as selection, crossover, and mutation, to search for the global optimum. The flowchart of a basic GA was shown in Fig. 2. The following is the explanation of several crucial steps of GA based on the proposed ANFIS model.

2.2.1. Chromosome Coding.
As mentioned before, ANFIS has two types of parameters that need to be trained, which are premise and consequence parts. In the premise part, the adjustable parameters are the center \((c_i)\) and the standard deviation \((\sigma_i)\) of the Gaussian function that is set as the membership function. Therefore, for the premise part of ANFIS structure with two inputs shown in Figure.1, the chromosome of the GA can be coded as follows:

\[
\begin{align*}
&C_{A1} \quad C_{A2} \quad \sigma_{A1} \quad \sigma_{A2} \\
&C_{B1} \quad C_{B2} \quad \sigma_{B1} \quad \sigma_{B2}
\end{align*}
\]

For the consequence part, according to the equation (5), the adjustable parameters are the coefficients of each polynomial that are used in defuzzification layer. The chromosome code for the consequent part in GA is:

\[
\begin{align*}
&p_1 \quad q_1 \quad r_1 \\
p_2 \quad q_2 \quad r_2 \\
p_3 \quad q_3 \quad r_3 \\
p_4 \quad q_4 \quad r_4
\end{align*}
\]

2.2.2. Fitness Function.
The purpose of the fitness function is to evaluate the quality of the candidate chromosomes (solutions) in the population [12]. In this proposed model, Root Mean Square Error (RMSE) method as Eq. (7) is adopted to calculate the fitness of an individual chromosome. RMSE is a frequently-used method for forecasting technique evaluation. Machine learning models usually minimize RMSE to obtain
appropriate parameter values during the training processes [13]. Therefore, the fitness of the chromosome should be inversely proportional to RMSE.

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (d_i - o_i)^2}
\]  

(7)

where the \( o_i \) is the output obtained by ANFIS and \( d_i \) is the desired output. \( N \) denotes the number of instance samples used in the application.

2.2.3. Selection.
The purpose of selection is to select superior individuals from the population with certain probability. Lots of different selection operations have been introduced in previous literatures, such as stochastic uniform, remainder, roulette and tournament [14]. The model of this study uses roulette wheel selection function, also called proportional selection method, which takes into account the fitness of each individual and a certain probability. In brief, candidate chromosomes with better fitness are more likely to be selected as the parents for the next generation.

2.2.4. Crossover.
Crossover is the process of genetic recombination between two parents by interchanging chromosome segments in relative positions. There are also lots of ways to implement crossover operation. Based on the chromosome code of the model mentioned above, this study adopts the multi-point crossover method. In brief, the algorithm selects two chromosomes as parents, and randomly set multiple crossover points, then carry out gene exchange, so as to get two different offspring chromosomes.

2.2.5. Mutation.
Once the crossover process is completed, there should be a certain probability of gene mutation to improve the diversity of the population. The mutation probability \( p_m \) is set very small, generally speaking, \( p_m \leq 0.05 \) [15]. The chromosome of the proposed model is coded by float number, which is suitable for Uniform Mutation. Uniform Mutation refers to replacing the original value of the chosen gene with random numbers conforming to the uniform distribution within a certain range for that gene. This operation is a two-step process. First, each gene in the individual chromosome code string is designated as the mutation point in turn. In the next step, for each mutation point, a random number is selected uniformly from the value range for that gene by probability \( p_m \) to replace the original gene value.

3. Dataset and Fuzzy Rule Generation

3.1. Dataset Description.
This study used the Cleveland dataset from UCI repository [16] as the experimental dataset, which contains 303 instances with 13 input attributes and one output field which refers to the class label. In the experiment of this study, the dataset is randomly divided into two sub-sets: training (253 instances) and testing (50 instances). Detailed information of the dataset attributes is shown in Table 1.
Table 1. Attributes' information of the Cleveland dataset.

| Attribute | Number of MF | Description |
|-----------|--------------|-------------|
| Age       | 3            | Age (Young, Mid-Aged, Old). |
| Sex       | 2            | Sex (0: Female; 1: Male). |
| CP        | 4            | Chest pain type (4 types). |
| Trestbps  | 4            | Resting blood pressure (Low, Medium, High, Very high). |
| Chol      | 4            | Serum cholesterol (Low, Medium, High, Very high). |
| Fbs       | 2            | Fasting blood sugar is greater than 120 mg/dl or not (0: false; 1: true). |
| Restecg   | 3            | Resting electrocardiographic results (3 types). |
| Thalach   | 3            | Maximum heart rate (Low, Medium, High). |
| Exang     | 2            | Exercise induced angina (0: no; 1: yes). |
| Oldpeak   | 3            | ST depression induced by exercise relative to rest (Low, Risk, Terrible). |
| Slope     | 3            | The slope of the peak exercise ST segment (3 types). |
| Ca        | 4            | Number of major vessels [0-3] colored by fluoroscopy |
| Thal      | 3            | The heart status (3 types) |

It is necessary to explain the column “Number of MF” in the table 1, which refers to the number of membership functions needed for a certain attribute in the fuzzification layer of the ANFIS model. For instance, in the original dataset, trestbps is the integer data in [94, 200], which is fuzzified into four categories as low, medium, high and very high by the calculation of membership function. Therefore, the input “trestbps” needs four nodes in the fuzzification layer.

3.2. Fuzzy Rule Generation
A reasonable fuzzy rule base directly determines the classification performance of a fuzzy inference system. When ANFIS is implemented with MATLAB, the rule base can be generated automatically or added manually. However, according to Table 1, this model has 13 inputs, and each input corresponds to several membership functions, thus the automatically generated rule base will be very large. Therefore, in order to improve the training speed and the accuracy of the model, it is necessary to extract effective rules. Referring to literature [17] and [18], and making a proper analysis of the dataset, 87 rules are adopted in this system. An example of these rules is shown below:

Rule 1: If (Age is Young) and (Trestbps is Low) and (Chol is Low) and (Restecg is Normal) and (Thalach is Low) and (Thal is Normal) and (Sex is Male) and (CP is Asymptomatic) and (Fbs is False) and (Exang is No) and (Oldpeak is Low) and (Slope is upsloping) and (Ca is 0) then (RESULT is healthy).

4. Experimental results
To implement the proposed heart disease diagnosis model, MATLAB version (7.12) is utilized. The work is carried out in windows 10 operation system that has Intel Core i7 processor with speed 2.7 GHz and 8GB RAM.

4.1. Evaluation metrics
In the experiment, three measure metrics are defined to evaluate the proposed model: specificity, sensitivity and accuracy:
Sensitivity = \( TP / (TP + FN) \)
Specificity = \( TN / (TN + FP) \)
Accuracy = \( (TN + TP) / (TN + TP + FN + FP) \)  

(8)

TP means True Positive, FP means False Positive, TN means True Negative and FN means False Negative.

4.2. Results

The experimental result for the training set (253 instances) the testing set (50 instances) achieved by the proposed ANFIS-GA model were shown as Table 2. A remarkable result can be seen that the experiment achieved 91.25% accuracy on the testing set.

|                | Sensitivity (%) | Specificity (%) | Accuracy (%) |
|----------------|----------------|-----------------|--------------|
| Training Set   | 94.35          | 92.78           | 93.27        |
| Testing Set    | 91.54          | 90.32           | 91.25        |

4.3. Performance comparison

The following is the performance comparison of several previous heart disease diagnosis system, which all used the Cleveland heart disease dataset. As shown in Table 3, the experimental results of the system proposed in this study was found to be satisfying based on comparison.

| Model                    | Accuracy |
|--------------------------|----------|
| Hybrid Neural Network [19]| 86.8%    |
| AGAFL [20]               | 90%      |
| ANN-Fuzzy_AHP [21]       | 91.1%    |
| Bagging-Fuzzy-GBDT [22]  | 87%      |
| Proposed Model           | 91.25%   |

5. Conclusion

This study proposes a soft computing method based on ANFIS and genetic algorithm to aid clinicians for early diagnosis of heart disease in patients. The heart disease diagnosis system proposed in this work has the following steps: first, this study uses the UCI Cleveland dataset and randomly divides it into two sub-sets: training (253 instances) and testing (50 instances). Second, the 13 attributes of the dataset are fed as input into ANFIS for fuzzification by the Gaussian function. Last, the premise and consequence parameters of the ANFIS are trained by genetic algorithm to achieve optimal results. The experimental results of the proposed technique achieved accuracy of 91.25% on testing set, which is better when compared with several previous methods that used the same heart disease dataset.

The future work may involve two ideas: first, applying fuzzy inference systems to diagnose more complex diseases, such as depression and Alzheimer's disease. Second, other or hybrid heuristic search algorithms can be applied to train the model parameters to achieve more optimal results.

Acknowledgments

This paper is supported by “the Fundamental Research Funds for the central universities of Northwest Minzu University” (Grant No. 31920160058).
References

[1] NHLBI, 2017. National Heart, Lung, and Blood Institute website. https://www.nhlbi.nih.gov/health/health-topics/topics/hf.

[2] WHO, 2020. World Health Organization, news, WHO reveals leading causes of death and disability worldwide: 2000-2019. https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019

[3] Anooj, P. . (2011) Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules. Central European Journal of Computer Science, 1(4), 482-498.

[4] Luo, M., & Zhao, R. (2018) A distance measure between intuitionistic fuzzy sets and its application in medical diagnosis. Artificial Intelligence in Medicine, 89, 34-39.

[5] Krashenyi, I., Popov, A., Ramirez, J., & Gorriz, J. M. . (2015) Application of fuzzy logic for Alzheimer's disease diagnosis. Signal Processing Symposium (SPSympo). IEEE.

[6] Jang, & J.-S., R. . (1993) Anfis: adaptive-network-based fuzzy inference system. IEEE Trans on Sme, 23(3): 665-685.

[7] Sungging, H.W., Sylvia Ayu, P., Santoso, M.Y., Arifin, S. . (2011) Application of Adaptive Neuro Fuzzy Inference System (ANFIS) for Lung Cancer Detection Software, Nominator TICA Cluster II.

[8] E. Derya Ubeyli, I. Guler. (2005) Adaptive neuro-fuzzy inference systems for analysis of internal carotid arterial doppler signals. Computers in Biology & Medicine, 35(8): 687-702.

[9] Lee, Tsair, Fwu, Wang, Chang-Yu, & Chou, et al. (2015) Predicting survival of individual patients with esophageal cancer by adaptive neuro-fuzzy inference system approach. Applied Soft Computing, 35: 583-590.

[10] Karaboga D, Kaya E. . (2018) Adaptive network based fuzzy inference system (anfis) training approaches: a comprehensive survey. Artificial Intelligence Review, 52(2): 1-31.

[11] Holland JH. (1975) Adaptation in natural and artificial systems. Ann Arbor MI: The University of Michigan Press.

[12] W. F. Mahmudy, R. M. Marian, and L. H. S. Luong. (2013) Hybrid Genetic Algorithms for Multi-Period Part Type Selection and Machine Loading Problems in Flexible Manufacturing System. IEEE Int. Conf. Comput. Intell. Cyberm. Yogyakarta, Indenes, pp. 126–130.

[13] Cárdenas, J. J., García, A., Romeral, J. L., & Kampouropoulos, K. (2011) Evolutive ANFIS training for energy load profile forecast for an IEMS in an automated factory. In ETFA2011 (pp. 1-8). IEEE.

[14] Goldberg, D. E. , & Deb, K. (1991) A comparative analysis of selection schemes used in genetic algorithms. Foundations of Genetic Algorithms, Vol. 1, pp. 69-93.

[15] Wahyuni, I., Mahmudy, W. F., & Iriany, A. (2017) Rainfall prediction using hybrid adaptive neuro fuzzy inference system (ANFIS) and genetic algorithm. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 9(2-8), 51-56.

[16] Robert Detrano & M.D & PhD, V.A. Medical Center, Long Each and Cleveland Clinic Foundation. Available: www.archive.ics.uc.edu/ml/datasets/Heart+Disease

[17] Adeli, A., & Neshat, M. (2010) A fuzzy expert system for heart disease diagnosis. In Proceedings of international multi conference of engineers and computer scientists, Hong Kong. Vol. 1, pp. 28-30.

[18] Zabeen, A., Utsav, A., & Lal, K. (2018) Detection of Heart Disease Applying Fuzzy Logics and Its Comparison with Neural Networks. In 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). IEEE, pp. 461-467.

[19] Kahramanli, H., & Allahverdi, N. (2008) Design of a hybrid system for the diabetes and heart diseases. Expert systems with applications, 35(1-2), 82-89.

[20] Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Rajput, D. S., Kaluri, R., & Srivastava, G. (2020) Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. Evolutionary Intelligence, 13(2), 185-196.
[21] Samuel, O. W., Asogbon, G. M., Sangaiah, A. K., Fang, P., & Li, G. (2017) An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. Expert Systems with Applications, 68, 163-172.

[22] Yuan, X., Wang, X., Han, J., Liu, J., Chen, H., Zhang, K., & Ye, Q. (2019) A High Accuracy Integrated Bagging-Fuzzy-GBDT Prediction Algorithm for Heart Disease Diagnosis. In 2019 IEEE/CIC International Conference on Communications in China (ICCC). IEEE. pp. 467-471.