Classification of gene expressions of lung cancer and colon tumor using Adaptive-Network-Based Fuzzy Inference System (ANFIS) with Ant Colony Optimization (ACO) as the feature selection

S Zainuddin¹, F Nhita² and U N Wisesty³
School of Computing, Telkom University, Bandung, Indonesia

E-mail:¹sulhijas@student.telkomuniversity.ac.id,²fhiranhita@telkomuniversity.ac.id,³untarinw@telkomuniversity.ac.id

Abstract. Cancer is one of the death causes in most countries. In 2015 the death count caused by cancer is reaching 8,8 million and in 2030 it is estimated that the death count reaches 13 million. Therefore, in this research conducted an expression classification of gene using Adaptive-Network-based Fuzzy Inference System (ANFIS) with Ant Colony Optimization (ACO) as the feature selection can help the process of early diagnosis to reduce mortality. The data used are colon tumor and lung cancer obtained from Kent Ridge Biomedical Data Set Repository. Accuracy results obtained are influenced by several factors such as data partition method, the number of ants, and the number of gene attributes. The best accuracy results obtained for colon tumor is 94,73% and lung data cancer is 100%.

1. Introduction
Cancer is a disease that occurs due to less controlled normal cells that divide and be able to attack other tissues that produce cancer cells. Cancer is currently one of the leading causes of death in most countries in the world [1]. In 2015 the death count caused by cancer is reaching 8,8 million and 70% occurred in low and average income countries [2]. In 2030 it is estimated that the death count by the cause of cancer reaches 13 million because of the growth of population and aging society [3]. Therefore, early diagnosis is needed to give the right treatment to patient to decrease the death risk [4].

In terms of diagnosis, it needs data taken from gene expressions in human body. Gene expressions data is high dimension data which has many attributes. Data that has many attributes, can spend a lot of time in diagnosis. Therefore, dimension decrease method is needed to choose the optimal attribute to help classification process and optimize the diagnosis result.

In the related research of classification microarray data has been done, among others ANFIS-GA [5], KNN-GA [5], SVM-GA [5], CART-GA [5], ANFIS-PSO [5], KNN-PSO [5], SVM-PSO [5], CART-PSO [5], SVM-ACO [6], Levenberg-Marquardt Backpropagation (LMBP)-Principal Component Analysis (PCA) [7], SVM-PCA [7], Random Forest-K-Means [8], and much more.

In the research refers to reference [5], it compared the four classification algorithms and two feature selection algorithms that have the best accuracy against the two data sets. Classification Algorithm used are Adaptive-Network-based Fuzzy Inference System (ANFIS), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Classification and Regression Trees (CART), while feature selection
algorithms are Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). Dataset used among others are colon tumor and lung cancer. The average accuracy result obtained from colon tumor data using GA as a feature selection for ANFIS is 100%, KNN about 98.41%, SVM about 98.41%, and CART about 98.41%. The average accuracy result obtained for colon tumor data by using PSO as a feature selection for ANFIS is about 96.83%, KNN about 95.24%, SVM about 93.65%, and CART about 96.83%. The average accuracy result obtained for lung cancer data using GA as a feature selection for ANFIS is 100%, KNN about 99.55%, SVM about 100%, and CART about 99.55%. The average accuracy result obtained for lung cancer data by using PSO as a feature selection for ANFIS is about 99.33%, KNN about 99.78%, SVM about 99.33%, and CART about 99.55%. From these results, it can be seen that ANFIS algorithm gives better results using feature selection.

In the research refer to reference [6] SVM algorithm was used as process of classification and ACO as a feature selection to five datasets, one of them is colon tumor. The accuracy of the results obtained is 91.5%. The process of feature selection using ACO delivers better results because it can improve the accuracy results from 90.3% to 91.5%.

Based on the results of the previously conducted researches, this research used ACO algorithm as a feature selection that is expected to increase accuracy in the classification process of ANFIS. Previously research, ACO and ANFIS algorithm has been used in the classification of breast cancer using data mammograms view [9], but for classification of gene expression the data has not yet been done. Therefore, the data being used in this study is the data for colon tumor and lung cancer. It is expected that this classification model can help the medical world.

2. Related work

In the research refers to reference [5], it compared the four classification algorithms and two feature selection algorithms that have the best accuracy against the two data sets. Classification algorithms used are Adaptive-Network-based Fuzzy Inference System (ANFIS), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Classification and Regression Trees (CART), while the feature selection algorithm are Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). Dataset used are colon tumor and lung cancer. The accuracy results can be seen in table 1.

| No | Dataset Name | Feature Selection Algorithms | Classification Algorithm | Accuracy Result |
|----|--------------|-------------------------------|--------------------------|----------------|
| 1  | Colon Tumor  | GA                            | ANFIS                    | 100%           |
| 2  |              |                               | KNN                      | 98.41%         |
| 3  |              |                               | SVM                      | 98.41%         |
| 4  |              |                               | CART                     | 98.41%         |
| 5  |              | PSO                           | ANFIS                    | 96.83%         |
| 6  |              |                               | KNN                      | 95.24%         |
| 7  |              |                               | SVM                      | 93.65%         |
| 8  |              |                               | CART                     | 96.83%         |
| 9  | Lung Cancer  | GA                            | ANFIS                    | 100%           |
| 10 |              |                               | KNN                      | 99.55%         |
| 11 |              |                               | SVM                      | 100%           |
| 12 |              |                               | CART                     | 99.55%         |
| 13 |              | PSO                           | ANFIS                    | 99.33%         |
| 14 |              |                               | KNN                      | 99.78%         |
| 15 |              |                               | SVM                      | 99.33%         |
| 16 |              |                               | CART                     | 99.55%         |
In the research refers to reference [6], the problem describing ACO algorithms to select gene as optimal with Support Vector Machine (SVM) used as a classification of five expression gene datasets, which are colon, leukemia, DLBCL, NC160, and brain. The accuracy results can be seen in table 2.

| No. | Dataset Name | SVM  | ACO-SVM |
|-----|--------------|------|---------|
| 1   | Colon        | 90.3% | 91.5% ± 1.5% |
| 2   | Leukemia     | 94.1% | 100%     |
| 3   | DLBCL        | -     | 100%     |
| 4   | NC160        | -     | 82.4% ± 1.9% |
| 5   | Brain        | -     | 90.7% ± 1.9% |

3. Research scheme
In this research, a system design to classify the gene expression of lung cancer data and colon tumor using ANFIS with ACO as the feature selection. The design can be seen in figure 1.

3.1. Dataset
In this research, the dataset using is gene expression data of cancer consisting of two datasets, consist of lung cancer and colon tumor. These datasets are taken from Kent Ridge Biomedical Data Set Repository [10]. The datasets can be seen in table 3.
Table 3. Expression gene dataset.

| NO | Dataset Name | Attribute | Samples | Class | Record |
|----|--------------|-----------|---------|-------|--------|
| 1  | Lung Cancer  | 12533     | 32      | Mesothelioma | 16 |
|    |              |           |         | ADCA   | 16 |
| 2  | Colon tumor  | 2000      | 62      | Positive | 22 |
|    |              |           |         | Negative | 40 |

3.2. Data normalization
This step is data normalization process. Normalization is a method to change the data value into certain scale. In this research the scale using [0,1] which is written in the equation (1) [4].

\[ X_{new} = \frac{x-x_{min}}{x_{max}-x_{min}} \]  

Where \( X_{new} \) is the result from certain range, \( x \) is the data before, \( x_{min} \) is the smallest data, and \( x_{max} \) is the maximum data.

3.3. Data partition
At this stage, data partition using 2 methods known as percentage split and K-fold Cross Validation. Data partition using percentage split, the dataset will be split into 2 parts, training data (70%), and testing data (30%). Data partition using K-fold Cross Validation with fold value equals 5. Data partition using the K-fold Cross Validation is done manually. Training data used to build a model from ANFIS algorithm, even testing data used to test the model that has been built.

3.4. Learning Adaptive-Network-based Fuzzy Inference System (ANFIS) and Ant Colony Optimization (ACO)
At this stage, there will be learning process of ANFIS and ACO. To overview the learning process, it can be seen in figure 2.

3.4.1. Feature Selection using Ant Colony Optimization (ACO)
At this stage, the feature selection is processed by using Ant Colony Optimization (ACO). ACO is one of artificial intelligence algorithms implements the behavior of ant colonies finding the food. Colonies of ants are able to get the shortest path by using their chemicals, called pheromone. Pheromones are useful to give signal to other ants along the path that has been passed [11,12].

When finding for food, for the first ants will go around the area of their nest randomly. After achieving food, the ants will bring portions of food to their lair and leave the pheromone on their way to home to signal the other ants to find food sources. The more food they eat, the stronger pheromone left in the path so that other pathways will further loses pheromone traces [11,12]. An illustration of old foraging ant colonies can be seen in figure 3.

Here are the stages of the feature selection process by using ACO. Firstly, insert the data that has been divided into training and testing data. The second is, defining the parameters, which are the number of ants, the number of iterations, the value of pheromones, and others. The third stage is, randomly choosing the gene attribute to the training and testing data for each ant by entering an integer value according to the attribute. Examples of chromosome representation on ants can be seen in figure 4.

The next stage, evaluating the ants using ANFIS. The accuracy result of each ant that has been obtained, will be used for the next stage. The best ants will be stored on a variable and compared with the previous ant results.

Next, the pheromones value is updated. Before proceeding to the next iteration, another random selection of attributes that has not been selected by the ant. Next, it will iterates until the maximum predefined iteration and resulting the best estimate accuracy.
Figure 2. Learning ANFIS and ACO.

Figure 3. Illustration of ants colony in search of food [13].
3.4.2. Classification using Adaptive-Network-based Fuzzy Inference System (ANFIS)

At this stage, the classification process used Adaptive-Network-based Fuzzy Inference System (ANFIS). Adaptive-Network-based Fuzzy Inference System (ANFIS) is a combination of Artificial Neural Networks (ANN) and Fuzzy Inference System (FIS) [14].

ANN is a network that mimics the workings of human neural networks in practice something. But this network is more simple than the real human neural network. ANN has willingness in learning process (learning), it caused ANN has a large network structure and spread in parallel, so it can solve large-scale problems [14].

Fuzzy Inference System (FIS) is a rule of fuzzy systems consists of three main processes which are, Fuzzification, Inference, and Defuzzification [14]. The Fuzzy System process can be seen in figure 5.

![Fuzzy Inference System](image)

**Figure 5.** Fuzzy Inference System [14].

Fuzzification serves to change the input whose true value is certain (crisp input) into a fuzzy form input using membership functions. Inference works to reasoning for fuzzy values input by using fuzzy route that has been defined so as to produce fuzzy output. Defuzzification serves to change fuzzy output into crisp value in the form of a real number [14].

To make it easier to explain, ANFIS architecture firstly can be seen in figure 5. In ANFIS, the model used is sugeno model. Here is the rule of sugeno model’s first order [14]:

IF \( x_1 \) is \( Y_1 \) AND \( x_2 \) is \( Y_3 \) THEN \( Z_1 = p_1 x_1 + q_1 x_2 + r_1 \)

IF \( x_1 \) is \( Y_2 \) AND \( x_2 \) is \( Y_4 \) THEN \( Z_2 = p_2 x_1 + q_2 x_2 + r_2 \)

In ANFIS has 5 layers, to further clarify the function and equation of each layer, here is the following explanation [14,15].

![ANFIS Architecture](image)

**Figure 6.** ANFIS architecture.
1. Layer 1:
Each node i on this layer is an adaptive node that generates linguistic membership values. This layer serves as a process fuzzification, x is the input and O_{1,i} is the output. The output of this layer is written in equation (2).
\[ O_{1,i} = \mu_{\gamma_i}(x_i) \text{ for } i = 1,2,...,n \text{ dan } j = 1,2,...,m \]  
(2)
Where \( \mu_{\gamma_i} \) & \( \mu_{\gamma_i} \) is a membership function of each node. In this research the membership function used is Gaussian [16] which written in equation (3).
\[ \mu_{\gamma}(x) = e^{\frac{(x-\mu)^2}{2\sigma^2}} \]  
(3)

2. Layer 2:
This layer is called the production process denoted by \( \Phi(x) \). Each node in this layer serves to calculate the activation power of each rule \( w_i \) as the products of all inputs. The output of this layer is written in equation (4).
\[ O_{2,i} = \prod_{j=1}^{m} \mu_{\gamma}(x_j) = w_i, \text{ i = 1,2,...,n} \]  
(4)

3. Layer 3:
This layer is called the normalization process which is denoted by \( N \). Each node in this layer is not adaptive. The output of this layer is the ratio calculation results between \( w_i \) to the total activation power of all the rules written in equation (5).
\[ O_{3,i} = \frac{w_i}{\sum_{j=1}^{m} w_j}, \text{ i = 1,2,...,n} \]  
(5)

4. Layer 4:
Each node i this layer is adaptive, which is described in equation (6).
\[ O_{4,i} = \frac{w_i Z_i}{\sum_{i=1}^{n} w_i Z_i} = \frac{w_i(p_i a + q_i b + r_i)}{\sum_{i=1}^{n} w_i} \]  
(6)
Where \( (p_i a + q_i b + r_i) \) is a function of the first order suggeno fuzzy model \( (Z_i) \) dan \( \{p_i, q_i, r_i\} \) is a consequent parameter.

5. Layer 5:
In this layer is called the model output layer denoted by \( \sum \), this layer serves to sum all incoming signals from layer 4 and produces a single output written in equation (7).
\[ O_{4,i} = \sum_{i=1}^{n} \frac{w_i Z_i}{\sum_{i=1}^{n} w_i} \]  
(7)
At this stage, the classification process begins with retrieving training and testing data that has been selected by the ACO algorithm. The second stage defines the parameters, including membership functions, split range, and others. The third stage is making a model using training data. In the fourth stage, after obtaining the model, data testing is entered to test the model that has been made. Finally, calculate the value of accuracy with equation (8).
\[ \text{Accuracy} = \frac{\text{Number of correct prediction}}{\text{Number of records}} \times 100\% \]  
(8)

4. Analysis and experimental results
At this stage, there will be a system testing by using partition data method and deciding parameters that can affect the research result. In ACO, the number of ant and the number of selected gene attributes will be the test parameters and ANFIS parameters have already been decided to use Gaussian membership function. This test takes 5 times observations.

4.1. Testing influence of number of ants with data partition percentage split method
This test is used to determine the effects of the number of ants to the result of accuracy. There is three category of the number of ants used, that are 5 ants, 10 ants, and 15 ants for both dataset. As for the value of other variables has been already determined. Number of gene attributes that are selected for colon tumor data as much as 5 and lung cancer data as much as 100. The result of the testing accuracy of the number of ants can be seen in figure 7 and figure 8.
Figure 7. The influence of number of ants against the accuracy results with data partition percentage split on colon tumor data.

Figure 8. The influence of number of ants against the accuracy results with data partition percentage split on lung cancer data.

Based on figure 7, the accuracy results obtained for colon tumor data differ. The best accuracy result is 89.47% with the number of ants is 10, though the best average accuracy results with the number of ants is 15. It can be concluded that the number of ants on the colon tumor data affects the accuracy result. The more the number of ants fed the better the accuracy result obtained. This is because when the number of ants increases then the pheromone traces are increasing on the path, and makes the pheromone stronger.

Based on figure 8, the accuracy results obtained for the lung cancer data are the same for the lung cancer data with a different number of ants. It can be concluded that the number of ants for lung cancer data does not affect the accuracy result. The accuracy results obtained is 100%.

4.2. Testing influence of number of selected gene attributes with data partition percentage split method. This test is used to determine the effect of the number of selected gene attributes in each set against the accuracy result. Number of selected gene attributes for colon tumor data is 5, 10, and 15. Number of selected gene attributes for lung cancer data is 100, 150, and 200. As for the value of other variables already determined. The number of ants as much as 5 for both data. The accuracy result of the tested number of selected gene attributes can be seen in figure 9 and figure 10.

Based on figure 9, the accuracy result obtained differs for colon tumor data. The best accuracy is about 94.73% with the number of gene attribute is 10. Though the best average accuracy result with the number of gene attribute is 15. It can be concluded that the number of selected gene attributes on colon
tumor data affected the accuracy result. The more the number of gene attribute used, the better the accuracy obtained. This is because in each of the gene attribute has a correlation one after another.

Based on figure 10, the accuracy result obtained is the same for the lung cancer data with a different number of gene attribute, it can be seen that the number of gene attribute does not affect the accuracy result. The accuracy results obtained is 100%.

4.3. Testing influence of number of ants with data partition k-fold cross validation method
This test is used to determine the effect of number of ants with accuracy result. The number of ants used are split into three categories, which are 5 ants, 10 ants, and 15 ants for both dataset. While for the other parameter value has already been determined. The number of selected gene attributes for colon tumor data are about 5 and the lung cancer data are about 100. The accuracy result of number of ants tested can be seen on figure 11 and figure 12.

Based on figure 11, the accuracy results obtained differs for colon tumor data. The best accuracy results is about 92.05% with a number of 10 ants, while the best average accuracy result are with a number of 15 ants. It can be concluded that the number of ants in colon tumor data affecting the accuracy result. The more the number of ants eat, the better the result obtained. This is because when the number of ants increasing, then the pheromone traces left in the path is increasing, so the pheromone traces getting stronger.
Based on figure 12, the accuracy results obtained differs for lung cancer data. The best accuracy result is 100% with the number of ants 5 and 15. The best average accuracy result with the number of ants 5 and 15. It can be concluded that the number of ants affects the accuracy result for lung cancer data.

**Figure 11.** The influence of number of ants against the accuracy results with data partition k-fold cross validation on colon tumor data.

**Figure 12.** The influence of number of ants against the accuracy results with data partition k-fold cross validation on lung cancer data.

4.4. **Testing influence of number of selected gene attributes with data partition k-fold cross validation.** This test is used to determine the effect of the number of selected gene attributes in each dataset against the result of accuracy. The number of selected gene attributes for colon tumor data is 5, 10, and 15. The number of selected gene attributes for lung cancer data is 100, 150, and 200. Whereas for the other parameter values have been determined. The number of ants is 5 for both data. The results of testing accuracy of the number of selected gene attribute can be seen in figure 13 and 14.

Based on figure 13, the accuracy results obtained differ for colon tumor data. The best accuracy result is 91.79% with the number of gene attributes 15. It can be concluded that the number of selected gene attributes in colon tumor data affects the accuracy result. The more number of gene attributes, then better the accuracy result. This is because each gene attribute had a correlation each other.

Based on figure 14, the results of accuracy differ for lung cancer data. The best accuracy results are 100% with the number of gene attributes 150 and 200. The results of the best average accuracy with the number of gene attributes are 150 and 200. It can be concluded that the number of gene attributes also affects the accuracy for lung cancer data.
Figure 13. The influence of number of selected gene attributes against the accuracy results with data partition k-fold cross validation on colon tumor data.

Figure 14. The influence of number of selected gene attributes against the accuracy results with data partition k-fold cross validation on lung cancer data.

Based on four experimental scheme that has been done, the best model for colon tumor data obtained by using a percentage split method, the number of ants is 15, and the number of genes is 10 with accuracy results is 94.73%. For lung cancer, the best model is obtained by using all the methods and parameters with accuracy results is 100%. However, for data partition method, percentage split gets better results than with k-fold cross-validation can be seen in figure 14. Because, k-fold cross-validation, produces overfitting effect against the accuracy results.

5. Conclusion

Based on result of study and test analysis using Ant Colony Optimization (ACO) as the feature selection and Adaptive-Network-based Fuzzy Inference System (ANFIS) as classification, it can be concluded that the system has been built is very good in classification for colon tumor data also lung cancer with the best accuracy reaches 94.73% for colon tumor data and 100% for lung cancer data. The Accuracy result affected partition data method and several ACO's parameters. Partition data used is percentage split and k-fold cross-validation. ACO's parameters that give effects are the number of ant and the number of gen attribute. Percentage split method with the number of ant is 15 and the number of gen is 15 attributes are the best model for colon tumor data and for lung cancer data. All the methods and parameters can result the best model.
References

[1] Wisesty U N, Warastri R S and Puspitasari S Y 2018 March leukemia and colon tumor detection based on microarray data classification using momentum backpropagation and genetic algorithm as a feature selection method 2018 Journal of Physics: Conference Series 971 no 1 p 012018 IOP Publishing

[2] World Health Organization 2018 Available Online: http://www.who.int/news-room/factsheets/detail/cancer Accessed: 03-09-2018

[3] American Cancer Society 2012 Available Online: https://www.cancer.org/research/cancer-facts-statistics/global.html. Accessed: 03-09-2018

[4] Ubaiddillah S H S A, Sallehuddin R and Ali N A 2013 Cancer detection using artificial neural network and support vector machine: A comparative study 2013 J. Teknologi 65(1) pp 73-81

[5] Mahmoudi S, Lahijan B S and Kanani H R 2013 ANFIS-based wrapper model gene selection for cancer classification on microarray gene expression data In Fuzzy Systems (IFSC) 2013 13th Iranian Conference on IEEE pp. 1-6

[6] Yu H, Gu G, Liu H, Shen J and Zhao J 2009 A modified ant colony optimization algorithm for tumor marker gene selection 2009 Genomics, proteomics & bioinformatics 7(4) pp 200-208

[7] Adiwijaya, Wisesty U N, Lisnawati E, Aditsania A, Kusumo D S 2018 Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification Journal of Computer Science 14(11).

[8] Ayadenta H and Adiwijaya 2018 A Clustering Approach for Feature Selection in Microarray Data Classification Using Random Forest Journal of Information Processing Systems 14(5).

[9] Thangavel K and Mohideen A K 2016 December mammogram classification using ANFIS with ant colony optimization based learning 2016 Annual Convention of the Computer Society of India (pp. 141-152) Springer Singapore

[10] A S Cano, A Masegosa and Moral ELVIRA Biomedical Data Set Repository 2005 Available Online: http://leo.ugr.es/elvira/DBCRepository/ Accessed: 05-Nov-2017

[11] Suyanto 2017 Swarm Intelligence Komputasi Modern Untuk Optimasi dan Big Data Mining (Bandung: Informatika Bandung)

[12] Purnomo H D 2014 Cara Mudah Belajar Metode Optimisasi Metaheuristik Menggunakan Matlab (Yogyakarta: Penerbit Gava Media)

[13] Fahrudin T M, Syarif I and Barakbah A R 2016 September Ant colony algorithm for feature selection on microarray data sets Electronics Symp. (IES) 2016 Int. (pp 351-356) IEEE

[14] Suyanto 2008 Soft Computing Membangun Mesin Ber-IQ Tinggi (Bandung: Informatika Bandung)

[15] Bahekar L, Mishal A, Bisen M, Koche D and Alone S 2017 Heart Valve Diseases Detection Using Anfis And Wavelet Transform Heart 3(2)

[16] Fatkurrozi B, Muslim M A and Santos D R 2012 Penggunaan Artificial Neuro Fuzzy Inference Sistem (ANFIS) dalam Penentuan Status Aktivitas Gunung Merapi Jurnal EECCIS 6(2) pp 113-118.