Classification and identification the most important features of cervical cancer based on the expression of microRNA gene with the random forest (RF) algorithm

E A Aziz, A Wibowo, P W Wirawan
Department of Informatics, Faculty of Science and Mathematics, Diponegoro University Jl. Prof Soedarto, SH, Ngesrep, Semarang 50275, Indonesia
E-mail: ekoabdulaziz@student.undip.ac.id

Abstract. Cervical cancer is the leading cause of death women in the world and number one in Indonesia. An effort that can be done for this case is early detection, for example, an IVA test (visual inspection test with acetic acid). However, the IVA test is not able to indicate patients who have potential cancer before cancer's physical characteristics are seen. Thus a new solution is needed for early detection of cervical cancer that can indicate patients who have potential cancer before cancer's physical characteristics are seen. In recent years, various types of miRNA that play a role in cancer malignancies have been identified and can be used as non-invasive biomarkers for cancer diagnosis and monitoring. The use of classification based on miRNA gene expression is a solution for early detection, but the use of high accuracy classification algorithms is something that must be considered. Random Forest (RF) algorithm is the solution to these problems because better generalization performance and is less susceptible to overfitting. In this study also identified important features that are very influential in the classification process. The results showed that the Random Forest algorithm was able to have 100% accuracy for classification and most important features supporting the cancer were miR-549c-5p, miR-183 and miR-515-5p.

1. Introduction
Cancer is a disease due to abnormal growth of cells of the body's tissues, and if allowed to develop, the cancer cells can spread to other parts of the body causing death [1]. Several types of cancer can be experienced by humans, for example cervical cancer. Prof. DR. dr. Aru Wicaksono as Chair of the Indonesian Cancer Foundation (YKI) explained that cervical cancer is the number one female killer disease in Indonesia because every year there are no less than 15,000 cases in Indonesia [2]. Obstetrician and gynecologist Dr. Laila Nuranna SpOG explained that the need for early detection such as IVA test (visual inspection test with acetic acid) because most patients with cervical cancer just realized that he had cancer when his condition was severe and this resulted in the treatment is less likely to recover [3]. However, with manual IVA testing indirectly shows the late in early detection because the physical characteristics of cancer have been seen. Thus a new solution is needed for early detection of cervical cancer that can indicate patients who have potential cancer before cancer's physical characteristics are seen. In recent years, various types of miRNA that have a role in cancer malignancy have been identified, so can be used as a non-invasive biomarker for cancer diagnosis and monitoring [4]. The use of classification based on miRNA gene expression is a solution that can solve the problem, but the use of high accuracy classification algorithms is something that must be considered. Also, not all miRNA gene expression can be used as important references/features that are very influential in
classification. In other words, it is necessary to identify the most important feature in cervical cancer classification based on gene expression of microRNA.

Various studies related to classification with Back Propagation Neural Network (BPNN) algorithm, Support Vector Machine (SVM), Neural Network (NN) and Decision Trees (DC) have been done quite a lot. In the study [5], classification of breast cancer with Support Vector Machine (SVM) algorithm. In the study [6], the classification for early prediction of heart disease with Back Propagation Neural Network (BPNN) algorithm. In the study [7], the classification of the handling of interference with the 20 KV electricity distribution network with the Neural Network (NN) algorithm. In the study [8], classification of breast cancer malignancy level with SVM algorithm. In the study [9], classification of diet nutrient selection for hypertensive patients with Decision Trees (DC) algorithm. The need for high accuracy in classification algorithms is needed here. Based on research [10] stated that the average Correct Rate (CR) in Back Propagation Neural Network (BPNN), Support Vector Machine (SVM) and Random Forest (RF) were 86.68%, 66.45%, and 99.07%. In the study [11] Random Forest (RF) has proven to outperform Neural Network (NN) and Support Vector Machine (SVM) regarding accuracy and stability. The research [12] after analyzing the classification performance with Random Forest (RF) algorithm is better than the classification with the J48 Decision Trees (DC) algorithm. Referring to the research [10], [11] and [12], the researcher proposes the Random Forest (RF) algorithm to be used in this case classification.

The use of the Random Forest (RF) algorithm for classification by utilizing microRNA gene expression has also been carried out in research [13]. In the study [13] it was stated that the classification with the Random Forest (RF) algorithm resulted in an accuracy of 93.48% for gastric cancer and 100% for ovarian cancer.

In this study, the classification of cervical cancer will be carried out based on the expression of the microRNA gene with the Random Forest Algorithm. Then the results of the classification will be used to identify important features that are most affected in the classification process. It is expected that knowing this important feature will facilitate researchers regarding classification.

2. Literature and Method Review

The tool used in this study is a software called Rapid Miner version 8.0. This research was conducted using the Auto Model feature on Rapid Miner. The following steps will be taken:

2.1 Select Data

The first step is to select data that has been stored in the Rapid Miners repository. If the data is not in the Rapid Miners repository, then import the data first.

2.2 Select Task

The second step is to choose what type of problem that want to solve. There are 3 types of choices, namely: prediction, clustering, and outliers. Predictions are selected if we want to predict a value for one of the selected columns. Clustering is chosen if we want to group data into clusters whose purpose is to find a set of data points that are close together. Outliers are selected if you want to find individual data points that are far from all other data points, perhaps because of errors in data collection or because of strange or unexpected behavior.

2.3 Prepare Target

This step will only exist if we choose the prediction in the previous step, which is select Task. A maximum of 10 classes will be displayed with the most data points. If there are only two classes, we can choose which main class we want.

2.4 Select Input

In this step focuses on data quality, especially the quality of each data column. It may be necessary to consider removing the data column (Attribute) which gives less value. Therefore Rapid Miner gives the percentage value of CISM as the consideration. (C) columns that are too close to the target
column, (I) column where almost all values are different, (S) columns where almost all values are identical, (M) columns with missing values.

Based on the CISM, data attributes/columns are categorized into 3 colors to make it easier to select inputs. The three colors are green, yellow, and red. Green indicates the attribute quality/column data is excellent. Yellow correlation is so low that it does not contribute much. Red indicates a bad quality column which in many cases must be removed from the data set.

Also, to simplify the sorting feature/column data with a large number of features, there are several input options available. The choice is to select all, deselect yellow, deselect red. Select all means choosing all features without exception. Deselect yellow means not choosing all the features that have a small contribution. Deselect red means not selecting all features that have poor quality.

2.5 Select Model
In this step, the algorithm that will be used is selected. There are several options, but in this case, the Random Forest (RF) algorithm will be selected. The Random Forest algorithm can be summarized in four simple steps [19]:

1. Draw a random bootstrap sample of size n (randomly choose n samples from the training set with replacement).
2. Grow a decision tree from the bootstrap sample. At each node:
   a. Randomly select d features without replacement.
   b. Split the node using the feature that provides the best split according to the objective function, for instance, maximizing the information gain.
3. Repeat the steps 1-2 k times.
4. Aggregate the prediction by each tree to assign the class label by majority vote.

The idea in Random Forest (RF) algorithm is to average multiple (rooted) decision trees with a high variance to build a robust model that has less susceptible to overfitting and better generalization performance [19]. In this paper [15] explains that Random Forest (RF) is an algorithm that can improve accuracy results because in generating child nodes for each node is done randomly. This method is used to construct a decision tree consisting of root nodes, internal nodes, and leaf nodes by taking attributes and data randomly according to the provisions applied. The root node is the node that is located at the top or commonly referred to as the root of the decision tree. The internal node is a branching node, where this node has a minimum output of two, and there is only one input. While the leaf node or terminal node is the last node that has only one input and does not have output, the decision tree begins by calculating the entropy value as a determinant of the impurity level of the attribute and the value of information gain. To calculate the entropy value, the formula is used as in equation 1, while the information gain value uses equation 2.

\[ \text{Entropy} (Y) = -\sum_i (c|Y) \log_2 (c|Y) \]  

(1)

Where Y is the set of cases and p (c | Y) is the proportion of Y values to class c.

\[ \text{Information Gain} (Y,a) = \text{Entropy}(Y) - \sum_{v \epsilon \text{Values}(a)} \frac{|Yv|}{|Ya|} \times \text{Entropy}(Y_v) \]  

(2)

Where Values (a) are all possible values in a case set a. Yv is a subclass of Y with class v associated with class a. Ya are all values that correspond to a.

2.6 Result
The final step, the classification results will be displayed in this step. The classification results will be categorized into several things, such as model, simulator, performance, life chart.

The model shows a graphical representation of the model. The simulator provides a real-time interface that is easy to use to convert inputs to models and see outputs and simultaneously show predictions, conferences, and explanations for these inputs. Performance lists the accuracy of the model and other performance criteria. Life chart shows the effectiveness of the model by calculating
the ratio between the results obtained with the model and the results obtained without the model. Only made for two class problems.

Please note that the simulator will also get a chart/diagram that will show the most important features that are very influential in classifying.

3. Result and Discussion
Classification with high accuracy algorithms is something that needs to be considered when you want to do classification. The Random Forest (RF) algorithm is proven to be better based on research [10-13]. In addition to cervical cancer classification based on the expression of the microRNA gene with the Random Forest (RF) Algorithm, there will also be the identification of the most important features that influence cancer in the classification process.

3.1. Dataset
The data used in this study is miRNA dataset from cervical cancer with 714 features and 58 samples, and a column named miRNA [14]. In the dataset, there are 29 normal classes and 29 classes of cervical cancer.

3.2. Implementations
The tool used in this study is a software called Rapid Miner version 8.0. This research was conducted using the Auto Model feature on Rapid Miner. The following steps will be taken: Select Data, Select Task, Prepare Target, Select Input, Select Model, Result.

3.3. Experiment
In this experiment has three experimental scenarios, which each scenario will be carried out three times. Each scenario has different input in the selection step (select input) but has the same steps. The first scenario is to select all the data input. The second scenario input data without including all features with bad data quality (deselect red). The third scenario, input data without including all features with bad data quality (deselect red) and at the same time all the features that contribute little (deselect yellow).

| Set | Select Input | Number of Input |
|-----|--------------|----------------|
| 1   | Select All   | 714            |
| 2   | Deselect Red | 578            |
| 3   | Deselect Red and Deselect Yellow | 565 |

3.4. Result
The result for scenario 1 (input select all) has an average accuracy of 91.7%, the sensitivity of 83.3%, specificity of 100 % and runtime of 35.33 s. More details can be seen in table 2.

| Test No | Accuracy | Sensitivity | Specificity | Runtime |
|---------|----------|-------------|-------------|---------|
| 1       | 91.7 %   | 83.3 %      | 100 %       | 35 s    |
| 2       | 91.7 %   | 83.3 %      | 100 %       | 34 s    |
| 3       | 91.7 %   | 83.3 %      | 100 %       | 37 s    |
| Average | 91.7 %   | 83.3%       | 100 %       | 35.33 s |

The result for scenario 2 (input deselect red) has an average accuracy of 83.3%, the sensitivity of 66.7%, specificity of 100 % and runtime of 26.67 s. More details can be seen in table 3.

| Test No | Accuracy | Sensitivity | Specificity | Runtime |
|---------|----------|-------------|-------------|---------|
| 1       |          |             |             |         |
| 2       |          |             |             |         |
| 3       |          |             |             |         |
| Average |          |             |             |         |
The result for scenario 3 (input deselect red and deselect yellow) has an average accuracy of 100%, the sensitivity of 100%, specificity of 100 % and runtime of 28.67 s. More details can be seen in table 2.

| Test No | Accuracy | Sensitivity | Specificity | Runtime |
|---------|----------|-------------|-------------|---------|
| 1       | 100 %    | 100 %       | 100 %       | 29 s    |
| 2       | 100 %    | 100 %       | 100 %       | 27 s    |
| 3       | 100 %    | 100 %       | 100 %       | 30 s    |
| Average | 100 %    | 100 %       | 100 %       | 28.67 s |

When conducting the test set, a search was also conducted to find out the most important features that were very influential in this classification process. Here are the results:

Figure 1 tells about the important features in the classification with Random Forest algorithm for scenario 1 (input select all). Furthermore, important features are supporting the normal class (not cancer) which are very influential in the classification process, miR542-5p, miR-497 * and miR-125b-1 *. Also, there are important supporting features of cervical cancer classes that are very influential in the classification process of miasm-miR-205, miR142-3p, miR-203, and Candidate-30.

| Important Factors for N |
|-------------------------|
| miR-205                 |
| miR-142-3p              |
| miR-542-5p              |
| miR-203                 |
| miR-497*                |
| miR-125b-1*             |
| Candidate-30            |

**Figure 1. Important Features Set 1**
Figure 2. Important Features Set 2

Figure 2 tells about the important features in the classification with Random Forest algorithm for scenario 2 (input select all). Furthermore, important features are supporting the normal class (not cancer) that are very influential in the classification process of miR-114 and miR-133b. Also, there are important supporting features of cervical cancer classes that are very influential in the classification process of miasm-miR-155, Candidate-5, Candidate-29-3p, miR-515-5p, and miR-801.

Figure 3. Important Features Set 3

Figure 3 tells about the important features in the classification with Random Forest algorithm for scenario 3 (input select all). Furthermore, it is important features supporting the normal class (not cancer) which are very influential in the classification process, miR-145*, miR-204, miR-126, and miR-768-5p. Also, it is important features supporting the class of cervical cancer which are very influential in the process of classification of miR-519c-5p, miR-183, and miR-515-5p.

3.5. Discussion

Because the highest accuracy is found in the third test experiment, it will identify important features are supporting the class of cervical cancer:
- miR-519c-5p with miR-19 and miR-345 serum levels predict adverse pathology in prostate cancer patients eligible for active surveillance [16].
• miR-183 has a potential oncogenic role through the regulation of 2 tumor suppressor genes, *EGR1* and *PTEN*, and the deregulation of this major miRNA regulatory network may be central to many tumor types [17].

• miR-515-5p controlling cancer cell migration through MARK4 regulation, where MARK4 inhibitors can represent new therapeutic agents to control the spread of cancer [18].

4. Conclusion
In this research we propose the use of classification based on miRNA gene expression is a solution for early detection. The classification algorithm used is the Random Forest (RF) algorithm as an attempt to optimize accuracy in classification. At the same time, there will also be the identification of the most important features that significantly influence the classification process of cervical cancer. Based on experiments carried out, classification with Random Forest (RF) algorithm based on miRNA gene expression can be used for early detection in cervical cancer, where the third test set experiment has an accuracy rate of up to 100% outperforming the first test set with 91.7% accuracy and the second experimental set with 83.3% accuracy. From the three scenarios, it is known that the selection of input for training data will significantly influence the accuracy and sensitivity. An important feature supporting the class of cervical cancer in experiments the third test set is miR-515c-5p, miR-183 and miR-515-5p. Thus an important feature in the third test set experiment can be used as a reference/biomarker for further research.

References
[1] Yayasan Kanker Indonesia No year *Tentang Kanker* [Online] Available at http://yayasankankerindonesia.org/tentang-kanker.
[2] Kumparan 2017 *Kanker Serviks Pembunuh Nomor 1 Perempuan Indonesia* [Online] Available at: https://kumparan.com/@kumparan news/kanker-serviks-pembunuh-nomer-1-perempuan-indonesia.
[3] CNN Indonesia 2017 *Jumlah Kasus Kanker Serviks Indonesia Tertinggi ke-2 di Dunia* [Online] Available at: https://www.cnnindonesia.com/gaya-hidup/20170827161551-255-237610/jumlah-kasus-kanker-serviks-indonesia-tertinggi-ke-2-di-dunia.
[4] Boedina K. 2011 *Indonesian Journal of Cancer* 5 3.
[5] Novianti F and Purnami S 2012 *Jurnal Sains dan Seni ITS* 1 1.
[6] Pramunendar R, Dewi I and Asari H 2013 *Semantik* 1 3.
[7] Somantri O, Sungkar M and Sasmito G 2015 *Neural Network Untuk Klasifikasi Penanganan Gangguan Jaringan Distribusi Listrik 20 KV*.
[8] Rachman F and Purnami S 2012 *Jurnal Sains dan Seni ITS* 1 1.
[9] Ningrat R and Santosa B 2012 *Jurnal Teknik ITS* 1 1.
[10] Liu M, Wang M, Wang J and Li D 2013 Sensors and Actuators B: Chemical 177 970.
[11] Han T, Jiang D, Zhao Q, Wang L and Yin K 2017 *Transactions of the Institute of Measurement and Control* 40 2681.
[12] Borges F, Fernandes R, Lucas A and Silva I 2015 *Comparison Between Random Forest Algorithm and J48 Decision Trees Applied to the Classification of Power Quality Disturbances* Proceedings of the International Conference on Data Mining (DMIN).
[13] Razak E, Yusof F and Raus R 2016 *Classification of miRNA Expression Data Using Random Forests for Cancer Diagnosis* International Conference on Computer and Communication Engineering (ICCCE).
[14] Witten D, Tibshirani R, Gu S, Fire A and Lui W 2010 *BMC Biology* 8 58.
[15] Nugroho Y and Emiliyawati N 2017 *Jurnal Teknik Elektro* 9 24.
[16] Wang S, Shiboski S, Belair C, Cooperberg M, Simko J and Stoppler H 2014 *PLOS One* 9 1
[17] Aaron L., Sarver, Lihuua Li and Subbaya Subramanian 2010 *American Association for Cancer Research* 70 9570
[18] Pardo O, Castellano L, Munro C, Hu Y, Mauri F, Krell J, Lara R, Pinho F, Choudhury T, Frampton A, Pellegrino L, Pshezhetskii D, Wang Y, Waxman J, Seckl M and Stebbing J 2016 *EMBO reports* 17 570.
[19] Raschka S and Mirjalili V 2017 *Hands-On Data Science and Python Machine Learning* (Birmingham: Packt Publishing)