Multiclass Classification of Brain Cancer with Multiple Multiclass Artificial Bee Colony Feature Selection and Support Vector Machine

S A A Kharis*, I Hadi and K A Hasanah

Department of Mathematics, Faculty of Mathematics and Science, Universitas Negeri Jakarta, Jl. Rawamangun Muka, Jakarta, 13220, Indonesia

* E-mail: sellyanas@yahoo.com

Abstract. A World Health Organization reported that the mortality rate due to brain cancer is the highest in the Asian continent. It is critical importance that brain cancer can be detected earlier so that the treatment process can be carried out more precisely and will be able to extend the life expectancy of brain cancer patients. Taking advantage of microarray data, machine learning methods can be applied to help brain cancer prediction according to its type. This problem can be referred to as a multiclass classification problem. Using the one versus one approach, there will be as many as \( \frac{k(k-1)}{2} \) two-class problems, where \( k \) indicates the number of classes. In this paper, Multiple Multiclass Artificial Bee Colony (MMABC) implemented as a feature selection method and Support Vector Machine (SVM) as a classification method. ABC algorithm proved successful in solving optimisation problems with high dimensionality, and SVM can produce accurate and robust classification results. The data obtained from Broad Institute data. The data consist of 7129 features and 42 samples. From the experiment, the accuracy of Multiple SVM using a feature selection based MMABC method reached 95.24% accuracy in usage 300 best features; this percentage slightly more superior than SVM method without feature selection.

1. Introduction
Cancer is a group of diseases characterised by the uncontrolled growth and spread of abnormal cells [1]. If it happens to the brain, then it is called ‘brain cancer.’ Brain cancer is commonly attacking young generation, from children to adults below 40 years old [2]. More than two-thirds of adults diagnosed with brain cancer will die within two years of diagnosis [3, 4]. Based on other data from National Cancer Institute, there were about 17.760 deaths in 2019 which were caused by brain cancer and 23.820 new cases [5]. Besides, A World Health Organization reported that mortality rate due to brain cancer was and is still the highest in Asia continent [6]. However, the attention and help which was given by the government to this brain cancer case were not as much as they did to the other types of cancer. Therefore, there should be researches in detecting and treating it.

On the other side, gens in certain bodies can be interpreted numerically, which is usually called microarray data. In detecting data, microarray data can be analysed by seeing the combination of some of the gens in patient’s body. The characteristics from the medical data have many features (gens), that the amount of them could reach hundreds or even thousands, and little number of samples, which were
fewer than one hundred, made it need a method that could help to detect cancer easier and in objective way. In the machine learning context, the problem of detecting cancer using microarray data is one of the classification problems. In a simple case, the classification problem is detecting whether cancer or not cancer (normal), so there will be two classes. However, this classification problem can be analysed further, which is detecting the types of cancer. Detecting the type of cancer is not restricted to two classes, but it can be more than that.

Some researchers have examined the multiclass problem. In 2015, Zhang developed the Multiple Support Vector Machine Recursive Feature Elimination algorithm [7]. This method is a development from the Multiple Support Vector Machine Recursive Feature Elimination (MSVM-RFE) concept, which was stated by Duan and Multiclass Support Vector Machine Recursive Feature Elimination (MCSVM-RFE) which was stated by Zhou. In Zhou’s research, the one-versus-rest method was used to form the two-classes problems from multiclass problems. In this research, another alternative is used, which is one-versus-one. Another research done by Azminuddin and his colleagues with the multi-class svm model uses 1V1 strategy to classify the wall-following robot navigation data [8]. Besides, the research of multiclass classification was done by Thirumalaimuthu Thirumalaippan, using the SVM design based on multi knowledge-based system (SMK) [9]. Other than the usage of SVM, another classifier can be applied to the multiclass classification, just like the one which was applied by Juntao Li in his research, by using multinomial regression [10]. In this research, multiclass classification technique was submitted with the optimal artificial bee colony (ABC) method. The usage of ABC is chosen based on the previous research: ABC algorithm performs better than Differential Evolution (DE), Particle Swarm Optimization (PSO), and an Evolutionary Algorithm (EA) for multi-dimensional and multimodal numeric problems with high dimensionality. Besides that, ABC algorithm is flexible and straightforward to use and robust the optimisation of the algorithm [11]. SVM as a classifier is chosen because it can produce accurate and robust classification results [12].

2. Materials and Method

2.1. Materials

The data used in this research from Broad Institute (http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi). It was actually obtained in .gct format. By using the R programming language, data was converted into .csv format to make it easier to be observed. The data consisted of four types of brain cancers, such as Medulloblastoma (MD), Malignant Glioblastoma (MG), Primitive Neuroectodermal Tumors (PNET) and Atypical Teratoid Rhabdoid Tumor (ATRT). The data also consisted of 7129 features and 42 samples with ten samples for MD, ten samples for MG, ten samples for ATRT, eight samples for PNET, and four samples for Normal. In brain cancer microarray data, the features are the gens and the labels are the types of brain cancer.

2.2. Method

2.2.1. Multiple Multiclass Artificial Bee Colony (MMABC)

MMABC is a concept development of Artificial Bee Colony for the multiclass problems. In this research, the approach used is one versus one. With this type of approach, each of probabilities of the pair’s counts, so there will be $\frac{k(k-1)}{2}$ types of ordering features on MMABC and classifier model, which k indicates the number of classes. The classification problem of brain cancer from Broad Institute is multiclass, so there will be ten types of binary class classification. Those pairs are MD and MG, MD and normal, MD and PNET, MD and ATRT, MG and normal, MG and PNET, MG and ATRT, normal and PNET, Normal and ATRT, PNET and ATRT. On each problem of the mention binary class problem, change the label of first-class into 1 and the second class into -1. Then, pick the number of features to be applied. Make limits to each of the binary class problems with the chosen features. After that, train the classifier for each of those binary class problems. The process above shown in Figure 1.
The MMABC algorithm with one versus one (OVO) approach, which k class of the data shown in Figure 2. First, divide the multiclass problems by using the OVO approach, then we will get $\frac{k(k-1)}{2}$ of binary class problems. Second, initialise done by scout bee by using this equation:

$$X^j_i = X^j_{\text{min}} + \omega (X^j_{\text{max}} - X^j_{\text{min}}), \quad 0 \leq \omega \leq 1.$$  

(1)

Annotation:

- $X^j_i$: Food source, $i = 1,2,...,SM$
- $X^j_{\text{min}}$: Minimum features index, $j = 1,2,...,D$
- $X^j_{\text{max}}$: Maximum features index, $j = 1,2,...,D$

On ABC method as the selection feature, the control parameter of X is $X^\text{min} = 1$ and $X^\text{max} = $ maximum features index $X^\text{max}$. After the initialisation has been done, count the value of fitness food source $X^i_i$, $i = 1,...,SM$, based on this equation (2).

$$\text{fit}(X^i_i) = \text{akurasi}(X^i_i)$$  

(2)

Third, the bees colony will be divided into three sections. They are employed bee, onlooker bee, and scout bee. The employed bee comes to the location of food source and looks for new food source which has the nectar more than the surroundings have in the same area. The searching for a new food source ($V^j_i$), is using this equation (3).

$$V^j_i = X^j_i + \Phi(X^j_i - X^j_k),$$  

(3)

$X_k$ is the chosen food source which is chosen randomly, $k = 1,2,...,SM$, $k \neq i$, and $-1 \leq \Phi \leq 1 - 1 \leq \Phi \leq 1$. Because the food sources consist of features index (with the value of 1 until the total of features from one matrix data), so the choice of population will be applied with this requirement:

- If $V^j_i < X^\text{min}V^j_i < X^\text{min}$, then
  $$V^j_i = (V^j_i \mod \text{Total Fitur}) + 1$$  

(4)

- If $V^j_i > X^\text{max}$, so
  $$V^j_i = (V^j_i \mod \text{Total Fitur}) + 1$$  

(5)

After that, count fitness value of each food sources and make the greedy selection: If the value of fitness $V^j_i$ is higher than fitness $X^j_i$, then $V^j_i$ will replace $X^j_i$. If the value of fitness $X^j_i$ is higher than fitness $V^j_i$, then $X^j_i$ will stay, and abandon food was added into 1.
After the greedy selection made, find the probabilities to find out which food source will be replaced on the next phase, by using equation (6).

\[ p_i = \frac{\text{fitness}_i}{\sum_{n=1}^{SM} \text{fitness}_n} \]  

(6)

which fitness\(_i\) is a quality of the food source towards-i on the employed bee phase.

If the searching process of food sources has finished, then the employed bee will share the information about the quality of the food sources and the probability to the onlooker bee. Onlooker bee chooses the new food source based on the given probabilities of the mentioned food. In this process, onlooker bee increases the probabilities of the new food sources based on the equation (3), and the value of fitness will be found out by using the equation (2). Then, do the greedy selection process.

---

**Figure 2. MMABC algorithm**

2.2.2. The procedure to solve the multiclass problem of brain cancer with MMABC and SVM

In this research, the usage of SVM as the classification method is combined with MMABC as the feature selection method to solve the multiclass classification problem of brain cancer. The learning process to solve the multiclass classification problem is shown in Figure 3.

---

**Figure 3. Feature Selection Flowchart**
The result of classification is running time and accuracy with or without using feature selection. Accuracy is one metric for evaluating classification models [13]. The accuracy value in percentage is between 0 and 100. The accuracy is measured by using matrix confusion. Table 1 shows the matrix confusion for the multiclass problem with the amount of k class of the examples [14].

| Actual Class | Predicted Class |
|--------------|-----------------|
|             | Class 1 | Class 2 | … | Class k |
| Class 1     | $f_{11}$  | $f_{12}$ | … | $f_{1k}$ |
| Class 2     | $f_{21}$  | $f_{22}$ | … | $f_{2k}$ |
| …           | …       | …       | … | …       |
| Class k     | $f_{k1}$  | $f_{k2}$ | … | $f_{kk}$ |

The accuracy of the multiclass problem can be measured by equation 7 [15].

$$\text{Accuracy} = \frac{\frac{\text{The number of true predictions}}{\text{The number of predictions}}}{\frac{\text{The diagonal values on matrix}}{\text{All values on confusion matrix}}} = \frac{\sum_{j=1}^{m} f_{jj}}{\sum_{k=1}^{m} \sum_{j=1}^{m} f_{jk}}$$

### 3. Results and Discussion

In this paper, there are four types of simulations done. They are brain cancer classification using SVM and linear kernel without feature selection, brain cancer classification using SVM and RBF kernel without feature selection, brain cancer classification using MMABC, and SVM with linear kernel, and brain cancer classification using MMABC and SVM with RBF kernel simulation. The simulation was done by using 3-cross validation, the data was divided into three parts, which each part of the data will be the test data and the average will be obtained. Based on the SVM, count the perpendicular distance of data with each of the hyperplane. If the distance of one sample on the test data is closer to hyperplane for class -1 than the hyperplane to the class 1, the sample is predicted in class -1. By doing this, on every binary class case, the prediction will be obtained from every sample of the test data.

To unite the results of the mentioned predictions, the class label for each of the binary class cases will be changed back into the multiclass label. With the max-wins principal, class, which commonly shows up is the last prediction of a sample on the test data. After all the samples on the data test have been successfully predicted, then make the confusion matrix based on Table 1. The next one is operating the accuracy with the equation (7). The first simulation, which will be done is doing the classification process using SVM with the linear kernel without feature selection. Table 2 shows the confusion matrix of the classification simulation’s results of brain cancer using SVM and linear kernel without feature selection.

| Actual Class | MD | MG | ATRT | Normal | PNET |
|--------------|----|----|------|--------|------|
| Class MD     | 8  | 1  |      |        | 1    |
| Class MG     | 1  | 5  | 3    | 1      |      |
| Class ATRT   |    |    | 10   |        |      |
| Class Normal |    |    |      | 4      |      |
| Class PNET   | 1  | 2  |      | 5      |      |

Based on the mentioned confusion matrix, the accuracy for overall classes is 76.19 %, which consist of 32 samples classified as true, and 10 samples which had error classifications, such as 1 data of MD.
classified as MG, 1 data of MD classified as PNET, 1 data of MG classified as MD, 3 data of MG classified as ATRT, 1 data of MG classified as PNET, 1 data of PNET classified as MD, 2 data of PNET classified as MG. The running time for simulation 1 was 20.488 seconds. After that, do the simulation 2 (classification process using SVM and RBF Kernel without feature selection). Table 3 shows the confusion matrix of the classification simulation’s results of brain cancer using SVM with RBF kernel without feature selection.

**Table 3.** Confusion matrix’s types of brain cancer classification using SVM and RBF kernel without feature selection

| Predicted Class | MD  | MG  | ATRT | Normal | PNET |
|-----------------|-----|-----|------|--------|------|
| Actual Class    |     |     |      |        |      |
| MD              | 9   | 1   | 1    | 1      | 0    |
| MG              | 1   | 9   | 1    | 1      | 0    |
| ATRT            | 10  |     |      |        |      |
| Normal          | 1   | 1   | 2    | 7      | 0    |
| PNET            | 1   |     |      |        |      |

Based on the mentioned confusion matrix, the accuracy for overall classes is 85.71%, which consist 36 samples classified as true, and 6 samples which had error classifications, such as 1 data of MD classified as ATRT, 1 data of MG classified of MD, 1 normal data classified as MG, 2 normal data classified as PNET, 1 data of PNET classified as MG. The running time for simulation 2 was 9.8114 seconds. Simulation 3 and 4 are the types of brain cancer classifications using MMABC and SVM with linear kernel and RBF, respectively. The features used were 10, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000. Table 4 shows the confusion matrix of the classification simulation’s results of brain cancer using MMABC and SVM with linear and RBF kernel.

**Table 4.** The accuracy result of brain cancer classifications using MMABC and SVM with linear kernel and RBF

| The number of features used | Linear Kernel | RBF Kernel |
|-----------------------------|---------------|------------|
|                            | The average accuracy | The average running time (seconds) | The average accuracy | The average running time (seconds) |
| 10                          | 88.05%        | 0.2669     | 85.71%        | 0.3716 |
| 50                          | 85.71%        | 0.2741     | 88.10%        | 0.3782 |
| 100                         | 85.71%        | 0.3371     | 92.86%        | 0.3865 |
| 200                         | 90.48%        | 0.3788     | 80.48%        | 0.4003 |
| 300                         | 92.86%        | 0.3898     | 95.24%        | 0.4152 |
| 400                         | 85.71%        | 0.3984     | 92.86%        | 0.4180 |
| 500                         | 88.10%        | 0.4343     | 92.86%        | 0.4228 |
| 600                         | 90.48%        | 0.4652     | 88.10%        | 0.4373 |
| 700                         | 80.95%        | 0.5091     | 90.48%        | 0.4457 |
| 800                         | 76.19%        | 0.5305     | 85.71%        | 0.4505 |
| 900                         | 90.48%        | 0.5489     | 80.95%        | 0.4536 |
| 1000                        | 85.71%        | 0.5764     | 90.48%        | 0.4790 |
| Mean                        | 86.70%        | 0.4258     | 88.65%        | 0.4216 |

The accuracy above can also be expressed in Figure 4. Based on the simulation result, the number of the most optimum of the usage of the linear kernel on MMABC-SVM is 300 features with the accuracy
92.86% meanwhile the number of the most optimum features for using RBF Kernel on MMABC-SVM is 300 features with 95.24%. The running time is shown in Figure 5.

![The Accuracy Graphic of Simulation with MMABC-SVM](image1)

**Figure 4.** The accuracy graphic of simulation with Linear Kernel and RBF for MMABC-SVM

![The Running Time Graphic of Simulation with MMABC-SVM](image2)

**Figure 5.** The running time graphic of simulation with Linear Kernel and RBF for MMABC-SVM

From four simulations, the highest accuracy is 95.24% by using MMABC-SVM with RBF kernel with a 0.4152 seconds amount of 300 features. The average of the accuracy using MMABC is higher if it is compared with the classification without the selection feature, for both linear kernel and RBF. On the graphic of the running time, it can be seen generally that the more features used, the length of running time needed. It also can be assumed that the importance of the selection feature process is to increase the accuracy and to reduce the running time.
4. Conclusion
The classification of brain cancer based on the data from Broad Institute can be solved by using machine learning. This problem is a multiclass classification problem. In this paper, the approach used is one versus one, so 1 multiclass classification problem became ten problems of binary class classification. The classification method used is SVM. 4 simulations have been done on this research. In the first simulation, the usage of SVM and linear kernel without feature selection resulted the accuracy 76.19%, with the running time 20.488 seconds. On the second simulation, the usage of SVM and RBF Kernel without feature selection resulted the accuracy 85.71 % with the running time 9.8114 seconds. In the third simulation, the usage of SVM and MMABC with linear kernel resulted in the highest accuracy 92.86 % and the running time 0.3898 with the number of usage 300 features. In the fourth simulation, the usage of SVM and MMABC with RBF Kernel resulted in the highest accuracy 95.24 % and 0.4152 of the running time with the number of usage 300 features. From all of those four simulations, the MMABC-SVM method with RBF Kernel is a method that reduces the highest average of the accuracy and the fastest of the running time.

Acknowledgements
This research is fully funded and supported by Lembaga Penelitian dan Pengabdian Masyarakat (Institute of Research and Community Services) Universitas Negeri Jakarta.

References
[1] American Cancer Society 2019 Cancer Facts & Figures (Atlanta: American Cancer Society) 1-72.
[2] Australian Institute of Health and Welfare 2017 Australian Cancer Incidence and Mortality (ACIM) (Canberra: AIHW 7)
[3] Gilbert M R, Dignam J J, Armstrong T S, Wefel J S, Blumenthal D T, Vogelbaum M A, Colman H, Chakravarti A, Pugh S, Won M, Jeraj R, Brown P D, Jaekle K A, Schiff D, Stieber V W, Brachman D G, Werner-Wasik M, Tremont-Lukats I W, Sulman E P, Aldape K D, Walter J. Curran J and Mehta M P 2014 N. Engl. J.Med 370 699-708
[4] Chinot O L, Wick W, Mason W, Henriksson R, Saran F, Nishikawa R, Carpentier A F, Hoang-Xuan K, Kavan P, Cernea D, Brandes A A, Hilton M, Abrey L and Cloughesy T 2014 N.Engl. J.Med 370 709-722
[5] National Cancer Institute 2019 Cancer Stat Facts: Brain and Other Nervous System Cancer (Rockville: National Cancer Institute)
[6] S Tandel G, Biswas M, Kakde O, Tiwari A, Suri H, Turk M 2019 Cancers, I 111
[7] Zhang L and Huang X 2015 International Joint Conference on Neural Networks (IJCNN) 1-6
[8] Azis A I S, Suhartono V and Himawan H 2017 J. Teknol. Inf. CyberKU 13 170
[9] Ramanathan T T and Sharma D 2017 Procedia Comput. Sci. 115 307
[10] Li J, Wang Y, Song X and Xiao H 2018 Comput. Biol. Med. 100 1
[11] Karaboga D and Basturk B 2007 J. Glob. Optim. 39 459
[12] Auria L, Moro R A 2008 DIW Discussion Papers, No. 811, (Berlin: Deutsches Institut für Wirtschaftsforschung)
[13] Hossin, M and Sulaiman, M N 2015 Int. J. Data Min. Knowl. Manag. Process 5 1
[14] Manliguez, C 2016 Generalized confusion matrix for multiple classes
[15] Santra, A K and Christy, C J 2012 Int. J. Comput. Sci. Issues 9 322