Accredited by National Journal Accreditation (ARJUNA) Managed by
Ministry of Research, Technology, and Higher Education, Republic Indonesia with Second Grade (Peringkat 2, Sinta 2)
since the year 2017 to 2021 according to decree No. 10/E/KPT/2019

Published online on the journal's webpage: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 5 No. 4 (2021) 794 - 801 ISSN Media Elektronik: 2580-0760

Cancer Detection based on Microarray Data Classification Using FLNN and Hybrid Feature Selection

Ghozy Ghulamul Afif[1], Adiwijaya[2] , Widi Astuti[3]
[1,2,3]Informatics, School of Computing, Telkom University
[1]jamessaldo@student.telkomuniversity.ac.id, [2]adiwijaya@telkomuniversity.ac.id, [3]astutiwidi@telkomuniversity.ac.id

Abstract

Cancer is one of the second deadliest diseases in the world after heart disease. Citing from the WHO's report on cancer, in 2018 there were around 18.1 million cases of cancer in the world with a total of 9.6 million deaths. Now that bioinformatics technology is growing and based on WHO's report on cancer, an early detection is needed where bioinformatics technology can be used to diagnose cancer and to help to reduce the number of deaths from cancer by immediately treating the person. Microarray DNA data as one of the bioinformatics technology is becoming popular for use in the analysis and diagnosis of cancer in the medical world. Microarray DNA data has a very large number of genes, so a dimensional reduction method is needed to reduce the use of features for the classification process by selecting the most influential features. After the most influential features are selected, these features are going to be used for the classification and predict whether a person has cancer or not. In this research, hybridization is carried out by combining Information Gain as a filtering method and Genetic Algorithm as a wrapping method to reduce dimensions, and lastly FLNN as a classification method. The test results get colon cancer data to get the highest accuracy value of 90.26%, breast cancer by 85.63%, lung cancer and ovarian cancer by 100%, and prostate cancer by 94.10%.

Keywords: cancer detection, microarray, information gain, genetic algorithm, hybrid

1. Introduction

Cancer is one of the second deadliest diseases in the world after heart disease. Citing from the WHO report on cancer [1], there will be at least one from six people who dies from cancer in the world. In 2018 it was recorded that around 18.1 million cancer cases in the world with the total death of 9.6 million people, WHO predicted that in 2040 there will be a possibility of an increase in cancer cases to 29.4 million cases with total deaths predicted to be almost double the death rate in 2018. Based on the presented data, the role of technology that is able to detect cancer early in order to reduce the number of cancer cases in the future is needed.

As time goes by bioinformatics technology is also getting more advanced, now microarray data becomes popular for use in the analysis and diagnosis of cancer in the medical world. DNA microarray data is often used to examine how large numbers of genes are expressed simultaneously at the same time. By utilizing the results of the analysis of gene expression, detecting whether a person is diagnosed with cancer will be more efficient than the traditional method where the medical team have to check the symptoms or signs of cancer of the patients [2].

DNA microarray data has an enormous dimensions that this can affect the level of accuracy when searching for informative genes in the DNA data [3]. A dimension reduction method is needed to identify informative genes that can be used to predict cancer. Mukesh Kumar et.al [4] conducted a research on the leukemia, ovarian, and breast cancer dataset using *t*-test dimension reduction method and Functional Link Neural Network classification. He explains based on the results of the research above, Legendre Polynomial is able to provide the best performance results compared to Functional Link Neural Network other three techniques and also he suggests the use of hybridization in dimensional reduction to reduce the complexity of the classification model. In that research, Kumar got an accuracy value of 97.22%, 98.42% and 85.57%. Putri Tsatsabilla Ramadhani et.al [5] with colon and leukemia cancer dataset got the accuracy value of 92.3% and 87.5%.

Accepted by the Editor: 29-07-2021 | Completed Revision: 22-08-2021 | Published Online: 26-08-2021

Bintang Peryoga et.al [6] used colon, prostate, lung, breast and ovarian cancer dataset got the accuracy value of 91,8%, 58,94%, 100%, 83,47% and 100%. Bisma Pradana et.al [7] conducted a research with colon, lung, ovarian cancer dataset got the accuracy value of 91,67%, 100%, 100%.

In this research, the author proposes an early cancer detection with the use of hybridization when reducing dimensions using IG-GA method and Functional Link Neural Network method based Legendre Polynomial to know how big the effect of differences during hybridization in reducing dimensions especially on the required computational time parameter and data classification's performance results along with the effect of Learning Rate parameter values on the FLNN method on the obtained performance value. The obtained results are expected to help the medical world to diagnose early symptoms or signs of cancer.

2. Research Method

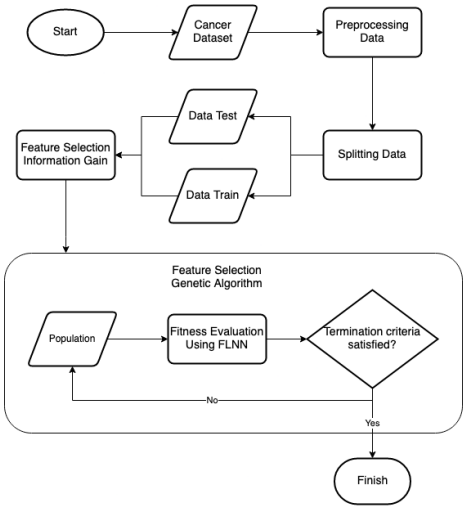

Figure 1. System Architecture

As in Figure 1, the system design that was built is divided into several stages of process, the first process is to pre-process the available dataset which will then be carried out a feature selection using Information Gain and Genetic Algorithm. The use of hybrid method in reducing the dimensions makes the features that will be used at the classification stage using the FLNN method less than the non-hybrid method.

2.1. Dataset

In this research, author uses microarray dataset from Kent-Ridge Biomedical. There are five datasets used, namely Breast Cancer, Ovarian Cancer, Prostate Cancer, Lung Cancer and Colon Cancer. Details of the five datasets are listed in the Table 1.

Table 1. Dataset Reference [16]

Data	Number of Features	Amount of Data	Number of Classes
Breast	24481	97	2 (51 non-relapse, 46 relapse)
Ovarian	15154	253	2 (91 Normal, 162 Cancer)
Prostate	12600	136	2 (59 Normal, 77 Cancer)
Lung	12533	181	2 (31 Mesothelioma, 150 ADCA)
Colon	2000	62	2 (40 Negative, 22 Positive)

2.2 Preprocessing Data

At the data pre-processing, there are two stages carried out by the author including solving the problem if missing values are found in the dataset along with standardizing the data. Solving the missing value problem is carried out in order to maintain good performance results, as for the used techniques are vary so there will be several scenarios/attempts to be carried out. Normalization will be done using Equation 1 as the min max scaler function.

$$X = \frac{Xi - \min(x)}{\max(x)} \quad (1)$$

2.3. Split Data

K-Fold Cross Validation is a method of dividing training data and test data. The proportion of the training data and test data distribution depends on the predetermined K value. In this research, author uses K with the value of five so that there will be five data partitions of four training data and one test data. During the process, the data that has been partitioned as training data and test data will be used for classification alternately and the classification results taken are the average results of partitions number. Figure 2 is the illustration of the K-Fold (K=5).

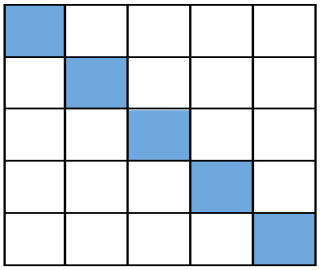

Figure 2. Illustration of K-Fold (k=5)

2.4. Information Gain

Feature selection is part of the dimension reduction process by selecting several features that are considered important for the classification process [11]. In this research author uses Information Gain as the filter method. Citing from [9] and [6] that filter method works without the influence of the classification technique/method. This explains that by ranking each feature, the feature selection is able to provide more

DOI: https://doi.org/10.29207/resti.v5i4.3352

efficient computing time. Equation 2 is the Information Gain's equation.

$$Gain(A) = Entropy(S) - Entropy_A(S) \quad (2)$$

$$Entropy(S) = \sum_{i=1}^{k} - P(C_i, S) * \log_2 (P(C_i, S)) \quad (3)$$

$$Entropy_A(S) = \sum_{i=1}^{v} - |S_i| / |S| * Entropy(S_i) \quad (4)$$

Information Gain is the subtraction of the values of *Entropy(S)* and $Entropy_A(S)$ where *Entropy(S)* is the parent entropy as seen in Equation 3 and $Entropy_A(S)$ is the child entropy as seen in Equation 4. *Entropy(S)* as parent entropy with $P(C_i, S)$ is the probability of class C_i on the S set. S_i is the number of cases in the i-th partition where A_I is the value of the attribute or feature of A.

2.5. Genetic Algorithm

According to Eric Cantu-Paz [12] Genetic Algorithm (GA) is a feature selection that is able to give good results and can produce higher performance results on certain datasets. In performing the feature selection using Genetic Algorithm, it is necessary to determine the proportion of training data and test data first. Referring to [5] the following are several stages of GA that have been adapted to the requirements of this study.

The first stage is Individual Representation. Where in this stage each individual will be represented as a binary number (0 or 1). Then initialization of the population based on the binary number is carried out randomly as much as the number of features and the size of the population. Feature selection is done by making each individual in the population as a representation of the to be selected feature. If a bit is equal to 0 then the feature will not be selected, whereas if the bit is equal to 1 then the feature will be selected.

After each individual is represented, Fitness Evaluation is conducted. In this second stage, the FLNN algorithm is used to produce performance results (F1-Score) as a function of the fitness as seen in Equation 5.

$$F1 = \frac{2}{\frac{1}{precission} + \frac{1}{recall}} = \frac{2*(precission*recall)}{precission+recall} \quad (5)$$

After the fitness value for each individual is obtained, the individual with the highest fitness value will be selected as The Elitism so that the fitness value does not disappear during the ongoing genetic operation.

The third stage is Genetic Operation. Where it has 3 sub-stages of which the first sub-stage is Parent Selection. Referring to [6], in selecting the parent for the next generation, two individuals with the highest fitness value from the last generation will be selected. The second sub-stage is Crossover, where it needs to be done on the chromosomes that have been selected as parents to get the offspring or commonly called children. Each offspring chromosome will have inherited genes from the parent chromosome. Referring to [5], the crossover probability used is 0.8. Lastly, mutations need to be done by generating offspring chromosomes randomly based on the predetermined mutation probability. Binary numbers that have been randomly generated will be checked whether they meet the criteria for less than the mutation probability, if they meet the criteria the binary numbers will be inverted. Referring to [5], the mutation probability used is 0.01.

Then, Survivor Selection is conducted for the fourth stage. Generational Replacement is needed as the survivor selection for the next generation where the next generations will contain new chromosomes resulting from crossover and mutations, as well as the best chromosomes that have been stored in The Elitism.

After all of the above stages are done, Criteria Termination is conducted as the last stage. Where the iteration in GA will end when it reaches the maximum generations or target that has been set.

Listed in Table 2 are the required parameters in the Genetic Algorithm implementation.

Table 2. Parameters of Genetic Algorithm

Parameter	Score
Mutation Rate	0.01
Crossover Rate	0.8
Population Size	10
Generation	5

2.6. Functional Link Neural Network

The next process that will be carried out after the dimension reduction is to classify the microarray data using the FLNN (Functional Link Neural Network) method with the Legendre Polynomial base function so that the results of microarray data classification into cancer classes are represented by a value of 1 and classified as negative with the value of 0. Functional Link Neural Network is an artificial neural network that has a single layer architecture, so that FLNN does not have a hidden layer [5]. Based on [8] from [5] when compared to neural networks that use hidden layers, it can be said that FLNN has more efficient and faster computation when compared to Multilayer Neural Network (MNN). This is supported in [4] which explains in his research that the Legendre Polynomial base function is able to provide the most optimal results compared to other FLNN base function in classifying microarray data. The following are the steps of the

DOI: https://doi.org/10.29207/resti.v5i4.3352

Functional Link Neural Network classification algorithm according to [8] in [11]:

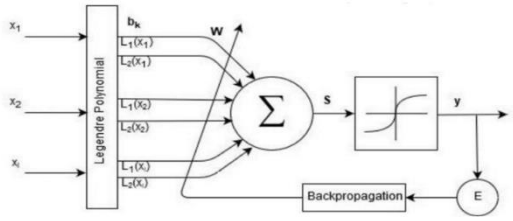

Figure 3. Illustration of Legendre Polynomial-based FLNN [8]

Based on Figure 3, the first step is to find the value of The Legendre Polynomial with Equation 6 as the base function.

$$L_{i+1} = \frac{1}{i+1}[(2i+1)xL_i - iL_{i-1}(x)] \quad (6)$$

Where L_i is the Legendre Polynomial, i is the order of polynomial and x is the original data input value.

The second step is to sum the value of the Legendre Polynomial as seen in the Figure 3 with Equation 7.

$$S_i = \sum_{i=1}^{n} w_i L_i(x) + b_i] \quad (7)$$

Where S_i is the linear sum value of Legendre Polynomial, w_i is the weight value, b_i is the bias value and n is the amount of data (feature) in one object.

Then, the obtained linear sum value will be activated by using the sigmoid activation in the third step with Equation 8.

$$F(s) = \frac{1}{1+e^{-s_i}} \quad (8)$$

The next step is to evaluate the obtained classification results using Equation 9 as the mean square error function.

$$E = \frac{1}{n}\sum_{i=1}^{n}[d_1 - y_i]^2 \quad (9)$$

Where d_1 is the prediction target value and y_i is the prediction results value.

Finally in the last step of FLNN, the backpropagation learning that is used by the algorithm has two calculation stages. The first calculation is a forward calculation to calculate the error between the prediction class with the target class, and then the second calculation is a backward calculation to propagate the error backwards to update the w value with Equation 10.

$$w_i = w_{i-1} - \eta \frac{\widehat{m_t}}{\sqrt{\widehat{v_t}}+e} \quad (10)$$

Where η is the learning rate, $\widehat{m_t}$ is the first momentum while $\widehat{v_t}$ is the second momentum and e is epsilon.

2.7. Performance Evaluation

The last step in this research is to evaluate the performance to find out how well the system that has been built uses hybrid in dimension reduction and FLNN as the classification method. The use of confusion matrix as the basis to determine the actual data and predicted data

Table 3. Confusion Matrix

Actual/Predicted	Predicted Positive	Predicted Negative
Positive	TP	FN
Negative	FP	TN

Based on Table 3, TP is the value for the system successfully classifying the data as positive for cancer according to the actual data, FP is the value for the system failing to classify the data as negative for cancer according to the actual data, FN is the value for the system failing to classify the data as positive for cancer according to the actual data, and TN is the value for the system successfully classifying the data as negative cancer according to the actual data.

Precision is the value of the match or compatibility between the requested information and the results provided by the system which can be obtained with Equation 11.

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

Recall is the value of the success of the system in finding back information which can be obtained with Equation 12.

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

F1-score is the average of precision and recall which can be obtained with Equation 13.

$$F1 - Score = \frac{2 * (Precision \times Recall)}{(Precision + Recall)} \quad (13)$$

Accuracy is the value of the system's success in predicting true positive and true negative compared to all data which can be obtained with Equation 14.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (14)$$

2.8. Test Scenario

In this study, two test scenarios were carried out on 5 cancer datasets including comparing the results of F1-score performance and accuracy for microarray data classification using FLNN with IG an FLNN with

DOI: https://doi.org/10.29207/resti.v5i4.3352

IG+GA to examine the effect of hybridization in the dimension reduction process. The next scenario is to examine the effect of the learning rate parameter on the FLNN classification model. The limitations in this study are the use of order 2-4 for the Legendre Polynomial and the use of GA parameter as the feature selection as shown in Table 2.

3. Result and Discussion

3.1. Test Result

Attached are the results of the test that have been carried out using the Information Gain and Genetic Algorithm hybridization feature selection method and using the FLNN classification method with predetermined parameters.

Table 4. Performance Results with Filtering Information Gain (LR = 0.6)

Data	Order	Features After IG	Avg. Accuracy	Avg. F1-Score	Avg. Comp. Time
Colon	2	100 Features	48.72%	32.11%	7.84s
	3		59.99%	37.17%	7.71s
	4		42.05%	29.11%	7.80s
Breast	2	100 Features	52.53%	36.03%	75.42s
	3		49.47%	34.68%	75.19s
	4		48.47%	32.63%	75.11s
Lung	2	100 Features	98.35%	97.02%	82.78s
	3		97.24%	94.42%	84.04s
	4		98.90%	98.09%	84.30s
Ovarian	2	100 Features	47.98%	36.24%	51.84s
	3		84.44%	75.67%	51.99s
	4		72.28%	68.35%	52.21s
Prostate	2	100 Features	48.57%	38.67%	35.51s
	3		54.39%	40.61%	35.41s
	4		52.17%	42.93%	35.27s

Attached in Table 4 experiments using the Information Gain dimension reduction method without the wrapping method and the learning rate of 0.6. Author obtains the optimal accuracy values for breast cancer data on order 2 of 52.53%, colon cancer, ovarian cancer, and prostate cancer data on order 3 of 59.99%, 84.44% and 54.39%, and for lung cancer data on order 4 of 98.90%.

Attached in Table 5 is the performance results after wrapping with the Genetic Algorithm method and the learning rate value is 0.6. Author obtains the optimal accuracy values for colon cancer and lung cancer data on all orders at 64.62% and 100%, on ovarian cancer data on order 4 with the value of 98.42%, and for breast cancer data the value for both order 2 and 3 is 53.58%, and prostate cancer data for order 2 is 61.69%.

Attached in Table 6, experiments using the Information Gain dimension reduction method without the wrapping method and the learning rate value of 0.001. Author obtains the optimal performance results for breast cancer data on order 2 with the accuracy value of 69.05%, for ovarian cancer data the optimal accuracy value for both order 2 and 3 is 99.61%, while for colon cancer and lung cancer data on order 4 the optimal accuracy values obtained are 84.10% and 99.44% and on prostate cancer data on order 3 of 91.19%.

Attached in Table 7 is the performance results after wrapping with the Genetic Algorithm method and the learning rate value of 0.001. Authors obtains the optimal accuracy value for colon cancer and prostate cancer data on order 4 of 90.26% and 94.10%, on breast cancer data on order 3 with the optimal value is 85.63%, while for lung cancer and ovarian cancer data on all orders with the values of 99.44% and 100%.

Table 5. Performance Results with Wrapping Genetic Algorithm (LR = 0.6)

Data	Order	Avg. Features After GA	Avg. Accuracy	Avg. F1-Score	Avg. Comp. Time
Colon	2	51 Features	64.62%	39.24%	160.09s
	3	55 Features	64.62%	39.24%	155.24s
	4	51 Features	64.62%	39.24%	153.40s
Breast	2	47 Features	53.58%	36.50%	250.87s
	3	49 Features	53.58%	36.50%	240.76s
	4	51 Features	52.58%	34.45%	240.06s
Lung	2	49 Features	100%	100%	157.94s
	3	53 Features	100%	100%	188.55s
	4	51 Features	100%	100%	223.15s
Ovarian	2	52 Features	97.22%	96.96%	218.12s
	3	53 Features	98.02%	97.84%	219.87s
	4	52 Features	98.42%	98.27%	228.57s
Prostate	2	54 Features	61.69%	56.30%	194.65s
	3	53 Features	57.99%	54.07%	203.85s
	4	48 Features	57.25%	54.33%	191.83s

Table 6. Performance Results with Filtering Information Gain (LR = 0.001)

Data	Order	Features After IG	Avg. Accuracy	Avg. F1-Score	Avg. Comp. Time
Colon	2	100 Features	84.10%	80.86%	8.82s
	3		84.10%	80.86%	8.34s
	4		84.10%	81.81%	8.59s
Breast	2	100 Features	69.05%	68.43%	65.13s
	3		68.05%	67.73%	65.25s
	4		65.99%	65.51%	65.03s
Lung	2	100 Features	98.89%	97.68%	65.31s
	3		98.89%	97.68%	67.25s
	4		99.44%	98.93%	65.12s
Ovarian	2	100 Features	99.61%	99.58%	50.16s
	3		99.61%	99.58%	50.22s
	4		98.82%	98.71%	51.75s
Prostate	2	100 Features	90.48%	90.34%	33.67s
	3		91.19%	91.09%	34.00s
	4		88.28%	88.17%	34.07s

Based on the results of the tests that have been conducted by author, where Table 4 and 6 are the experiments using only the Information Gain as the dimension reduction, while Table 5 and 7 are experiments using the

DOI: https://doi.org/10.29207/resti.v5i4.3352

hybridization method of Information Gain and Genetic Algorithm as dimension reduction. Overall, colon cancer data got the highest accuracy value of 90.26%, breast cancer of 85.63%, lung cancer and ovarian cancer of 100%, and lastly prostate cancer of 94.10%.

Table 7. Performance Results with Wrapping Genetic Algorithm (LR = 0.001)

Data	Order	Avg. Features After GA	Avg. Accuracy	Avg. F1-Score	Avg. Comp. Time
Colon	2	49 Features	88.72%	87.47%	167.79s
	3	53 Features	88.72%	87.83%	128.87s
	4	50 Features	90.26%	89.53%	135.77s
Breast	2	52 Features	82.53%	82.48%	207.93s
	3	49 Features	85.63%	85.58%	208.40s
	4	53 Features	79.26%	79.20%	228.48s
Lung	2	49 Features	99.44%	98.93%	135.59s
	3	53 Features	99.44%	98.93%	143.12s
	4	52 Features	99.44%	98.93%	126.56s
Ovarian	2	51 Features	100%	100%	86.05s
	3	49 Features	100%	100%	83.41s
	4	56 Features	100%	100%	83.83s
Prostate	2	53 Features	93.36%	93.27%	137.74s
	3	54 Features	92.65%	92.58%	160.49s
	4	50 Features	94.10%	94.02%	144.55s

3.2. Effect of Hybridization Method on Performance Results

There have been tests on five cancer datasets used by author and the use of predetermined parameters. The parameter values used refer to the author's research reference. Attached are the results of the tests that have been conducted.

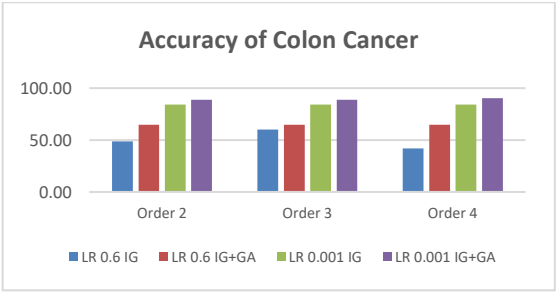

Figure 4. Accuracy of Colon Cancer

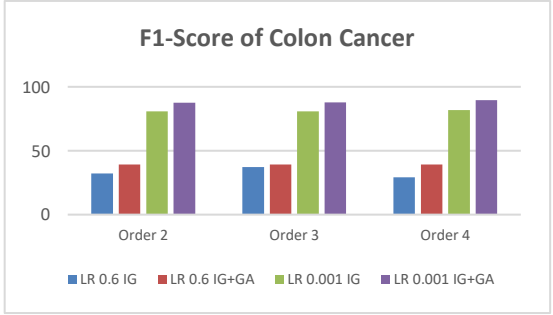

Figure 5. F1-Score of Colon Cancer

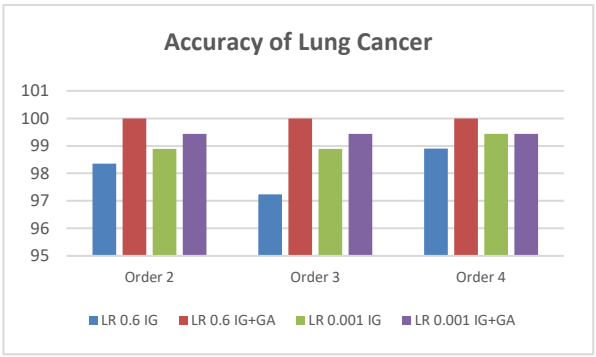

Figure 6. Accuracy of Lung Cancer

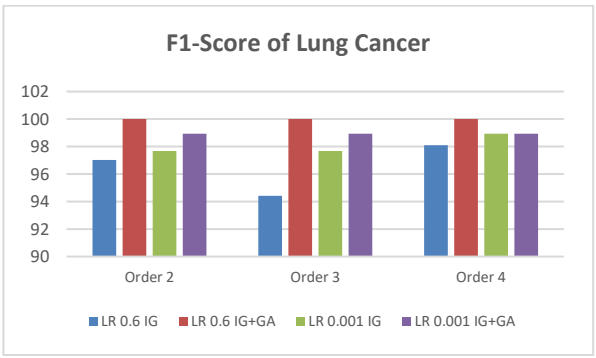

Figure 7. F1-Score of Lung Cancer

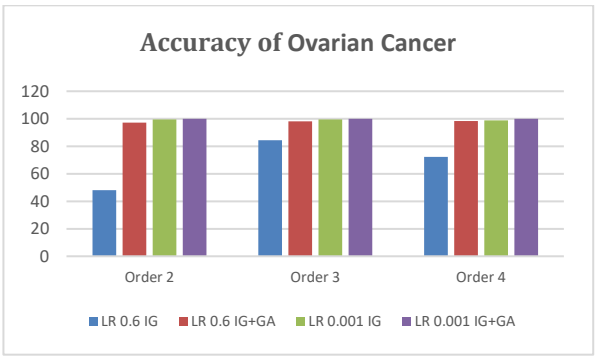

Figure 8. Accuracy of Ovarian Cancer

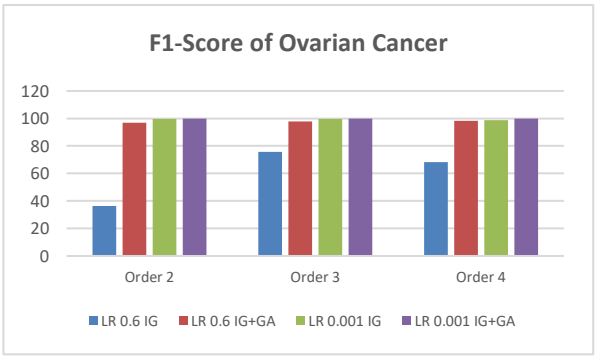

Figure 9. F1-Score of Ovarian Cancer

DOI: https://doi.org/10.29207/resti.v5i4.3352

Based on Figure 4 to 13, it can be seen that the use of Information Gain and Genetic Algorithm as hybridization method is able to increase the accuracy value and F1-score in almost all scenarios for each dataset. This is because the use of the Genetic Algorithm method as the wrapper method in the dimension reduction process is able to optimize the FLNN model which is used as the fitness function in the Genetic Algorithm process. Those can be seen on the colored bars where compared to the blue bars, the red bars will always be higher and goes the same for green bars compared to purple bars. As for citing from [9] in [6], the use of Genetic Algorithm as the wrapper has a weakness of inefficient computation time due to taking hypotheses model into training and testing on the used feature space. Increasing the order value of the FLNN can also affect the computation time due to the increase in the input space so it will require a longer computation time.

3.3. The Effect of Learning Rate on Performance Results

In the FLNN classification method, there is a Learning Rate parameter which is very influential on the performance results of a dataset. Referring to the research conducted by Putri [5], Putri explains that the Learning Rate parameter had a role during the training process where Putri used the Learning Rate parameter values of 0.6 and 0.01. Based on the experiment, Putri explains that the Learning Rate parameter of 0.6 is able to provide more optimal performance results on colon cancer and leukemia cancer data.

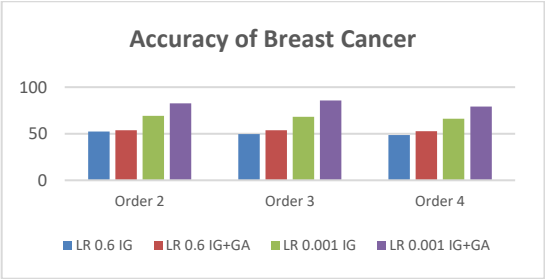

Figure 10. Accuracy of Breast Cancer

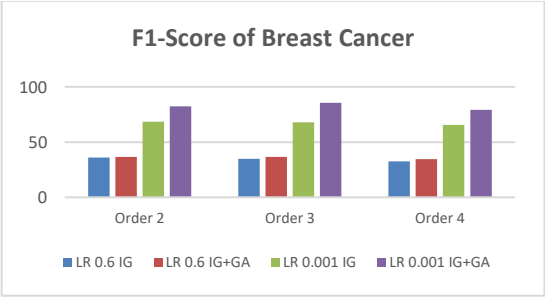

Figure 11. F1-Score of Breast Cancer

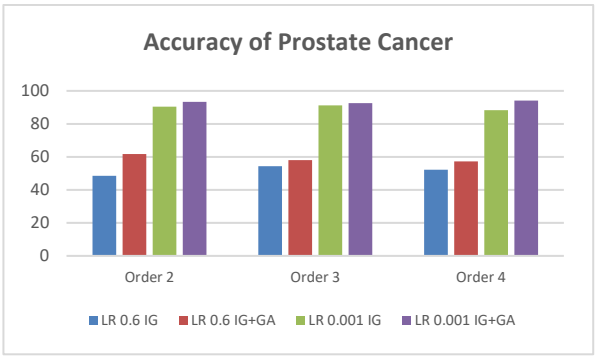

Figure 12. Accuracy of Prostate Cancer

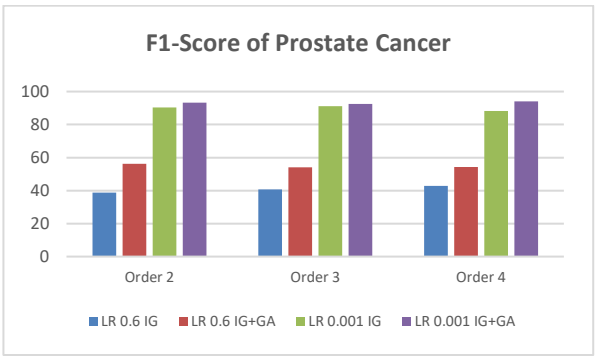

Figure 13. F1-Score of Prostate Cancer

In contrast to the test conducted by Putri, author uses five datasets which include colon cancer, breast cancer, lung cancer, ovarian cancer, and prostate cancer. Based on various tests on the five datasets, author gets different results like what Putri gets [5] where colon cancer tends to have more optimal performance results using the LR of 0.6 compared to the LR of 0.01. In the colon, breast, ovarian, and prostate cancer datasets, author gets the optimal performance results using the LR parameter of 0.001 as respectively seen in Figure 4 and 5, Figure 8 and 9, Figure 10 and 11, and lastly Figure 12 and 13. As for lung cancer, it tends to be more optimal with the use of the LR parameter of 0.6 as seen in Figure 6 and 7. The difference in results on the colon cancer dataset with Putri [5] in the colon cancer dataset can be caused by differences in the used parameters in the used FLNN algorithm. Attached in Figures 12 and 13, the performance results from prostate cancer data are able to increase significantly with the use of the LR of 0.001 compared to the LR of 0.6. It is the same for breast cancer and ovarian cancer data, although the increase is not as significant as the prostate cancer data. According to [5], the difference in performance results obtained can occur due to differences in the characteristics possessed by each dataset so that determining the value of the LR parameter is one aspect that should be considered in the FLNN classification model because it greatly affects the

DOI: https://doi.org/10.29207/resti.v5i4.3352

performance of the neural network in achieving the expected results. Determining the value of LR will have an impact on the performance of backpropagation learning where LR is the parameter used in the process of updating the weights for each input. If the LR value is too small, the training process will take longer because the steps to reach the minimum point of the loss function will be smaller, while if the LR is too large, the training process will be divergent. The LR value that is too large can also cause a very large weight change so that the optimizer can worsen the loss value.

4. Conclusion

Based on the test that have been conducted on the five datasets that the author uses, the author is able to obtain a cancer detection using the hybridization method when reducing dimensions where Information Gain and Genetic Algorithm can optimize the performance results and the required time consumption. The use of Information Gain serves to optimize the consumption of computational time by taking the best 100 features based on the ranking that has been done, while the use of Genetic Algorithm functions to optimize the results of data performance that has been previously selected by Information Gain. From a series of test scenarios that have been conducted, author finds that the value of the Learning Rate parameter has a major influence on performance results where LR of 0.6 is able to provide optimal values for lung cancer data with the highest accuracy values of 100%. In contrast to lung cancer, colon cancer, breast cancer, ovarian cancer and prostate cancer datasets have the highest accuracy values of 90.26%, 85.63%, 100% and 94.10%. The increase in the Legendre Polynomial's order referring to [5] and [11] cannot guarantee that it will increase the performance value and tends to increase the input space so that it will require a longer computation time. From the obtained results, it can be concluded that the use of hybridization method is able to optimize the performance result of FLNN model and the consumption of computational time whereas Learning Rate is a hyperparameter which the optimal value can be obtained by trying different values and see which one gives the best loss without sacrificing speed of training model.

In future research, it can be done by changing the combination of dimensional reduction methods used during hybridization like t-test and Genetic Algorithm as recommended by Mukesh Kumar [4] or by optimizing the parameters used in the FLNN algorithm.

References

[1] WHO, *Who report on cancer: setting priorities, investing wisely and providing care for all*. 2020.

[2] Adiwijaya., Wisesty, U.N., Lisnawati, E., Aditsania, Annisa., & Kusumo, D.S. (2018). "Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification". *Journal of Computer Science*.

[3] Adiwijaya. (2018). "Deteksi Kanker Berdasarkan Klasifikasi Microarray Data". *Media Informatika Budidarma*.

[4] M. Kumar, S. Singh, and S. K. Rath, "Classification of Microarray Data using Functional Link Neural Network," *Procedia Comput. Sci.*, vol. 57, no. March 2016, pp. 727–737, 2015, doi: 10.1016/j.procs.2015.07.463.

[5] Ramadhani, P.T., Wisesty, U.N., Aditsania, Annisa. "Deteksi Kanker Berdasarkan Klasifikasi Data Microarray Menggunakan Functional Link Neural Network dengan Seleksi Fitur Genetic Algorithm". *Indones. J. Comput.*, vol. 2, no. 2, p. 11, 2017, doi: 10.21108/indojc.2017.2.2.173.

[6] B. Peryoga, A. Adiwijaya, and W. Astuti, "Deteksi Kanker Berdasarkan Data Microarray Menggunakan Metode Naïve Bayes dan Hybrid Feature Selection," *J. Media Inform. Budidarma*, vol. 4, no. 3, p. 486, 2020, doi: 10.30865/mib.v4i3.2096.

[7] B. Pradana and A. Aditsania, "Implementasi Minimum Redundancy Maximum Relevance (MRMR) dan Genetic Algorithm (GA) untuk Reduksi Dimensi pada Klasifikasi Data Micorarray Menggunakan Functional Link Neural Network (FLNN),". *e-Proceeding of Engineering*., vol. 6, no. 2, pp. 8966–8977, 2019.

[8] Sahoo, D.M., Chakraverty, S. (2017). "Functional Link Neural Network Approach to Solve Structural System Identification Problems". *The Natural Computing Applications Forum* 2017.

[9] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data". *Adv. Bioinformatics*, 2015.

[10] Erick Cantu-Paz. "Feature subset selection, class separability, and genetic algorithms". In *Genetic and Evolutionary Computation – GECCO* 2004, pages 959–970. Springer, 2004.

[11] Priyono, Iyon., Adiwijaya, & Aditsania, Annisa. "Cancer Detection based on Microarray Data Classification Using Principal Component Analysis and Functional Link Neural Network". *Journal of Data Science and Its Application*, 2020.

[12] W. Astuti and A. Adiwijaya, "Principal Component Analysis Sebagai Ekstraksi Fitur Data Microarray Untuk Deteksi Kanker Berbasis Linear Discriminant Analysis". *J. MEDIA Inform. BUDIDARMA*, 2019.

[13] N. Almugren and H. Alshamlan, "A survey on hybrid feature selection methods in microarray gene expression data for cancer classification," *IEEE Access*. 2019.

[14] H. Salem, G. Attiya, and N. El-Fishawy, "Classification of human cancer diseases by gene expression profiles," *Appl. Soft Comput. J.*, 2017.

[15] C. Arun Kumar, M. P. Sooraj, and S. Ramakrishnan, "A Comparative Performance Evaluation of Supervised Feature Selection Algorithms on Microarray Datasets," in *Procedia Computer Science*, 2017.

[16] Jinyan Li, Kent-ridge bio-medical data repository, School of Computer Engineering Nanyang Technology University, Singapore, Downloaded at January from URL: https://leo.ugr.es/elvira/DBCRepository/.

DOI: https://doi.org/10.29207/resti.v5i4.3352