OPEN ACCESS
J DATA SCI APPL, VOL. 3, NO. 2, PP.048-056, JULY 2020

E-ISSN 2614-7408
DOI: 10.34818/JDSA.2020.3.62

JOURNAL OF DATA SCIENCE AND ITS APPLICATIONS

Comparative Analysis of Support Vector Machine-Recursive Feature Elimination and Chi-Square on Microarray Classification for Cancer Detection with Naïve Bayes

Talitha Kayla Amory[1], Adiwijaya[2], Widi Astuti[3]

School of Computing, Telkom University
Bandung, Indonesia

[*]adiwijaya@telkomuniversity.ac.id

Received on 25-04-2020, revised on 05-08-2020, accepted on 09-06-2021

Abstract

Cancer is a world-famous deadly disease. According to the World Health Organization (WHO), cancer is the second leading cause of death globally and is responsible for an estimated 9.6 million deaths in 2018. One well-known technique for cancer detection is the DNA microarray technique. DNA microarray technology provides an opportunity for researchers to analyze thousands of gene expression profiles at the same time to determine whether a person has cancer or not. However, one of the problems in DNA microarray data is the large number of features that require feature selection. To overcoming these problems, this study will use the feature selection Support Vector Machine-Recursive Feature Elimination (SVM-RFE) and Chi-Square and use the Naïve Bayes classification method. The accuracy between using the two feature selection methods will be compared to find which feature selection method is better when combined with the Naïve Bayes classification method and become the best solution for the problem. The best accuracy results obtained were 100% lung cancer data with SVM-RFE, 99.6% ovarian cancer with SVM-RFE, 93.7% breast cancer with SVM-RFE, and 90% colon cancer with SVM- RFE.

Keywords: Cancer, Microarray, Feature Selection, Support Vector Machine-Recursive Feature Elimination (SVM-RFE), Chi-Square, Naïve Bayes.

I. INTRODUCTION

CANCER is a world-famous deadly disease. According to the World Health Organization (WHO), cancer is the second leading cause of death globally and is responsible for an estimated 9.6 million deaths in 2018 [1]. Fast and accurate cancer detection is needed so that cancer can be handled and handled appropriately.

One of the well-known techniques for cancer detection is the DNA microarray technique. DNA microarray is a collection of thousands of microscopic DNA in the form of DNA fragments that are placed on a chip. The DNA microarray technology provides the opportunity for researchers to analyze thousands of gene expression profiles simultaneously. By analyzing gene expression profiles through a classification process, it can be seen whether a person has cancer or not. The obstacle experienced in the processing of gene expression profiles is that DNA microarray data contain many insignificant and irrelevant features that affect the classification

process and results. Meanwhile, getting an accurate model requires sample data and informative features [2]. Therefore, the relevant features must first be determined [3]. One way to overcome this is by performing feature selection on the data. Feature selection is made to select important and relevant features of the data and remove features that do not affect the cancer classification process.

This study will analyze the effect of feature selection on DNA microarray data to determine whether feature selection can improve gene expression classification accuracy and to find the best combination method. The feature selection method that will be used is the Support Vector Machine-Recursive Feature Elimination (SVM-RFE) and Chi-Square. A classification method was also chosen to classify the DNA microarray gene expression, namely Naïve Bayes. The selection of these methods is based on previous studies which state that these methods have a good performance in the classification of DNA microarray gene expression. The data used are colon cancer, breast cancer, lung cancer, and ovarian cancer from Kent-Ridge Biomedical [4]. The cancer classification performance test compares data classification accuracy with and without feature selection in the classification method. The accuracy between using the two feature selection methods will also be compared to find which feature selection method is better when combined with the Naïve Bayes classification method. To get an overall picture of the performance comparison, this study also considers precision, recall, and F1-score.

The rest of the paper is organized as follows. The literature review is presented in Sect. 2. The research method is presented in Sect. 3. Section 4 presents the results and discussion. Section 5 concludes the paper.

II. LITERATURE REVIEW

Several studies on DNA microarray data for cancer classification implement the Chi-Square and SVM-RFE feature selection methods. Research conducted by Omara et al. [5] aims to explain the effect of feature selection on the accuracy of classification methods widely used in cancer classification research. This study evaluates the selection methods for Information Gain, Chi-Square, mRMR, Linear Correlation, and Random Forest Selector; and evaluating their effects on the classification methods SOM, K-NN, K-means, and Random Forest. This study found that when Chi-Square is combined with K-means, it will produce 100% accuracy for lymphoma cancer data. An excellent accuracy also occurs when using the Chi-Square method combined with the Random Forest method, which is 98.63% for leukemia data.

Babu and Sarkar [6] conducted a study using the Chi-Square and SVM-RFE feature selection methods. This study aimed to develop a system for classifying cancer more accurately using DNA microarray data. This experiment compares the accuracy of the T-test, Chi-Square, Information Gain, Relief-F, SVM-RFE, and mRMR feature selection methods combined with the SVM and K-NN classification methods. The results obtained are that the SVM-RFE, combined with the SVM and K-NN methods, produces an accuracy of 90% in colon cancer data. In the leukemia data, the highest accuracy value is 97% using the Chi-Square method combined with SVM.

Research conducted by Rustam and Kharis [7] aims to compare the accuracy results between the Gaussian Kernel and SVM-RFE feature selection methods combined with the SVM classification method. From the test results, it can be seen that the accuracy of using the SVM-RFE method reaches 96.7899% for lung cancer data. This percentage is slightly superior to the SVM method without using feature selection and SVM using the Gaussian Kernel.

Several studies have also used the Naïve Bayes classification method for cancer classification using DNA microarray data. Sharbaf et al. [8] researched by combining the Fisher Criterion, CLA, and ACO feature selection methods combined with the K-NN, SVM, and Naïve Bayes classification methods. The accuracy rate obtained when using the Naïve Bayes classification for leukemia and prostate cancer data is 100%.

Research conducted by Nurviarelda et al. [9] aims to carry out the classification process using the feature selection method of the Daubechies4 family of Discrete Wavelet Transform (DWT) and the Naïve Bayes classification method, which will later be compared using the mRMR feature selection method. The highest accuracy results from the db4 method are found in ovarian cancer data, which is 98.41%.

III. RESEARCH METHOD

A. System Description

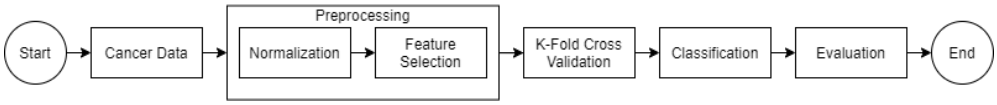

Fig. 1. Flowchart of The System

Judging from the flowchart above, the first step is the preparation of cancer data. The second stage of this final project is preprocessing. At this stage, two preprocessing processes will be carried out, namely data normalization and feature selection. Feature selection will be carried out using the Support Vector Machine-Recursive Feature Elimination (SVM-FRE) and Chi-Square methods to reduce the number of features in the data and select features that affect the dataset. The third stage of this final project is the classification process. Data that has gone through the feature selection process will then be classified using Naïve Bayes. The cancer classification performance test compares data classification accuracy with and without feature selection in the classification method. The accuracy between using the two feature selection methods will also be compared to find which feature selection method is better when combined with the Naïve Bayes classification method. The precision, recall, and F1-score results will also be analyzed to get a picture of the overall performance.

B. Cancer Data

The data used in this final project is DNA microarray data from Kent Ridge Biomedical Dataset [4]. The DNA microarray data consists of ovarian cancer, lung cancer, breast cancer, and colon cancer. The dataset specifications used can be seen in Table I.

TABLE I
DNA MICROARRAY DATASET SPECIFICATIONS [4]

Data	Number of Classes	Number of Features	Number of Samples
Ovarian cancer	2 (normal, cancer)	15154	252
Lung cancer	2 (mesothelioma, ADCA)	12533	181
Breast cancer	2 (relapse, non-relapse)	24482	97
Colon tumor	2 (negative, positive)	2000	62

C. Preprocessing

The preprocessing process corrects problems that arise in the data processing. Such as data with too many features and the high difference in range values in each feature. These problems can cause the results of data processing to be less good or not optimal.

1) Normalization

Normalization is carried out so that the data training process becomes faster or improves the classification model's performance and helps the model understand the classification process's data. The method used in this final project is Min-Max Normalization. The Min-Max Normalization method is a method of changing complex data without eliminating the contents, making it easier to process [10]. Min-Max Normalization will provide a feature value between 0 and 1. The formula used in normalization can be seen in equation (1).

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}(newX_{max} - newX_{min}) + newX_{min} \quad (1)$$

Based on this equation, X 'is the new feature value in the normalization domain, X is the feature value before the normalization process, X_{min} is the lowest feature value in the normalized data, and X_{max} is the highest feature value in the normalized data. $newX_{max}$ is the highest range value and $newX_{min}$ is the lowest range value.

2) Feature Selection with SVM-RFE

Support Vector Machine-Recursive Feature Elimination (SVM-RFE) is a method of reducing data dimensions. This algorithm was first introduced by jokes in his research entitled "Gene Selection for Cancer Classification using Support Vector Machines" [11]. SVM-RFE proved to be one of the best feature selection methods, ranking features by practicing the SVM classification method and recursively eliminating it with the least weight in each iteration [12]. The feature weight can be determined by calculating the weight w shown in equation (2).

$$\boldsymbol{w} = \sum_{k=0}^{n} \alpha_k y_k \boldsymbol{x_k} \quad (2)$$

Where α_k is the SVM classification result from the training data, y_k is the label class, and x_k is the training data. SVM has four kernels, namely linear, polynomial, RBF, and sigmoid. This study uses SVM with a linear kernel because it works best on small and large features compared to other kernels [13].

3) Feature Selection with Chi-Square

Another feature selection option that is commonly used is the Chi-Square (X^2) statistical method. Chi-Square is a statistical test to determine two events' dependence (in feature selection, the two events are feature occurrence and class occurrence). The process consists of a Chi-Square calculation between each feature X and labels Y. If dependent, this feature will be used in the training model [14]. Chi-Square calculations can be seen in equation (3).

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{3}$$

Where Oi is the observed frequency, and E_i is the expected frequency. E_i can be calculated by equation (4).

$$E_i = \frac{row\ total\ x\ column\ total}{sample\ size} \tag{4}$$

A higher Chi-Square value indicates that the feature is more informative, so it should be chosen for the training model [5].

D. K-Fold Cross Validation

This study will use the K-Fold Cross Validation technique in dividing the dataset used into train and test data. The split train and test data will start with the first part being the test data and the other part being the train data. The next split data will use the second part for test data and the other part for train data. This split is repeated until the k-part becomes the test data and the other part becomes the train data. This study uses K = 10, which is usually referred to as 10-Fold Cross-Validation [15].

E. Classification with Naïve Bayes

Naïve Bayes is a classification method based on probability and the Bayesian Theorem, assuming that each variable X is independent (independence). In general, the Bayesian Theorem is in equation (5).

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \tag{5}$$

The above equation (5) is used to calculate each class C's probability-based on condition X, which is the probability of each particular class feature.

The three popular types of Naïve Bayes are Gaussian, Multinomial, and Bernoulli. The Naïve Bayes type used in this study is Gaussian, which is better for data with continuous feature values [16]. The likelihood value can be measured using equation (6).

$$P(X_j|C = C_i) = \frac{1}{\sigma_{ji}\sqrt{2\pi}} exp\left(-\frac{(X_j - \mu_{ji})^2}{2\sigma^2{}_{ji}}\right) \tag{6}$$

The value of μji is the mean of X_j with class = c_i, and σ_{ji} is the standard deviation of X_j with class = c_i [8].

F. System Evaluation

In this study, the feature selection method was introduced to reduce the dimensions in gene expression DNA microarray data classification into the specified cancer class types. The proposed performance measures are calculated using accuracy. Accuracy refers to how often a classifier predicts the correct class [17]. To obtain

accuracy, the test data prediction results from the classification model will be compared with the actual test data label [18]. The number of correct predictions will be divided by the total number of data and multiplied by 100%.

The accuracy results obtained from the proposed method will be compared to see whether feature selection is better than the classification without using feature selection. The accuracy between the two feature selection methods will also be compared to determine which feature selection method produces better performance when combined with the Naïve Bayes classification method. To get an overall picture of the performance comparison, this study also considers precision, recall, and F1-score. Precision refers to the presentation of how many of the correct predictions are positive of the total predictions. Recall refers to how the model predicts positive, and in fact, it is indeed positive data. The F1-score represents a balance between precision and recall.

IV. RESULTS AND DISCUSSION

This study uses four datasets, as shown in Table I. Cancer data will go through the feature selection stage using SVM-RFE and Chi-Square to reduce the number of features in the data. After feature selection is carried out, the classification model will be implemented using the Naïve Bayes method with the Gaussian type. The K-Fold Cross Validation technique will be used to share training and testing data. This research will use cross-validation with K = 10.

A. Performance of The Proposed Method

1) Performance of SVM-RFE and Naïve Bayes

In this scenario, the SVM-RFE feature selection will be used, which will later be combined with the Naïve Bayes classification method. The selection of the number of features is based on previous research [19], where the number of features tested is 5, 10, 50, 100, 200, 500, and 1000. After that, K-Fold Cross Validation will be carried out with K's number to be tested based on the research previously [15], namely K = 10.

TABLE II
ACCURACY RESULTS OF MICROARRAY DATA USING SVM-RFE AND NAÏVE BAYES

Data	Accuracy						
	Features = 5	**Features = 10**	**Features = 50**	**Features = 100**	**Features = 200**	**Features = 500**	**Features = 1000**
Ovarian	99.2%	99.6%	98.4%	98.8%	96.8%	97.6%	97.4%
Lung	97.3%	98.3%	100%	100%	100%	99.4%	99.4%
Breast	79.3%	85.6%	93.7%	92.7%	92.8%	86.3%	73.3%
Colon	88.3%	90%	88.3%	86.7%	80.2%	72.4%	62.4%

TABLE III
PRECISION RESULTS OF MICROARRAY DATA USING SVM-RFE AND NAÏVE BAYES

Data	Precision						
	Features = 5	**Features = 10**	**Features = 50**	**Features = 100**	**Features = 200**	**Features = 500**	**Features = 1000**
Ovarian	70%	70%	70%	70%	70%	70%	70%
Lung	100%	100%	100%	100%	100%	99.4%	99.4%
Breast	55%	52.7%	60%	60%	60%	56.7%	56.7%
Colon	71.7%	71.7%	71.7%	70%	55.8%	50.7%	43%

TABLE IV
RECALL RESULTS OF MICROARRAY DATA USING SVM-RFE AND NAÏVE BAYES

Data	Recall						
	Features = 5	**Features = 10**	**Features = 50**	**Features = 100**	**Features = 200**	**Features = 500**	**Features = 1000**
Ovarian	69.6%	69.6%	68.4%	68.8%	66.8%	67.6%	67.2%
Lung	96.8%	98.1%	100%	100%	100%	100%	100%
Breast	45.7%	50.7%	55.7%	57%	53.7%	53.7%	41.7%
Colon	60%	65%	61.7%	61.7%	65%	65%	61.7%

TABLE V
F1- SCORE RESULTS OF MICROARRAY DATA USING SVM-RFE AND NAÏVE BAYES

Data	F1-Score						
	Features = 5	**Features = 10**	**Features = 50**	**Features = 100**	**Features = 200**	**Features = 500**	**Features = 1000**
Ovarian	69.8%	69.8%	69.2%	69.4%	68.3%	68.7%	68.5%
Lung	98.3%	99%	100%	100%	100%	99.7%	99.7%
Breast	49.2%	50.9%	57.5%	58.4%	56.4%	55%	47.1%
Colon	63.3%	66.7%	64.7%	63.3%	58.9%	55.4%	49.1%

The accuracy of the DNA microarray data using the SVM-RFE and Naïve Bayes methods is shown in Table II. Based on Table II, it was found that the lung cancer and ovarian cancer dataset obtained better accuracy values compared to other datasets on each number of features. An accuracy value of 100% is obtained by lung cancer data using 50, 100, and 200 features. In the ovarian cancer data, the greatest accuracy value was 99.6% obtained when using ten features. Colon cancer data gets the highest accuracy value of 90% when using ten features. In the breast cancer data, the greatest accuracy value is 93.7% obtained when using 50 features.

The highest precision results are also owned by lung cancer data with a precision level of up to 100%. Ovarian cancer data has a stable precision value from the use of 5 features to 1000 features, which is 70%. The best recall and F1-Score results are held by lung cancer data which can reach 100%

2) Performance of Chi-Square and Naïve Bayes

In this scenario, the Chi-Square feature selection will be used, which will later be combined with the Naïve Bayes classification method. The selection of the number of features is based on previous research [19], where the number of features tested is 5, 10, 50, 100, 200, 500, and 1000. After that, K-Fold Cross Validation will be carried out with K's number to be tested based on the research previously [15], namely K = 10.

TABLE VI
ACCURACY RESULTS OF MICROARRAY DATA USING CHI-SQUARE AND NAÏVE BAYES

Data	Accuracy						
	Features = 5	**Features = 10**	**Features = 50**	**Features = 100**	**Features = 200**	**Features = 500**	**Features = 1000**
Ovarian	96.5%	95.7%	96.5%	96.9%	96.5%	92.5%	92.2%
Lung	99.4%	97.8%	98.3%	98.9%	100%	99.4%	98.9%
Breast	56.6%	58.8%	60.8%	62.9%	61.9%	62.9%	62.9%
Colon	85%	86.7%	85%	81.9%	79.3%	69.3%	62.6%

TABLE VII
PRECISION RESULTS OF MICROARRAY DATA USING CHI-SQUARE AND NAÏVE BAYES

Data	Precision						
	Features = 5	**Features = 10**	**Features = 50**	**Features = 100**	**Features = 200**	**Features = 500**	**Features = 1000**
Ovarian	70%	70%	70%	70%	70%	70%	70%
Lung	100%	100%	100%	100%	100%	98%	96%
Breast	40%	40%	40%	40%	40%	40%	40%
Colon	70%	70%	64.2%	56.7%	54.7%	50.3%	44.3%

TABLE VIII
RECALL RESULTS OF MICROARRAY DATA USING CHI-SQUARE AND NAÏVE BAYES

Data	Recall						
	Features = 5	**Features = 10**	**Features = 50**	**Features = 100**	**Features = 200**	**Features = 500**	**Features = 1000**
Ovarian	68%	67.2%	68.4%	68.8%	68.4%	63.8%	64.2%
Lung	99.4%	97.7%	98.3%	98.9%	100%	100%	100%
Breast	9.30%	9.30%	14.7%	15.7%	14.7%	15.7%	15.7%

Colon	56.7%	61.7%	65%	70%	73.3%	70%	66.7%

TABLE IX
F1-SCORE RESULTS OF MICROARRAY DATA USING CHI-SQUARE AND NAÏVE BAYES

Data	F1-Score						
	Features = 5	**Features = 10**	**Features = 50**	**Features = 100**	**Features = 200**	**Features = 500**	**Features = 1000**
Ovarian	69%	68.5%	69.2%	69.4%	69.2%	66.3%	66.5%
Lung	99.7%	98.8%	99.1%	99.4%	100%	98.9%	97.8%
Breast	14.8%	15%	20.4%	21.7%	20.1%	21.2%	21.2%
Colon	59%	62.3%	61.9%	61.1%	61.4%	56.6%	51.6%

The accuracy of the microarray DNA data using the Chi-Square and Naïve Bayes methods is shown in Table VI. Based on Table VI, it is found that the lung cancer dataset gets a better accuracy value compared to other datasets on each feature count test. An accuracy value of 100% is obtained by lung cancer data when using 200 features. In the ovarian cancer data, the greatest accuracy value is 96.9% obtained when using 100 features. Colon cancer data gets the highest accuracy score of 86.7% when using ten features. However, the model still does not provide good accuracy in breast cancer data. The Chi-Square Feature Selection cannot handle breast cancer data where this data has the highest data complexity compared to other data.

The highest precision results are also owned by lung cancer data with a precision level of up to 100%. Ovarian cancer data has a stable precision value from the use of 5 features to 1000 features, which is 70%. The best recall and F1-Score results are held by lung cancer data which can reach 100%.

3) Performance Naïve Bayes without Feature Seelction

In this scenario, DNA microarray data classification does not use the feature selection method. The type of Naïve Bayes to be used is Gaussian, which, based on research [18], will produce better performance when the data type is continuous compared to other Naïve Bayes types. Like other methods, K-fold Cross-Validation will be performed with the number of K to be tested based on previous research [15], namely K = 10.

TABLE X
RESULTS OF MICROARRAY DATA USING NAÏVE BAYES WITHOUT FEATURE SELECTION

Data	Accuracy	Precision	Recall	F1-Score
Ovarian	88.6%	70%	63.3%	63.3%
Lung	97.3%	94.1%	99.4%	99.4%
Breast	54%	35%	15.3%	15.3%
Colon	56.2%	38.8%	55.8%	55.8%

The accuracy of the DNA microarray data test using the Naïve Bayes method without feature selection is shown in Table X. Based on Table X, it is found that the lung cancer dataset has a better accuracy value than other datasets, namely 97.3%. In the data on ovarian cancer, the accuracy value obtained is 88.6%. Colon cancer data got an accuracy value of 56.2%, and breast cancer data got an accuracy value of 54%.

The highest precision results are also owned by lung cancer data with a precision level of up to 94.1%. The best recall and F1-Score results are held by lung cancer data which can reach 99.4% for recall and 96.3% for F1-Score.

B. Performance Analysis of The Proposed Method

From the data listed above, the accuracy has increased and decreased with each use of different features. However, this does not necessarily make this experiment unsuccessful. In [20] research, it was stated that feature selection is the process of selecting a small part of the features that are ideally needed and sufficient to produce optimal performance. From this statement, it can be concluded that the number of features is the optimal number of features when the accuracy increases. When the accuracy decreases, the number of features is less than optimal. The number of features that get the highest accuracy value in the proposed methods also depends on how the feature's influence level is selected on the cancer data.

The accuracy results shown in Table II, IV, and X shows the significant effect of feature selection on the classification. The better accuracy results can be seen compared to the classification without using the feature selection method. From that table, we can observe that all data can get the best accuracy results when using the SVM-RFE feature selection method combined with the Naïve Bayes classification. It should be noted that the suitability between feature selection and classification methods affects the amount of accuracy.

All feature selection works very well in the ovarian cancer data, even though the SVM-RFE has slightly greater accuracy than Chi-Square. In lung cancer data, Chi-Square got better results than SVM-RFE on the use of 5 features. In the breast cancer data, only SVM-RFE was able to achieve more than 70% accuracy. Chi-Square is still unable to handle breast cancer data. This data has the highest data complexity than other data because the microarray-based breast cancer dataset consists of a small number of samples with a much larger number of features than other cancer datasets. [21]. The microarray-based breast cancer dataset had a very small sample because it was very time-consuming and expensive [22]. Apart from being very time-consuming and expensive, the sample size is further reduced when predicting clinical outcomes due to missing clinical factors for some of the microarray dataset samples [23]. Even though SVM-RFE reaches 90% for ten features in the colon cancer data, the accuracy is still somewhat lower than other datasets. This is due to the smaller number of colon data compared to other data.

In conclusion, lung cancer data can produce the best accuracy by using the SVM-RFE method with an accuracy of 100% (100% precision, 100% recall, and 100% F1-Score) on the use of 50, 100, and 200 features; and using the Chi-Square method with an accuracy of 100% (100% precision, 100% recall, and 100% F1-Score) using 200 features.

V. Conclusion

Based on the results of the research conducted, the conclusions that can be drawn from this study are as follows. The SVM-RFE and Chi-Square feature selection methods combined with the Naïve Bayes classification method can classify DNA microarray data for cancer detection. SVM-RFE and Chi-Square can affect the classification accuracy of the Naïve Bayes method. SVM-RFE and Chi-Square both obtained 100% accuracy results in lung cancer data for 200 features. However, for breast cancer data, Chi-Square has not handled the data's complexity maximally. Based on the study results, five of the seven features of each cancer dataset get the best accuracy from the Naïve Bayes model combined with SVM-RFE, so it can be concluded that Naïve Bayes combined with SVM-RFE produces better performance than Naïve Bayes combined with Chi. -Square. This research was developed by comparing SVM-RFE and Chi-Square and their effect on Naïve Bayes' accuracy. For further research, we can compare other feature selection methods or other classification methods to determine which method can produce the best accuracy level.

ACKNOWLEDGMENT

The author would like to thank Allah S.W.T, my family who supports me endlessly, the people involved in this research, such as lecturers and friends. The author would also thank all authors whose papers are used as references in this paper.

REFERENCES

[1] "Cancer," World Health Organization. [Online]. Available: https://www.who.int/health-topics/cancer. [Accessed: 17-Nov-2020.
[2] R. K. Singh and M. Sivabalakrishnan, "Feature selection of gene expression data for cancer classification: A review," *Procedia Comput. Sci.*, vol. 50, pp. 52–57, 2015.
[3] H. Ayadenta and Adiwijaya, "A clustering approach for feature selection in microarray data classification using random forest," *Journal of Information Processing Systems*, vol. 14, pp. 1167–1175, Jan. 2018.
[4] Elvira biomedical Dataset Repository, "Kent Ridge Biomedical Data Set Repository," 2005. [Online]. Available: http://leo.ugr.es/elvira/DBCRepository/. [Accessed: 17-Dec-2020].
[5] H. Omara, M. Lazaar, and Y. Tabii, "Effect of Feature Selection on Gene Expression Datasets Classification Accuracy," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 5, p. 3194, 2018.
[6] M. Babu and K. Sarkar, "A comparative study of gene selection methods for cancer classification using microarray data," *2016 Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 2016.
[7] Z. Rustam and S. A. A. Kharis, "Comparison of Support Vector Machine Recursive Feature Elimination and Kernel Function as feature selection using Support Vector Machine for lung cancer classification," *Journal of Physics: Conference Series*, vol. 1442, p. 012027, 2020.
[8] F. V. Sharbaf, S. Mosafer, and M. H. Moattar, "A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization," *Genomics*, vol. 107, no. 6, pp. 231–238, 2016.
[9] R. Nurviarelda., Adiwijaya., and A. A. Rohmawati., "Klasifikasi Data Microarray Menggunakan Discrete Wavelet Transform dan Naives Bayes Classification," *e-Proceeding Eng.*, vol. 5, no. 1, p. 1536, 2018.
[10] A. Manik, A. Adiwijaya, and D. Q. Utama, "Classification of electrocardiogram signals using Principal Component Analysis and Levenberg Marquardt Backpropagation for detection Ventricular Tachyarrhythmia," *J. Data Sci. Appl.*, vol. 2, no. 1, pp. 78–87, 2019.
[11] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, vol. 46, no. 1/3, pp. 389–422, 2002.
[12] Z. Li, W. Xie, and T. Liu, "Efficient feature selection and classification for microarray data," *Plos One*, vol. 13, no. 8, 2018.
[13] A. Goel and S. K. Srivastava, "Role of kernel parameters in performance evaluation of SVM," *2016 Second International Conference on Computational Intelligence & Communication Technology (CICT)*, 2016.
[14] H. Zhang et al., "Informative gene selection and direct classification of tumor based on Chi-square test of pairwise gene interactions," *Biomed Res. Int.*, vol. 2014, p. 589290, 2014.
[15] C. S. R. Annavarapu, S. Dara, and H. Banka, "Cancer microarray data feature selection using multi-objective binary particle swarm optimization algorithm," *EXCLI J.*, vol. 15, pp. 460–473, 2016.
[16] B. M. and C. P., "An automated technique using Gaussian naïve Bayes classifier to classify breast cancer," *Int. J. Comput. Appl.*, vol. 148, no. 6, pp. 16–21, 2016.
[17] M. Nuruddin Qaisar Bhuiyan, M. Shamsujjoha, S. H. Ripon, F. H. Proma, and F. Khan, "Transfer learning and supervised classifier based prediction model for breast cancer," in *Big Data Analytics for Intelligent Healthcare Management*, Elsevier, 2019, pp. 59–86.
[18] R. B. Purnomoputra, A. Adiwijaya, and U. Novia Wisesty, "Sentiment analysis of movie review using Naïve Bayes method with Gini index feature selection," *Journal of Data Science and Its Applications*, vol. 2, no. 2, pp.85-94. 2019.
[19] N. Cilia, C. De Stefano, F. Fontanella, S. Raimondo, and A. Scotto di Freca, "An experimental comparison of feature-selection and classification methods for microarray datasets," *Information (Basel)*, vol. 10, no. 3, p. 109, 2019.
[20] Kira, Kenji, and Larry A. Rendell. "The feature selection problem: Traditional methods and a new algorithm." *Aaai*, vol. 2, 1992.
[21] A. Saini, J. Hou, and W. Zhou, "Breast cancer prognosis risk estimation using integrated gene expression and clinical data," *Biomed Res. Int.*, vol. 2014, p. 459203, 2014.
[22] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 15, pp. 5923–5928, 2006.
[23] M. Shi and B. Zhang, "Semi-supervised learning improves gene expression-based prediction of cancer recurrence," *Bioinformatics*, vol. 27, no. 21, pp. 3017–3023, 2011.