Implementation of Support Vector Machine - Recursive Feature Elimination for MicroRNA Selection in Breast Cancer Classification

by Adi Wibowo

https://jurnaleeccis.ub.ac.id/
p-ISSN : 1978-3345, e-ISSN(Online): 2460-8122

Jurnal EECCIS Vol. 14, No. 1, April 2020
pp 1-5

Implementation of Support Vector Machine - Recursive Feature Elimination for MicroRNA Selection in Breast Cancer Classification

Ratih Permatasri [1], Adi Wibowo [2]
[1,2] Departemen Ilmu Komputer / Informatika, Fakultas Sains dan Matematika, Universitas Diponegoro
Email: permataranit@gmail.com, bowo.adi@live.undip.ac.id

***Abstract*— Breast cancer is the most frequent cancer caused death among women. An attempt to reduce death cases caused by breast cancer, was to detect cancer cells when it still in early stage. MicroRNA is one of the biomarkers for cancer that can be used to detect cancer cell even in its early stage. However, MicroRNA data tends to have thousand types of expression which required a lot of costs if it examined one by one thoroughly. Feature selection method can be used to extract important MicroRNAs that support classification process between normal people and people with breast cancer. Support Vector Recursive Feature Elimination (SVM-RFE) is one of the feature selection method that can be used to select MicroRNA data. This research aims to produce the best smallest subset that contains selected MicroRNA expressions using the SVM-RFE as feature selection method. This experiment result showed that the best selected subset was able to provide 99% classification accuracy with only 3 MicroRNA expressions, where 2 from 3 selected MicroRNA hold potential as a biomarker of breast cancer.**

***Index Terms*— Breast Cancer, Feature Selection, MicroRNA, SVM-RFE.**

***Abstrak*-- Kanker payudara merupakan kanker penyebab kematian terbanyak pada wanita. Salah satu upaya untuk mengurangi jumlah kasus kematian akibat kanker payudara yaitu dengan mendeteksi sel kanker pada saat stadium awal. MicroRNA merupakan salah satu biomarker kanker yang dapat digunakan untuk mendeteksi sel kanker bahkan pada kanker stadium awal. Namun, data MicroRNA cenderung memiliki ribuan jenis ekspresi sehingga membutuhkan biaya yang banyak apabila ditelusuri satu persatu. Metode seleksi fitur dapat digunakan untuk mengekstrak MicroRNA penting yang dapat membantu proses klasifikasi antara orang normal dan orang yang menderita kanker payudara. Support Vector Machine Recursive Feature Elimination (SVM-RFE) merupakan salah satu metode seleksi fitur yang dapat digunakan untuk menyeleksi data MicroRNA. Penelitian ini bertujuan untuk menghasilkan subset kecil terbaik ekspresi MicroRNA yang sudah terseleksi dengan metode seleksi fitur SVM-RFE. Hasil penelitian menunjukkan subset MicroRNA terbaik hasil seleksi mampu memberikan akurasi klasifikasi sebesar 99% dengan hanya 3 ekspresi MicroRNA saja, di mana 2 dari 3 ekspresi MicroRNA tersebut memiliki potensi sebagai biomarker penyakit kanker payudara.**

***Kata Kunci*— Kanker Payudara, Seleksi Fitur, MicroRNA, SVM-RFE.**

I. INTRODUCTION

Breast cancer is one of the most common cancer among women. Every year, there is 2.1 million women affected by this disease. Breast cancer is also included in the greatest number of cancer-caused deaths among women [1]. One of the attempts to suppress death rates caused by breast cancer is to identify early staged cancer cells. This is due to the most critical point for best prognosis is to detect cancer cells when it still in early stage [2].

Many studies were developed to detect early staged cancer cells. One of the studies used MicroRNA expressions as a tool to diagnose early-stage breast cancer [3]. Since its discovery, MicroRNA have proven to be an important and essential layer in gene regulation, especially in post transcriptional regulation. MicroRNAs carry information about patho-physiological state of a human, and so can be employed as biomarkers [4]. Numerous studies have demonstrated that MicroRNAs are not only found intracellularly, but also can easily found in outside cells, including various body fluids as example: serum, plasma, saliva, urine, breast milk, and tears [5]. However, amongst the benefit of MicroRNA as a promising candidate for biomarker, MicroRNA has a downside which is it has a thousand of expressions so it will cost a lot if it is examined thoroughly.

Feature Selection (FS) is a systematical process to reduce the dimensionality of dataset thus can produce an optimal subset for classification purpose [6]. In cancer classification, Feature Selection can be used to extract important MicroRNAs (called MicroRNA marker) that effectively have the impact on classification accuracy. Irrelevant or redundant MicroRNA expressions will be eliminated by Feature Selection thus increase the performance of classification model [7].

In 2002, [8] exploit Support Vector Machine (SVM) method to select genes which would be used to classify cancer. The developed method uses ranking criterion from SVM coefficient to evaluate gene expressions and recursively eliminate genes that is not satisfying the criterion. Later, this method is called Support Vector Machine Recursive Feature Elimination (SVM-RFE). In this research we applied Support Vector Machine - Recursive Feature Elimination (SVM-RFE) for selecting MicroRNA expressions which is used in breast cancer classification.

Manuscript submitted on September 2019, accepted and published on April 2020

II. METHODS

A. Min-Max Normalization

Variables tend to have a diverse range. Some of the variables even have big gap ranges from each other. Such big differences in the ranges might cause a problem for some classification algorithms. It will lead to a tendency for the variable with greater range to have undue influence on the results [9]. Therefore, researchers should normalize their numerical variables, to standardize the scale of effect each variable has on the results [9]. There are several techniques for normalization, and one of the most widely used is Min Max Normalizations. The said normalization techniques preserves the relationship among the original data values [10]. In this research we used [0,1] as interval value, so min max normalization is calculated by the following formula [10]:

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

where:
X' = normalized value
X = original value
min(X) = the fewest value in original range
max(X) = the largest value in original range

B. Sequential Support Vector Machine (Sequential SVM)

Solving Quadratic Programming (QP) problem in SVM tend to get very complex, time consuming, and prone to numerical instabilities. In 1999, [11] propose a sequential learning method for SVM. [11] modifies the formulation of the bias from SVM to generate a fast and simple implementation of SVM which optimize margin for QP problem in high-dimensional feature space. This method will be referred as Sequential SVM in the future discussion. Steps for Sequential SVM can be done as follows [11]:

1) Initialize values for learning rate (γ), lambda (λ), cost constant (C), error target (ε), and maximum *epoch* (max *epoch*).
2) Initialize $\alpha_i = 0$ for i = [1,2, 3,,,.. l];
3) Compute matrix D_{ij}.
$$D_{ij} = y_i y_j \left(K(x_i, x_j) + \lambda^2\right); \quad (2)$$
4) Do step (a), step (b), and step (c) in a row:
   a) $E_i = \sum_{j=1}^{l} \alpha_j D_{ij}$; (3)
   b) $\delta\alpha_i = \min(\max(\gamma(1 - E_i), -\alpha_i), C - \alpha_i)$; (4)
   c) $\alpha_i = \alpha_i + \delta\alpha_i$; (5)
5) If the training has converged which is achieved when $\max(\delta\alpha_i) < \varepsilon$ or when *epoch* = max *epoch*, then stop else go back to step 4.

Notes:
y = target data
x = training data
K(x_i,x) = *kernel function*
α = *Lagrange multiplier*
l = the total amount of data

Testing process is done subsequently after training process is complete. The classifier function for Sequential SVM as bellow [11]:

$$f(x) = sign\left(\sum_{i \in SV} \alpha_i y_i K(x_i, x) + \alpha_i y_i \lambda^2\right) \quad (6)$$

where:
x_i = training data
x = testing data
SV = the total amount of *support vector*

C. Support Vector Machine - Recursive Feature Elimination (SVM-RFE)

SVM-RFE is one of the examples for backward elimination procedure implementation. A subset consist of features will be selected by removing one feature variable that least important at a time [8]. At each step, the coefficients of the weight vector of SVM are used to compute the feature ranking score. The fewest feature ranking score $c = (w_i)^2$, where w_i represents the corresponding component in the weight vector w, will be eliminated. More clearly, steps for SVM-RFE can be done as follows [12]:

1) Start: ranked feature $list\ r = [\,]$; and $list\ s = [1, \ldots, d]$;
2) Repeat until all features are ranked or *list s* = []:
   a) Train a linear SVM with features in *list s* as input variables $\alpha_i = SVMTrain(x)$; (7)
   b) Compute the weight vector w
   $$w = \sum_{k=1}^{n} \alpha_k y_k x_k; \quad (8)$$
   c) Compute the ranking scores for features in *list s*
   $$c = (w_i)^2; \quad (9)$$
   d) Find the feature with the smallest ranking score
   $$f = argmin(\boldsymbol{c}); \quad (10)$$
   e) Update *list r*
   $$r = [s(f), r]; \quad (11)$$
   f) Eliminate feature with the smallest ranking score from *list s*
   $$s = s - [s(f)]; \quad (12)$$
3. Output: ranked feature *list r*.

Notes:
x = data with adjusted feature from *list s*
d = the original total amount of feature in the dataset

D. K-Fold Cross-Validation

K-Fold Cross-Validation is the basic form of Cross-Validation. Cross-Validation is a statistical method used to evaluate learning algorithms. Cross-Validation divides dataset into two sections, one used to train a model and the other used to validate the model [13]. The training and validation sets must cross-over in successive rounds such that each data point has a chance of being the test set [13]. In K-Fold Cross-Validation, the dataset divided into K equally (or nearly equally) sized subset or folds. Then K iterations of training and validation are performed by treating the Kth fold as the validation set on the Kth iteration while the remaining folds are played as training set.

In data mining and machine learning 5-fold cross-validation (k = 5) and 10-fold cross-validation (k = 10) is the most common [14]. Figure 1 demonstrates an example with k = 10. The stripes blocks are subset data used for testing while the solid blocks are used for training.

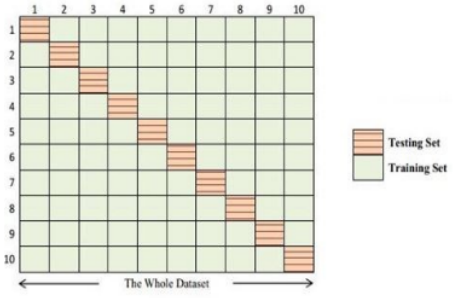

Fig. 1. Illustration for K-Fold Cross Validation with k = 10

E. Confusion Matrix and ROC Curve

Confusion matrix can be used to evaluate the correctness of classification. Confusion matrix for binary case shown in Table I is constructed by True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN). True Positives is the number of correctly recognized examples from positive class while True Negatives is the number of correctly recognized examples from negative class. False Positives is the number of incorrectly assigned examples to the positive class (where it should be assigned to the negative class) and False Negatives is the number of incorrectly assigned example to the negative class (where it should be assigned to the positive class) [15].

TABLE I
CONFUSION MATRIX FOR BINARY CASE

Actual	Prediction / Classification	
	Positive	Negative
Positif	TP	FN
Negatif	FP	TN

Accuracy, Sensitivity, and Specificity are some measurement that can be derived from the confusion matrix:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \; x \; 100\% \quad (13)$$

$$Sensitivity = \frac{TP}{TP+FN} \; x \; 100\% \quad (14)$$

$$Specificity = \frac{TN}{TN+FP} \; x \; 100\% \quad (15)$$

Area Under the Curve (AUC) and Receiver Operating Characteristic (ROC) curve is a performance measurement for classification problem at various thresholds settings [16]. It tells how much model is capable of distinguishing between classes. Higher the AUC means better the model is at predicting *Xs* as *Xs* and *Ys* as *Ys* [15].

III. RESULT AND DISCUSSION

The dataset used in this paper are gathered from National Cancer Institute Genomic Data Commons that can be accessed from URL http://gdc.cancer.gov/. The dataset consists of MicroRNA expression quantification from normal solid tissue and primary tumor sample in a breast cancer case. There are 1881 feature profiles of 248 samples divided into two classes namely cancer and normal tissue. Because of the large amounts of MicroRNA data, it would take a lot of time and cost to detect the most important expression. Feature selection can remove MicroRNA expression that are not too important, to improve the accuracy and reduce the complexity of the model.

A. Experiment

Scenario 1

This scenario aims to get optimal parameter values which will be used in feature selection process. The parameters that will be tested include learning rate (γ), lambda (λ), and C. Each test was performed by using classification method, Sequential SVM, and K- fold Cross Validation method to calculate the model's accuracy with error target (ε) = 0.000001, epoch maximum = 500, and K = 10. Testing process is done by using a range of values which is 0.0005, 0.005, 0.05, and 0.5. Based on the range of values and the number of parameters used, this scenario will have total 64 combination of values. In each value tested on each parameter will has 16 combinations. For example, to evaluate learning rate (γ) parameter at the value of 0.5, it will have learning rate (γ) = 0.5, lambda (λ) with the range of value [0.0005, 0.005, 0.05, 0.5], and C with the range of values [0.0005, 0.005, 0.05, 0.5]. To determine the accuracy of learning rate (γ) at the value of 0.5, it would take the average value of the 16-accuracy generated in any combination. The average accuracy of each value in each parameter showed in Table II, Table III and Table IV.

TABLE II
AVERAGE ACCURACY RESULTS FOR LEARNING RATE

No	Learning Rate (γ)	Mean Accuracy
1	0.5	60.75%
2	0.05	60.75%
3	0.005	60.75%
4	0.0005	61.81%

TABLE III
AVERAGE ACCURACY RESULTS FOR LAMBDA

No	Lambda (λ)	Mean Accuracy
1	0.5	60.875%
2	0.05	61.063%
3	0.005	61.063%
4	0.0005	61.063%

TABLE IV
AVERAGE ACCURACY RESULTS FOR C

No	C	Mean Accuracy
1	0.5	50.063%
2	0.05	51.094%
3	0.005	50.000%
4	0.0005	92.500%

Based on the results shown in Table II, Table III and Table IV, the best accuracy was gained when parameter C at the value of 0.0005. Parameter C plays a role in the update of alpha (α), thus giving a significant effect on the accuracy rate. C is the parameter that controls tradeoff between margins and the classification error [17]. For large values of C, the optimization will choose hyperplane that does an excellent job of getting all the training data classified correctly, even if that hyperplane has relatively smaller margin. Conversely, a very small value of C will cause the optimizer to look for a larger-margin hyperplane even though the selected hyperplane misclassified some training data.

Table II implies the best accuracy for learning rate at the value of 61.81% when $\gamma = 0.0005$. Meanwhile it can be seen from Table III the best accuracy in lambda is 61 063% which gained in several value that is $\lambda = 0.05$, $\lambda = 0.005$, and $\lambda = 0.0005$. Study by [11] observed the influence of lambda's magnitude toward hyperplane quality. In the study, it was found that the larger value of lambda the better hyperplane quality produced, however too large value of lambda will normally lead to slower convergence speeds and instability in the learning process [11]. Thus, in this research we will select $\lambda = 0.05$ as the considered appropriate value for the learning process.

From the explanation we can conclude a value for each parameter that is $\gamma = 0.0005$, $\lambda = 0.05$, and $C = 0.0005$. The values of these parameters will be used in further scenarios.

Scenario 2

This scenario aims to get the best smallest subset from MicroRNA dataset using SVM-RFE feature selection method. Scenario 2 used the original dataset from MicroRNA with 1881 expression. SVM-RFE method will be performed with parameter values from previous scenario that is: $\gamma = 0.0005$, $\lambda = 0.05$, $C = 0.0005$, $\varepsilon = 0.000001$, and max *epoch* = 500. The outcome from SVM-RFE is a list of ranked expressions. After feature selection process, the outcome will be evaluated. The first evaluation step is to make new subsets in accordance with the ranked expression. As example, the first subset will be filled with data from 1st rank only. Subsequently, the second subset will be filled with data from 1st and 2nd rank. This procedure will be continued until the last subset which includes data from all ranks. The new subsets then will be processed one by one to obtain its accuracy rate by using Sequential SVM. Each process will be using the same parameter values as mentioned early with addition $k = 10$ for K-Fold Cross-Validation. Table V contains evaluation results from subset with the best accuracy which in this scenario is 99%.

TABLE V
BEST EVALUATION RESULTS FROM SCENARIO 2

Subset	Accuracy	Sensitivity	Specificities
3	99%	99%	99%
45	99%	100%	98%
46	99%	100%	98%
47	99%	100%	98%
48	99%	100%	98%

As shown in Table V, we know there is a difference at sensitivity and specificity rate at subset 3. This difference caused a confusion to which result is the subset, thus we converted evaluation results from Table V to ROC curves as shown in Figure 2. ROC curves for subset 46, 47, and 48 was represented by subset 45 because they have the similar shape towards each other.

Based on ROC curves we can calculate AUC for each subset. The AUCs is used to compare the performance of model's classification. The best model will have the highest AUC value. We applied this concept to determine the best subset in our research. The AUC value for subset 3 is 0,9891 while subset 45, 46, 47, 48 has the same AUC that is 0,98. Regardless the AUCs from all subsets, Subset 3 concluded as the best subset from scenario 2.

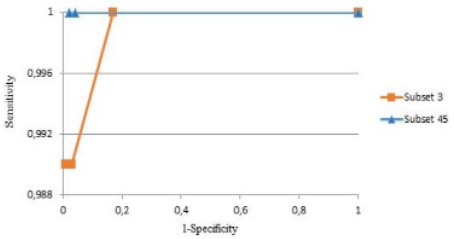

Fig. 2. ROC Curves for Best Subsets in Scenario 2

Scenario 3

Much of the microarray data contains *missing values* [18] so does in our original MicroRNA dataset. In MicroRNA dataset, zero (0) value is considered as a *missing value* [19]. To be useful for classification purposes, the dataset needs to undergo preprocessing, in the form of data cleaning and data transformation.

A common data cleaning method of handling *missing values* is simply to omit the records or fields with *missing values* from the analysis [20]. Our data cleaning procedure is done by calculating the average of each MicroRNA expression then eliminate expression with the average value that is less than 10. Through data cleaning procedure we obtained a new MicroRNA dataset with only 315 expression.

Afterwards, this new dataset was processed in the same feature selection & evaluation procedure with the exact same parameter values as in Scenario 2. In Table VI contains best evaluation results from Scenario 3.

TABLE VI
BEST EVALUATION RESULTS FROM SCENARIO 3

Subset	Accuracy	Sensitivity	Specificities
115	99%	100%	98%
147	99%	100%	98%
148	99%	100%	98%
149	99%	100%	98%
150	99%	100%	98%
151	99%	100%	98%
152	99%	100%	98%
153	99%	100%	98%
155	99%	100%	98%
156	99%	100%	98%
157	99%	100%	98%
158	99%	100%	98%
165	99%	100%	98%
166	99%	100%	98%

As shown in Table VI each subset has the same accuracy, sensitivity, and specificity rate. Therefore, for this scenario the best subset would be chosen according to the amount of expression contained in each subset, the lesser the better. Subset with fewest expression was announced as the best subset which in this scenario was subset 115.

B. Experiments Discussions

Either scenario 2 or scenario 3 give their own best result, scenario 2 with subset 3 and scenario 3 with subset 115. Each result is constructed by different expressions. Judging by the number of expressions, subset 115 resulted in scenario 3 is considered still too large compared to scenario 2. So, we excluded subset 115 from future discussion. Scenario 2 gave its best result with

subset constructed by only 3 expression that is miR-7705, miR-21, and miR-10b.

Many papers have considered miR-21 and miR-10b as a promising biomarker for breast cancer. A research by [21] found the levels of circulating miR-155, miR-21, and miR-10b were significantly up-regulated in Breast Cancer patients compared with healthy participants. [21] further evaluated 3 selected expressions with ROC curves and AUC values and figured miR-21 had the highest sensitivity of 77.4% meanwhile miR-10b had the highest specificity of 75.5%.

MiR-21 has been identified as one of the most protruding oncogenic microRNAs and has been proved upregulated in various human cancers [22]. MiR-21 regulates the expression of several cancer-correlated genes [number]. It is hypothesized that up-regulated miR-21 could be used as a potential biomarker for human cancer diagnosis [22].

MiR-10b was highly expressed in metastatic breast cancer cells and positively regulated cell migration and invasion [number]. MiR-10b inhibits translation of the mRNA encoding homeobox D10, leading to increased expression of RHOC (a well-characterized premetastatic gene. Therefore, in [number] was hypothesized that increased expression of miR-10b might be correlated with metastasis of breast cancer [23]. In another research found that the level of miR-10b expression was correlated with the patient survival status, stage of breast cancer tumor, and tumor size [24].

In the contrary of mir-21 and mir-10b, up until recent date there are still no published journals regarding the influence of miR-7705 towards breast cancer. Meanwhile in another papers, miR-7705 was reported has a correlation with Lung adenocarcinoma [25] and bladder cancer [26]. Mir-7705 still need further exploration regarding to its possibility as a potential biomarker for breast cancer.

IV. Conclusion

Our research was carried out by performing several scenarios. Scenario 1 aims to get the optimal values that would be used for classification and feature selection processes in further scenarios. Scenario 1 gave the final optimal parameter values that is $\gamma = 0.0005$, $\lambda = 0.05$, and $C = 0.0005$. From scenario 1 we learned that parameter C gave a significant effect on the accuracy rate cause its role on the update of alpha (α) and parameter C controls the tradeoff between margins and the classification error. Scenario 2 and 3 aimed to obtain their best smallest subset that later would be compared.

Selected MicroRNA subset obtained from scenario 2 gives a better result with only 3 expression that is miR-7705, miR-21, and miR-10b. MiR-21 and miR-10b have considered as a promising biomarker for breast cancer by many papers. While miR-7705 was discovered has a correlation with Lung Adenocarcinoma and bladder cancer but the said miRNA still need to be explored as a potential biomarker for breast cancer.

References

[1] World Health Organization (WHO), 2018. *Breast Cancer*. [Online]. Available at: https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en. [Accessed July 6th, 2018].

[2] Wang, L., 2017. Early diagnosis of breast cancer.

[3] Schrauder, M., Strick, R., Schulz-Wendtland, R., Strissel, P., Kahmann, L., Loehberg, C., et al., 2012. Circulating micro-RNAs as potential blood-based markers.

[4] Madhavan, D., Cuk, K., Burwinkel, B., & Yang, R., 2013. Cancer diagnosis and prognosis decoded by blood based.

[5] Zhu, H., & Fan, G.-C., 2011. Extracellular/circulating microRNAs and their potential role in cardiovascular disease.

[6] George, G. S., & Raj, D. V., 2011. Review on feature selection techniques. *International Journal of Computer Science & Engineering Survey*.

[7] Singh, R. K., & Sivabalakrishnan, D. M., 2015. Feature selection of gene expression data for cancer classification: A Review. *Procedia Computer Science 50,* 52-57.

[8] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V., 2002. Gene selection for cancer classification using support vector machines.

[9] Larose, D. T., 2005. Discovering knowledge in data: An introduction to data mining. New Jersey: John Willey & Sons, Inc.

[10] Jain, Y. K., & Bhandare, S. K., 2014. Min Max Normalization based data perturbation method for privacy protection.

[11] Vijayakumar, S., & Wu, S., 1999. Sequential Support Vector Classifiers and Regression.

[12] Duan, K.-B., Rajapakse, J. C., Wang, H., & Azuaje, F., 2005. Multiple SVM-RFE for gene selection in cancer classification with expression data.

[13] Refaeilzadeh, P., Tang, L., & Liu, H., 2009. Cross Validation. In L. Liu, & M. T. Özsu, *Encyclopedia of Database Systems*. Boston: Springer.

[14] Lever, J., Krzywinski, M., & Altman, N., 2016. Model selection and overfitting. Nature Methods, 13(9), 703-704.

[15] Sokolava, M., & Lapalme, G., 2009. A systematic analysis of performance measure for classification tasks. Information Processing and Management, 45, 427-437.

[16] Gorunescu, F., & Belciug, S., 2016. Boosting backpropagation algorithm by stimulus-sampling: Application. Journal of Biomedical Informatics, 63, 74-81.

[17] Nugroho, A. S., 2003. Support Vector Machine: Theories and application in bioinformatics. Support Vector Machine: Teori dan Aplikasi dalam bioinformatika.

[18] Yang, Y., Xu, Z., & Song, D., 2016. Missing value imputation for microRNA expression data by using a GO-based similarity measure.

[19] Eminaga, S., Christodoulou, D. C., Vigneault, F., Church, G. M., & Seidman, J., 2013. Quantification of microRNA Expression with Next-Generation Sequencing.

[20] Marabita, F., Candia, P. d., Torri, A., Tegnér, J., Abrignani, S., & Rossi, a. R., 2015. Normalization of circulating microRNA expression data obtained by quantitative real-time RT-PCR.

[21] Zhang, J., Jiang, C., Shi, X., Hua Yu, Lin, H., & Peng, Y., 2016. Diagnostic value of circulating miR-155, miR-21, and miR-10b

[22] Gao, Y., Dai, M., Liu, H., He, W., Lin, S., Yuan, T., et al., 2016. Diagnostic value of circulating miR-21: An update meta-analysis in various cancers and validation in endometrial cancer.

[23] Chen, W., Cai, F., Zhang, B., Barekati, Z., & Zhong, X. Y., 2012. The level of circulating miRNA-10b and miRNA-373 in detecting lymph node metastasis of breast cancer: potential biomarkers.

[24] Zhang, J., Yang, J., Zhang, X., Xu, J., Sun, Y., & Zhang, P., 2018. MicroRNA-10b expression in breast cancer and its clinical association.

[25] Hsu, Y.-L., Hung, J.-Y., Lee, Y.-L., Chen, F.-W., Chang, K.-F., Chang, W.-A., et al., 2017. Identifcation of novel gene expression signature in lung adenocarcinoma by using next-generation sequencing data and bioinformatics analysis. Oncotarget.

[26] Hu, Y., Cheng, C., Hong, Z., & Shi, Z., 2017. Independent prognostic miRNAs for bladder urothelial carcinoma. Oncology Letters.

Implementation of Support Vector Machine - Recursive Feature Elimination for MicroRNA Selection in Breast Cancer Classification

ORIGINALITY REPORT

17%	12%	11%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	j-ptiik.ub.ac.id Internet Source	1%
2	Shengqiang Xu, Seyedmehdi Hossaini Nasr, Daoyang Chen, Xiaoxian Zhang, Liangliang Sun, Xuefei Huang, Chunqi Qian. "MiRNA Extraction from Cell-Free Biofluid Using Protein Corona Formed around Carboxyl Magnetic Nanoparticles", ACS Biomaterials Science & Engineering, 2018 Publication	1%
3	Mohammad Abdul Haque Farquad. "Rule extraction from support vector machines: a hybrid approach for solving classification and regression problems", International Journal of Information and Decision Sciences, 2011 Publication	1%
4	mafiadoc.com Internet Source	1%

Submitted to University of Southern California

5	Student Paper	1 %
6	patents.justia.com Internet Source	1 %
7	repository.tudelft.nl Internet Source	1 %
8	tessera.spandidos-publications.com Internet Source	1 %
9	Submitted to RMIT University Student Paper	<1 %
10	openreview.net Internet Source	<1 %
11	downloads.hindawi.com Internet Source	<1 %
12	www.oncotarget.com Internet Source	<1 %
13	A. Diaf, B. Boufama, R. Benlamri. "Non-parametric Fisher's discriminant analysis with kernels for data classification", Pattern Recognition Letters, 2013 Publication	<1 %
14	Submitted to University of Lancaster Student Paper	<1 %
15	Submitted to University of Nizwa Student Paper	<1 %

16	indico.bnl.gov Internet Source	<1 %
17	Submitted to University of Lincoln Student Paper	<1 %
18	www.igi-global.com Internet Source	<1 %
19	fkp2tn.onesearch.id Internet Source	<1 %
20	Li Ma. "Therapeutic silencing of miR-10b inhibits metastasis in a mouse mammary tumor model", Nature Biotechnology, 03/28/2010 Publication	<1 %
21	Submitted to University of Bradford Student Paper	<1 %
22	conf.unnes.ac.id Internet Source	<1 %
23	datatron.com Internet Source	<1 %
24	personalpages.manchester.ac.uk Internet Source	<1 %
25	Experientia Supplementum, 2015. Publication	<1 %
26	Submitted to Universiti Teknologi Malaysia Student Paper	<1 %

27	Duman, E.. "Comparing alternative classifiers for database marketing: The case of imbalanced datasets", Expert Systems With Applications, 201201 Publication	<1 %
28	Qiangxin Huang, Qian Song, Weixian Zhong, Yalan Chen, Ludong Liang. "MicroRNA-10b and the clinical outcomes of various cancers: A systematic review and meta-analysis", Clinica Chimica Acta, 2017 Publication	<1 %
29	Shi, L.. "Discriminative analysis of skull morphology in adolescent idiopathic scoliosis patients: Comparative study with normal controls", Pattern Recognition, 200809 Publication	<1 %
30	Submitted to Universiti Malaysia Pahang Student Paper	<1 %
31	worldwidescience.org Internet Source	<1 %
32	www.dovepress.com Internet Source	<1 %
33	www.springerprofessional.de Internet Source	<1 %
34	Submitted to Universiti Kebangsaan Malaysia Student Paper	<1 %

35	iranarze.ir Internet Source	<1%
36	library.unisel.edu.my Internet Source	<1%
37	www.research.ed.ac.uk Internet Source	<1%
38	Mikio Kawamura, Yuji Toiyama, Koji Tanaka, Yasuhiro Inoue, Yasuhiko Mohri, Masato Kusunoki. "Can Circulating MicroRNAs Become the Test of Choice for Colorectal Cancer?", Current Colorectal Cancer Reports, 2014 Publication	<1%
39	Nova Gerungan. "HUBUNGAN TINGKAT PENGETAHUAN DENGAN PERILAKU PERIKSA PAYUDARA SENDIRI (SADARI) PADA MAHASISWI FAKULTAS KEPERAWATAN UNKLAB", Jurnal Skolastik Keperawatan, 2019 Publication	<1%
40	interscience.in Internet Source	<1%
41	theses.hal.science Internet Source	<1%
42	Ce Zheng. "Prediction of beta-turns at over 80% accuracy based on an ensemble of	<1%

predicted secondary structures and multiple alignments", BMC Bioinformatics, 2008
Publication

---

43 Irimie-Aghiorghiesei, Pop-Bica, Pintea, Braicu, Cojocneanu, Zimța, Gulei, Slabý, Berindan-Neagoe. "Prognostic Value of MiR-21: An Updated Meta-Analysis in Head and Neck Squamous Cell Carcinoma (HNSCC)", Journal of Clinical Medicine, 2019 <1 %
Publication

---

44 Yi Zhang, Li-Juan Wang, He-Quan Yang, Rong Wang, Hua-Jun Wu. "MicroRNA-10b expression predicts long-term survival in patients with solid tumor", Journal of Cellular Physiology, 2018 <1 %
Publication

---

45 "Multi-disciplinary Trends in Artificial Intelligence", Springer Science and Business Media LLC, 2017 <1 %
Publication

---

46 Apexa Raval, Jigna Joshi, Franky Shah. "Significance of metastamiR-10b in breast cancer therapeutics", Journal of the Egyptian National Cancer Institute, 2022 <1 %
Publication

---

47 G A Pradnyana, I G M Darmawiguna, D K S Suditresna Jaya, A Sasmita. "Performance analysis of support vector machines with <1 %

polynomial kernel for sentiment polarity identification: A case study in lecturer's performance questionnaire", Journal of Physics: Conference Series, 2021
Publication

48 Hala Ahmed, Hassan Soliman, Mohammed Elmogy. "Early detection of Alzheimer's disease using single nucleotide polymorphisms analysis based on gradient boosting tree", Computers in Biology and Medicine, 2022 <1%
Publication

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

Implementation of Support Vector Machine - Recursive Feature Elimination for MicroRNA Selection in Breast Cancer Classification

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5