# Implementation of Genetic Algorithm-Support Vector Machine on Gene Expression Data in Identification of Non-Small Cell Lung Cancer in Nonsmoking Female

Tasya Nurfauziah Ramadhani Rohimat
*School of Computing*
*Telkom University*
Bandung, Indonesia
tasyanfauziah@student.telkomuniversity.ac.id

Fhira Nhita
*School of Computing*
*Telkom University*
Bandung, Indonesia
fhiranhita@telkomuniversity.ac.id

Isman Kurniawan
*School of Computing*
*Telkom University*
Bandung, Indonesia
ismankrn@telkomuniversity.ac.id

***Abstract*— Lung cancer is the leading cause of death in the world. There are two types of lung cancer, i.e., non-small cell lung cancer and small cell lung cancer. The major cause of lung cancer is smoking. However, there are several cases of non-small lung cancer, with 7% of women with lung cancer in Taiwan having a smoking history. Early detecting of cancer will help it go faster and save lives every year. Nowadays, the technology being used are very helpful in the medical field because it uses microarray technology which can help detect cancer in early phase by analyzing DNA and RNA. In this study, we utilized GA combined with SVM for the classification of Non-Small Cell Lung Cancer in non-smoking female with microarray data. Hyperparameter tuning is performed to improve model performance. We discovered that SVM with a linear kernel performs better than alternative kernels with accuracy and F1-score values of 0.91 and 0.91 respectively.***

**Keywords—** ***Non-small Cell Lung Cancer (NSCLC), Microarray, Genetic Algorithm, Support Vector Machines.***

## I. INTRODUCTION

Lung cancer is a main causes of cancer death worldwide [1]. The different types of lung cancer are small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). NSCLC accounts for 80% to 85% of all lung cancers, while SCLC accounts for 15% to 20%. Smoking is known as a major factor in lung cancer. Only 7% of female patients with lung cancer in Taiwan had a smoking history [2]. Another factor that causes lung cancer in non-smoker women is environmental exposure or heredity. However, there are no molecular mechanisms of NSCLC in women who do not smoke, although several genes, including EML4-ALK, EGFR, TP53.9, and PIK3CA, are linked to lung cancer in smokers [3].

In the case of lung cancer, a chest X-ray and a Computerized Tomography Scan (CT-Scan) are usually used for the diagnosis and prognosis. However, this method can only detect malignant cells in lung cancers in their advanced stages [4][5]. With advances in technology in the field of molecular biology, especially in microarray technology, information about DNA, RNA, and protein will be obtained for early detection of tumor formation [6]. One application of microarray technology is to analyze thousands of DNA, RNA, and protein samples at the same time.

Currently, machine learning on microarrays and bioinformatics analysis are frequently used to identify potential cancer biomarkers, particularly genes related to the prognosis of lung cancer.[2]. The use of microarrays produces more complete information about variations in cancer molecules as well as obtains more accurate classification results [7]. In the past decade, machine learning methods have been widely used to medical data analysis [8]. Cancer prognosis is commonly predicted using machine learning techniques, especially Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Bayesian networks [6]. Also microarray data has been implemented to identify a various cases used ensemble method [9].

Identification of lung cancer with microarrays using machine learning methods has been done several times in previous studies. In 2017, Nurul et al. improved cancer classification using the Multi Support Vector Machines (MSVM) method and the Recursive Feature Elimination (RFE) method as feature selection with an accuracy value of 98.6% [10]. In 2017, Wu and Zhao conducted a study that detected SCLC using a novel neural network algorithm with the entropy degradation method (EDM) and the accuracy value reached 77.8% [4]. In 2019, Shrikant and Pawar conducted a web-based study to classify cancer types from microarray of Gene Expression Data (DEGs) using MAS 5.0 and Robust Multi-Array (RMA) as feature selection and SVM method for classification with an accuracy value of 95% [11]. Then in 2019, José et al. conducted research on predicting radiation pneumonitis in advanced stages II-III in NSCLC using machine learning. Using machine learning algorithms, carried out an analysis of contributing factors in the development of radiation pneumonitis to uncover previously unidentified criteria. In multivariate analysis, Random Forest has an accuracy value of 66% [12].

Several other studies have also been carried out previously. In 2020, Yu classified the types of NSCLC using Convolutional Neural Networks with a validated prediction with an accuracy value was 86.4% [13]. In 2020, Pankaj et al. conducted research the hybrid algorithm for the classification of lung cancer using SVM and Neural Network by detecting CT-Scan images with an accuracy value of 98.08% [14]. Reinel and Co conducted research in 2020 comparing machine learning with deep learning to classify cancer types based on gene expression microarray data. The accuracy of the identification result is 90.6% while using Logistic Regression and 94.43% while using Convolutional Neural Network (CNN) [15]. In 2021, Sunil et al. compared the performance of lung cancer classification using the Swarm Intelligence technique with the best results shown when the test was classified with the

Decision Tree classifier for one hundred genes, and the highest classification accuracy value was 99.1% [16]. To the best of our knowledge, the studies of NSCLC identification for non-smoker woman is very rare.

In this study, we aimed to predict the potential occurrence of NSCLC in non-smoker women using the GA-SVM methods. This research used a Genetic Algorithm (GA) for feature selection because the evolution operators in a Genetic Algorithm (GA) make this algorithm very effective in global search and this algorithm has high flexibility to be hybridized with other search methods to make it more effective. Also GA has been implemented to develop prediction model on a various cases [17][18]. For the development of the prediction model using the Support Vector Machines (SVM) method because Support Vector Machines (SVM) offer high accuracy results and work well with high dimensional spaces. The Support Vector Machines (SVM) classifier uses a subset of train points so that the result uses very little memory. Several research of SVM has also been done implemented to identify various cases [19][20].

## II. DATASET AND METHODOLOGY

### A. Dataset

In this study, a microarray dataset was used from Geo Datasets with the GSE19804 [2]. The dataset consisted of 54,675 lung tissue cell gene expressions consisting of 120 samples with two classes, 60 with lung cancers and 60 with normal lungs. Then the dataset is split into train and test sets with a ratio of 70:30. The comparison of the number of samples for each class on the train and test is shown in Figure 1. In the train set, the number of cancer samples is 42 and the number of normal samples is 42. In the test set, the number of cancer samples and the number of normal samples are 18 dan 18, respectively. The comparison from Figure 1 show that the used dataset are balanced.

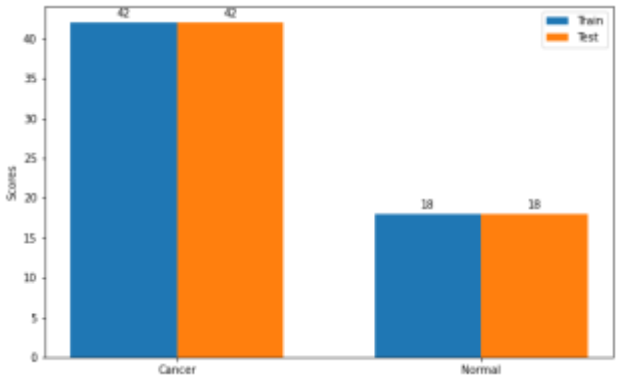


Figure 1. The number of samples in the test and train set

### B. Feature Selection

The variance threshold is a basic method for selecting features. It removes all features whose variance is below a defined threshold. The variance threshold eliminates all features whose variance values are less than 0.5. The features with higher variance is a more useful features.

A Genetic Algorithm is used to perform the feature selection stage. GA is designed to find a good and selected subset of genes. The usage of archives to record the quality subset of genes that would be identified during the gene subset selection stage is one of the most essential characteristics of the GA established in this study. This archive is then evaluated to find several frequently occurring genes utilized in the final stage of classification. The concept of archiving is not a new approach, as it has been implemented in several multi-objective algorithm evolutions. However, a method of utilizing archive data to find prognostic genes might be valuable. The GA stages will be shown in Figure 2 below:

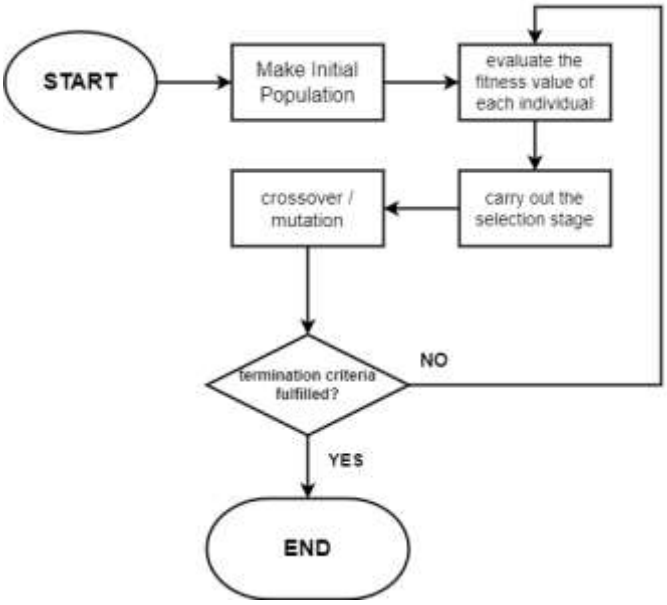


Figure 2. Flowchart of Genetic Algorithm

Based on the genes that have been selected by GA, the SVM method will be used to classify and select the final set of genes from the test data. There are several parameters, i.e., number of generations, population, mutation probability, parent selection and survivor selection with the values in Table I.

TABLE I. FEATURE SELECTION PARAMETER OF GA

| No | Parameter | Value |
|---|---|---|
| 1 | Generation | 100 |
| 2 | Population | 20 |
| 3 | Mutation Probability | 0,1% |
| 4 | Parent Selection | Roulette Wheel Selection |
| 5 | Survivor Selection | Fitness-based selection |

### C. Model Development

SVM developed by Cortes & Vapnik in 1995 for binary classification, is currently a hot issue in machine learning theory and one of the most effective algorithms for microarray data classification. The basic concept of SVM for classification is to find an optimal dividing hyperplane between two classes by maximizing the distance between their nearest points [21][7]. Figure 3 shows the illustration of optimal classifier.

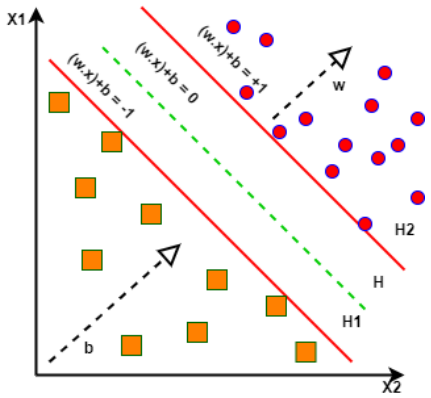


Figure 3. Optimal Classifier

For example, there is a data set consisting of attributes $x$ and label $y$ with $y \in \{-1, 1\}$. For the case of a two-dimensional hyperplane, it can be expressed in equation (1):

$$w.x + b \tag{1}$$

where $w$ where w is the vector's weight and $b$ is its bias. After the hyperplane is obtained, the sample will be classified as class -1 if $w.x + b \leq -1$ and will be classified as class +1 if $w.x + b \geq +1$. Meanwhile, the distance between the two margins is formulated as $\frac{2}{\|w\|}$ [19].

Using kernel functions, the input from SVM will be transferred to a higher-dimensional representation to create a linear model. Equation (2) represents the linear function utilized by SVM:

$$f(x) = (\omega, x) + b \tag{2}$$

where $\omega$ the coefficient of each feature's weight value. During processing, this value will be modified. The algorithm will attempt to determine the weight of the value that generates the hyperplane with the maximum margin. There are several kernel functions commonly used in SVM, such as linear, polynomial, and RBF which are formulated as in Table II.

TABLE II. EQUATION OF EACH KERNEL

| Kernel | Equation |
|---|---|
| Linear | $k(X_i, X_j) = X_i^T . X_j$ |
| Polynomial | $k(X_i, X_j) = (\gamma X_i^T X_j + r)^p, \gamma > 0$ |
| RBF | $k(X_i, X_j) = exp\left(-\gamma \left\|X_i, X_j\right\|\right)^2$ |

Schematic hyperparameter tuning is done to improve model performance. This scheme compares the performance of all potential parameter combinations to determine the optimal combination. In this study, the linear, polynomial, and RBF will be used for modeling. The Grid Search cross-validation was used to carry out the tuning procedure [18]. Grid Search cross-validation is a hyperparameter tuning procedure that constructs and evaluates a model for each grid-specified parameter method combination. The Support Vector Machine parameters and values are shown in Table III.

TABLE III. MODEL DEVELOPMENT PARAMETERS

| No | Parameter | Range of Values |
|---|---|---|
| 1 | C | [0.1, 1, 10, 100, 1000] |
| 2 | Degree | [0, 1, 2, 3, 4, 5, 6] |
| 3 | Gamma | [‘scale’, ‘auto’] |

The parameter C is a parameter that serves to minimize errors in the classification results. This parameter will control the boundary between the decision boundary and the correct prediction result. The degree is a polynomial degree, this parameter is used to find a hyperplane to divide the data. Gamma is a parameter for a hyperplane that is non-linear. The higher the gamma value, the better the dataset to use.

### *D. Model Validation*

Optimized SVM model parameters were used to create the prediction. Furthermore, the model validation stage will validate the model by calculating the validation parameters for the classification problem. This includes true positive (TP), false positive (FP), true negative (TN), false negative (FN), sensitivity (SE), specificity (SP), precision (PR), total prediction accuracy (Q), and the Matthew correlation coefficient (MCC). At this stage, the accuracy value is a quality parameter used to select the optimal model. In Table IV, the confusion matrix will be shown.

TABLE IV. CONFUSION MATRIX

| | | **Actual Class** | |
|---|---|---|---|
| | | **Cancer** | **Normal** |
| **Predicted Class** | **Cancer** | True Positive (TP) | False Positive (FP) |
| | **Normal** | False Negative (FN) | True Negative (TN) |

The value of the confusion matrix can determine the performance value of several validation parameters, among others, Recall or Sensitivity (SE), Specificity (SP), Precision (PR), Accuracy (Q) and Matthew correlation coefficient (MCC). Evaluation of these parameters is carried out using equations (3) – (7):

$$SE = \frac{TP}{TP+FN} \tag{3}$$

$$SP = \frac{TN}{TN+FN} \tag{4}$$

$$PR = \frac{TP}{TP+FP} \tag{5}$$

$$Q = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

$$MCC = \frac{(TP\ x\ TN)-(FPxFN)}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \tag{7}$$

The model was evaluated by visualizing the receiver operating characteristic (ROC) curve. This curve describes the success and failure of the predictions observed in the development model. The ROC curve was plotted by taking the true-positive and false-positive values on the y-axis and x-axis, respectively. Predictive accuracy can be easily identified by analyzing the characteristics of the curve. This curve calculates additional parameters, including under the ROC curve, commonly known as AUC. AUC measures the ability of the model to distinguish between the two classification groups and represents the accuracy of the prediction [17].

## III. Result and Discussions

### A. *Feature Selection*

In this study, the data tested are the gene expression features of lung tissue cells by using the GA and SVM methods. The features were selected using the variance threshold method and genetic algorithm (GA). After the variance threshold stage, the number of features is reduced from 54,675 to 546 features [17].

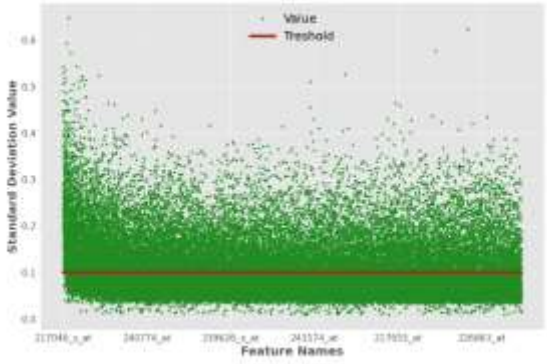


Figure 4. Distribution of variance

After selecting the features on the variance threshold, then the features are selected using GA for each kernel. The number of selected features and the accuracy fitness value of each kernel is determined by default parameters. To retrieve the accuracy values and selected features, several experiments were carried out 20 times to compare the selected features and the best accuracy fitness values for each process. The number of selected features and the best accuracy fitness value for each kernel are shown in Table V. The average fitness accuracy result of the best process of 100 iterations for each kernel, which is shown in figure 5.

TABLE V. SUMMARY OF SEVERAL EXPERIMENTS FOR EACH KERNELS

| Kernel | Selected Features | Fitness Accuracy | Avg | STD |
|---|---|---|---|---|
| Linear | 164 | 0.969 | 0.953 | 0.016 |
| Polynomial | 179 | 0.932 | 0.953 | 0.060 |
| RBF | 163 | 0.970 | 0.957 | 0.015 |

Table V illustrates the disparity in fitness accuracy between different numbers of selected features, where a smaller number of features, i.e. 163 or less, produces a better accuracy score than larger numbers of features, i.e. 164 and 170. However, the number of features used is less than the number of features that fulfill the threshold (564), but the accuracy achieved is close to the maximum value (1.00).

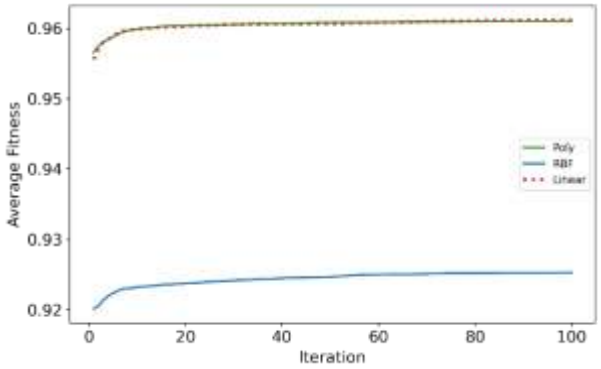


Figure 5. Average fitness accuracy from Feature Selection GA

In general, linear, and polynomial kernels have higher fitness value than RBF kernels. The linear kernel and polynomial kernel have almost the same fitness value. the three kernels begin to show a stable fitness value in the 60th iteration to the 100th iteration.

From the Figure 6-8, the skewness value on the kernels value shows negative skewness when these values are higher between 0.94 to 0.98. A negative Skewness means that the tail of the distribution is on the left, indicating that most of the values are on the right side of the curve.

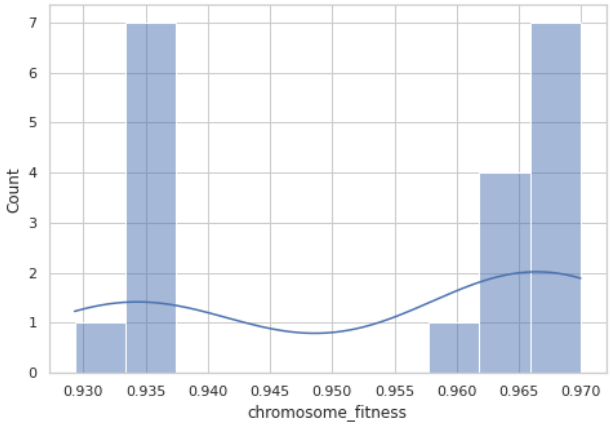


Figure 6. Distribution of Linear Kernel

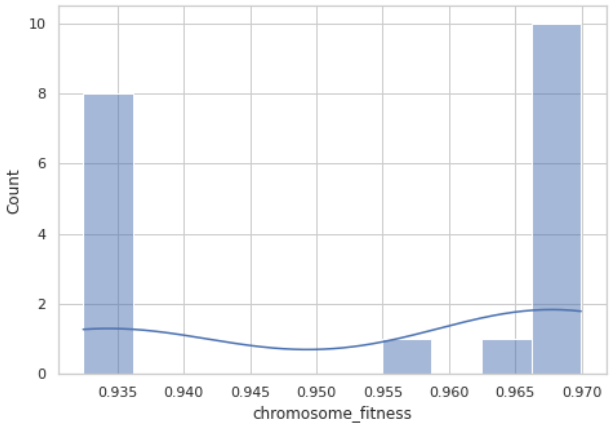


Figure 7. Distribution of Polynomial Kernel

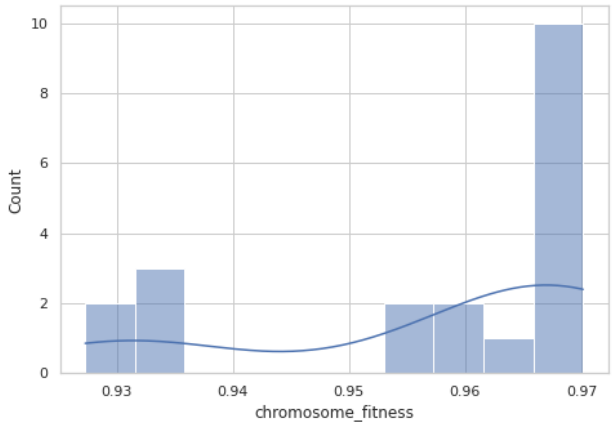


Figure 8. Distribution of RBF Kernel

### B. *Model Development*

Hyperparameter tuning is used to get the best parameters for all kernels in the model. By optimizing hyperparameters, the performance of the SVM model is improved. Results of hyperparameter tuning with their parameters are shown in Table VI. According to Table VI, the polynomial kernel is a kernel with different parameters from other kernels.

TABLE VI. BEST PARAMETER OF EACH KERNELS

| Parameter | Kernel | | |
|---|---|---|---|
| | Linear | Polynomial | RBF |
| C | 0.1 (1) | 0.1 (1) | 0.1 (1) |
| Degree | - | 2 (3) | - |
| Gamma | - | Auto (Scale) | Scale (Scale) |

*The value contained in the bracket is the model's default parameter value of SVM

### C. *Model Validation*

We consider the F1 score as an overall measure to determine the best model. The results of the model validation are presented in Table VII. Testing on linear kernels produced the best performance, hyperparameter tuning gave a less significant improvement for all kernels because the values were stable. Table VII displays the performance of each kernel in the test. The results of the model performance validation are presented in Table VII.

In terms of training data, we found that all models could accurately predict the target values, which were indicated by high PR, RC, F1, and AUC_ROC values. However, linear kernels seem to produce higher values compared to other kernels. Using a linear kernel, the PR, RC, F1, and AUC_ROC values are 1.00, 1.00, 1.00, and 1.00, respectively. The values using the polynomial kernel obtained for the PR, RC, F1, and AUC_ROC values are 0.97, 1.00, 0.98, and 0.98, respectively. For the RBF kernel, the PR, RC, F1, and AUC_ROC values are 0.97, 1.00, 0.98, and 0.98, respectively. This shows that the linear kernel predicts the target in the training data with better results.

The test set validation turned out to produce a lower validation parameter value than the training set validation. Based on the results, it was found that the prediction of the test set using the linear kernel was more accurate than other kernels. This is indicated by the validation parameter value which is higher than other kernels. The PR, RC, F1, and AUC_ROC values obtained from the use of the kernel are 0.94, 0.88, 0.91, and 0.91 respectively. the linear kernel is the best kernel for predicting accuracy in this model compared to the polynomial kernel and RBF kernel.

TABLE VII. SUMMARY OF VALIDATION RESULT

| **Train** | **Linear** | Polynomial | RBF |
|---|---|---|---|
| TP | **42** | 41 | 41 |
| FP | **0** | 1 | 1 |
| TN | **42** | 42 | 42 |
| FN | **0** | 0 | 0 |
| PR | **1.00** | 0.97 | 0.97 |
| RC | **1.00** | 1.00 | 1.00 |
| Q | **1.00** | 0.98 | 0.98 |
| F1 | **1.00** | 0.98 | 0.98 |
| AUC_ROC | **1.00** | 0.98 | 0.98 |
| **Test** | **Linear** | Polynomial | RBF |
| TP | **17** | 15 | 15 |
| FP | **1** | 3 | 2 |
| TN | **16** | 16 | 16 |
| FN | **2** | 2 | 2 |
| PR | **0.94** | 0.84 | 0.84 |
| RC | **0.88** | 0.89 | 0.89 |
| Q | **0.91** | 0.86 | 0.86 |
| F1 | **0.91** | 0.86 | 0.86 |
| AUC_ROC | **0.91** | 0.86 | 0.86 |

From table VII, the best accuracy of the training model and test model is related to the value of the characteristic curve, which can be calculated from the value of the area under the curve (AUC). The linear kernel is the kernel with the best accuracy because the AUC value is the highest value among other kernels.

The implication of the experimental result shows that GA plays an important role in feature selection and classification, where the best model will be used for model validation. Features Selection specifies important features based on the best value in GA to produce a good model for classifying.

## IV. CONCLUSION

A predictive model was developed to classify patients with NSCLC in non-smoker women using a Genetic Algorithm (GA) and Support Vector Machine (SVM). Feature selection using best fitness search and tested on three SVM kernels, i.e., linear, polynomial and RBF. Three kernels were used to evaluate the contribution of feature numbers to model performance. According to the result we found that the best model obtained from SVM with linear kernel that produced accuracy and F1-score are 0.91 and 0.91, respectively. Hence, this model can be the alternative method to predict NSCLC. This prediction model can be developed in the future using feature selection and classification with other methods.

## REFERENCES

[1] Y. H. Lai, W. N. Chen, T. C. Hsu, C. Lin, Y. Tsao, and S. Wu, "Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning," *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, 2020, doi: 10.1038/s41598-020-61588-w.

[2] G. Yang, Q. Chen, J. Xiao, H. Zhang, Z. Wang, and X. Lin, "Identification of genes and analysis of prognostic values in nonsmoking females with non-small cell lung carcinoma by bioinformatics analyses," *Cancer Manag. Res.*, vol. 10, pp. 4287–

4295, 2018, doi: 10.2147/CMAR.S174409.

[3] T. P. Lu *et al.*, "Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women," *Cancer Epidemiol. Biomarkers Prev.*, vol. 19, no. 10, pp. 2590–2597, 2010, doi: 10.1158/1055-9965.EPI-10-0332.

[4] Q. Wu and W. Zhao, "Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm," *Proc. - 2017 Int. Symp. Comput. Sci. Intell. Control. ISCSIC 2017*, vol. 2018-Febru, pp. 88–91, 2018, doi: 10.1109/ISCSIC.2017.22.

[5] F. Ramadhaniah, A. Mulawarman, E. Suzanna, and L. R. Andalucia, "Gambaran Kanker Paru Karsinoma Bukan Sel Kecil dengan Efusi Pleura Characteristic of Non Small Cell Lung Cancer with Pleural Efusion," *J. Respirologi Indones.*, vol. 36, no. 2, pp. 60–66, 2016.

[6] H. Azzawi, J. Hou, Y. Xiang, and R. Alanni, "Lung cancer prediction from microarray data by gene expression programming," *IET Syst. Biol.*, vol. 10, no. 5, pp. 168–178, 2016, doi: 10.1049/iet-syb.2015.0082.

[7] P. Guan, D. Huang, M. He, and B. Zhou, "Lung cancer gene expression database analysis incorporating prior knowledge with support vector machine-based classification method," vol. 7, pp. 1–7, 2009, doi: 10.1186/1756-9966-28-103.

[8] Y. Wu *et al.*, "Preoperative Prediction of Lymph Node Metastasis in Patients With Early-T-Stage Non-small Cell Lung Cancer by Machine Learning Algorithms," *Front. Oncol.*, vol. 10, no. May, pp. 1–8, 2020, doi: 10.3389/fonc.2020.00743.

[9] D. N. Purba, Fhira Nhita, and Isman Kurniawan, "Implementation of Ensemble Method in Schizophrenia Identification Based on Microarray Data," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 1, pp. 64–69, 2022, doi: 10.29207/resti.v6i1.3788.

[10] N. N. M. Hasri, N. H. Wen, C. W. Howe, M. S. Mohamad, S. Deris, and S. Kasim, "Improved support vector machine using multiple SVM-RFE for cancer classification," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 7, no. 4–2 Special Issue, pp. 1589–1594, 2017, doi: 10.18517/ijaseit.7.4-2.3394.

[11] S. Pawar, *Web-Based Application for Accurately Classifying Cancer Type from Microarray Gene Expression Data Using a Support Vector Machine (SVM) Learning Algorithm*, vol. 11466 LNBI. Springer International Publishing, 2019.

[12] J. M. Luna *et al.*, "Predicting radiation pneumonitis in locally advanced stage II–III non-small cell lung cancer using machine learning," *Radiother. Oncol.*, vol. 133, pp. 106–112, 2019, doi: 10.1016/j.radonc.2019.01.003.

[13] K. H. Yu *et al.*, "Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks," *J. Am. Med. Informatics Assoc.*, vol. 27, no. 5, pp. 757–769, 2020, doi: 10.1093/jamia/ocz230.

[14] P. Nanglia, S. Kumar, A. N. Mahajan, P. Singh, and D. Rathee, "A hybrid algorithm for lung cancer classification using SVM and Neural Networks," *ICT Express*, vol. 7, no. 3, pp. 335–341, 2020, doi: 10.1016/j.icte.2020.06.007.

[15] R. Tabares-Soto, S. Orozco-Arias, V. Romero-Cano, V. S. Bucheli, J. L. Rodríguez-Sotelo, and C. F. Jiménez-Varón, "A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data," *PeerJ Comput. Sci.*, vol. 2020, no. 4, pp. 1–22, 2020, doi: 10.7717/peerj-cs.270.

[16] S. K. Prabhakar, H. Rajaguru, and D. O. Won, "A Holistic Performance Comparison for Lung Cancer Classification Using Swarm Intelligence Techniques," *J. Healthc. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/6680424.

[17] I. Kurniawan, D. Tarwidi, and Jondri, "QSAR modeling of PTP1B inhibitor by using Genetic algorithm-Neural network methods," *J. Phys. Conf. Ser.*, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012059.

[18] M. Fajar Rizqi, R. Rendian Septiawan, and I. Kurniawan, "Implementation of Simulated Annealing-Support Vector Machine on QSAR Study of Indenopyrazole Derivative as Anti-Cancer Agent," *2021 9th Int. Conf. Inf. Commun. Technol. ICoICT 2021*, no. Icidm, pp. 662–668, 2021, doi: 10.1109/ICoICT52021.2021.9527416.

[19] E. Wicaksana, D. T. Murdiansyah, and I. Kurniawan, "Classification Model of Consumer Question about Motorbike Problems by Using Naïve Bayes and Support Vector Machine," *Indones. J. Comput.*, vol. 6, no. 2, pp. 1–10, 2021, doi: 10.34818/indojc.2021.6.2.561.

[20] A. Nabilah 'Izzaturrahmah, F. Nhita, and I. Kurniawan, "Implementation of Support Vector Machine on Text-based GERD Detection by using Drug Review Content," *2021 Int. Conf. Adv. Data Sci. E-Learning Inf. Syst. ICADEIS 2021*, 2021, doi: 10.1109/ICADEIS52521.2021.9702073.

[21] O. F.Y, A. J.E.T, A. O, H. J. O, O. O, and A. J, "Supervised Machine Learning Algorithms: Classification and Comparison," *Int. J. Comput. Trends Technol.*, vol. 48, no. 3, pp. 128–138, 2017, doi: 10.14445/22312803/ijctt-v48p126.