**4th Bioinformatics and Biodiversity Conference**
Volume 2023
http://dx.doi.org/10.11594/nstp.2023.3701

NST Proceedings
OPEN ACCESS

**Conference Paper**

# Feature Reduction of Lung Cancer Microarray Data Using Mutual Information Selection and PyCaret-Supported Recursive Feature Elimination

Andrew Jonathan Brahms Simangunsong*, Valha Tsabita Hidayat

Universitas Indonesia, Depok 16911, Indonesia

*Corresponding author:
E-mail:
andrewjonathanbs@gmail.com

ABSTRACT

Lung cancer remains a leading cause of cancer-related mortality worldwide, and Indonesia's ever-increasing amount of pollution signals an urgency for improvement in lung cancer early detection. One of the methods to detect lung cancer is molecular diagnosis using DNA microarray, which has been proven to be effective. However, the complexity of microarray data with a vast number of features hinders the timely and accurate detection of lung cancer. This study seeks to optimize the features of the data to improve classification performance. Our approach combines Mutual Information Feature Selection with Recursive Feature Elimination, leveraging the PyCaret library to train and evaluate machine learning models. The process involves initial feature reduction using Mutual Information to enhance computational efficiency, followed by training machine learning models with PyCaret. The two best-performing models for each dataset are used to perform recursive feature elimination to search for the most optimal feature. A support vector machine is also used for comparison. The final output will be three subsets of features and another subset that consists of combined features of the rest of other subsets. Finally, PyCaret will be utilized again to train machine learning models with all feature subsets. The study shows that other models can select fewer features compared to the Support Vector Machine and still maintain a powerful predictive power with high accuracy (95% - 98%). In conclusion, our research offers a new approach to selecting optimal features for microarray analysis, with implications for more effective and timely cancer diagnosis.

*Keywords: Lung cancer, microarray data, feature reduction, mutual information feature selection, recursive feature elimination, PyCaret*

## Introduction

Air pollution has been a globally wide concern, especially because of its impact on people's health. The World Health Organization (WHO) established air quality guidelines (AQG) as a way for quantitative assessment of air quality. Several kinds of pollutants measured are chosen based on each of their relation in concentration-response to health outcomes, such as $PM_{2.5}$, $PM_{10}$, nitrogen oxide, sulfur dioxide, and carbon monoxide. Annual air quality report by IQAir (2021) specifically use PM---$_{2.5}$, particulate matter with an aerodynamic diameter equal to or less than 2.5 μm, which represents many and various sources of pollutant of different chemical components, such as sulfates, carbons, sulfates, and ammonium. IQAir (2022) Reports showed the mean level of $PM_{2.5}$ in Indonesia in the year of 2022 stands at 30.4 μg/m-$^3$, 34.3 μg/m$^3$ in the year of 2021, and 40.8 μg/m$^3$ in the year of 2020, whereas the standard of WHO's AQG is 5 μg/m$^3$ for annual mean. Not only is the $PM_{2.5}$ level the highest in Southeast Asia, but it also exceeds WHO $PM_{2.5}$ guidelines by 5-7 times.

*How to cite:*
Simangunsong, A. J. B., & Hidayat, V. T. (2023). Feature reduction of lung cancer microarray data using mutual information selection and PyCaret-supported recursive feature elimination. *4th Bioinformatics and Biodiversity Conference*. NST Proceedings. pages 1-6. doi: 10.11594/ nstp.2023.3701

$PM_{2.5}$ has been associated with various diseases, including cardiopulmonary problems, such as lung cancer, airway inflammation, chronic obstructive pulmonary disease (COPD), asthma, and ischemic heart disease. With lung cancer as the variety of cancer with the highest mortality rate (13.2%) in Indonesia as of 2020 according to Global Burden of Cancer (GLOBOCAN), it is of the utmost importance to take a focused approach to early detection of lung cancer.

One of the methods to detect lung cancer early is by performing DNA microarray analysis. However, microarray analysis is a challenging task because microarray data have many features (Mao et al., 2019). The complexity of the data will hinder the evaluation process because it will take a longer amount of time and cost just to get one person's analysis. This evaluation process efficiency can be increased by selecting the best features of the data to be analyzed. In this research, we propose a combination of mutual information selection and recursive feature elimination to reduce the features of microarray data. Mutual information selection is used to support the process of recursive feature elimination because recursive feature elimination is a heavy computational process and by decreasing the features first, the process will be computationally lighter. In this research, we also utilize the PyCaret library to select the best model to perform recursive feature elimination. This will make the remaining features more robust because the features will be obtained from various models with various mechanisms and it will help further research on important genes to be observed to detect lung cancer.

## Material and Methods

### *Materials*

The data used in the study are microarray analysis data from lung cancer patients that were acquired from CuMiDA database which contains various clean cancer microarray data (Feltes et al., 2019). Three datasets were used in this study: GSE18842, GSE19804, and GSE27262 (Table 1). All three datasets were obtained in a CSV format where the features are represented in columns and the values are represented in rows.

Table 1. Lung Cancer Dataset Overview

| Dataset | Total Samples | Total Features | Total Class |
|---|---|---|---|
| GSE18842 | 90 | 54676 | 2 |
| GSE19804 | 114 | 54676 | 2 |
| GSE27262 | 48 | 54676 | 2 |

### *Mutual information selection*

In information theory, mutual information is a useful technique for calculating the degree to which one random variable contains information about another. Because it offers a quantifiable way to assess a feature's significance, this idea is frequently used in feature selection. A feature is more significant for the task at hand if it has a higher mutual information score since it carries more discriminatory information.

Information theory employs two fundamental concepts: entropy and conditional entropy. Entropy (H) quantifies the uncertainty associated with a random variable. Higher entropy suggests that each event within the variable is equally likely to occur, while lower entropy implies varying probabilities among events.

$$H(X) = -\sum_{x \in dom(X)} p(x) \log p(x) \quad (1)$$

Conditional entropy, denoted as H(X|Y), assesses the uncertainty of X given the knowledge of Y.

$$H(X|Y) = -\sum_{y \in dom(Y)} p(y) \sum_{x \in dom(X)} p(x|y) \log p(x|y) \quad (2)$$

Mutual information, on the other hand, quantifies the amount of information that two random variables, X and Y, share. It is a measure of the degree of correlation between these two variable sets, and it is particularly useful for feature selection (Wang et al., 2019).

$$\begin{aligned} MI(X,Y) &= H(X) - H(X|Y) \\ &= \sum_{x \in dom(X)} \sum_{y \in dom(Y)} p(xy) \log \frac{p(xy)}{p(x)\,p(y)} \end{aligned} \quad (3)$$

Due to its simplicity and efficiency, mutual information is widely regarded as one of the most effective feature selection techniques. It can be applied to select features for various machine learning models because of its' model agnostic nature (Zhongxin et al., 2016). In this research, the authors performed mutual information selection first to make the recursive feature elimination process faster and more efficient by reducing the number of features to be eliminated.

### *PyCaret*

PyCaret is a powerful open-source Python library created by Moez Ali. PyCaret enables us to create various classification and regression machine learning models with low code and automate various activity such as feature selection, cross-validation, and hyperparameter tuning. The library also helps calculate various metrics to evaluate the model performance (Whig et al., 2023).

### *Recursive feature elimination*

Recursive Feature Elimination (RFE) is a feature selection method that uses a machine learning model's criterion to select the most important features from a dataset. To perform recursive feature selection, first, we train a model to the target dataset to evaluate the importance of the features in the dataset. Importance scores will be given to each feature in the dataset. Features will be ranked, and the least important features will be removed. The steps will be iterated until the optimal number of features that maximize the model performance is met (Misra et al., 2018).

Usually, the model used to perform RFE is a Support Vector Machine. However, in this study, other machine learning models that achieve high metrics will also be used to create more robust data. RFE is used to find the most optimal minimum number of features in the dataset based on several models' weights (Huang et al., 2018). Three subsets of features will be produced for each dataset. After that, a fourth subset will be created by combining the three subsets generated for each dataset.

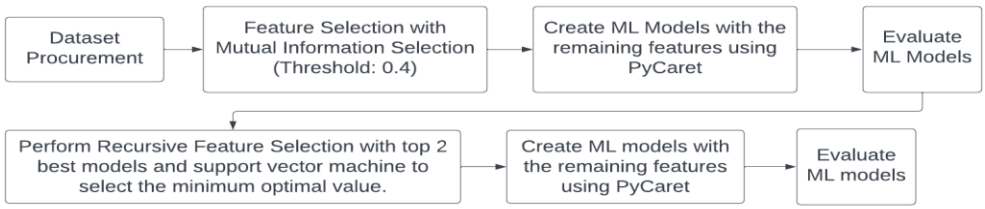


Figure 1. Research framework

## Results and Discussion

The research was conducted on the Google Collab platform. Mutual information selection was performed on the dataset with a 0.4 threshold. Usually, mutual information selection is performed by selecting the top 50 – 200 features with the highest score (Zhongxin et al., 2016). However, we choose a threshold approach because many of the data in the dataset have similar high score and

selecting the top 200 features may result in loss of valuable data. It can be observed from the table below that the dataset has more information compared to the other.

Table 2. Remaining features after mutual information selection

| Dataset | Total Remaining Features |
|---|---|
| GSE18842 | 2953 |
| GSE19804 | 342 |
| GSE27262 | 3974 |

***Recursive feature elimination result***

PyCaret was used to create machine learning models based on the data with remaining features after mutual information selection. PyCaret generates 16 classification machine learning models complete with the evaluation metrics. After that, two of the best models selected by PyCaret and Support Vector Machine were selected to perform RFE. RFE was performed with 10 cross-validations using balanced accuracy as the metric to search for the most minimal feature with the best prediction power. Finally, PyCaret will be used again to train machine learning models based on the data with remaining features after RFE, and the top model is selected to evaluate the performance of each feature subset. The results of models trained with the remaining features can be observed in all the Figures.

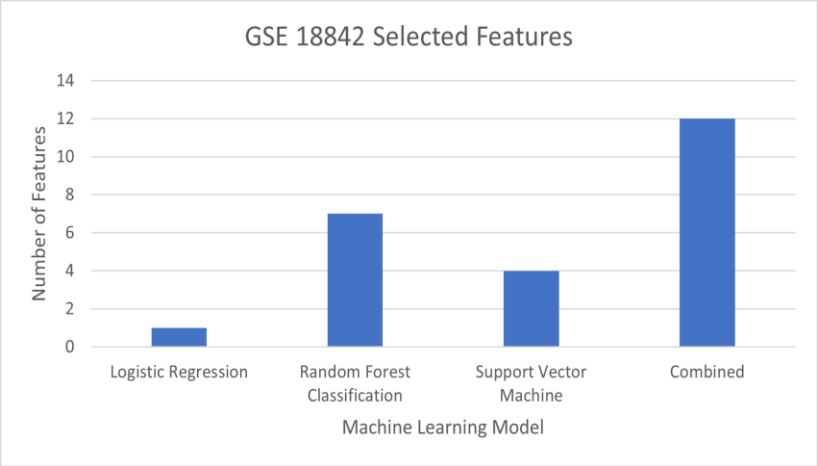


Figure 2. Remaining features in GSE18842 After Recursive Feature Elimination

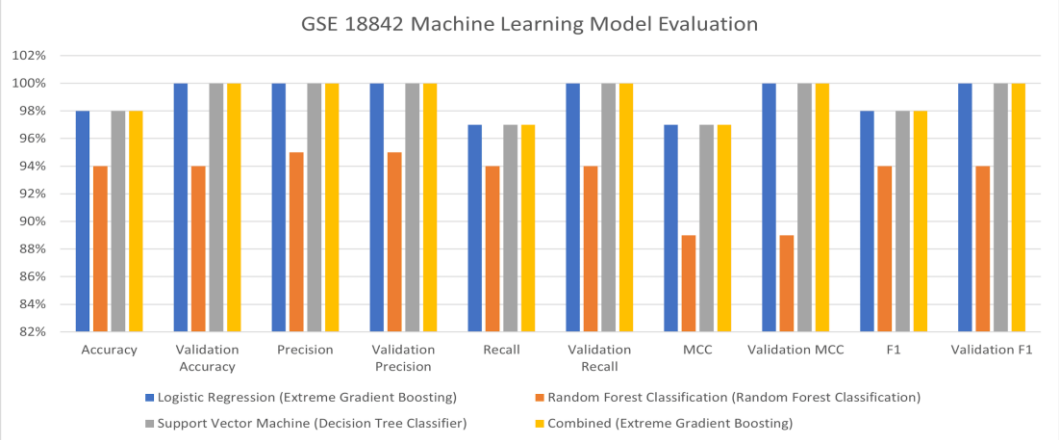


Figure 3. Evaluation of models trained with remaining GSE18842 features

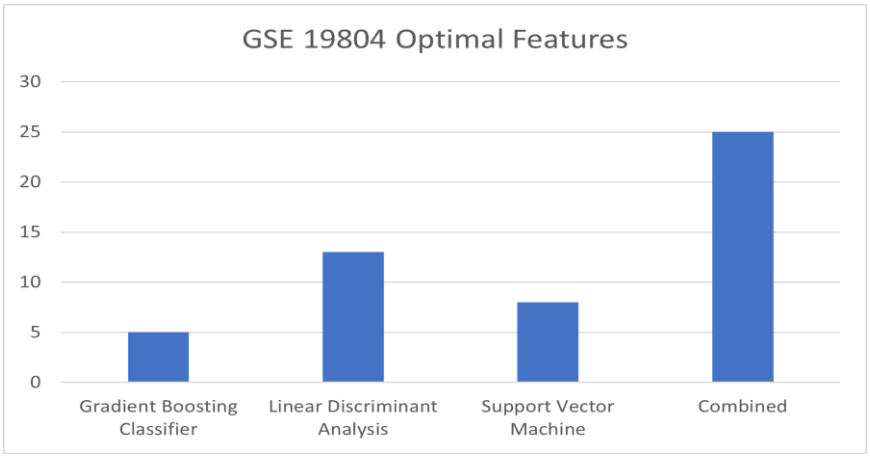


Figure 4. Remaining features in GSE19804 after recursive feature elimination

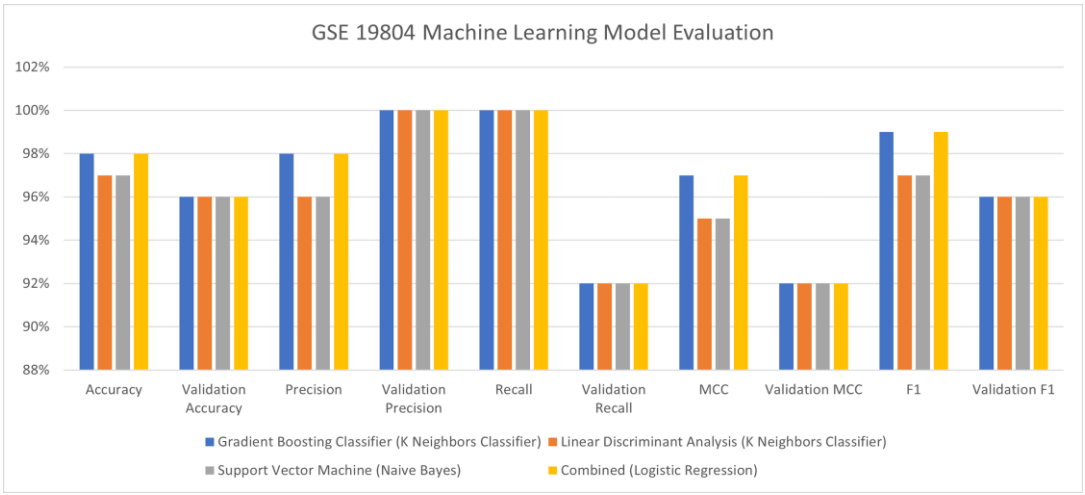


Figure 5. Evaluation of models trained with remaining GSE19804 features

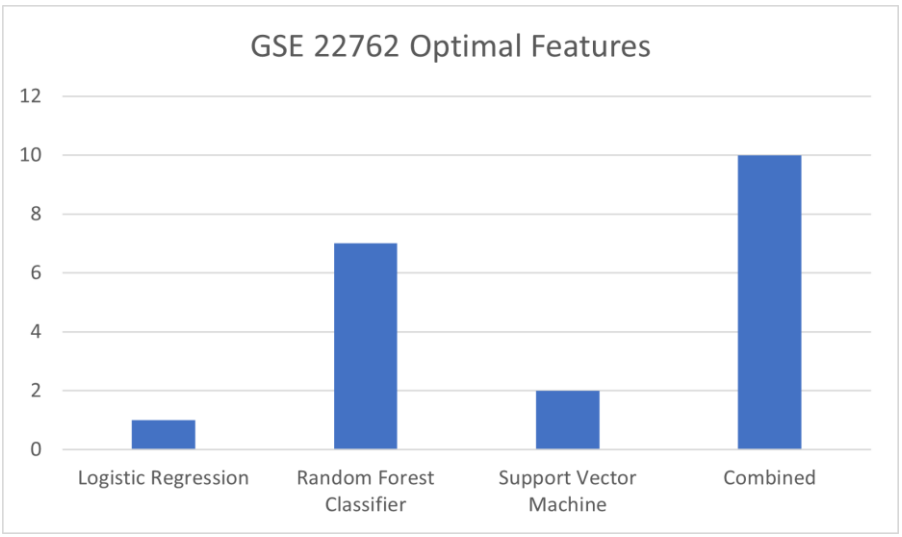


Figure 6. Remaining features in GSE22762 after recursive feature elimination

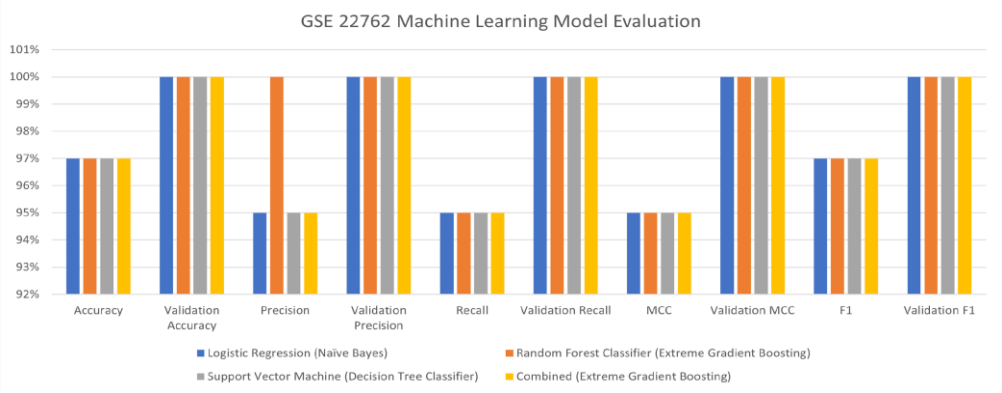


Figure 7. Evaluation of models trained with remaining GSE22762 features

Results indicate that the model can identify the core feature of the dataset and yield models with good results. At first, we were skeptical of the result because data leakage might be present. However, after rechecking, we realized that based on the dashboard in the CuMiDa dataset, it was proven that other people have reached 100% accuracy with another type of model. So, the result is not too good to be true.

## Conclusion

Results indicate that this method can be used to enhance research data by creating more robust research and identifying the least amount of features in the lung cancer microarray dataset that still has a powerful predictive power. The number of dataset features can be reduced from around 50000 to just 25 features and all the machine learning models yield promising results with all metrics achieving scores beyond 95%. Overall, this method can potentially be used in further research on microarray analysis. For the next research, the remaining features can be studied further to know more about how the features impacted the model performance.

## References

Feltes, B. C., Chandelier, E. B., Grisci, B. I., & Dorn, M. (2019). CuMiDa: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. *Journal of Computational Biology*, *26*(4), 376–386. https://doi.org/10.1089/cmb.2018.0238

Huang, X., Zhang, L., Wang, B., Li, F., & Zhang, Z. (2018). Feature clustering based support vector machine recursive feature elimination for gene selection. *Applied Intelligence*, *48*(3), 594–607. https://doi.org/10.1007/s10489-017-0992-2

IQAir. (2021). *2021 World Air Quality Report: Region & City PM2.5 Ranking*. IQAir. https://www.iqair.com/world-most-polluted-cities/world-air-quality-report-2021-en.pdf

IQAir. (2022). *2022 World Air Quality Report: Region & City PM2.5 Ranking.* IQAir. https://www.iqair.com/world-air-quality-report

Mao, Y., Xue, P., Li, L., Xu, P., Cai, Y., Chu, X., … Zhu, S. (2019). Bioinformatics analysis of mRNA and miRNA microarray to identify the key miRNA-gene pairs in small-cell lung cancer. *Molecular Medicine Reports*, *20*(3), 2199–2208. https://doi.org/10.3892/mmr.2019.10441

Misra, P., & Yadav, A. S. (2020). Improving the classification accuracy using recursive feature elimination with cross-validation. *International Journal on Emerging Technologies*, *11*(3), 659–665.

Wang, Y., Yang, X. G., & Lu, Y. (2019). Informative gene selection for microarray classification via adaptive elastic net with conditional mutual information. *Applied Mathematical Modelling*, *71*, 286–297. https://doi.org/10.1016/j.apm.2019.01.044

Whig, P., Gupta, K., Jiwani, N., Jupalle, H., Kouser, S., & Alam, N. (2023). A novel method for diabetes classification and prediction with Pycaret. *Microsystem Technologies*. https://doi.org/10.1007/s00542-023-05473-2

Zhongxin, W., Gang, S., Jing, Z., & Jia, Z. (2016). Feature Selection Algorithm Based on Mutual Information and Lasso for Microarray Data. *The Open Biotechnology Journal*, *10*(1), 278–286. https://doi.org/10.2174/1874070701610010278