UTM Computing Proceedings
Innovations in Computing Technology and Applications
Volume: 6 | Year: 2024 | ISBN No. : 2976-2278

A Combination of Principal Component Analysis and SVM-RFE on Multi-Omics Lung Cancer Classification

Muhammad Nawiyuddin Bin Nunhati, Zuraini Binti Ali Shah
Faculty of Computing
Universiti Teknologi Malaysia
Johor Bahru, Malaysia
nawiyuddin@gradute.utm.my, aszuraini@utm.my

***Abstract*—Classifying lung cancer is a difficult diagnostic since it takes a long time to determine what form of lung cancer the patient has. As the technology became more extensively used, predictive analysis based on previous patient data became available. The omics data of lung cancer are examined as multi-omics in this study for better outcome in identifying lung cancer. The goal of this research is to investigate the machine learning method in the context of solving biological problems via SVM-RFE as feature selection method and Principal Component Analysis as feature extraction method. The lung cancer omics dataset obtained included four omics datasets, three of which are omics data (gene expression, DNA methylation, and miRNA expression) and one clinical data (survival patient). SVM-RFE performed on the pre-processed three omics before is goes for data integration and the pre-processed omics integrated without undergoing the SVM-RFE. So, there are two categories of data, which is with the SVM-RFE and without SVM-RFE. After that the selected features data, it goes for PCA feature extraction before it undergoes classifying. There are three data categories; complete data (multi-omics data without SVM-RFE and PCA), selected data (multi-omics data with SVM-RFE but without PCA) and selected and extracted data (multi-omics data with SVM-RFE and PCA), that went for four classifiers, which is, SVM, Random Forest, Naïve Bayes and Decision Tree. On this research, the selected and extracted data with SVM classifier have the highest accuracy with 0.9688.**

Keywords-Multi-omics; SVM-RFE; PCA; lung cancer; classification.

I. INTRODUCTION

Multi-omics analysis has grown in popularity in biomedical research as various omics, each with its own unique set of characteristics, are combined utilizing effective integration methodologies to get usable data for solving biological problems.

Lung cancer is a deadly cancer that starts in the cells of the lungs. This cancer is dangerous because it can jeopardize someone's life if it is not treated. Males are more likely than females to develop this cancer. Integrating omics data into multi-omics data with the help of technology can aid medical institutions in categorizing diseases. Lung cancer categorization will be done utilizing deep learning feature selection and feature extraction in this study.

II. MATERIALS AND METHODS

The data used is the lung cancer dataset and it consists of three different types of omics data which are gene expression, DNA methylation, microRNA, and patient clinical data. The dataset was obtained from http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html. The main steps involve in conducting the study are data preparation and pre-processing, feature selection, data integration, feature extraction and classifying the lung cancer dataset with SVM, Random Forest, Naïve Bayes and Decision and finally testing and evaluating the performance of all classification models.

A. Data pre-processing

A few tasks have been added, including data transposition, missing value checking, duplicate data removal, and data normalization. To prevent data duplication, it is necessary to check the dataset for missing values and duplicate data and remove them. To improve the precision of the dataset's prediction, data normalization has next been implemented by scaling the numeric attributes into the ranges 0 and 1.

CONFUSION MATRIX TO BE USED FOR THE CLASSIFIERS PERFORMANCE.

Data type	**Number of patients**	**Number of features**
Gene expression	553	20351
DNA methylation	413	5000
miRNA expression	388	1046
Survival data	626	3

B. Feature Selection

One of the involved processes in developing a predictive model is feature selection because it can reduce the number of input variables and the computational burden of modelling, both of which enhance the model's performance. Therefore, the highly correlated features can be chosen using feature selection techniques. SVM-RFE is the feature selection technique used in this research study out of all those that are available. A model is built using this common wrapper method, and features with low weights are then eliminated.

C. Data Integration

It will be possible for researchers to gain a thorough understanding of the mechanisms that underlie biological behavior through the integration of a multi-omics dataset, which will offer significant insights into the flow of biological information at various levels. Integration techniques can also assist the researcher in bridging the gap between various omics and examining the relationship between them. As a result, the pre-processing continues with the concatenation-based integration of the omics and the survival data. Table 2 shows 344 samples and 10243 features to be formed after data integration.

DATASET AFTER DATA INTEGRATION.

Dataset	Number of Samples	Number of Features
Multi-omics	344	10243

D. Feature Extraction

After integration, the dimension of the features is improved using PCA, which showed in Figure 1. We will list the top genes from the combined dataset and PCA dataset that account for 90% of PCs. The dataset must be loaded, the data scaled, the variance ratios for each PC explained, and the top genes that contribute to PCs for data integration listed in order to achieve the desired result in this type of feature extraction.

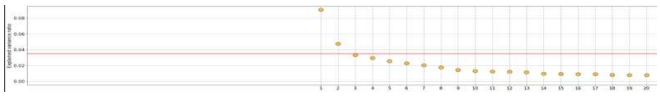

Figure 1 PCA Plot Graph.

E. Classification

The data will be classified into three groups using all four classifiers, SVM, Random Forest, Naive Bayes, and Decision Tree: without SVM-RFE and PCA, without PCA, and with PCA. first importing the dataset. Dataset was gathered from three different sources based on their previously described methods. After that, the data is split into train and test groups. There will be more features in train data than test data. With a train size of 0.7 or 70% of the data, the data will be split into two datasets.

F. Performance Measurement

Confusion matrix will show us whether the classifier can classify the lung cancer patient who survives or dies with the cancer as true or false data. Figures 2, 3, 4, and 5 are the confusion matrices for all classifiers.

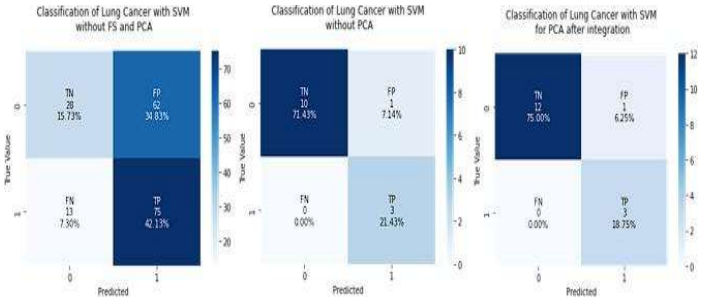

Figure 2 SVM classifier.

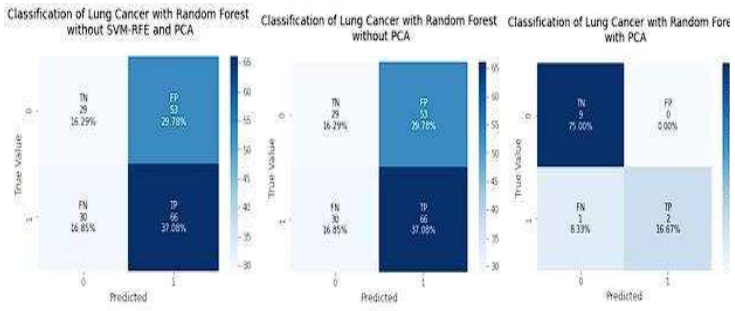

Figure 3 Random Forest classifier

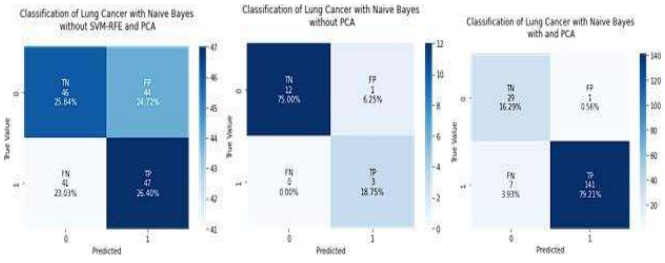

Figure 4 Naive Bayes classifier

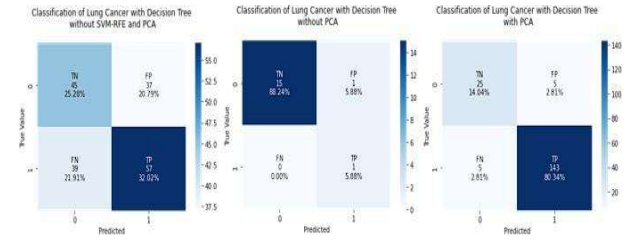

Figure 5 Decision Tree classifier

III. RESULTS AND ANALYSIS

A. Accuracy Results

Table 3 shows the accuracy results of all four classifiers, which are, SVM, random forest, naïve bayes and decision tree and with three methods; complete dataset which are only multi-omics dataset without SVM-RFE and PCA, SVM-RFE data and SVM-RFE with PCA data.

TABLE III ACCURACY RESULTS OF FOUR DIFFERENT CLASSIFIERS WITH THREE APPROACHES IN THE MULTI-OMICS DATASET.

Methods (feature selection and feature extraction)	**Classifiers**			
	SVM	**Random Forest**	**Naïve Bayes**	**Decision Tree**
Complete dataset (without SVM-RFE and PCA)	**0.5787**	0.5337	0.5225	0.5730
SVM-RFE (without PCA)	0.9285	0.9091	0.9375	**0.9412**
SVM-RFE and PCA	**0.9688**	0.9167	0.9551	0.9438

IV. DISCUSSION

Based on the accuracy results in Table 3, result of without SVM-RFE and PCA have the lowest values between the other categories. For without PCA, Decision Tree classifier have the highest accuracy with 0.9412 followed by SVM with 0.9285, Naïve Bayes and Random Forest with 0.9375 and 0.9092 respectively. And the last categories which is with PCA, SVM have the highest accuracy with 0.9688 and Random Forest have the lowest accuracy with 0.9167, the Naïve Bayes and Decision Tree accuracy are 0.9551 and 0.9438 respectively. Here we can say that the difference between without SVM-RFE and with SVM-RFE are contrast. And difference between with and without PCA are not so far yet it increases the accuracy.

V. CONCLUSION

In conclusion, based on the result, SVM-RFE feature selection does increase the accuracy for classifying lung cancer as well as the PCA feature extraction. Based on accuracy, the suggested and best workflow for classifying lung cancer data is by do the combination of SVM-RFE and PCA on multi-omics dataset.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to Faculty of Computing and Universiti Teknologi Malaysia (UTM) for their invaluable support and resources, which played a crucial role in the successful research.

REFERENCES

[1]‘Arif, M., Hassan, H., Nasien, D. and Haron, H. (2015) ‘A Review on Feature Extraction and Feature Selection for Handwritten Character Recognition’, International Journal of Advanced Computer Science and Applications, 6(2), pp. 204–212.

[2] Ayesha, S., Hanif, M. K. and Talib, R. (2020) ‘Overview and comparative study of dimensionality reduction techniques for high dimensional data’, Information Fusion, 59.

[3] Bommert, A., Sun, X., Bischl, B., Rahnenführer, J. and Lang, M. (2020) ‘Benchmark for filter methods for feature selection in high-dimensional classification data’, Computational Statistics and Data Analysis, 143.

[4] Gárate-Escamila, A. K., Hajjam El Hassani, A. and Andrès, E. (2020) ‘Classification models for heart disease prediction using feature selection and PCA’, Informatics in Medicine Unlocked, 19.

[5] Harikumar Rajaguru and Sannasi Chakravarthy S R. (2019) ‘Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer’, Asian Pac J Cancer Prev, 3777-3781.

[6] Mia Huljanah, Zuherman Rustam1, Suarsih Utama1 and Titin Siswantining. (2019). ‘Feature Selection using Random Forest Classifier for Predicting Prostate Cancer’, IOP Conf. Ser.: Mater. Sci. Eng. 546 052031.

[7] Siblini, W., Kuntz, P. and Meyer, F. (2021) ‘A Review on Dimensionality Reduction for Multi-Label Classification’, IEEE Transactions on Knowledge and Data Engineering, 33(3),