RISK CLASSIFICATION FOR NSCLC SURVIVAL USING MICROARRAY AND CLINICAL DATA

[1]RUSSUL ALANNI, [2]JINGYU HOU, [3]HASSEEB AZZAWI, [4]YONG XIANG

[1,2,3,4]School of Information Technology, Deakin University, Burwood, VIC, Australia
E-mail: {ralanni, jingyu.hou, hazzawi, yong.xiang}@deakin.edu.au

Abstract— Lung cancer is one of the most common cancers in the world, and Non-Small Cell Lung Cancer (NSCLC) is the most dangerous and common type of lung cancer. Therefore, it is of paramount importance to predict NSCLC survival, so that suitable treatments can be sought. Nonetheless, conventional methods of risk classification of cancer survival rely solely on histopathology data and predictions are not reliable in many cases. In this paper, we proposed a risk classification model using high-throughput gene expression data and clinical data to predict NSCLC survival. We used Gain Ratio (GR) and Improved Gene Expression Programming (IGEP) algorithms for attribute selection. For classification, we used Support Vector Machine (SVM) alongside with 10-fold cross validation. The results demonstrated the effectiveness of the proposed model with the average accuracy 90.7% which is higher than other representative models. We obtained three gens LCK, DUSP6 and ERBB3 with T_stage and N_stage clinical factors that can get good prediction results.

Index Terms- Lung cancer, Risk classification, microarray dataset, clinical data.

I. INTRODUCTION

Lung cancer is the second prevalent tumor after breast cancer and the most common leading cause of death. In Australia alone 12,741 newly diagnosed lung tumor cases were reported in only one year (2018) and about 9,198 people died from this disease [1]. Lung cancer has many types and subtypes, and each type/subtype has different characteristics. In terms of cell size, lung cancer can be classified into two main types: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). The latter is the most prevalent type of lung cancer, into which about 90% of lung cancer cases are classified [1].

Currently, the most commonly used treatment for NSCLC is the complete surgical resection. However, this treatment approach is improper for all cases and causes many side effects. Therefore the NSCLC surgical resection must be carefully selected and alternative treatments should always be considered [2]. Prognosis prediction of NSCLC could play a very important role in assisting practitioners in selecting proper treatments.

The prognosis assessment for cancer patients who have a specific type/sub-type of cancer depends on several variables in relevant laboratory and clinical data. In medical research, the survival analysis is usually used to analyze time-to-event data. The major output is known as “survival time” which is often calculated as the time length from the time when the cancer is diagnosed to death. The traditional method of assessing whether a treatment is suitable for a patient is to predict a risk score based on which the patient is classified as either low risk or high risk [3-5].

A patient with a low risk score should receive the complete surgical resection, while a patient with a high-risk score should receive either the interrupted treatment or the placement therapy. Recently several cancer survival studies for different kinds of cancer have analyzed clinical factors with high throughput gene expression data such as next-generation sequencing and microarray data [3, 6, 7]. These integrated data can help identify the relationship between different features and cancer. Microarray technology is widely used in cancer classification. Numerous studies have combined microarray gene expression data with clinical data from NSCLC patients to classify patients into different groups for different survival outcomes [3-5]. However, the results are not reliable in some cases. One of the most important reasons is the use of small samples with many parameters. There are currently no standard methods for processing and analyzing cross-laboratory data. Thus, a robust dataset is necessary for such research. We used a large microarray NSCLC dataset and common protocols for pre-processing and analyzing data to establish classification models. In this paper, we propose a method for risk classification of NSCLC survival from microarray data and clinical data.

During the past years, many machine learning approaches have been applied to process and analyze microarray gene expression data [8, 9] to extract different biological meanings. In recent years, a new evolutionary algorithm known as gene expression programming (GEP) [10] was proposed and widely used in classification and decision making. GEP has many advantages that make it easy to achieve an optimal solution to many classification problems. Many studies have shown that GEP provides promising results in cancer prediction [11-15, 29]. Improved Gene Expression Programming (IGEP) algorithm was firstly proposed in [13] to extract the most related genes. The algorithm was is applied to ten microarray datasets and the results showed the effectiveness of the method in selecting small number of related genes while providing higher classification accuracies compared with other existing feature selection methods.

Considering the advantages of IGEP, in this paper we

propose a classification model based on IGEP and SVM. We assessed the prediction of survival among cancer patients, using microarray data and clinical data obtained from multiple laboratories.

This paper is organized as follows. In Section 2, we briefly present the background knowledge of this study. In Section 3, the proposed model is described. The experimental results of our proposed model and the comparison with other representative classifiers on the same datasets are presented in Section 4. Finally, Section 5 concludes our research work.

II. BACKGROUND

A. Gain Ratio

Gain Ratio (GR) is a filtering method which is widely used for reducing bias and noise attributes in a dataset [16]. This algorithm, when being applied to a gene expression dataset, works to rank each gene/feature in the dataset based on its informative information. A gene with high informative information has a higher rank. More details about GR method can be found in [16-18]. In this paper, we calculate the GR values of all the genes in a microarray dataset by using WEKA software [19] with the package called "GainRatioAttributeEval". All the genes are ranked based on their GR values and the genes with a non-zero GR value are selected. The outputs of the GR method and the clinical factor are the input values of the IGEP algorithm.

B. Gene Expression Programming

GEP is an evolutionary algorithm introduced by Ferreira in 2002 [20]. GEP is capable of emulating biological evolution on the basis of computer programming which inherits the merits from Genetic Algorithm (GA) [21] and Genetic Programming (GP) [22]. GEP structure consists of two parts. The first part is phenotype which is a characteristic linear chromosome. Each chromosome in GEP may have many genes and each gene is composed of two parts: head (h) and tail (t). The head contains functional elements such as (+, −, seq, ×, /) and terminal elements which represent the dataset attributes, while the tail contains terminal elements only. In this paper, the terminal elements represent the genes selected by the GR method and the clinical factors. To avoid the possible confusion between the gene in GEP chromosome and the gene in microarray datasets, we use the term feature to represent the element in the terminal set. The size of the chromosome (ch) is based on the length of gene and the number of genes in each chromosome, and it is calculated as follows:

$ch = \text{gene length} * \text{number of genes}$

$\text{gene length} = h + t$

$t = h(n - 1) + 1$

where n represents the maximum number of parameters needed in the function set. When the GEP process starts, the number of genes and h should be determined. Expression tree (ET) or Genotype is the second part of GEP, which is created by converting the phenotype into a tree structure using a special language (Karva notation) invented by the GEP author. GEP process includes several steps. It starts with initializing the GEP parameters such as the terminal set, the function set, and the number of genes in each chromosome. The second step is the coding and initializing the population. In step three each individual (chromosome) is evaluated using a suitable fitness function. All the exiting fitness functions defined by GEP are not suitable for gene selection problem, so in this paper we propose a new fitness function (see Sect. 2.4). In step four, the fitness function result is evaluated based on the GEP conditions. If the condition is satisfied the GEP will stop, and the best individual is selected, otherwise GEP process will continue with step five which applies the genetic operations to modify the individuals. Many genetic operations are defined in GEP such as mutation, root transposition, transposition, gene transposition and recombination. The last step is to prepare the next generation from the modified individuals. The GEP process steps are shown in Fig. 1.

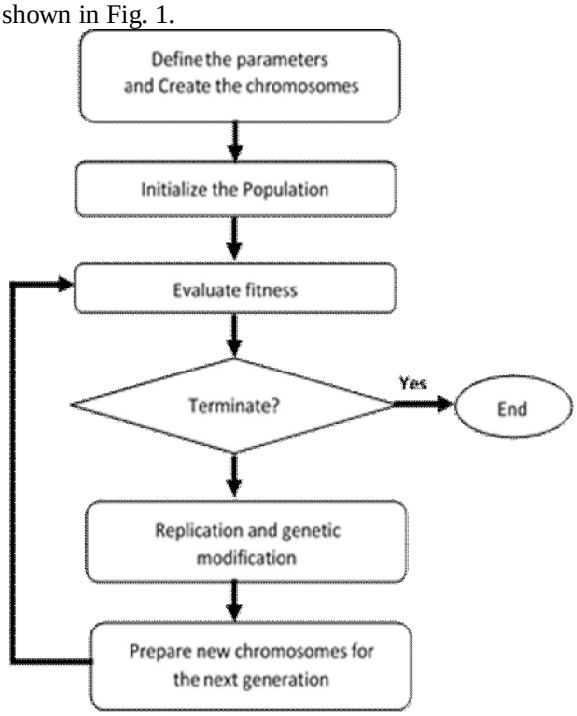


Fig. 1 The flowchart of the GEP process steps.

III. METHOD

C. Data Preparation and Preprocessing

We used R Language [23] to prepare the required data from the original datasets by removing those records that have unknown survival time and duplicated data. We did normalization on microarray data to avoid the effects from variation in the technology rather than from biological differences between the printed probes or between the RNA samples. We used MAS5 function in the microarray package of R for the normalization process. We also used the quintile numeric conversion to reduce individual differences in

gene expression. All the genes expressions are converted into quintile values of 1–5 (very low, low, normal, high and very high).

D. Hybrid GR-IGEP

In this research paper, we use Gain Ratio (GR) as a filtering method to filter the noise and unrelated genes. GR is used on the microarray data only. All the genes in microarray data are assigned GR values and sorted according to their rank. A gene with a higher rank indicates a higher discrimination of this feature compared to the others, as well as more useful information it contains.

The filtering method GR is used as the first stage to decrease the noise by decreasing the number of irrelevant genes/features. Since our results on the microarray data showed that most GR values were zero, the gene selection threshold in this study is then set to zero. If the GR value of a feature/gene is higher than the threshold, the feature is selected. Then we use the Improved Gene Expression Programming (IGEP) to extract the most related features from the microarray and genes and the clinical factors.

The outputs of GR method and the clinical features (age, T_stage and N_stage) are the IGEP input variables. To start the IGEP process, we need to identify the parameters and initialize the population. The parameters used in our IGEP model in this study are listed in Table 1.

Table 1: Parameters used in the proposed IGEP

Parameter	Setting
Terminal set	GR selected genes from the microarray dataset and the clinical factors/data (age, T_stage and N_stage)
Function set	+, -, ×, ÷, sqr
The number of genes	5
Linking functions	Addition
Head size	8
Tail size	9
Gene length	17
Number of chromosomes	200
The number of generations	2000
Genetic operators	
Mutation	0.044
Gene transposition	0.1

After initializing the population, all chromosomes are evaluated based on their fitness value generated by the fitness function. The fitness function is predefined based on the feature selection problem. In our research, we define a new fitness function as follow for feature/gene selection:

$$i = (1 - w) * AC(i) + w * \frac{(f - s_i)}{f}$$

This function consists of two main parts. The first part is based on the accuracy result AC(i) of chromosome i which is generated by support vector machine (SVM) [24] classifiers using 10 fold cross validation. The second part of the fitness function is based on the number of selected features s_i from the total number of features f in the dataset from the microarray and clinical data. Since the AC is more important than the number of selected features, we give a higher weight (w) to AC than the one given to the number of selected features. Specifically w is a random value within the range (0.1, 0.4) giving more importance to the accuracy.

If the terminal condition of IGEP is not satisfied, the process of the method is repeated. We define two terminal conditions: the expected fitness value and the number of generations. The result of this method is the chromosome with the best fitness value. After 10 times of training and testing, the average AC is calculated, while the selected features are selected from the best run. Flow chart of the proposed method is shown in Fig. 2.

IV. EXPERIMENTAL RESULTS

In our experiments, we evaluated the performance of our proposed model for NSCLC risk/recurrence prediction on the integrated microarray data and clinical data. We also compared our model in terms of classification performance with four representative machine learning methods. The settings of our method for building a classifier are presented as follows.

E. Data Preparation and Normalization

In this research, we used microarray and clinical datasets from various platforms for our experiments. Integrated dataset from four institutes was used for training/testing our models.

NSCLC patient's raw data was downloaded from NCBI [40]. This dataset contained the data of 442 NSCLC patients collected from four institutes: University of Michigan Cancer Center (UM) 177 patients, Memorial Sloan-Kettering Cancer Center (MSK) 104 patients, Dana-Farber Cancer Institute (DF) 82 patients, and Moffitt Cancer Center (MC) 79 patients. The clinical data contained information of the NSCLC patients like age, diagnosis stage, treatment, and survival time. The number of records was reduced to 280 because we removed all records that are duplicated or do not contain any information about survival time.

F. Prediction Evaluations

We divided the patients into high-risk/low-risk groups based on their survival time. We used the median survival time as a threshold which was 40 months. The performance of our risk prediction model was evaluated by Accuracy (AC), Positive Predictive Value (PPV), Negative Predictive Value (NPV), Specificity (SP), Sensitivity (SN) and Area Under Curve (AUC). In order to properly evaluate our mode, we used 10-fold cross validation, i.e., we repeated the whole process 10 times on different training and testing datasets and recorded the average results of

these runs. The results of the risk prediction model for ten runs are provided in Table2.

We repeated training and testing for ten times on microarray dataset then we obtained three probes (3 genes). Combining these genes with T_stage and N_stage from clinical data, we can get good prediction results. These genes were 204891_s_at (LCK), 208893_s_at (DUSP6) and 202454_s_at (ERBB3). Table 3 gives a summary of these selected genes.

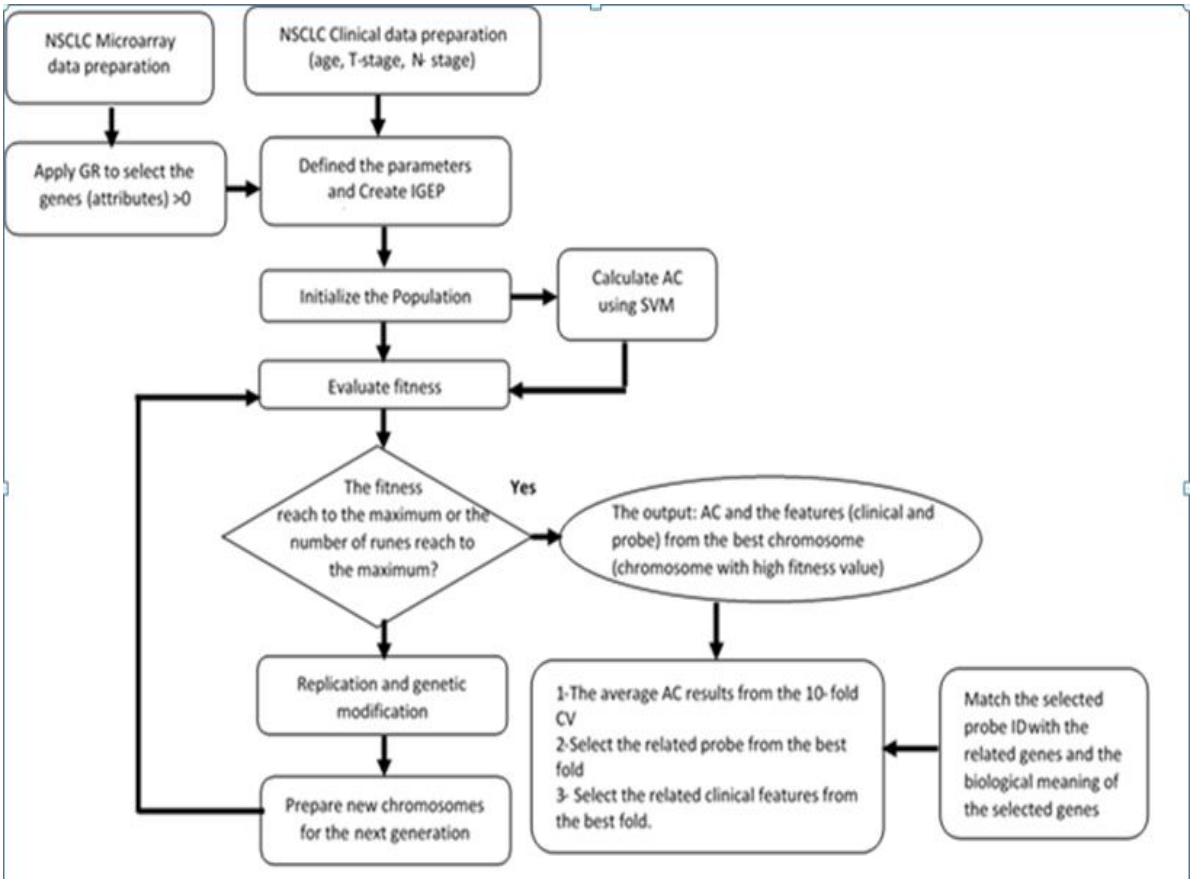


***Fig. 2:** Flow chart of the proposed method. For training and testing the model we use 10-fold cross-validation.*

Table 2: The results of the risk prediction model for ten runs with the average (avg) and the standard deviation (SD)

	AC	PPV	NPV	SN	SP	AUC
1	92.1	91.5	92.7	91.5	92.7	0.9
2	87.7	85.3	90.5	85.3	90.5	0.8
3	87.7	85.3	90.5	85.3	90.5	0.8
4	92.1	91.5	92.7	91.5	92.7	0.9
7	93.7	94.4	92.9	94.4	92.9	0.9
6	93.7	94.4	92.9	94.4	92.9	0.9

G. The competition results

We also compared our model with four representative machine learning methods on the integrated dataset. These four methods were General Regression Neural Network (GRNN) [25], Support Vector Machine (SVM) [26], Decision Tree (DT) [27] and Naive Bayes (NB) [28]. The evaluation values in terms of Accuracy (AC), Positive Predictive Value (PPV), Negative Predictive Value (NPV), Sensitivity (SN), Specificity (SP) and Area Under Curve (AUC) for GR-IGEP, SVM, GRNN, DT, and NB are presented in Fig. 3 and 4.

CONCLUSION

This study introduces a new effective model on microarray and clinical datasets to predict the risk of NSCLC survival. Our proposed model, with other proposed techniques, is able to select informative genes (LCK, DUSP6 and ERBB3) of NSCLC and the T_stage and N_stage from clinical data to construct an effective prediction model which can achieve a higher accuracy. This study combines gene expression data and clinical survival data to predict the risk of NSCLC survival. However, the study did not consider all clinical factors such as the impact of radon, smoke, and history of adult pneumonia on patients. This will be our further research in the near future.

Table 3: The final selected genes from the feature selection model

Probe ID	Gene symbol	Gene Name	Chr.	NCBI UniGene number	Summery	Differential expression (NSCLC vs normal)*	Ranked
204891_s_at	LCK	lymphocyte-specific protein tyrosine kinase	1	3932	The encoded protein is a key signaling molecule in the selection and maturation of developing T-cells	Down-regulated	0.1996
208893_s_at	DUSP6	dual-specificity phosphatase6	12	1848	is gene inactivates (ERK2), resulting in tumor suppression and apoptosis	Down-regulated	0.1954
202454_s_at	ERBB3	v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 3	12	2065	This gene encodes a member of the epidermal growth factor receptor (EGFR) family of receptor tyrosine kinases which are often aberrantly expressed and/or activated in human cancers	Up-regulated	0.1954

* NCBI UniGene number with more information about the genes can be found from NCBI website [42]

REFERENCES

[1] C. Austuralia. (2018, 20 November). Lung cancer in Australia. Available: https://lung-cancer.canceraustralia.gov.au/statistics

[2] W. R. Smythe, A. L. Estrera, S. G. Swisher, K. W. Merriman, G. L. Walsh, J. B. Putnam Jr, et al., "Surgical resection of non–small cell carcinoma after treatment for small cell carcinoma," The Annals of thoracic surgery, vol. 71, pp. 962-966, 2001.

[3] H.-Y. Chen, S.-L. Yu, C.-H. Chen, G.-C. Chang, C.-Y. Chen, A. Yuan, et al., "A five-gene signature and clinical outcome in non–small-cell lung cancer," New England Journal of Medicine, vol. 356, pp. 11-20, 2007.

[4] K. Shedden, J. M. Taylor, S. A. Enkemann, M.-S. Tsao, T. J. Yeatman, W. L. Gerald, et al., "Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study," Nature medicine, vol. 14, p. 822, 2008.

[5] Y.-C. Hsu, S. Yuan, H.-Y. Chen, S.-L. Yu, C.-H. Liu, P.-Y. Hsu, et al., "A four-gene signature from NCI-60 cell line for survival prediction in non–small cell lung cancer," Clinical Cancer Research, pp. 1078-0432. CCR-09-1572, 2009.

[6] H. K. Dressman, A. Berchuck, G. Chan, J. Zhai, A. Bild, R. Sayer, et al., "An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer," Journal of Clinical Oncology, vol. 25, pp. 517-525, 2007.

[7] Y.-C. Chen, W.-C. Ke, and H.-W. Chiu, "Risk classification of cancer survival using ANN with gene expression data from multiple laboratories," Computers in biology and medicine, vol. 48, pp. 1-7, 2014.

[8] F. Zhang, H. L. Kaufman, Y. Deng, and R. Drabier, "Recursive SVM biomarker selection for early detection of breast cancer in peripheral blood," BMC medical genomics, vol. 6, p. S4, 2013.

[9] D. L. Tong and A. C. Schierz, "Hybrid genetic algorithm-neural network: Feature extraction for unpreprocessed microarray data," Artificial intelligence in medicine, vol. 53, pp. 47-56, 2011.

[10] C. Ferreira and U. Gepsoft, "What is gene expression programming," ed, 2008.

[11] R. Al-Anni, J. Hou, R. D. a. Abdu-aljabar, and Y. Xiang, "Prediction of NSCLC recurrence from microarray data with GEP," IET systems biology, vol. 11, pp. 77-85, 2017.

[12] H. Azzawi, J. Hou, Y. Xiang, and R. Alanni, "Lung cancer prediction from microarray data by gene expression programming," IET systems biology, vol. 10, pp. 168-178, 2016.

[13] R. Alanni, J. Hou, H. Azzawi, and Y. Xiang, "New Gene Selection Method Using Gene Expression Programing Approach on Microarray Data Sets," in International Conference on Computer and Information Science, 2018, pp. 17-31.

[14] H. Azzawi, J. Hou, R. Alanni, Y. Xiang, R. Abdu-Aljabar, and A. Azzawi, "Multiclass Lung Cancer Diagnosis by Gene Expression Programming and Microarray Datasets," in International Conference on Advanced Data Mining and Applications, 2017, pp. 541-553.

[15] H. Azzawi, J. Hou, R. Alnnni, and Y. Xiang, " A Hybrid Neural Network Approach for Lung Cancer Classification with gene expression dataset and prior biological knowledge," in 2018 International Conference on Machine Learning for Networking (MLN'2018), 2018.

[16] R. P. Priyadarsini, M. Valarmathi, and S. Sivakumari, "Gain ratio based feature selection method for privacy preservation," ICTACT J Soft Comput, vol. 1, pp. 2229-6956, 2011.

[17] A. G. Karegowda, A. Manjunath, and M. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," International Journal of Information Technology and Knowledge Management, vol. 2, pp. 271-277, 2010.

[18] J. Dai and Q. Xu, "Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification," Applied Soft Computing, vol. 13, pp. 211-221, 2013.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," ACM SIGKDD explorations newsletter, vol. 11, pp. 10-18, 2009.

[20] C. Ferreira, "Gene expression programming in problem solving," in Soft computing and industry, ed: Springer, 2002, pp. 635-653.
[21] D. E. Golberg, "Genetic algorithms in search, optimization, and machine learning," Addion wesley, vol. 1989, p. 102, 1989.
[22] J. R. Koza, Genetic programming: on the programming of computers by means of natural selection vol. 1: MIT press, 1992.
[23] R. C. Team, "R: A language and environment for statistical computing," 2014.
[24] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in European conference on machine learning, 1998, pp. 137-142.
[25] D. F. Specht, "A general regression neural network," IEEE transactions on neural networks, vol. 2, pp. 568-576, 1991.
[26] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," 2003.
[27] M. Pal and P. M. Mather, "An assessment of the effectiveness of decision tree methods for land cover classification," Remote sensing of environment, vol. 86, pp. 554-565, 2003.
[28] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy," Applied and environmental microbiology, vol. 73, pp. 5261-5267, 2007.
[29] H. Azzawi, J. Hou, R. Alnnni, and Y. Xiang, "SBC: A New Strategy for Multiclass Lung Cancer Classification Based on Tumour Structural Information and Microarray Data," in 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), 2018, pp. 68-73.

★★★