Lung Cancer Prediction Using Machine Learning Classifiers

Ankush Kumar Gulia [a], Rajat Bhatt [b]

[a],[b]Computer Science and Engineering,SRM Institute of Science and Technology ,Kattankulathur, India ᶜ Siirt University, Education Faculty, Siirt/Turkey
E-Mail: [a]aa1161@srmist.edu.in, [b]rv8014@srmist.edu.in

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 23 May 2021

Abstract: Lung Cancer is the major cause of mortality which are cancer-related. Therefore, the diagnosis, prediction and detection of this disease has become very important. Machine Learning techniques have been in use for the medication of such conditions because of their high accuracy results. For the prognosis and analysis of lung cancer in the healthcare sector, various machine learning algorithms have been utilized such as Artificial Neural Networks (ANN), Naive Bayes, Logistic Regression and Support Vector Machines (SVM). The causes of lung cancer and machine learning algorithms' applications are discussed in this review. The advantages and disadvantages of these algorithms are also discussed. Experimental study proved that the proposed model is a highly optimized model of existing machine learning algorithms..

Keywords: Lung Cancer Prediction, Machine Learning, Random Forest, Support Vector Machines (SVM), Naïve Bayes, Disease Prediction, Dataset

1. Introduction

A large number of deaths are caused globally due to this harmful disease called lung cancer. Lung cancer needs to be encountered necessarily to reduce the rate at which patients suffer death. Thus detection and diagnosis of lung cancer is a huge hurdle encountered by researchers and doctors every year. Medical imagery like MRI scans, chest X-rays, computed tomography, etc., can be used for the detection of lung cancer. The important characteristics of vast lung cancer datasets are what are recognized by the approaches of ML. Machine Learning algorithms such as Naive Bayes, Decision Trees, K-Nearest Neighbours and so on have had a profound impact on health care.

Region	*Population*	*Cancer Cases*	*Deaths*
Asia	60%	66%	57.3%
Europe	9.0%	23.4%	20.3%
America	13.3%	21.0%	14.4%
Africa	17%	9-10%	7.3%

Figure 1: Lung Cancer Cases and Deaths over the globe

Depending on the location and tumor size, the symptoms are categorized. Due to the lack of pain and symptoms in the early stages, it is very difficult to detect in some cases. Cellular breakdown in the lungs analyzed patients may endure Cough, Chest torment, Shortness of breath, Wheezing, Hemoptysis. For example hacking up blood, Pancoast condition (shoulder torment), Hoarseness (loss of motion of vocal ropes), Weight misfortune, Weakness, and Fatigue. Kinds of cellular breakdown in the lungs are pictorially appeared in figure 2.

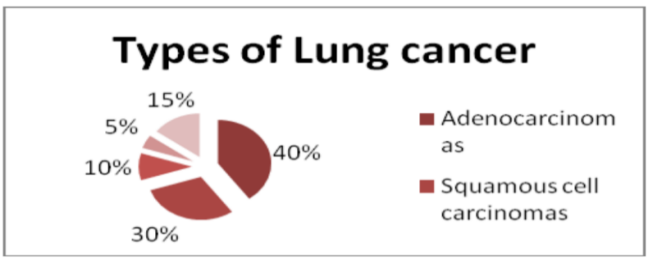

Figure 2: Lung Cancer Classification

90% of it is initiated because of smoking. Impregnation of tobacco smoke likewise causes cellular breakdown in the lungs for example known as uninvolved smoking. Another factor for cellular breakdown in the lungs is heredity. Vehicular contamination, ventures, the admission of destructive gases, for example, Radon remains at the second situation in causing the passings from cellular breakdown in the lungs.

Doctors distinguish the presence and phase of cancer by inciting different tests, for example, MRI sweeps, CT examine, X-beam, bone outputs, and PET sweeps. NSCLC is fanned into four stages dependent on seriousness: Stage 1 Limited to lungs, Stage 2 Limited to the chest, and Stage 3 Limited to the chest anyway in the midst of bigger and major forceful tumors and Stage 4 spread to various areas in the body. SCLC type prescription is controlled by the two-layered model: Limited stage SCLC, Extensive stage (ES) SCLC.

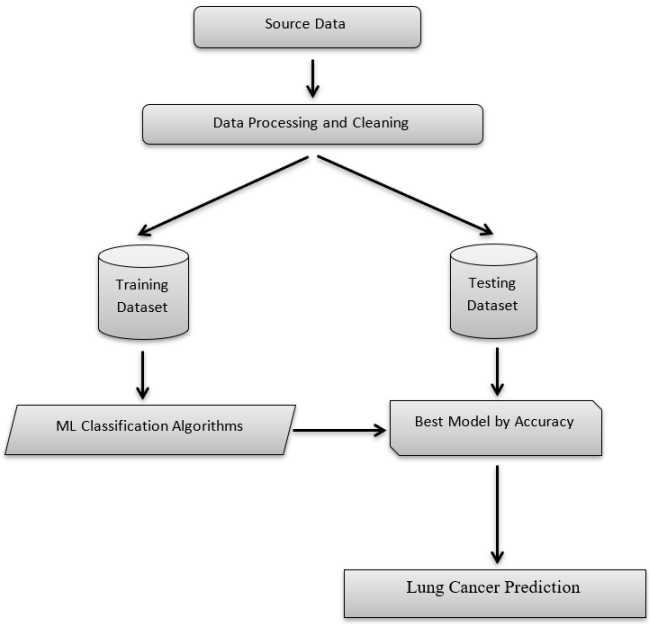

Figure 3: Workflow Diagram

2. State Of The Art (Literature Survey)

Sanjukta Rani Jena [1]; said Cellular breakdown in the lungs is the most hazardous infection, treatment of which should be the essential objective all through logical exploration. The early acknowledgement of malignancy can be useful in restoring infection totally. There are various strategies found in writing for the identification of cellular breakdown in the lungs. A few agents have contributed their realities for malignancy forecasts. These papers generally agree about winning cellular breakdown in the lungs identification strategies that are reachable in the writing. A number of procedures have been begun in disease identification approaches to advance the proficiency of their discovery. Various applications like neural organizations, support vector machines, picture handling procedures are broadly utilized for malignant growth discovery which is expounded in this work.

Jue Jiang [2]; said that they created two different goal-excessively associated network plans called steady MRRN and thick MRRN. Their organizations all the while joined highlights across different picture goals and highlight levels through leftover associations with recognized and fragment lung tumours. They assessed their technique on an aggregate of 1210 non-little cell (NSCLC) lung tumours and knobs from three datasets comprising of 377 tumours from the open-source Cancer Imaging Archive (TCIA), 304 progressed stage NSCLC treated with hostile to PD-1 designated spot immunotherapy from inside foundation MSKCC dataset, and 529 lung knobs from the Lung Image Database Consortium (LIDC). The calculation was prepared utilizing the 377 tumours from the TCIA dataset and approved on the MSKCC and tried on LIDC datasets. The division exactness contrasted with master depictions was assessed by registering the Dice Similarity Coefficient (DSC), Hausdorff distances, affectability and accuracy measurements. Their best performing gradual MRRN strategy delivered the most noteworthy DSC of 0.74±0.13 for TCIA, 0.75±0.12 for MSKCC and 0.68±0.23 for the LIDC datasets. There was no critical distinction in the assessments of volumetric tumour changes figured utilizing the gradual MRRN strategy contrasted and master division proposed two neural organizations to fragment lung tumours from CT pictures by adding various lingering surges of shifting resolutions. Their outcomes unmistakably show the improvement in division precision across numerous datasets.

Naji Khosravan [3]; said that early recognition of lung knobs is critical in the screening of cellular breakdown in the lungs. Numerous CAD frameworks, which are utilized as malignancy identification apparatuses, produce a ton of false positives (FP) and require a further FP decrease step. Moreover, rules for early finding and therapy of cellular breakdown in the lungs comprise of various shape and volume estimations of anomalies. To help this theory they proposed a 3D profound perform multiple tasks CNN. They tried their framework on the LUNA16 dataset and accomplished a normal dice likeness coefficient of 91% as division exactness and a score of almost

92% for FP decrease. As proof of their speculation, they showed enhancements of division and FP decrease errands more than two baselines. They proposed a 3D profound perform multiple tasks CNN for all the while performing division and FP decrease. They showed that sharing some fundamental highlights for these assignments and preparing a solitary model utilizing shared highlights can improve the outcomes for the two undertakings, which are basic for the screening of cellular breakdown in the lungs. Moreover, They showed that a semi-supervised approach can improve the outcomes without the requirement for a huge number of named data sets in the preparation.

Nidhi S. Nadkarni [4]; This paper presents a computerized approach for the location of cellular breakdown in the lungs in CT examine pictures. The calculation for the cellular breakdown in the lungs discovery is proposed utilizing techniques, for example, middle separating for picture preprocessing followed by division of lung district of interest utilizing numerical morphological tasks. The framework for programmed recognition of cellular breakdown in the lungs in CT pictures was effectively evolved utilizing a picture handling strategy. The received procedure performs well in upgrading, fragmenting and separating highlights from CT pictures. The middle separating method was compelling in killing motivation commotion from the pictures without obscuring the picture. Numerical morphological activities empower the exact division of lung and tumour locale.

Tianle Shen [5]; said that they led a review study to assess adequacy and security in cellular breakdown in the lungs patients with persistent silicosis, particularly zeroing in on the occurrence of radiation pneumonitis (RP). Cellular breakdown in the lungs patients with constant silicosis who had been treated with radiotherapy from 2005 to 2018 in their medical clinic were selected for this review study. RP was reviewed by the National Cancer Institute's Common Terminology Criteria for Adverse Events (CTCAE), variant 3.0. Of the 22 patients, ten (45.5%) created RP ≥2. Two RP-related passings (9.1%) happened within 3 months after radiotherapy. Dosimetric factors V 5 , V 10 , V 15 , V 20 and mean lung portion (MLD) were fundamentally higher in patients who had RP >2 ($P < 0.05$). The middle by and large endurance times in patients with RP ≤2 and RP>2 were 11.5 months and 7.1 months, individually. Radiotherapy is related to unnecessary and lethal pneumonic harmfulness in cellular breakdown in the lungs with constant silicosis.

Saeed S Alahmari [6]; said that they directed examinations to evaluate the presentation of fusing delta with regular (non delta) highlights utilizing AI to foresee lung knob threat. They tracked down the best improved territory under the collector working trademark bend was 0.822 when delta highlights were joined with traditional highlights versus an AUC 0.773 for customary highlights as it were. Generally speaking, this examination showed the significant utility of joining delta radiomics highlights with customary radiomics highlights to improve execution of models in the cellular breakdown in the lungs screening setting.

Yanbo Wang [7]; said that they report a non-Gaussian graphical model to remake the quality connection network utilizing two recently distributed quality articulation datasets. Their graphical model plans to specifically distinguish net primary changes at the degree of quality cooperation between organizations. Their technique is widely approved, showing great strength, just as the selectivity and particularity anticipated depend on their natural experiences. In synopsis, quality administrative organizations are still moderately stable during probably the beginning phase of neoplastic change. Strangely, their strategy can likewise recognize early secluded changes, with the ALDH3A1 and its related communications being emphatically embroiled as an expected early marker, whose enactments seem to adjust LCN2 module just as its collaborations with the significant TP53-MDM2 hardware. Their methodology utilizing the graphical model to reproduce quality communication work with naturally propelled limitations represents the significance and magnificence of science in building up any bio-computational methodology.

3.Proposed Work

Machine learning supervised classification algorithms will be used to give a dataset as the input and then extract patterns, which would, in turn, help in predicting how likely it is that the patient is affected. Thereby, helping to make better decisions in future.

- ❑ The dataset is split into a test set and training set. Generally, a 3:7 ratio is applied to split the test set and training set.
- ❑ Using the following algorithms:

1. Naive Bayes
2. Logistic Regression
3. Random Forest
4. SVM

The accuracy of each model is calculated and the model with the highest accuracy is selected.

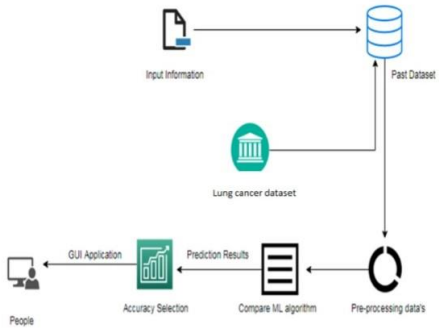

Figure 4: Business Diagram / System Architecture

4.Implementation

4.1.Data Validation and preprocessing:

Machine learning data validation techniques are used to obtain the error rate of the ML model. The error rate obtained by validation is quite close to the true error rate of the dataset. However, if the data volume is large, so much so that it can represent the population then one might not need the validation techniques. But in the real-world we need to work with samples of data that may not be a true representative of the population of given dataset.

```
In [11]: #Checking datatype and information about dataset
         df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1298 entries, 0 to 1297
Data columns (total 11 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Name                 1298 non-null   object
 1   Member_ID            1298 non-null   int64
 2   Diagnosis            1298 non-null   object
 3   Age                  1298 non-null   int64
 4   Smokes               1298 non-null   int64
 5   Smokes (years)       1298 non-null   object
 6   Smokes (packs/year)  1298 non-null   object
 7   AreaQ                1298 non-null   int64
 8   Alkhol               1298 non-null   int64
 9   family history       1298 non-null   int64
 10  Result               1298 non-null   int64
dtypes: int64(7), object(4)
memory usage: 121.7+ KB
```

Figure 5: Checking data type and information about dataset

Whenever the data is gathered from different sources, it is in raw form and is not feasible for analysis. Pre-processing is used to convert this raw data into much more cleaner dataset. This transformation, that the raw data undergoes before feeding it to the algorithm is called pre-processing.

4.2.Data Visualization:

Data visualization is a very crucial process in machine learning and applied statistics. Statistics focuses on quantitative descriptions and estimations of data. Data visualization provides a set of important tools for gaining a qualitative understanding. It can be helpful while exploring and understanding a dataset and can help in identifying corrupt data, patterns, outliers and much more. With a little domain knowledge, data visualizations can be used for demonstrating and expressing key relationships in charts and plots.

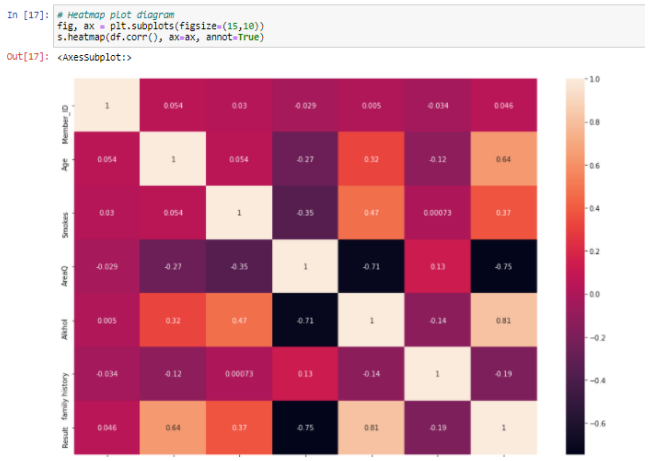

Fig 6: Heatmap plot diagram of correlation between features

Data visualization and exploratory data analysis are fields of study in themselves and it will require a deeper dive into some books. Sometimes data does not make much sense until it is in a visual form, such as with charts and plots. Being able to quickly visualize data samples and others is a crucial skill both in applied statistics and in machine learning. It discovers the many types of plots that you will need to know when visualizing data in

Python and how to use them to better understand your own data.

4.3.Algorithms:

4.3.1.Naive Bayes Algorithm:

It is a supervised learning algorithm based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset.

```
Accuracy result of Naive Bayes is: 98.4047619047619

Confusion Matrix result of Naive Bayes is:
 [[99  0]
 [ 4 86]]

Sensitivity :  1.0

Specificity :  0.9555555555555556
```

Figure 7: Implementation Results of Naive Bayes

It is one of the simple and most effective classification algorithms which helps in building fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

4.3.2.Logistic Regression:

It is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

```
Classification report of Logistic Regression Results:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00        99
           1       1.00      1.00      1.00        90

    accuracy                           1.00       189
   macro avg       1.00      1.00      1.00       189
weighted avg       1.00      1.00      1.00       189

Cross validation test results of accuracy:
[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]


Accuracy result of Logistic Regression is: 100.0

Confusion Matrix result of Logistic Regression is:
 [[99  0]
 [ 0 90]]

Sensitivity :  1.0

Specificity :  1.0
```

Figure 8: Implementation Results of Logistic Regression

The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. In other words, the logistic regression model predicts P(Y=1) as a function of X.

4.3.3. Support Vector Machine (SVM)

A classifier that categorizes the data set by setting an optimal hyperplane between data. This classifier is incredibly versatile in the number of different kernel functions that can be applied and this model can yield a high predictability rate. Support Vector Machines are perhaps one of the most popular and talked about machine learning algorithms.

```
Classification report of Support Vector Machines Results:

              precision    recall  f1-score   support

           0       0.96      1.00      0.98        99
           1       1.00      0.96      0.98        90

    accuracy                           0.98       189
   macro avg       0.98      0.98      0.98       189
weighted avg       0.98      0.98      0.98       189

Cross validation test results of accuracy:
[0.88888889 1.         1.         1.         1.         0.88888889
 1.         1.         1.         1.         1.         0.88888889
 1.         1.         1.         1.         0.88888889 1.
 1.         1.         1.         1.         1.         0.88888889
 1.         1.         1.         1.         1.         1.
 0.88888889 1.         1.         1.         1.         1.
 1.         0.88888889 1.         1.         1.         1.
 1.         1.         0.88888889 1.         1.         1.
 1.         1.         1.         0.88888889 1.         1.
 1.         1.         1.         1.         0.88888889 1.
 1.         1.         1.         1.         1.         0.88888889
 1.         1.         1.         1.        ]


Accuracy result of Support Vector Machines is: 98.25396825396825

Confusion Matrix result of Support Vector Machines is:
 [[99  0]
 [ 4 86]]

Sensitivity :  1.0

Specificity :  0.9555555555555556
```

Figure 9: Implementation Results of SVM

We use Kernelized SVM for non-linearly separable data. The kernel function in a kernelized SVM tells you, given two data points in the original feature space, what the similarity is between the points in the newly transformed feature space. There are various kernel functions available, but two of are very popular :

i) Radial Basis Function Kernel (RBF)

ii) Polynomial Kernel

4.3.4.Random Forests:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or the same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e.

multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

```
Classification report of Random Forest Results:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00        99
           1       1.00      1.00      1.00        90

    accuracy                           1.00       189
   macro avg       1.00      1.00      1.00       189
weighted avg       1.00      1.00      1.00       189

Cross validation test results of accuracy:
[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]


Accuracy result of Random Forest is: 100.0

Confusion Matrix result of Random Forest is:
 [[99  0]
 [ 0 90]]

Sensitivity :  1.0

Specificity :  1.0
```

Figure 10: Implementation Results of Random Forests

The following are the basic steps involved in performing the random forest algorithm:

1. Pick N random records from the dataset.
2. Build a decision tree based on these N records.
3. Choose the number of trees you want in your algorithm and repeat steps 1 and 2.
4. In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output). The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote.

5.Results discussion

Lungancer prediction is designed using Python and its important libraries such as Pandas for reading CSV files for using the classification algorithms.

Algorithm	**Accuracy (%)**
Naive Bayes	98.4 %
Logistic Regression	100.0 %
Random Forests	100.0 %
Support Vector Machines	98.2 %

The first phase is Data Validation where we clean our dataset so that it can be used for ML algorithms. The second phase is Data Pre-Processing where some transformations are applied to the data set before it is fed to the algorithm. The third phase is to check the accuracy of the algorithms mentioned, NB, LR, RF and SVM.

The below figure shows the accuracy of the algorithms tested.

Table 1: Experimental Results

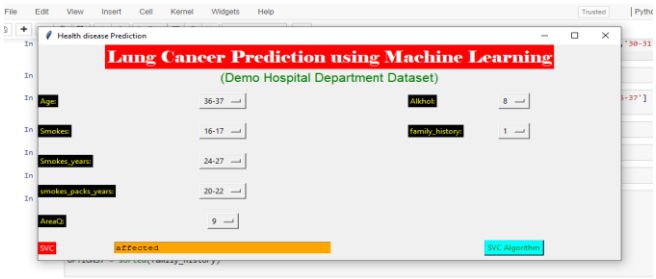

Figure 11: Result through Graphical User Interface

6.Conclusion

In this paper, we have shown that the Logistic Regression and Random Forests algorithms had the highest accuracy among the four ML algorithms when applied to our dataset. As we already know, Logistic Regression is an algorithm that is specifically used for binary classification problems. And since our problem statement is a binary classification problem, Logistic Regression is preferred over Random Forests in general. So in the analysis it can be predicted that with suitable feature selection methods and an integrated approach with other supervised learning processes and modified functional approach in Logistic Regression, accuracy will be further improved with respect to other datasets.

References

1. Sanjukta Rani Jena, Dr. Thomas George and Dr. Narain Ponraj; “Feature Extraction and Classification Techniques for the Detection of Lung Cancer”, 2019 International Conference on Computer Communication and Informatics
2. Jue Jiang, Yu-chi Hu, Chia-Ju Liu, Darragh Halpenny, Matthew D. Hellmann, Joseph O. Deasy, Gig Mageras and Harini Veeraraghavan; “Multiple Resolution Residually Connected Feature Streams For Automatic Lung Tumor Segmentation From CT Images”, IEEE Transactions on Medical Imaging Jan 2019
3. Naji Khosravan and Ulas Bagci; “Semi-Supervised Multi-Task Learning for Lung Cancer Diagnosis”, Annual International Conference of IEEE Engineering in Medicine and Biology Society 2018
4. Nidhi S. Nadkarni and Prof. Sangam Borkar; “Detection of Lung Cancer in CT Images using Image Processing”, International Conference on Trends in Electronics and Informatics 2019
5. Tianle Shen, Liming Sheng, Ying Chen , Lei Cheng and Xianghui Du; “High incidence of radiation pneumonitis in lung cancer patients with chronic silicosis treated with radiotherapy”, Journal of Radiation Research, Dec 2019
6. Saeed S Alahmari, Dmitry Cherezov, Dmitry Goldgof, Lawrence Hall, Robert J Gillies and Matthew B Schabath; “Delta Radiomics Improves Pulmonary Nodule Malignancy Prediction in Lung Cancer Screening”, IEEE Access 2018
7. Yanbo Wang, Weikang Qian and Bo Yuan; “A Graphical Model of Smoking-Induced Global Instability in Lung Cancer”, IEEE/ACM Transactions on Computational Biology and Bioinformatics 2016