A Prediction Model for Early-Stage Detection of Lung Cancer

Chand Basha P[1], Anjan Babu G[2]

Dept of Computer Science, SV University, Tirupati

***Abstract*—** *Researchers have widely used statistical and machine learning techniques to construct prediction models in several domains such as prediction of software faults, spam detection, disease diagnosis, and financial fraud identification. The prediction of patients prone to lung cancer can help doctors in their decision making regarding their treatments. In this regard, this research paper attempts to evaluate the discriminative power of several predictors in the study to increase the efficiency of lung cancer detection through their symptoms. A number of classifiers including Logistic Regression, Support Vector Machine (SVM), Decision tree, K-Nearest Neighbour (KNN), and Naïve Bayes (NB) and Random Forest Classification are evaluated on a benchmark dataset obtained from UCI repository. The performance is also compared with well-known ensembles such as Random Forest and Majority Voting. Based on performance evaluations, it is observed that ROC Curve outperformed all other individual as well as ensemble classifiers and achieved 98% accuracy*.

I. INTRODUCTION

In 2012, a survey was conducted, which reports 1.6 million deaths and 1.8 million new cases of lung cancer patients. Lung cancer is common in both gender of US and reported as more dangerous as compared to other types of cancer. Only 15% of cases are detected at the early stage. The most common symptom, i.e., smoking, is reported for lung cancer patients but not all patients involve this symptom. However, several symptoms of lung cancer patients such as their smoking rate can help to detect the lung cancer patient at the early stage. Though, the research community has used certain machine learning techniques such as, Fuzzy deep learning, and SVM. Wender et al. reported that lung cancer as a serious killer disease in the world mainly in America and East Asia. Moreover, the authors presented that lung cancer patients are 25% higher than patients of other cancer types such as breast cancer and blood cancer. The tumor movements are divided into two parts intra-fractional variation and interfractional variation. Intra-fractional works in single treatment sessions and inter-fractional arises between different sessions. Consequently, we evaluate the discriminative power of certain predictors used to improve the accuracy of the prediction model.

The dataset was retrieved from the UCI repository. Firstly, the effectiveness of Naïve Bayes (NB), Random Forrest (RF), Support Vector Machine (SVM), and MLP is assessed in terms of accuracy and f-measure. Secondly, a majority voting-based ensemble of top-3 best performing classifiers is constructed to predict lung cancer.

The major contributions of this research are as follows

- Identification of well-known classifiers and ensemble approaches utilized for lung cancer prediction
- Computation of results over benchmark dataset obtained from UCI repository
- Proposed a majority voting-based ensemble based on top-3best performing individual classifiers
- Comparison and evaluation of results which show that Gradient-boosted Tree outperformed all another individual as well as ensemble classifiers
- Achieved 90% accuracy for lung cancer identification the remaining part of the study is structured in four sections.

II. LITERATURE REVIEW

Lung Cancer Classification Tool Using Microarray Data and Support Vector Machines

Authors: Jennifer Cabrera; Abigaile Dionisio; Geoffrey Solano - 2019

Lung cancer is one of the deadliest types of cancer around the world. Epidemiologic studies have shown that genetic variability is among the factors that affect a person's susceptibility to lung cancer. A recent study conducted by a team of researchers from the United States National Cancer Institute among 14,000 Asian women found out that Asian women, whether smokers or not, are more prone to developing cancer due to their genetic variations. This study proposes a system that utilizes gene expression data from oligonucleotide microarrays to predict the presence or absence of lung cancer, predict

the specific type of lung cancer should it be present, and determine marker genes that are attributable to the specific kind of the disease. The proposed system would help in the faster diagnosis and serve as a reliable adjunct approach to current lung cancer classification methods.

Early-Stage Lung Cancer Diagnosis by Deep Learning-Based Spectroscopic Analysis of Circulating Exosomes

Authors: Hyunku Shin, Seunghyun Oh, Soonwoo Hong – 202

Lung cancer has a high mortality rate, but an early diagnosis can contribute to a favorable prognosis. A liquid biopsy that captures and detects tumor-related biomarkers in body fluids has great potential for early-stage diagnosis. Exosomes, nanosized extracellular vesicles found in blood, have been proposed as promising biomarkers for liquid biopsy. Here, we demonstrate an accurate diagnosis of early-stage lung cancer, using deep learning-based surface-enhanced Raman spectroscopy (SERS) of the exosomes. Our approach was to explore the features of cell exosomes through deep learning and figure out the similarity in human plasma exosomes, without learning insufficient human data. The deep learning model was trained with SERS signals of exosomes derived from normal and lung cancer cell lines and could classify them with an accuracy of 95%. In 43 patients, including stage I and II cancer patients, the deep learning model predicted that plasma exosomes of 90.7% patients had higher similarity to lung cancer cell exosomes than the average of the healthy controls. Such similarity was proportional to the progression of cancer. Notably, the model predicted lung cancer with an area under the curve (AUC) of 0.912 for the whole cohort and stage I patients with an AUC of 0.910. These results suggest the great potential of the combination of exosome analysis and deep learning as a method for early-stage liquid biopsy of lung cancer.

Exploratory Study on Classification of Lung Cancer Subtypes Through a Combined K-Nearest Neighbor Classifier InBreathomics

Authors: Chunyan Wang, Yijing Long, Wenwen Li - 2020

Accurate classification of adenocarcinoma (AC) and squamous cell carcinoma (SCC) in lung cancer is critical to physicians' clinical decision-making. Exhaled breath analysis provides a tremendous potential approach in non-invasive diagnosis of lung cancer but was rarely reported for lung cancer subtypes classification. In this paper, we firstly proposed a combined method, integrating K-nearest neighbor classifier (KNN), borderline2-synthetic minority over-sampling technique (borderlin2-SMOTE), and feature reduction methods, to investigate the ability of exhaled breath to distinguish AC from SCC patients. The classification performance of the proposed method was compared with the results of four classification algorithms under different combinations of borderline2-SMOTE and feature reduction methods. The result indicated that the KNN classifier combining borderline2-SMOTE and feature reduction methods was the most promising method to discriminate AC from SCC patients and obtained the highest mean area under the receiver operating characteristic curve (0.63) and mean geometric mean (58.50) when compared to others classifiers.

The result revealed that the combined algorithm could improve the classification performance of lung cancer subtypes in breathomics and suggested that combining non-invasive exhaled breath analysis with multivariate analysis is a promising screening method for informing treatment options and facilitating individualized treatment of lung cancer subtypes patients.

A Hybrid Algorithm for Lung Cancer Classification Using SVM and Neural Networks

Authors: Pankaj Nanglia, Sumit Kumar- 2020

The present research article focused on the factual findings of the potential usage of the combinational Feed-Forward Back Propagation Neural Network as a judgment making for lung cancer. In this context, Support Vector Machine is integrated with Feed-Forward Back Propagation Neural Network to create a hybrid algorithm that further helps in reducing the computation complexity of the classification. A set of 500 images are utilized in which 75% data is used for the training purpose and the rest 25% is used to achieve the classification. In the view of forgoing, a three-block mechanism is proposed for the classification in which the first block pre-processes the dataset, the second block extracts the features via the SURF technique followed by the optimization using Genetic Algorithm and the terminal block is for the classification via FFBPNN.

The hybrid classification algorithm is named as Kernel Attribute Selected Classifier and the overall classification accuracy of the proposed algorithm is 98.08%. Herein, the objective of the study is to enhance the classification accuracy by applying a hybrid classification algorithm.

Lung Cancer Prediction Using Stochastic Diffusion Search (SDS) Based Feature Selection and Machine Learning Methods

Authors: S. Shanthi, N. Rajkumar – 2020

The symptoms of cancer normally appear only in the advanced stages, so it is very hard to detect resulting in a high mortality rate among the other types of cancers. Thus, there is a need for early prediction of lung cancer for the purpose of diagnosing and this can result in better chances of it being able to be treated successfully. Histopathology images of lung scan can be used for classification of lung cancer using image processing methods. The features from lung images are extracted and employed in the system for prediction. Grey level co-occurrence matrix along with the methods of Gabor filter feature extraction are employed in this investigation. Another important step in enhancing the classification is feature selection that tends to provide significant features that helps differentiating between various classes in an accurate and efficient manner. Thus, optimal feature subsets can significantly improve the performance of the classifiers. In this work, a novel algorithm of feature selection that is wrapper-based is proposed by employing the modified stochastic diffusion search (SDS) algorithm. The SDS, will benefit from the direct communication of agents in order to identify optimal feature subsets. The neural network, Naïve Bayes and the decision tree have been used for classification. The results of the experiment prove that the proposed method is capable of achieving better levels of performance compared to existing methods like minimum redundancy maximum relevance, and correlation-based feature selection.

Problem Statement

- Automatic detection of small lung nodules on Computed tomography utilizing a local density maximum algorithm, it is old model and provides poor detection.
- In some approach's uniformity, connectivity, and position features were extracted.
- Lung cancer detection by using artificial neural network and fuzzy clustering methods, presents two segmentation method and Lung Cancer Detection Using and Early Detection and Prediction of Lung Cancer Survival using Neural Network Classifier have been developed but they provide poor detection and identification.

Disadvantages

- In the existing method they use neural network but the accuracy is poor.
- They used Artificial Neural Network based Classification and detection system of lung cancer

III. PROPOSED WORK

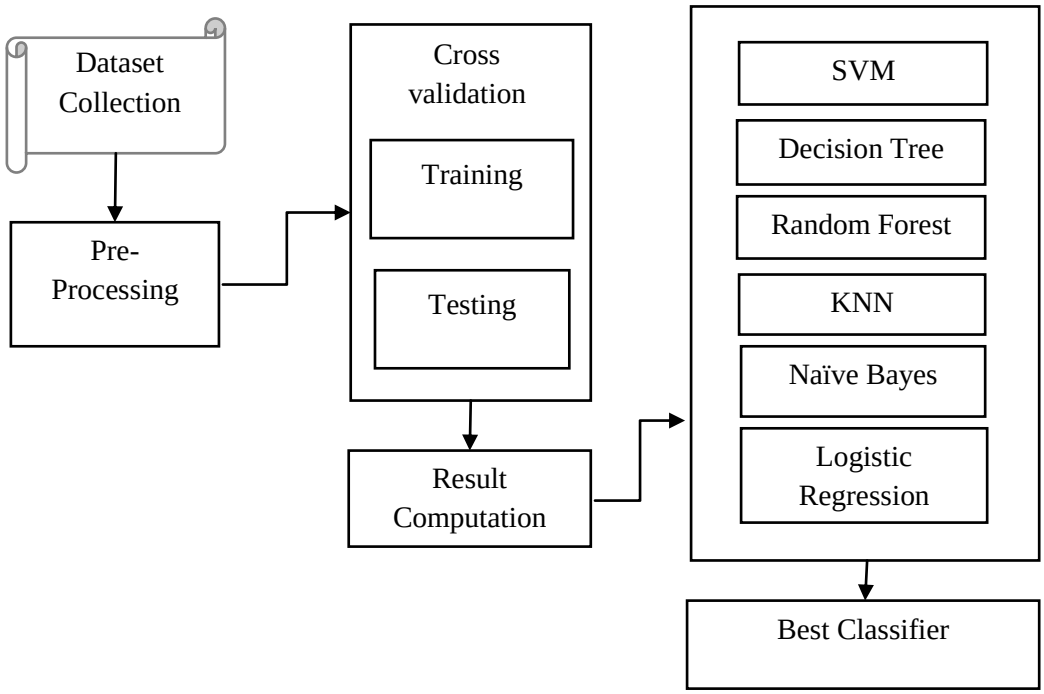
The proposed methodology starts with data acquisition which is followed by pre-processing. The selected classifiers are then trained and tested on the benchmark dataset using standard 10- fold cross-validation approach. The results are computed and evaluated to identify the best methodology for lung cancer detection.

- Identification of well-known classifiers and ensemble approaches utilized for lung cancer prediction
- Computation of results over benchmark dataset obtained from UCI repository
- Proposed a majority voting-based ensemble based on top-3 best performing individual classifier
- Comparison and evaluation of results which show that Gradient-boosted Tree outperformed all another individual as well as ensemble classifiers
- Achieved 90% accuracy for lung cancer identification

Advantages

- We construct our approach via the use of a widely use ensemble technique namely majority voting.
- The voting ensemble technique is a common example of the multi expert approach, which helps to combine the classifiers in a parallel fashion.
- High accuracy prediction.

IV. SYSTEM ARCHITECTURE



The main objective of our proposed study is to investigate and measure the effectiveness of different base level predictors

Step-1: Extraction of datasets through an online repository.

Step-2: Application of pre-processing for data cleaning.

Step-3: Standard cross validation is applied for training and testing.

Step-4: Computation of results for all individual classifiers.

Step-5: Select top-3 classifiers based on the performance measure such as accuracy and compose majority voting-based ensemble.

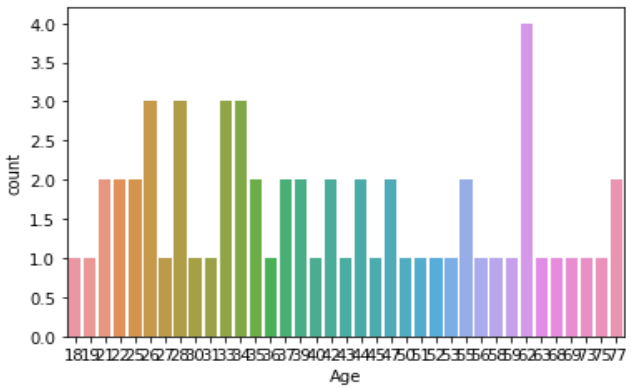
Step-6: Compute results for Majority Voting based ensemble.

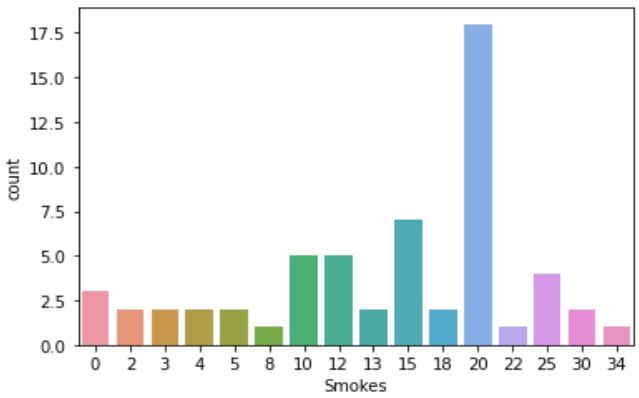
Step-7: Performance comparison is conducted for all individual as well as ensemble classifiers in order.

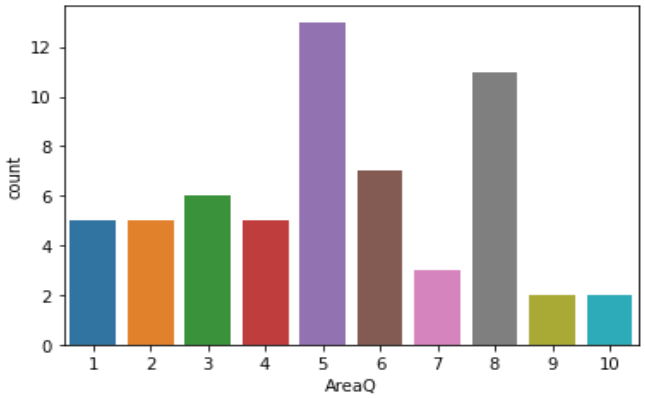
TABLE 1
DATASET

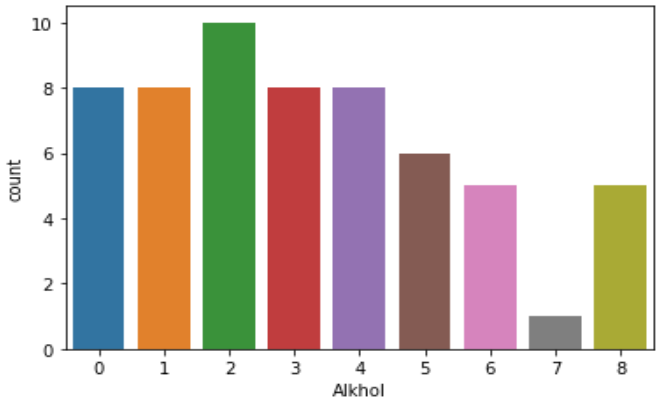
	Name	Surname	Age	Smokes	Area Q	Alkhol	Result
0	John	Wick	35	3	5	4	1
1	John	Constantine	27	20	2	5	1
2	Camela	Anderson	30	0	5	2	0
3	Alex	Telles	28	0	8	1	0
4	Diego	Maradona	68	4	5	6	1
5	Cristiano	Ronaldo	34	0	10	0	0
6	Mihail	Tal	58	15	10	0	0
7	Kathy	Bates	22	12	5	2	0
8	Nicole	Kidman	45	2	6	0	0
9	Ray	Milland	52	18	4	5	1
10	Fredric	March	33	4	8	0	0
11	Yul	Brynner	18	10	6	3	0
12	Joan	Crawford	25	2	5	1	0
13	Jane	Wyman	28	20	2	8	1

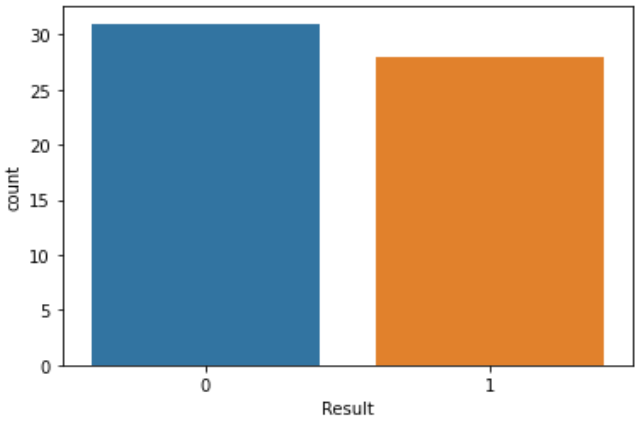
14	Anna	Magnani	34	25	4	8	1
15	Katharine	Hepburn	39	18	8	1	0
16	Katharine	Hepburn	42	22	3	5	1
17	Barbra	Streisand	19	12	8	0	0
18	Maggie	Smith	62	5	4	3	1
19	Glenda	Jackson	73	10	7	6	1
20	Jane	Fonda	55	15	1	3	1
21	Maximilian	Schell	33	8	8	1	0
22	Gregory	Peck	22	20	6	2	0
23	Sidney	Poitier	44	5	8	1	0
24	Rex	Harrison	77	3	2	6	1
25	Lee	Marvin	21	20	5	3	0
26	Paul	Scofield	37	15	6	2	0
27	Rod	Steiger	34	12	8	0	0
28	John	Wayne	55	20	1	4	1
29	Gene	Hackman	40	20	2	7	1
30	Marlon	Brando	36	13	5	2	0
31	Jack	Lemmon	56	20	3	3	1
32	Jack	Nicholson	47	15	1	8	1
33	Peter	Finch	62	25	3	4	1
34	Richard	Dreyfuss	26	10	7	2	0
35	Dustin	Hoffman	25	20	8	2	0
36	Henry	Henry	59	20	3	4	1
37	Robert	Duvall	62	15	5	5	1
38	Ellen	Burstyn	33	25	8	2	0
39	Faye	Dunaway	37	10	5	3	0
40	Diane	Keaton	50	20	2	4	1
41	Jane	Fonda	47	12	8	0	0
42	Sally	Field	69	20	5	4	1
43	Sissy	Spacek	63	20	4	5	1
44	Jessica	Lange	39	15	7	2	0
45	Gwyneth	Paltrow	21	20	8	3	0
46	Halle	Berry	31	20	9	4	0
47	Nicole	Kidman	28	10	4	1	0
48	Charlize	Theron	53	20	6	3	1
49	Katharine	Hepburn	62	20	5	6	1
50	Katharine	Hepburn	42	12	6	2	0
51	Barbra	Streisand	44	30	1	6	1
52	Maggie	Smith	26	34	1	8	1
53	Glenda	Jackson	35	20	5	1	0
54	Ernest	Borgnine	26	13	6	1	0
55	Alec	Guinness	77	20	5	4	1
56	Charlton	Heston	75	15	3	5	1
57	Gregory	Peck	43	30	3	8	1
58	Sidney	Poitier	51	25	9	0	0

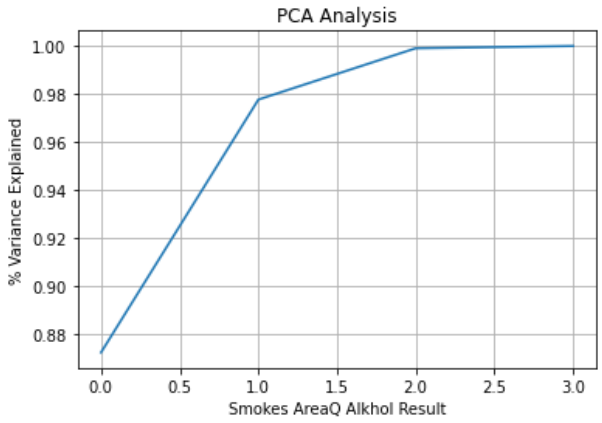












Classifier	Accuracy
Naive Bayes	1.0
Decision Tree	1.0
KNN	0.66
Logistic Regression	1.0
SVM	1.0
Random Forest	1.0

V. CONCLUSION

The utilization of SVM and Logistic Regression in an optimized way has been investigated in-depth to create an evaluation of early prediction for lung cancer classification. The results of the classification are tested by varying number of neurons and the final selection is made on the basis of regression value. In this context, the author has extended the previous version of this paper in which SURF and SVM have been implemented. Here, the implemented algorithm has been named as SVM, KNN, Decision Tree, Naïve Byes, Logistic Regression and Random Forest Tree. The performance of six algorithm has been evaluated and compared in the form of parameters like precision, accuracy, recall and f-measure for training samples ranging

from 100 to 500. The implementation of this unique Cross validation technique is due to the combination of efficient feature extraction technique. This demonstrates a higher value of average precision, accuracy, recall and f-measure those are 98.17%, 98.08%, 96.5% and 97%, respectively for the lung cancer classification over 500 Samples data. Therefore, there are several future aspects of this research work and it would be interesting to see the effect of Machine Learning and improvement in the classification accuracy.

VI. APPLICATION AND FUTURE WORK

In this research, we also found some threats. The first threat is related to a generalization of results since we performed our experiments on a single dataset. Consequently, the result may vary if we consider several experiments with different datasets. The secondly related threat to the selection of one ensemble technique. We report the results according to the functionality of the majority voting method.

The focus of this research is an evaluation of machine learning classifiers as well as ensembles for lung cancer detection. For this purpose, individual classifiers including SVM, KNN, Decision Tree, Naïve Bayes, Naïve Bayes, and Random Forest Tree are assessed.

In the future, we plan to evaluate other lung cancer and different disease datasets. Similarly, other ensemble technique like Stacking, Adaboost, and Bagging will be analysed.

REFERENCES

[1] Nanglia, Pankaj, Sumit Kumar, Aparna N. Mahajan, Paramjit Singh, and Davinder Rathee. "A hybrid algorithm for lung cancer classification using SVM and Neural Networks." *ICT Express* (2020).

[2] Shin, H., Oh, S., Hong, S., Kang, M., Kang, D., Ji, Y.G., Choi, B.H., Kang, K.W., Jeong, H., Park, Y. and Hong, S., 2020. Early-stage lung cancer diagnosis by deep learning-based spectroscopic analysis of circulating exosomes. *ACS nano*, *14*(5), pp.5435-5444.

[3] Maleki, N., Zeinali, Y. and Niaki, S.T.A., 2021. A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection. *Expert Systems with Applications*, *164*, p.113981.

[4] Wu, Yijun, Jianghao Liu, Chang Han, Xinyu Liu, Yuming Chong, Zhile Wang, Liang Gong et al. "Preoperative prediction of lymph node metastasis in patients with early-T-stage non-small cell lung cancer by machine learning algorithms." *Frontiers in Oncology* 10 (2020): 743.

[5] Jenipher, V. Nisha, and S. Radhika. "A Study on Early Prediction of Lung Cancer Using Machine Learning Techniques." In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pp. 911-916. IEEE, 2020.

[6] Maleki, Negar, Yasser Zeinali, and SeyedTaghiAkhavanNiaki. "A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection." *Expert Systems with Applications* 164 (2021): 113981.

[7] Wang, Chunyan, Yijing Long, Wenwen Li, Wei Dai, ShaohuaXie, Yuanling Liu, Yinchenxi Zhang et al. "Exploratory study on classification of lung cancer subtypes through a combined K-nearest neighbor classifier in breathomics." *Scientific reports* 10, no. 1 (2020): 1-12.

[8] Xie, Ying, Wei-Yu Meng, Run-Ze Li, Yu-Wei Wang, Xin Qian, Chang Chan, Zhi-Fang Yu et al. "Early lung cancer diagnostic biomarker discovery by machine learning methods." *Translational oncology* 14, no. 1 (2021): 100907.

[9] Luo, Shengda, Jiahui Xu, Zebo Jiang, Lei Liu, Qibiao Wu, Elaine Lai-Han Leung, and Alex Po Leung. "Artificial intelligence-based collaborative filtering method with ensemble learning for personalized lung cancer medicine without genetic sequencing." *Pharmacological Research* 160 (2020): 105037.

[10] Pradhan, Kanchan, and Priyanka Chawla. "Medical Internet of things using machine learning algorithms for lung cancer detection." *Journal of Management Analytics* 7, no. 4 (2020): 591-623.