IJBPAS, July, 2023, 12(7): 3088-3096

ISSN: 2277–4998

International Journal of Biology, Pharmacy and Allied Sciences (IJBPAS)

'A Bridge Between Laboratory and Reader'

www.ijbpas.com

A NOVEL METHOD FOR EXPLORATION AND PREDICTION OF LUNG CANCER USING SVM

VIJAYA G

Professor, Department of CSE, Sri Krishna College of Engineering and Technology, Coimbatore – 641008, India

*Corresponding Author: Dr. G.Vijaya: E Mail: vijayag@skcet.ac.in

Received 10th July 2022; Revised 15th Sept 2022; Accepted 19th Oct. 2022; Available online 1st July 2023

https://doi.org/10.31032/IJBPAS/2023/12.7.7244

ABSTRACT

Owing to enormous increase of Lung Cancer Cases Worldwide, primary diagnosis is not a feasible one for developing countries like India, which in turn increases the mortality rate. Early detection of lung cancer is also a challenging task for the clinicians, as there are no preliminary symptoms. Computer Aided Detection & Diagnostic (CADe & CADx) system acts as a supporting tool for lung cancer detection & diagnosis **[1]**. In this paper, a supporting diagnostic tool with 'dbest' Feature Selection method based on Support Vector Machine (SVM) model was presented. The parameters used in this study are: cross validation score, Randomized search and Grid search methods. By comparing the performance metrics of both Grid and Randomized Search, the Grid model outperforms the Randomized in terms of precision, recall and F1 score.

Keywords: Cross – validation – score, Randomized Search, Grid Search, precision, recall, F1 Score

INTRODUCTION:

According to World Health Organization (WHO), of all the deadliest disease, Lung Cancer is one among the topmost cause of death Worldwide **[2]**. As lungs are internal organs which are not visible to our naked eyes, in general Lung Cancer can be identified only at later stages, which in turn increase the mortality rate **[3]**. Early prophecy of Lung Cancer is possible after the development of machine learning techniques **[4-7]**. Owing to the enormous increase of CAD algorithms make

radiology diagnosis easy and designers have sought their models commercially. As a result of the divergent software platforms developed by each team, their results were not reproducible and the software knowledge of the Radiologists are not up to the ground truth. In order to overcome this difficulty, automatic exploration of data from the source came into picture **[8-11]**.

Literature Review:

A number of medical research had been carried for the past two decades based on machine learning approaches because, machine learning is one of the powerful tools for the diagnosis and prediction of cancer related diseases. **[12]** proposed a novel Chaos Particle Swarm Optimization (CPSO) based on SVM Classifier is used for feature selection and ANN is used for classification. In **[13]**, SVM regression algorithm along with cubic kernel was used to predict the malignant cancer, which ranges from 0 to 1 based on the severity level. The paper **[14]** analysis different machine learning algorithms for detecting lung cancer using IoT. The authors in **[15]** analyzed the classification algorithms such as Naïve Bayes, SVM, Decision Tree and Logistic Regression for lung cancer prediction. In **[16]**, eXtreme Gradient Boost (XGBoost) model is used to detect lung cancer and forecast the patients' recovery after surgery. In the first stage of **[17]**, image processing techniques were used to extract lung regions and the feature extraction and classification have to be performed by various machine learning approaches. The authors in **[18]** introduces a new technique called "Wilcoxon Signed – Generative Deep Learning" method for the detection of lung cancer. In this paper they used Wilcoxon Signed Ranking Model for pre-processing and Deep Learning for classification of features. The feature extraction techniques used in **[19]** is: UNet and ResNet models and the classification techniques used are: XBoost and Random Forest. For clear identification and measuring the performance CT scan images are used **[20]** and it has been proved that Watershed segmentation outperforms all other segmentation in terms of accuracy.

Proposed Methodology:

Machine learning based algorithms need not require much pre-requisites, because most of the existing machine learning approaches are automized. The selection of the appropriate feature is one of the major challenges faced by CAD developers. To handle this situation, this study proposes a novel method for feature selection and the work flow of the proposed method is depicted in **Figure 1**. The data set was downloaded from the public repository: Kaggle. The raw data set was fed into the system for pre-processing. Pre-processing is the first step of all CAD system, in which un-available and irrelevant data are

dropped out for further processing. After the completion of pre-processing, features have to be extracted from the data and fed into the next process, feature selection. The main purpose of using feature selection is to reduce the input variable and using only relevant data. In this study, only those features whose feature score is greater than 200 will be considered for 'dbest' features. Those features are undergone training and testing by means of SVM algorithm and the performance metrics such as: precision, recall and F1 Score have to be computed for Randomized Search and Grid Search models.

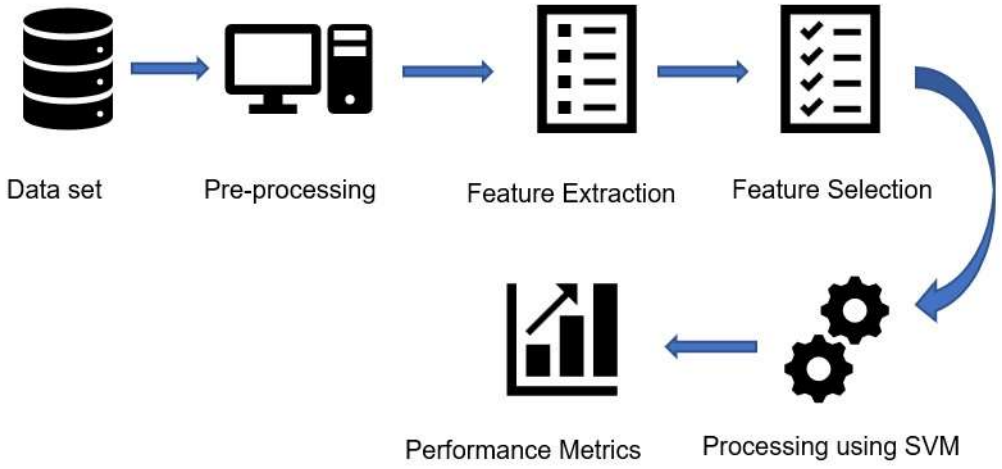

Figure 1: Diagrammatic representation of the workflow

RESULTS AND DISCUSSION:

Figure 2 depicts the first five columns of Lung Cancer Dataset, which is readily available in the public repository: Kaggle. From the figure 2, it is clearly known that the first 23 columns are input or independent variables and the last 24th column is the output or dependent variable. In this data set, the lung cancer is classified as three levels namely: Low, Medium and High, based on the severity of the cancer.

Figure 3 – 8 represents the graphical view of Levels of Cancer vs Smoking, Dust Allergy, Alcohol usage, Air pollution, Age and Gender. Only those features which are responsible for playing a vital role in deciding the levels of lung cancer, is considered for representation. Smoking and Dust Allergy, are the major cause of High-level cancer than the medium level. In contrast to this, Alcohol usage will give impact on both high and low levels of cancer. Even though, Air pollution have good impact on high level, it also affects medium and low level of cancers too. But in the case of Age factor, the age group

around 40+ having more chances to g*et al*l levels of lung cancer. Male gets more chance of having high-level lung cancer than Female mainly due to Smoking. Because the male and female count in low – level, is more are less same.

The 'dbest' features have to be selected by means of SelectKbest class which is available in Scikit - learn. SelectKbest method means, hand-picking the highest score for best features. Here, the threshold value for K (the highest score) is 300. The intention of using SelectKBest, is to select the features according to the priority. The motivation behind the feature selection is to reduce the training time by eliminating unnecessary data from the data set, whenever we are having a large data set. **Figure 9** shows the graphical representation of all the features based on their scores. From the figure, it is clearly shown that, Obesity is the major cause of cancer, followed by Coughing of Blood. Surprisingly, when compared with the active smokers, passive smokers have higher chance of getting lung cancer. Only those features whose threshold value is greater than or equal to 300 will be considered for further processing. The 'dbest' features are depicted in **Figure 10**.

Subsequently, the final procedure is to categorize the lung cancer based on their levels. One of the best and all – time favorite Support Vector Machine (SVM) is used for classification. The data set was split up into training and testing in the ratio of 80:20. The parameters used for validation are: cross validation score, Randomized search & Grid search. The performance metrics computed are: Precision, Recall & F1 Score. The precision is the ratio of relevant occurrences with the retrieved occurrences. While the recall is the ratio between relevant occurrences with all the relevant elements. F1 score has to be computed from the precision and recall. The formulae for computing Precision, Recall and F1 Score is described below:

$$\boldsymbol{Precision} = \frac{\boldsymbol{rl}}{\boldsymbol{rt}} \quad (\mathbf{1})$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{rl}}{\boldsymbol{re}} \quad (\mathbf{2})$$

Where rl →relevant occurrences

$rt \rightarrow$ retrieved occurrences

$re \rightarrow$ all the relevant elements

$$\boldsymbol{F1\ Score} = \mathbf{2} \times \frac{\boldsymbol{Precision} \times \boldsymbol{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}} \quad (\mathbf{3})$$

From **Table 1** it is clearly shown that the performance measure for Grid Search method outperforms the Randomized Search in terms of Precision, Recall & F1 Score. After eliminating the unwanted features by means of 'dbest' feature selection, Grid Search method will yield 100% performance.

	0	1	2	3	4	5	6	7	8	9	...	15	16	17	18	19	20	21	22	23	24
0	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	...	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring	Level
1	P1	33	1	2	4	5	4	3	2	2	...	3	4	2	2	3	1	2	3	4	Low
2	P10	17	1	3	1	5	3	4	2	2	...	1	3	7	8	6	2	1	7	2	Medium
3	P100	35	1	4	5	6	5	5	4	6	...	8	7	9	2	1	4	6	7	2	High
4	P1000	37	1	7	7	7	7	6	7	7	...	4	2	3	1	4	5	6	7	5	High

5 rows × 25 columns

Figure 2: Kaggle Lung Dataset

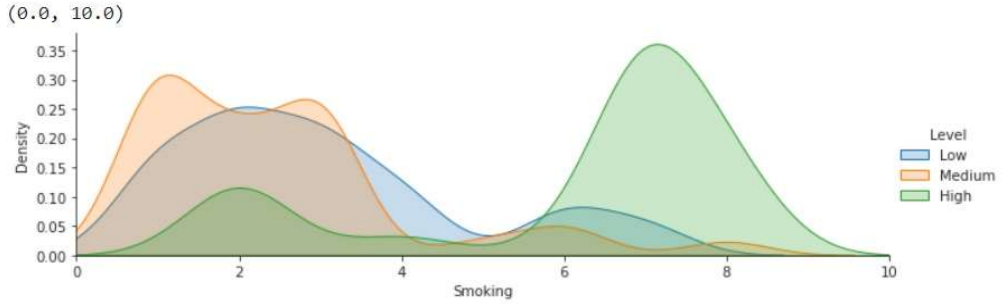

Figure 3: Level vs Smoking

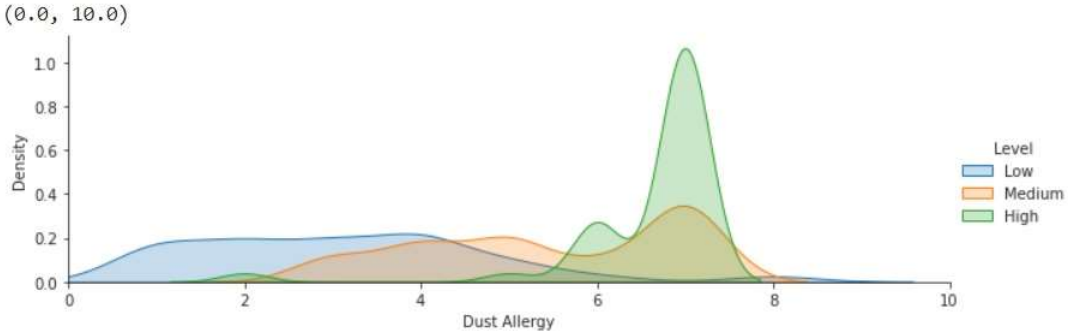

Figure 4: Level vs Dust Allergy

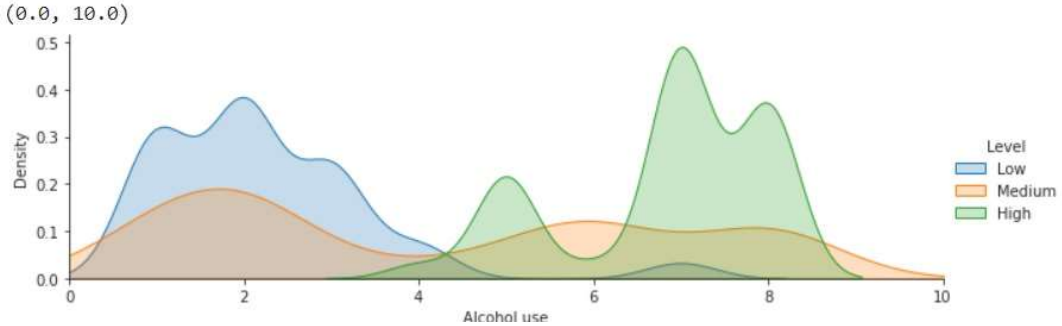

Figure 5: Level vs Alcohol use

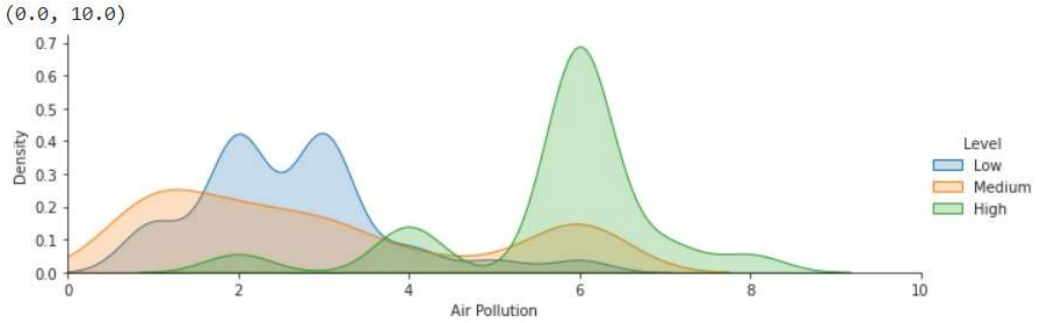

Figure 6: Level vs Air pollution

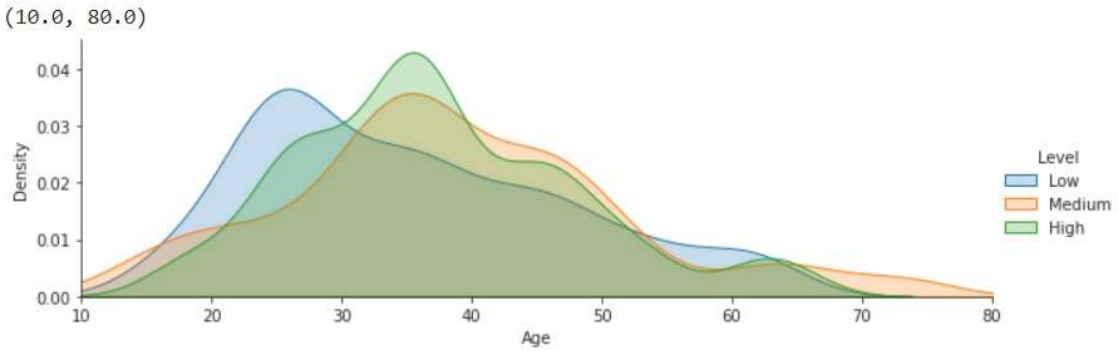

Figure 7: Level vs Age

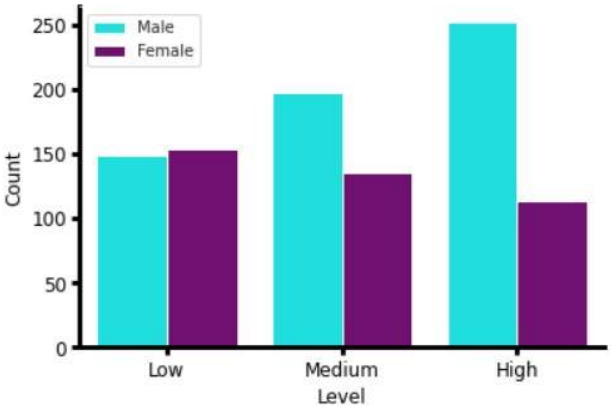

Figure 8: Level vs Gender

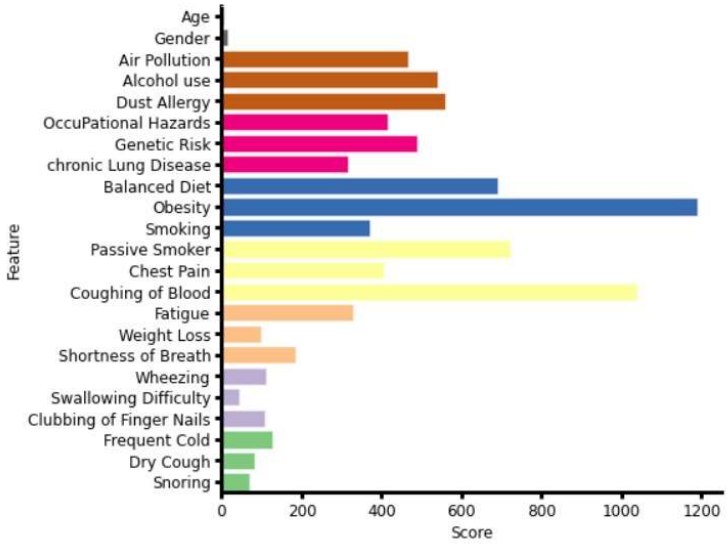

Figure 9: Features vs Score

	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Obesity	Smoking	Passive Smoker	Chest Pain	Coughing of Blood	Fatigue	Level
1	2	4	5	4	3	2	2	4	3	2	2	4	3	Low
2	3	1	5	3	4	2	2	2	2	4	2	3	1	Medium
3	4	5	6	5	5	4	6	7	2	3	4	8	8	High
4	7	7	7	7	6	7	7	7	7	7	7	8	4	High
5	6	8	7	7	7	6	7	7	8	7	7	9	3	High

Figure 10: Feature Score >=300

Table 1: Performance Measure for Randomized & Grid Search in terms of Precision, Recall & F1 Score

Performance Measure	Randomized Search	Grid Search
Precision	0.89	1.0
Recall	0.75	1.0
F1 Score	0.81	1.0

CONCLUSION:

As almost 80% of the lung cancer can be identified only at the later stages, the life – span of the patients' is reduced. After the invention of CAD system, the life expectancy was improved. The main objective of this paper, is to reduce the dimensionality of the data set to be used, by means of selecting only the appropriate features, which is necessary for further processing. 'dbest' is the feature selection algorithm used, based on the ranking method and those features whose score value will be grater than 300 will only be considered for further processing. The most trending and efficient classification algorithm SVM is utilized and the parameters used are: cross – validation, Grid search and Random Search techniques. The Grid search model outperforms Random search in terms of precision, recall and F1 score.

References:

[1] Ayman El-Baz, Garth M. Beache, Georgy Gimel'farb, Kenji Suzuki, Kazunori Okada, Ahmed Elnakib, Ahmed Soliman and Behnoush Abdollahi, "Computer – Aided Diagnostic Systems for Lung Cancer: Challenges and Methodologies", International Journal of Biomedical Imaging, Volume 2013 |Article ID 942353. https://doi.org/10.1155/2013/942353

[2] https://www.who.int/news-room/fact-sheets/detail/cancer#:~:text=The%20most%20 common %20 causes%20of, lung %20(1.80%20million%20deaths)%3B

[3] https://www.nature.com/articles/d41586-020-03157-9

[4] Michael K Gould, Brian Z Huang, Martin C Tammemagi, Yaron Kinar, Ron Shiff, "Machine Learning for Early Lung Cancer Identification Using Routine Clinical and Laboratory Data", American Journal of Respiratory and Critical Care Medicine, 204(4), 2021, 445 – 453. doi: 10.1164/rccm.202007-2791OC.

[5] Ying Xie, Wei – Yu Meng *et al*., "Early lung cancer diagnostic biomarker discovery by machine learning methods", Translational Oncology, 14(1), 2021, 100907. https://doi.org/10.1016/j.tranon.2020.100907

[6] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," Cancer Inform, 2, 2007, 59-77. [Online]. https://www.ncbi.nlm.nih.gov/pubmed/19458758.

[7] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Comput Struct Biotechnol J, 13, 2015, 8-17. doi: 10.1016/j.csbj.2014.11.005.

[8] Furqan Shaukat, Gulistan Raja, Ali Gooya, Alejandro F. Frangi, "Fully automatic detection of lung nodules in CT images using a hybrid feature set", Medical Physics, 44(7), 2017, 3615 – 3629

[9] Lukui Shi, Hongqi Ma & Jun Zhang, "Automatic Detection of Pulmonary nodules in CT images based on 3D Res – I Network", The Visual Computer, 37, 2021, 1343 – 1356

[10] Omnia Elsayed, Khaled Mahar, Mohamed Kholief & Hatem A Khater, "Automatic Detection of the Pulmonary Nodules from CT images", SAI Intelligent Systems Conference, 2015, 742- 746

[11] Jing Gong, Ji – yu Liu, Li – jia Wang, Xi – wen Sun, Bin Zheng, Sheng – dong Nie, "Automatic detection of pulmonary nodules in CT images by incorporating 3D tensor filtering with local image feature analysis", Physica Medica, 46, 2018, 124 - 133

[12] C.Thinkal Dayana, K.S.Mithra, S.Sanjith, "An unconventional SVM Classification using Chaos PSO Optimization for lung cancer discovery", Indian Journal of Science and Technology, 14(6), 2021, 527 – 533. https://doi.org/10.17485/IJST/v14i6.1810

[13] Timor Kadir, Fergus Gleeson, "Lung cancer prediction using machine learning and advanced imaging techniques", Translational Lung Cancer Research, 7(3), 2018. doi: 10.21037/tlcr.2018.05.15

[14] Kanchan Pradhan, Priyanka Chawla, "Medical Internet of Things using Machine Learning Algorithms for Lung cancer detection, Journal of Management Analytics, 7(4), 2020, 591 – 623

[15] Radhika P.R., Rakhi A.S. Nair, Veena G, "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms", 2019 IEEE International Conference on Electrical, Computer and Communication

Technologies.
DOI: 10.1109/ICECCT.2019.8869001

[16] Rana Dhia'a Abdu-aljabar and Osama A. Awad, (2021) "A comparative analysis study of lung cancer detection and relapse prediction using XGBoost classifier", IOP Conference Series: Materials Science and Engineering 1076, 2021, 012048.
doi:10.1088/1757-899X/1076/1/012048

[17] Meraj Begum Shaikh Ismail, "Lung Cancer Detection and Classification using Machine Learning Algorithm", Turkish Journal of Computer and Mathematics Education. 12(13), 2021, 7048- 7054

[18] O. Obulesu, Suresh Kallam, Gaurav Dhiman, Rizwan Patan, Ramana Kadiyala, Yaswanth Raparthi, and Sandeep Kautish, "Adaptive Diagnosis of Lung Cancer by Deep Learning Classification using Wilcoxon Gain and Generator", Journal of Health Care Engineering, Volume 2021. https://doi.org/10.1155/2021/5912051

[19] Siddharth Bhatia, Yash Sinha and Lavika Goel, Soft Computing for Problem Solving, Advances in Intelligent Systems and Computing 817, 2019. https://doi.org/10.1007/978-981-13-1595-4_55

[20] Dakhaz Mustafa Abdullah & Nawzat Sadiq Ahmed, "A Review of most Recent Lung Cancer Detection Techniques using Machine Learning," International Journal of Science and Business, 5(3), 2021,159-173.