JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Data Driven Lung Cancer Detection Using Machine Learning Algorithms

Nalin Pawar, Akhilesh Kumar Bhardwaj

Researcher, Guide

Department of Computer Science and Engineering

SKIET, Delhi, India

Abstract : Lung cancer is one of the leading causes of death from cancer. The overlapping of cancer cells makes it difficult to spot it early. Identifying genetic and environmental variables is critical for developing better preventative approaches. The importance of time in detecting anomalies in target photos cannot be overstated. The significant patterns with their related weightage and score are explored in this paper using a decision tree algorithm for lung cancer prediction. When lung most cancers is located early, there are fewer remedy alternatives available, a decrease danger of invasive surgery, and a better risk of survival. As a result, a simple, cost-effective, and time-saving method for lung most cancers screening and prediction will yield encouraging results. Image processing algorithms have excelled at detecting lung cancer in a variety of high-end jobs. This paper discusses numerous category structures for predicting lung most cancers in its early stages. Machine studying algorithms are used to stumble on whether or not lung tumours are malignant or benign. The overall performance of system studying classifications which includes logistic regression, SVM (Support Vector Machine), Naive bayes, Decision tree, Random forest, and K-Nearest Neighbor category is evaluated in phrases of accuracy, sensitivity, and specificity. In this study, the CNN technique with a restrained dataset achieves the best accuracy of ninety six percentage whilst in comparison to different methodologies, while EDM has the lowest accuracy of 77.8%.

1. INTRODUCTION

Lung most cancers, additionally called lung carcinoma, is a malignant tumour that reasons out of control mobileular proliferation withinside the lungs. It is crucial to deal with this on the way to save you the most cancers from metastasizing to different elements of the body. The maximum not unusualplace sort of lung most cancers is carcinoma. The fundamental sorts of lung carcinoma are small-mobileular lung carcinoma and nonsmall-mobileular lung carcinoma [1]. Long-time period tobacco use is the essential purpose of lung most cancers in eighty five percentage of people [2]. In the ones who've by no means smoked, air pollution, secondhand smoking, asbestos, and radon fueloline purpose 10–15 percentage of cases.As a result, it's miles vital to increase a novel, dependable technique for diagnosing lung most cancers at an in advance stage [3]. The modern-day have a look at used 20 lung imaging samples and 4 algorithms to examine them. The mixture of an adaptive median filter, adaptive histogram equalisation, and a confident convergence particle swarm optimization (GCPSO)-based definitely approach has been established to deliver more accurate results, among specific things.

Smoking, the environment, alcohol, obesity, continual lung disease, a balanced diet, highbrow trauma, radiation therapy, tobacco, and hereditary danger are all elements considered in this suggested early detection and prediction approach [4]. Because cigarette smoke consists of over 4,000 chemicals, lots of that have been recognized as carcinogenic, it's miles one of the maximum not unusualplace reasons of lung most cancers (approximately 90%). Excessive alcohol use and environmental pollution, mainly air pollution, are different elements that make contributions to lung most cancers. Someone who smokes a couple of percent of cigarettes each day has a 20-25 instances better chance of lung most cancers than a person who does now no longer smoke [5].As a end result of immoderate mobileular growth, lung most cancers develops in a single or each lungs. Visual loss and weak point

on one facet of the frame may arise if lung most cancers spreads to the brain. Coughing up blood, chest pain, and shortness of breath are all signs of number one lung most cancers [6]. Chest radiography (x-ray), computed tomography (CT), magnetic resonance imaging (MRI scan), and sputum cytology are all superior techniques for detecting lung most cancers. Many patients, however, can't manage to pay for those therapies, which might be additionally time consuming. The majority of the strategies indexed above can simplest diagnose lung most cancers in its superior degrees, reducing the patient`s probabilities of survival. A new approach for detecting lung most cancers in its early degrees is desperately needed. As a result, image processing techniques can help to improve manual analysis[7].

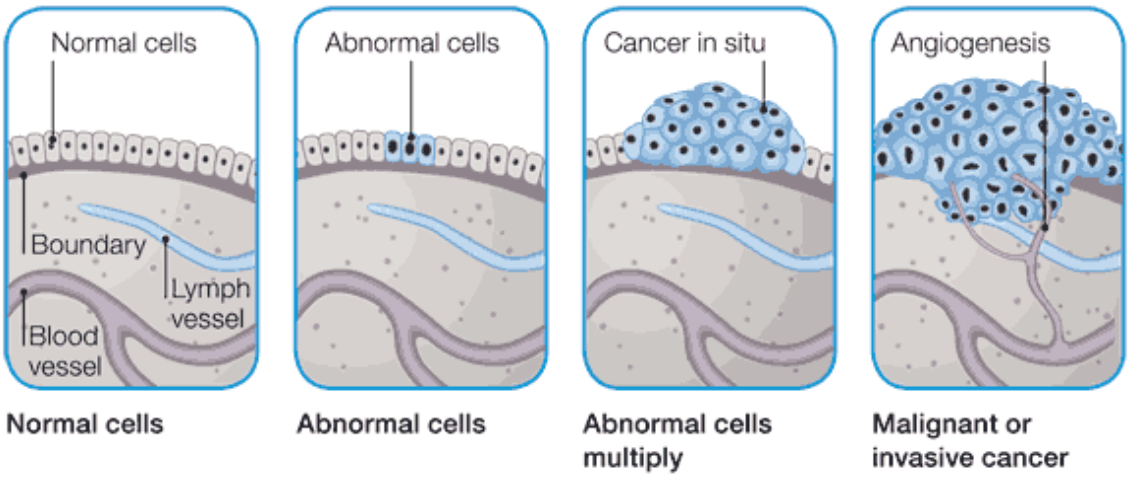


Fig.1: the starting stages of cancer[8]

This lung cancer risk prediction system, which employs significant pattern prediction algorithms, will aid the detection procedure. As a result, early prediction should be a key component of both the diagnosis and the formulation of a successful preventative strategy. Current lung cancer diagnosis techniques, as previously noted, are both costly and time consuming for many people. It also detects cancer at an advanced stage, reducing patients' chances of survival [9]. As a result, the suggested method for predicting lung cancer in its early stages is based on a few characteristics and thresholding. Because the number of testing rules in this system is less, the amount of time and money spent on unnecessary medical tests is reduced.Another advantage of the proposed system is that it will be web-based, allowing patients in remote places to contact directly with clinicians.

2. METHODOLOGY

The following is a discussion of a lung cancer detection system (Fig. 2) [10] that uses machine learning classification methods such logistic regression, SVM (Support Vector Machine), Naive bayes, Decision tree, Random forest, and K-Nearest Neighbor classification to classify chest CT data.

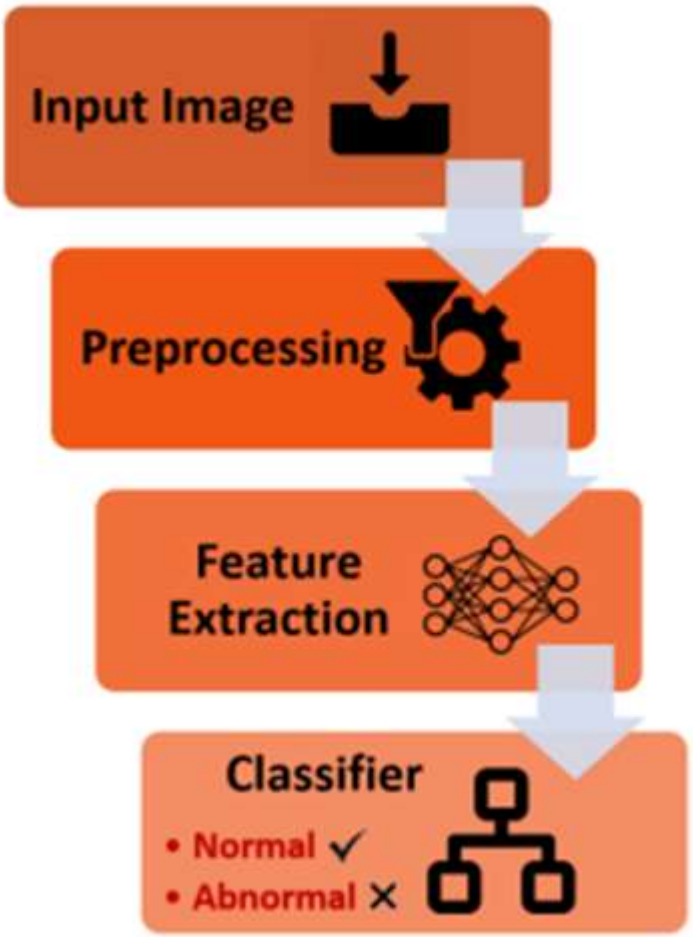


Fig. 2 block diagram of the lung cancer detection system

Lung CT images are preprocessed in the first step using a median filter to reduce degradation during acquisition. The lung areas are then extracted from the CT imaging scans. Tumors are identified by segmenting each slice. The classifier uses the segmented tumours as input to determine if the tumour in a patient's lung is malignant or non-cancerous [11]. Figure 3 shows noncancerous and cancerous lung pictures. Figure 4 shows the median filtered photos.

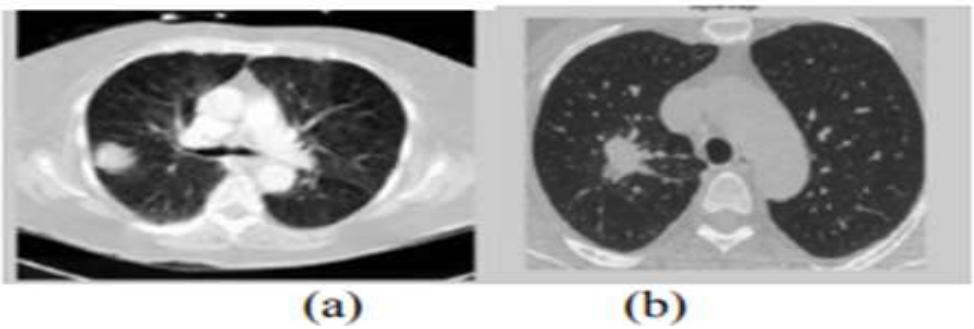


Fig. 3. (a) non-cancerous (b) cancerous images

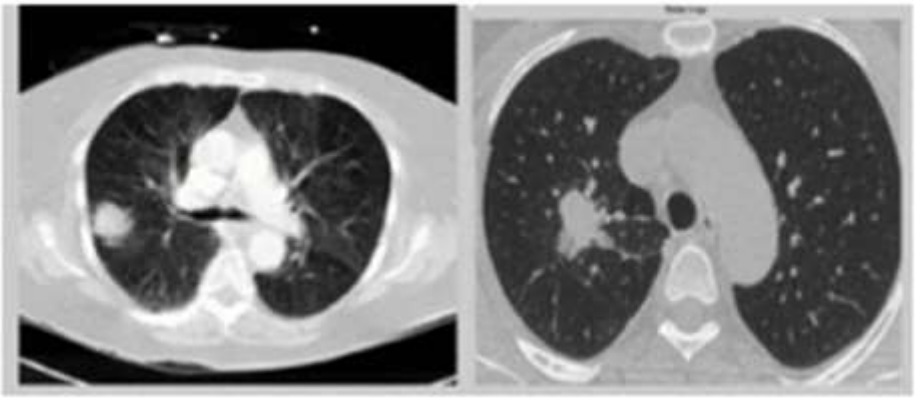


Fig. 4. median filtered images

1. Logistic regression classification

Binary classification problems can be solved using the logistic regression technique. The sigmoid function, additionally referred to as the logistic function, is an S-formed curve that could assign any actual price range to a price among zero and 1, however in no way precisely inside the ones borders. As a result, the default elegance possibility (the risk that an input (X)(X) belongs to the default elegance (Y=1)(Y=1) (P(X)=P(Y=1|X) is modelled[12] the use of logistic regression. The possibility forecast is created the use of the logistic function, which permits us to derive the log-odds or the probit. As a consequence, the version is a linear mixture of the inputs, however it's miles associated with the log-odds of the default elegance. To begin, create a model instance and set the default defaults. Number ten requires the inverse of the regularisation strength. Before being applied to the test data, the logistic regression model was trained on the training data.

2. SVM (Support Vector Machine) classification

SVM was first introduced by Vapnik [13], and it quickly gained popularity because to its great accuracy. This method separates the training data using an optimal separating hyperplane (OSH). Maximum margin classifiers are a supervised getting to know method that labels the schooling statistics with the output class, permitting the empirical threat to be minimised on the equal time. The mastering problem putting for SVMs corresponds to an unknown and nonlinear dependency (mapping, function) y=(x)y=f(x) amongst some high-dimensional input vector xx and scalar output yy amongst some high-dimensional input vector xx and scalar output yy amongst some high-dimensional input vector xx and scalar output yy amongst some high-dimensional input vector xx and scalar output yy. A unfastened distribution getting to know method need to be utilised due to the fact no records at the joint opportunity capabilities is provided.He was in charge of learning techniques[14].

In terms of classification, the goal of SVM is to find a hyperplane in an N-dimensional space that clearly categorises the data points. As a result, hyperplanes serve as decision boundaries for data classification. Data points on either side of the hyperplane can be allocated to different classifications[15].

3. Naive bayes classification

The nave Bayesian classifier is a probabilistic classifier based on Bayes' theorem that assumes great feature independence. Using Bayes theorem, we can determine the likelihood of XX occurring[16]:

(P(X|Y)=P(Y|X)P(X)P(Y))(P(X|Y)=P(Y|X)P(X)P(Y))(P(X|Y)=P(Y|X)P(X)P(Y))(P(X|Y)=P(Y|X)P(X)P(Y))(P(The proof is YY, and the speculation is XX. The assumption is that the lifestyles of 1 function has no bearing at the presence of another (the predictors/capabilities are independent). As a result, it's miles stated to as naïve. In this situation, we`ll think that the values are drawn from a Gaussian distribution, subsequently we're going to use a Gaussian Naive Bayes model[17].

4. Decision tree classification

A decision tree is a tree shape that resembles a flowchart, with every leaf node representing the conclusion, an internal node indicating a feature, and a department representing a choice rule[18]. The intention of a choice tree is to examine a hard and fast of information a good way to assemble a hard and fast of guidelines or questions that may be used to expect a class, i.e., the

intention of a decision tree is to create a version that predicts the price of a goal variable through gaining knowledge of easy choice guidelines inferred from information features. In this way, the decision tree chooses the most appropriate alternative[19].

5. Random forest classification

Random forest is a supervised learning technique that, based on a prior classification procedure, generates a forest at random. This forest is made up of decision trees that have been trained using the bagging method. Bagging's primary notion is to reduce variation by averaging a large number of noisy but roughly impartial models[20]. Each tree is constructed using the following algorithm:

- Assume there are NN test cases and MM classifier variables.
- Let mm denote the number of input variables that will be used to determine a node's decision; mMmM
- For this tree, choose a training set and estimate the mistake the use of the ultimate take a look at cases.

- At every node of the tree, pick mm variables to base the selection on at random. Determine the excellent training set division using the millimetre variables[21].

Each time a forecast is required, a new case is moved down the tree. After that, the label of the terminal node wherein it ends is applied. This approach is repeated with the aid of using all the bushes withinside the assembly, and the label with the maximum incidences will become the forecast. The variety of bushes withinside the wooded area is counted in 100s.

6. K-Nearest Neighbor classification

K-Nearest Neighbors is a storage and class method for brand new instances primarily based totally on a similarity metric (e.g., distance functions). Because no assumptions approximately the distribution of the underlying records are made, this method is non-parametric, and it's also lazy due to the fact no education records factor version is constructed. All of the education records become used withinside the take a look at phase[22]. This will increase the velocity of education whilst slowing and elevating the fee of testing. With this method, if the range of instructions is 2, the range of neighbours ok is normally an atypical range. To discover the nearest comparable locations, calculate the gap among them the use of distance metrics like Euclidean distance, Hamming distance, Manhattan distance, and Minkowski distance[23].

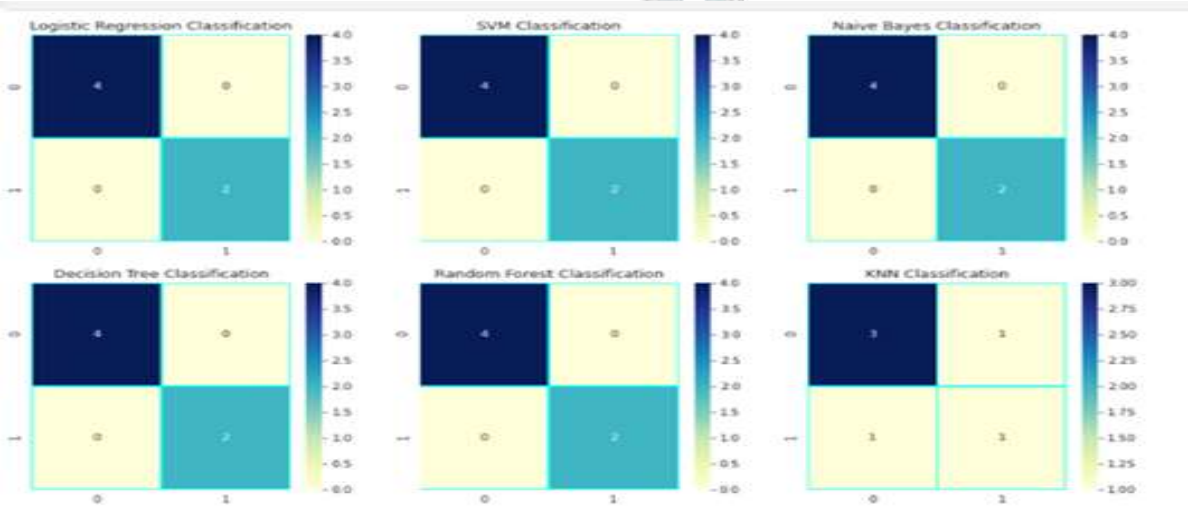
S. No	Column	Non-Null Count	Dtype
1.	Name	59 non-null	object
2.	Surname	59 non-null	object
3.	Age	59 non-null	int64
4.	Smokes	59 non-null	int64
5.	AreaQ	59 non-null	int64
6.	Alkhol	59 non-null	int64

Predictor variable use in classifying lung cancer:Age,Smokes,AreaQ and Alkhol.

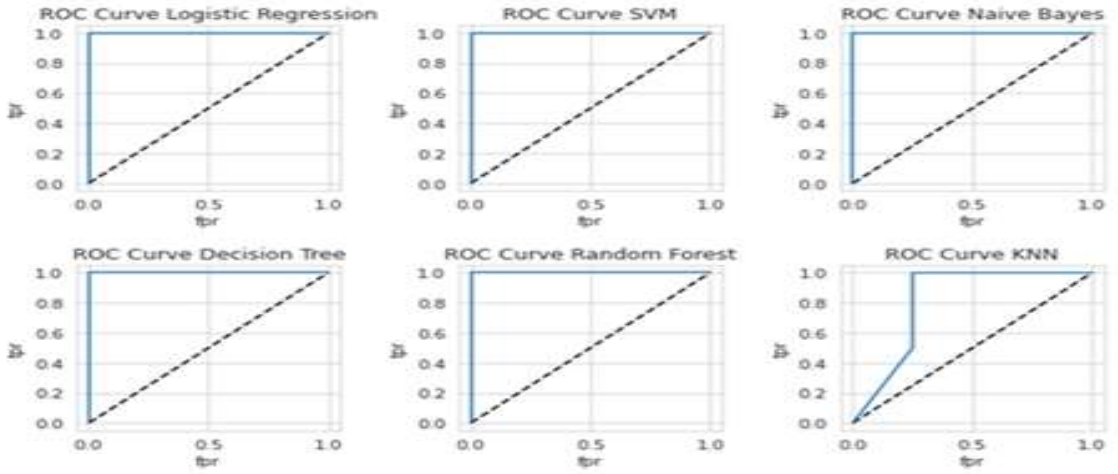
The confusion matrix

Table Comparison of classification techniques

S. No	Classification techniques	Values
1.	Logistic Regression Score	1.000000
2.	Support Vector Machine Score	1.000000
3.	Naive Bayes Score	1.000000
4.	Decision Tree Score	1.000000
5.	Random Forest Score	1.000000
6.	K-Nearest Neighbour Score	0.666667



ROC curve



CONCLUSION

The overall performance of numerous system getting to know procedures hired withinside the type of lung tumours, together with CNN, SVM, ANN, MLP, KNN, EDM, and RF, is tested in phrases of accuracy, sensitivity, and specificity on this work. With a take a look at facts of 30 photos, the CNN device changed into capable of categorise benign and cancerous cells with a excessive accuracy of ninety six percentage. The CNN-primarily based totally device`s sensitivity and specificity also are in comparison to different procedures. SVM classifier additionally has a higher accuracy of ninety six percentage the usage of take a look at facts of 32 images, so it can be utilised to assist radiologist distinguish among malignant and benign pulmonary lung nodules, in addition to for destiny enhancement. With big datasets, ANN, MLP, KNN, and RF can also additionally reap excessive accuracy of 92.sixty eight percentage, 98.31 percentage, 98.30 percentage, and 89.ninety percentage, respectively. The EDM technique has the lowest accuracy, at 77.8%. EDM necessitates a whole lot of room for improvement.

REFERENCES

[1] N. Sudhir Reddy, V Khanaa, Detection and Prediction of Lung Cancer Using Different Algorithms in International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8, Issue-6S3, September 2019

[2] Annangi, Pavan, Sheshadri Thiruvenkadam, Anand Raja, Hao Xu, XiWen Sun, and Ling Mao. "A region based active contour method for x-ray lung segmentation using prior shape and low level features." In 2010 IEEE international symposium on biomedical imaging: from nano to macro, pp. 892-895. IEEE, 2010.

[3] Armato, Samuel G., Maryellen L. Giger, Catherine J. Moran, James T. Blackburn, Kunio Doi, and Heber MacMahon. "Computerized detection of pulmonary nodules on CT scans." Radiographics 19, no. 5 (1999): 1303-1311.

[4] Besbes, Ahmed, and Nikos Paragios. "Landmark-based segmentation of lungs while handling partial correspondences using sparse graph based priors." In Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on, pp. 989-995. IEEE, 2011.

[5] K. Senthil Kumar , K. Venkatalakshmi and K. Karthikeyan, Lung Cancer Detection Using Image Segmentation by means of Various Evolutionary Algorithms in Hindawi Computational and Mathematical Methods in Medicine Volume 2019, Article ID 4909846, 16 pages https://doi.org/10.1155/2019/4909846

[6] Chen, Hui, Yan Xu, Yujing Ma, and Binrong Ma. "Neural network ensemble-based computer-aided diagnosis for differentiation of lung nodules on CT images: clinical evaluation." Academic radiology 17, no. 5 (2010): 595-602.

[7] A. A. Brindha, S. Indirani, and A. Srinivasan, "Lung cancer detection using SVM algorithm and optimization techniques," Journal of Chemical and Pharmaceutical Sciences, vol. 9, no. 4, 2016.

[8] M. Kurkure and A. +akare, "Introducing automated system for lung cancer detection using Evolutionary Approach," International Journal of Engineering and Computer Science, vol. 5, no. 5, pp. 16736–16739, 2016.

[9] P. Bhuvaneswari and A. Brintha +erese, "Detection of cancer in lung with K-NN classification using genetic algorithm," International Conference on Nanomaterials and Technologies, vol. 10, pp. 433–440, 2014.

[10] Konstantina KourouThemis P. ExarchosKonstantinos P. ExarchosMichalis V. KaramouzisDimitrios I. Fotiadis, Machine learning applications in cancer prognosis and prediction in **Computational and Structural Biotechnology Journal** Volume 13, 2015, Pages 8-17

[11] N. Camarlinghi, "Automatic detection of lung nodules in computed tomography images: Training and validation of algorithms using public research databases", Eur. Phys. J. Plus, vol. 128, no. 9, p. 110, Sep. 2013.

[12] K. Mohanambal, Y. Nirosha, E. Oliviya Roshini, "Lung Cancer Detection Using Machine Learning Techniques", IJAREEIE, vol.8, no 2, pp.266-271, February 2019.

[13] F. Taher, N. Prakash, A. Shaffie, A. Soliman, A. El-Baz, An Overview of Lung Cancer Classification Algorithms and their Performances in IAENG International Journal of Computer Science, 48:4, IJCS_48_4_19

[14] N. Panpaliya, N. Tadas, S. Bobade, R. Aglawe, A. Gudadhe, "A survey on early detection and prediction of lung cancer", International Journal of Computer Science and Mobile Computing, pp.175-184, 2015.

[15] Q. Song, L. Zhao, X. Luo, and X. Dou, "Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images", Journal of Healthcare Engineering, vol.2017, pp.1-7, Article ID 8314740. [5] S. Sasikala, M. Bharathi, B.R. Sowmiya, "Lung Cancer Detection and Classification using Deep CNN", IJITEE, vol.8, no. 2S, pp. 259-262, December 2018

[16] Vapnik, V.N. Statistical Learning Theory; Wiley: New York, NY, USA, 1998

[17] S. Khan, S. Hussain, S. Yang, K. Iqbal, "Effective and Reliable Framework for Lung Nodules Detection from CT Scan Images", Nature, pp.1-14,2019.

[18] K. Mohanambal, Y. Nirosha, E. Oliviya Roshini, "Lung Cancer Detection Using Machine Learning Techniques", IJAREEIE, vol.8, no 2, pp.266-271, February 2019.

[19] N. Panpaliya, N. Tadas, S. Bobade, R. Aglawe, A. Gudadhe, "A survey on early detection and prediction of lung cancer", International Journal of Computer Science and Mobile Computing, pp.175-184, 2015.

[20] Q. Song, L. Zhao, X. Luo, and X. Dou, "Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images", Journal of Healthcare Engineering, vol.2017, pp.1-7, Article ID 8314740.

[21] S. Sasikala, M. Bharathi, B.R. Sowmiya, "Lung Cancer Detection and Classification using Deep CNN", IJITEE, vol.8, no. 2S, pp. 259-262, December 2018.

[22]] N. S. Reddy, V Khanaa, "Detection and prediction of lung cancer using different algorithms", International Journal of Engineering and Advanced Technology (IJEAT), pp.2088-2093, vol. 8, September 2019.

[23] M. Glatzer, A. Rittmeyer, J. Müller, I. Opitz, "Treatment of limited disease small cell lung cancer: the multidisciplinary team", Thoracic Oncology, pp.1-10, 2017.