*Prediction of Lung Cancer using Machine Learning and Deep Learning*

Meghashree.B.V   
*Dept. of computer science*   
*Dayananda Sagar College of Engineering   
(Affiliated to VTU)*Bengaluru, India.  
meghashree1298@gmail.com

Sneha.B.S  
*Dept. of computer science*   
*Dayananda Sagar College of Engineering   
(Affiliated to VTU)*Bengaluru, India.  
snehasureshabs@gmail.com

Nalme.R.G   
*Dept. of computer science*   
*Dayananda Sagar College of Engineering   
(Affiliated to VTU)*Bengaluru, India.  
nalmerg98@gmail.com

Asnika Shridhar   
*Dept. of computer science*   
*Dayananda Sagar College of Engineering   
(Affiliated to VTU)*Bengaluru, India.  
asnikashridhar@gmail.com

Pushpa.S.N  
*Dept. of computer science*   
*Dayananda Sagar College of Engineering   
(Affiliated to VTU)*Bengaluru, India.  
pushpasn1234@gmail.com

***Abstract***— **The technology is been increasing rapidly in finding several solutions to the day-to-day problems. It provides several techniques to predict and prevent most of the dreadful diseases, among them cancer being the top most disease that has killed most human life and still being suffering from that disease. Many techniques are available in predicting the lung cancer at its 4th stage, which is the final stage of lung cancer. At that time, since the tumor cells has abruptly increased its size, preventing them is an impossible task. Hence, prediction and detection of lung cancer is very important to notice in the early stage itself. Thus, in our project the data or the inputs are obtained on processing the CT scan images, which are most appropriate compared to X-ray scan images. Hence, processed images gets converted to binary data and stored by using bit scan technique. These are then used to later prediction process. Using machine learning techniques, the classifiers are first trained by the existing datasets and then prediction the cancerous datasets. It is not assumed that this method give the accurate results. There are some outliers which fails this technique in the prediction. Thus, it is essential to define another technique which is then compared with both the results to predict the accurate algorithm used in the prediction. CNN is another knowledge learning technique that is it is a deep learning technique, which produces most accurate results and even optimized results. This project mainly focuses on these domains.**

Keywords— lung cancer prediction, Deep learning, machine learning, feature selection, feature abstraction.

# Introduction

The most common and widely spread disease over the world with increasing the mortality rate is lung cancer. Identification of this disease is done by knowing the symptoms from the patients who are suffering from it. Diagnosis and the treatment depend on the age, gender, data of family members suffered and from those who are still experiencing lung cancer, based on the percentage of tobacco consumption and the fatal environment that they are staying in. There are few treatments available to remove the disorder such as physiotherapy, surgeries, chemotherapy, etc. But these measures will not provide appropriate results in most of the cases.

The disease Lung cancer arises due to the growth of unwanted tissues in the lung leading to tumor cell. These tumor cells divide abnormally and uncontrollably thus, detection of lung cancer in the early stages is very important and essential too. Small Cell Lung Cancer increases faster by forming a large tumor that spreads throughout the body. There are yet more number of cancer types which are even considered in the datasets.

Preventing the lung cancer is another tedious task. There are several prevention methods available such as prevention of smoking and tobacco, modifying diet plans and chemoprevention. Carcinogens present in the environment results in damaging the genome present in human beings. These carcinogens are harmful and they must be eradicate by taking some measures. Since, it is the most commonly affecting factor.

The knowledge discovery in databases that is used in the prediction of lung cancer is machine learning and deep learning. In machine learning, classifiers are used to obtain the results. Classifiers are trained by huge number of well labelled datasets that yields accurate results. In deep learning, CNN class are used in order to precise existing results.

# LITERATURE SURVEY

[H.Bharathi][T.S.Arulananth][1]Early prediction of lung cancer is done using data mining and clustering methods. One of the easy to implement data visualization technique is SOM (Self Organizing Map). SOM is a technique that identifies the concealed pattern in the lung disorder CT image and reduces high-dimension data using Self organizing neural networks. SOM is unsupervised leaning technique, that is the class label of the input data is unknown. This type of learning is very essential in real practices.

The major drawback is SOM gets trained using k-means clustering, where k-means is not robust in nature. Since they are sensitive to the outliers. Data mining is applicable only for structured data. But huge amount of unstructured data are stored in the healthcare databases. This kind of data is not considered as the datasets, thus giving inaccurate results.

[S. Perumal][T.Velmurugan][2]Initially, they used CT scan images which found to be effective in lung cancer prediction rather than using X-ray images. Artificial Bee colony is an algorithm which mainly help in the estimation of percentage that the people suffering from lung cancer. The main advantage of this mechanism is that the input used in this technique is CT scanned images due to histogram adaptation, which improves image contrast and helps in the removal of noisy data.

Most of the times the ABC algorithm provides accurate results, but sometimes it suffers from premature convergence, resulting in suboptimal conclusions. Another disadvantage is a huge amount of time is required to implement this technique.

[Ankit Agrawal][Sanchit Misra][3]SEER is a program to collect the cancerous data from all in-situ cancer. The main advantage of this work is using random forest, decision tree algorithm and support vector machines to check the accuracy among the results obtained from these algorithms.

Only fewer attributes of datasets are considered in this mechanism. Thus, this technique throws a drawback of insufficient data collection.

[Dr. S. Senthil][B. Ayshwarya][4]This paper is an integrated framework for predicting lung cancer introduced using Neural Network with Particle Swarm Optimization. ANN is advantageous with long training time, high computational cost, and adjustment of weight.

The backpropagation that is sensitive to the noisy data. The datasets must be free from noisy data without altering the other data in the dataset. But this technique does not provide this mechanism.

[Roberto F. Machado][Daniel Laskowski][5] Exhaled breath data samples collected using sensor array, sensor is the one that senses the data from the lung, later these data accesses and extracted to provide the best data sets. A difficult yet effect technique to collect samples. Applying the algorithms and obtaining the results.

Just collecting the data from exhaled breathe is not suitable. They would also considered the attributes of family background (family people who are suffering from lung cancer), age and their gender in the dataset.

[Animesh Hazra][Nanigopal Bera][Avijit Mandal][6] In this paper we can see that it done on large cancerous cell patients. This large cancerous cell is more complicated and difficult to cure. We can notice that there are more contrast ways for detecting and diagnosing the small cancerous cell and large cancerous cell. Among those different ways of detecting lung cancer, one of the method to obtain good result is SVM and LR algorithms for training the classifiers and to predict the proper outcomes. It’s been seen that the accurate results among the stages of cancer is made by using SVM algorithm and thus leading to the better results.

The LR classifier provides the accuracy range of about 77.40% which is then compared to the accuracy range obtained by support vector machine that would results in 76.20%.

The only issue over this mechanisms is that it requires more time and lot of memory spaces for performing operations on healthcare datasets.

[Neha Panpaliya1][Neha Tadas2][Surabhi Bobade3][Rewti Aglawe4][Akshay Gudadhe5][7] For feature extracting process and preprocessing of images they have used histogram equalization which helps in determining the cancerous attack about a human, that is whether that person is suffering from lung cancer or not. Based on these predicted results we can conclude the outlive rate of a person.

In order to make benefit for the poor people who will not be having lacks of money spent on prediction. This technique is made very beneficial for them. Advanced level of estimating the lung cancer is of no use. Even this will fall under this category, where the survival rate of the patient is very low. By knowing all this we can say that there is need to identify and cure the lung cancer at its initial stages itself.

It is better to sum up both the techniques used in the paper. They are GLCM and a classifier, where they will increase cost as well as time efficiency.

The main and a powerful fact about this paper is that it is a open source software, that is available in the internet and it is of free of cost. Thus, this will find very useful for the patients who are suffering from cancer but due to lack of money they may not go for the proper solutions. But this software provides some benefits at least in the prediction purpose.

[SumitaMishra1][Naresh Kumar Chaudhary2][Pallavi Asthana1][Anil Kumar1][8] As CT SCAN being the most easiest and accurate method to collect the proper datasets that are required for the accurate results after processing them. It is a technique in which the part of CT SCAN image is obtained from the 3D image and later it is trained in such a way that it resulted in the 2D output .

The nodules meaning the carcinogens in the image is obtaining by making several trials of collecting the parts or the slices of CT images. Its been identified that each part is considered as a whole image similar to appearance based detection and then this would result for loss of data in the boundaries of the image in each part. Thus, the mechanism used to solve this kind of problems by 3D-CNN which helps in determining the nodules (the cells that are cancerous in nature).

Thus this kind of classification will yield a good and better results by obtaining 94,8% of accuracy. Being the deep learning which makes the subject clearer and confident by understanding the results.

Its been seen that several researchers have found various ways to obtain the results but the point is they could not yield this much of accuracy. Sing CNN is highly complex, it requires more time to obtain the results for large dimensional datasets.

[Dr.T.ChristopherP1][P, J.Jamera][9] Data mining is a mechanism that only focuses the data that are similar but in many healthcare department very non-similar data is obtained due to some ethical and social issues. Since such data are simply ignored or discarded from the databases and treats only the structured data. This arises inaccuracy in the obtained results and thus leading to improper results. Thus care should be taken in order to consider those data as well. Thus this kind of data must be first converted to the machine understandable binary data.

WEKA, a tool that is used to convert these unstructured data into binary ones thus resulting the accurate prediction of the subject. Later we consider some classification algorithm, that would results in good outcomes. Naive Bayes algorithm based classifier yields accurate results and thus leading to the most efficient algorithms rather than others. Then the obtained results are then compared with the other kinds of classification such as J48 and Bayesian networks.

In the paper it’s been mentioned that the prediction can be continued by applying more newly obtained technologies to get best results.

Thus, considering various algorithms and computing them with the proper inputs, and comparing the various results would help in understanding the algorithms very well and it is essential to obtain the results that are very accurate in order to save many lives.

[Panayiotis Petousis, Audrey winter][10]The percentage of cases of people suffering with cancer and people with out cancer is time independent that means everything gets changed in a fraction of seconds. It is found that while providing the initial steps in the training process ,the above problem is least bothered by most of the well-known researchers. But this situation is considered and framing it is the most important perspective to be accepted. The model used in this technique would result in the continuous values rather than obtaining the obtaining the discrete ones.

The disadvantage of this proposed technique is that it is found to detect the cancerous patients over a long period of time. This causes drawback because it is better to detect the presence of lung cancer once and then those patients must immediately undergo few treatments rather than wasting time in just predicting the life expectancy. This is of course analyzed by less educated people.

Here they use DBN to generate observations for patients unlike others where they used a static regression model ,when theses two are compared we observe two important advantages for using a dynamic model – one is this model keeps changing with the inputs provided by the user. Other advantage is that this model is even more helpful it frequently capturing the results on predicting the patients biotic disorder.

The advantages over this technique would be this set of results can be used as samples in order to check or compare it with other samples and helps in increasing the survival expectancy.

[Patrice Monkam,Chen LI][11]We can use a single primary learner instead of an ensemble learner which can reduce the computational cost(advantage). Primary learner in the sense the model that understands the data by training it efficiently, one among them is that CNN. Listening to the various methods CNN being the most effective algorithm in updating the weights of the input data samples.

Though the primary learners are used they would not always produce the effective outputs, as expected. Hence another technique is used , here they use another way to produce a multiscale features which do not use multiple networks that is they use multiple-crop pooling strategy, by doing this we reduce the complexity in finding the outputs. But the training time is little longer in single learner than that of ensemble learner. Accuracy of prediction is estimated to be more in this technique other than many other methodologies used to do the same work. Thus it is mostly expected in providing the better way of person survival rate.

Using this technique we can understand both the uses of ensemble and primary learners and giving the best possible results out of them.

[Oguz Akin,Yingli Tian][12]The main intention of this project is that comparing the model that are well trained. This system is both highly sensitivity and also has high specificity .In comparison with the other models trained only on the LUNA16 data set, this framework with the pre-trained model exhibited a significant improvement in all false positive levels across all the datasets. Especially on a 1∕8 false positive per scan, that is in other words for 8 false positives per scan, the accuracy is equivalent to that of LUNA16.

The significant improvement in performance shows the robustness of applying the self-supervised pre-trained model across different datasets without requiring additional requirements. By using the HS2 network and treating each and every CT scan as a video, false positives are significantly reduced which are compared with the patterns that created or identified the cancerous cells or the nodules in the lung.

By introducing a self-supervised feature learning schema, spatiotemporal features of CT scans can be best learned from large-scale CT scans without using any additional labels. The learned features can very well improve the robustness of the proposed framework across different clinic datasets. The highly recommended framework amazingly exceeds other futuristic techniques and has attained both high specificity and high sensitivity and is robust to all the data from multiple CT scanner manufacturers which so happens to have great future in a routine day to day clinical practice.

[ Munimanda Prem Chander1][M. Venkateshwara Rao2][T. V. Rajinikanth3][13] They proposed a solution for the problem in predicting the topic. It is seen the a different technique been applied by the scientist by using CAD, since this is a technique that are usually used in the platform where automatic response is obtained from the algorithm proposed from this technique. This will help in identifying the carcinogens in the lung by X-ray images and then separating the data and it is seen that they are not in the linear structure but looks to be in helical structure. By using this kind of algorithms the author built best way to obtain the results that are almost accurate and found essential in most of the cases.

It is seen that same technology CT scan is used in order to efficiently obtain the datasets by removing the outliers which are called to be the noisy data in it.

Hence in many situations it is better to use the CT scan approach to get the most proper results. With the help of this scanning technique , one can identify the nodules that are present in the lymph node of each cancerous patient and thus help in providing the essential results. Which would lead to the good results on applying different algorithms on them.

The main drawback seen in this paper is that they are of high cost leading to less usage of this kind of technique. They are not compatible with other models that is the models that are trained other than from the CAD would provide insufficient data.

Hence, this kind of technique is more suitable only if there is sufficient amounts of bidding on these projects and improves the efficiency by using them.

[Gurpreet Kaur1][Harpreet Singh2][14]developed a good solution by using powerful algorithms. It is seen that CT scanned images were taken as the sample data and then they are processed using image processing techniques through which we obtain the binary data which are considered as the datasets and they are still helped in the identification of particular disorder.

Thus after obtaining that data they still used good algorithms such as SVM and Back propagation algorithms in order to provide best and comfortable results. Thus using the back propagation algorithm we can train the model by providing various known datasets and helping in providing the best results for the data with unknown outputs.

The network is developed for the further processing. This is known to be the best comparison approach , by comparing SVM and BPN we obtain the improved results in the proposed work.

At first the data is collected and sent to the BN network, later the results obtained from this network is fed as input to the next algorithm that is SVM. Thus, leading to efficient hybrid results.

But the only problem in this technique is that is the outliers are present in the input data then there would be some bad results obtained after treating the BNN later which is hybridized with SVM leading to really different results that are not suitable for later predictions.

[Xiao-Zheng Chen1][ Lian-Hua Cui2][Hong-Zong Si3][Shi-Hai Liu4 ][15] Its is seen that the paper mainly focuses on the biomarkers. Biomarkers are the tools that are used in the discovery of some biological information, in this case it is seen that biomarkers are used to check the carcinogens in the lung of the patients suffering from the lung cancer disease. Thus providing good measure of obtaining the inputs. Serum marker which is type of biomarker that is used in obtaining the proper inputs that are required for the computation. Genes are the ones that are mostly helpful in the spread of cancerous cells thus helping in obtaining the datasets. Later once the data is collected, it is then computed on neural networks.

By then the results obtained from such algorithms will help in the identification of lung cancer.

The only disadvantage about this way of getting the output is that predicting a single tumor cells which are present in the patients is very difficult and also time consuming but provides accurate results at the end.

# CONCLUSION

Based on the literature survey made many paper focused on several disadvantages that are gradually causing the techniques to reject. Thus in our project the most good techniques are used which overcomes from most of the disadvantages mentioned above.

Primarily we use textual mining in order to obtain the data that have similar attributes from that of the data that are not having similar attributes. Thus this would lead to perfect datasets with all conditions are considered. Later it is processed by the best known algorithm such as classifier algorithms thus leading to the most accurate outputs. The obtained outputs using several algorithm will provide the detailed view of cancerous and non-cancerous patients datasets. We have come across various algorithms, in which several unpredictable disadvantages are present in each algorithms. In order to conclude the best algorithm used in this estimation technique is comparing each of the results obtained from various algorithms, which is perform on various data sets .This techniques is cost-effective and helps in many biomedical fields.

##### References

1. H Bharathi and 1TS Arulananth “ A Review of Lung cancer Prediction System using Data Mining Techniques and Self Organizing Map (SOM)” International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 10 (2017) pp. 2190-2195.
2. S. Perumal and T.Velmurugan “Lung cancer detection and classification on CT scan images using enhanced artificial bee colony optimization”. International Journal of Engineering & Technology, 7 (2.26) (2018) 74-79.
3. Ankit Agrawal and Sanchit Misra “Lung cancer survival prediction using ensemble data mining on SEER data”. Scientific Programming 20 (2012) 29–42 29 DOI 10.3233/SPR-2012-0335.
4. Dr. S. Senthil and B. Ayshwarya ”Lung Cancer Prediction using Feed Forward Back Propagation Neural Networks with Optimal Features” International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 1 (2018) pp. 318-325.
5. Roberto F. Machado and Daniel Laskowski “Detection of Lung Cancer by Sensor Array Analyses of Exhaled Breath” Am J Respir Crit Care Med Vol 171. pp 1286–1291, 2005.
6. Animesh Hazra ,Nanigopal Bera and Avijit Mandal.” PREDICTING LUN CANCER SURVIVABILITY USING SVM AND LOGISTIC REGRESSION ALGORITHMS” International Journal of Computer Applications (0975 – 8887) Volume 174 – No.2, September 2017.
7. Animesh Hazra ,Nanigopal Bera and Avijit Mandal.” PREDICTING LUN CANCER SURVIVABILITY USING SVM AND LOGISTIC REGRESSION ALGORITHMS” International Journal of Computer Applications (0975 – 8887) Volume 174 – No.2, September 2017.
8. Sumita Mishra1,Naresh Kumar Chaudhary2,Pallavi Asthana1 and Anil Kumar1” DEEP 3D CONVOLUTIONAL NEURAL NETWORK FOR AUTOMATED LUNG CANCER DIAGNOSIS” 1 Amity School of Engineering and Technology, Amity University, Lucknow Campus , India smishra, 2 Dr. Ram Manohar Lohia Awadh University, Faizabad, U.P., India
9. Dr.T.ChristopherP1 and P, J.Jamera banuP2 ” STUDY OF CLASSIFICATION ALGORITHM FOR LUNG CANCER PREDICTION.”1.PPG and Research Department of Computer Science, Government Arts College,Coimbatore, TN, India, chris.hodcs@gmail.com 2.PM.Phil Scholar, Department of Computer Science, Government Arts College, Udumalpet, TN, India.
10. PANAYIOTIS PETOUSIS, AUDREY WINTER, WILLIAM SPEIER , DENISE R. ABERLE, WILLIAM HSU , AND ALEX A. T. BUI“Using Sequential Decision Making to Improve Lung Cancer Screening Performance”,v7 p119413-119417,September 6, 2019.
11. Baihua zhang,Shoulding QI,Patrice mokam,chen li,fan yang,yu-dong yao,(Fellow,IEEE),and wei qian”ensemble learners of multiple deep CNNs for pulmonary Nodules classification using CT images”in IEEEAccess. v7 p110360,110366,110369,August 22, 2019.
12. Jingya Liu, Liangliang Cao, Oguz Akin and Yingli Tian“Accurate and Robust Pulmonary Nodule Detection by 3D Feature Pyramid Network with Self-supervised Feature Learning”Medical Image Analysis , p11-14, July 25, 2019.
13. Munimanda Prem Chander, M. Venkateshwara Rao, T.V.Rajinikanth (2017). “Detection of Lung Cancer Using Digital Image Processing Techniques: A Comparative Study”, International Journal of Medical Imaging 2017; 5(5): 58-62.
14. Gurpreet Kaur1, Harpreet Singh. “An Intelligent System for Lung Cancer Diagnosis Using Fusion of Support Vector Machines and Back Propagation Neural Network” , International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.
15. Zhuang Yu1&, Xiao-Zheng Chen1&, Lian-Hua Cui2, Hong-Zong Si3, Hai-Jiao Lu1, Shi-Hai Liu4. “Prediction of Lung Cancer Based on Serum Biomarkers by Gene Expression Programming Methods”.