A Review of Machine Learning Algorithms for Lung Cancer Prediction

Ravi Choubey[1], Pratima Gautam[2],
[1]Research Scholar, RNTU, Bhopal (M.P.) India.
[2]Dean, Dept. of CS and IT, RNTU, Bhopal (M.P.) India.

ABSTRACT

The Machine learning is the subset of Artificial intelligent which can imitate like human intelligence and it can process the large information. The classification is process in machine learning based on predictive modeling, which can predict future using current information. The Lung cancer prediction is the method which can predict future Lung Cancer possibilities based on the collection of previous dataset. There are many Lung Cancer prediction models are reviewed in this paper. It is analyzed that Ensemble Classification algorithms are the best predictive and efficient algorithms for the Lung Cancer prediction.

Keywords: Lung Cancer Prediction, Machine learning, Ensemble classification.

I INTRODUCTION

The Lung Cancer is the most deadliest disease in human life. Lung Cancer (LC) cause major deaths due to late diagnosis and it has affected a number of people worldwide [1]. Lungs are responsible for exchange oxygen and carbon dioxide between human body and environment. Lung's structure are like spongy and can easily trap oxygen from environment. Cancer in lungs are Condition where lungs cells grow uncontrollable. In recent times, lung cancer becomes the major reason of death. among people of any age groups. Due to today's unhealthy life style, smoking and pollution all are main reason for lung cancer. Therefore, the improvement in predicting the lung cancer is required in the health sector with the help of different ML methods. The commonly seen symptoms of lung cancer are breath shortness, new cough that doesn't go away. chest pain , coughing up blood, hoarseness, shortness of breath.

Researchers build a solution for discovering an effective models so that the lung cancer can be detected because the existing methods of lung cancer are shown poor efficiency in early time for identification of cancer in lung's cell. There are many reasons behind this for less accuracy and long execution time for detection. The accurate and fast detection of Lung Cancer illness is depend on the earlier knowledge and data regarding the pathological events[2], So there is a need of monitoring certain body metrics like lungs CT Scan ,Smoking habits, blood in cough , air pollution etc. in all characteristics of the patient who suffers from Lung Cancer. All variables are independent and assist us for selecting the best AI and ML algorithms.

The common methods face too many difficulties in the analysis of large size of medical data which is acquired from the healthcare devices due to its big size and complexity. Machine Learning technique is used to analysis of the primary unknown patterns and trends in databases and uses that information for building the predictive models. The hidden patterns and relationships are mined from the huge databases by integrating the Data mining, statistical analysis and database technology in Machine learning. The Data Mining schemes helps the Doctors in produce the future prediction. These schemes and predictive frameworks are adaptable to carry out the prediction of related disease in patient in near future with accurately. lung cancer predictive framework planned on the basis of Data mining has included a number of stages. The primary stage of selecting the data have involved the collection of data from various sources in order to perform predictive analysis in lung cancer. Certain diagnostic variables are included in the data set from which the person is classified as normal or having Lung cancer. The collected dataset often have many outliers and missing values [3]. Availability of missing values and outliers in the collected dataset lays a great impact on the accuracy and efficiency of the classification.

The removing outlier and missing values from the collected data is not so easy task. Also in Image dataset CT scans Computed Tomography images is not easy to read. Many problems come with images and require preprocessing. CT scan images also need more computing power for prediction of Lung Cancer. It is time consuming process but it gives us best result, its accuracy is near to symptomatic prediction. The data preprocessing is the process where we transform raw data into understandable format. It is a main step in machine learning as we cannot work with raw data because it has missing values and outliers. Always it requires that the quality of data must be checked before applying any machine learning algorithms. The process used in data preprocessing are Data Cleaning, Data integration, Data Reduction Data Transformation and final is Feature selection.

(a) **Data cleaning** is first process which removes incorrect data or incomplete data from the dataset. it also remove inaccurate data and replaces the missing values in dataset. There are some data cleaning techniques which are handle missing values and remove noisy data from dataset.

(b) **Data integration**, it is a process of joining the multiple sources into single dataset. The Data integration process is the one of the main component in data preprocessing. Data reduction, this process help us in the volume of data which makes easy to analysis and produces the almost the same result. in this data reduction also helps to reduce storage space in secondary memory. There are some of the main techniques in data reduction are Data compression, numerosity reduction, Dimensionality reduction.

(c) **Data transformation** is the change in the format or structure of the dataset. Based on requirement this step can be simple or complex. There are some methods in the data transformation are Data smoothing and Aggregation.

(d) **Feature extraction** this step essential for the machine learning process because occasionally redundant features influence the classification efficiency of the machine learning classifier. [4] This step optimizes the classification accuracy and decreases the execution time of the applied classifier model. Some commonly used feature selection algorithms are Relief Feature Selection Algorithm, Minimal-Redundancy-Maximal-Relevance Feature Selection Algorithm (mRMR), and Least Absolute Shrinkage and Selection Operator.

II LITERATURE REVIEW

Qing Wu et.al (2017) In this paper they proposed novel neural network based algorithm which called Entropy Degradation Method with vectorized histogram features to detect small cell lung cancer. [5] They clustered images on the basis of histogram of image and classified that image as it is cancer positive or negative. In this study the EDM algorithm achieved accuracy 77.80% accuracy at earlier time of cancer. EDM treated as weak classifier and Adaboost used as base classifier in this study.

O.Gunaydin et.al (2019) in this study they Compared several machine learning algorithms for predicting lung cancer disease. [6] The alrotithm used in this paper was KNN, Support Vector Machines, Naïve Bayes, Decision Trees and Artificial Neural Networks. Dataset used for the study has 12 CT scan Images. 6 image was negative and 6 with positive labeled. Accuracy and Sensitivity calculated for better outcomes. The final result showed that Artificial Neural Network achieved 82% accuracy with image processing and Decision Tree achieved 93% accuracy without image processing.

Muhammad Imran Faisal et.al (2018) in this study they worked on text based dataset which has 56 attributes and 1 class labeled.[7] They Analyzed both ML algorithm and Ensemble algorithm. Supervised learning algorithms used in this study. Dataset divided into train and test with 10 fold cross validation. Both algorithm used for prediction of lung cancer. Finaly they Concluded that the Ensemble algorithms are best for predication of lung cancer on text based dataset instead of CT scan images.Gradient-boosted algorithm achieved 90% accuracy.

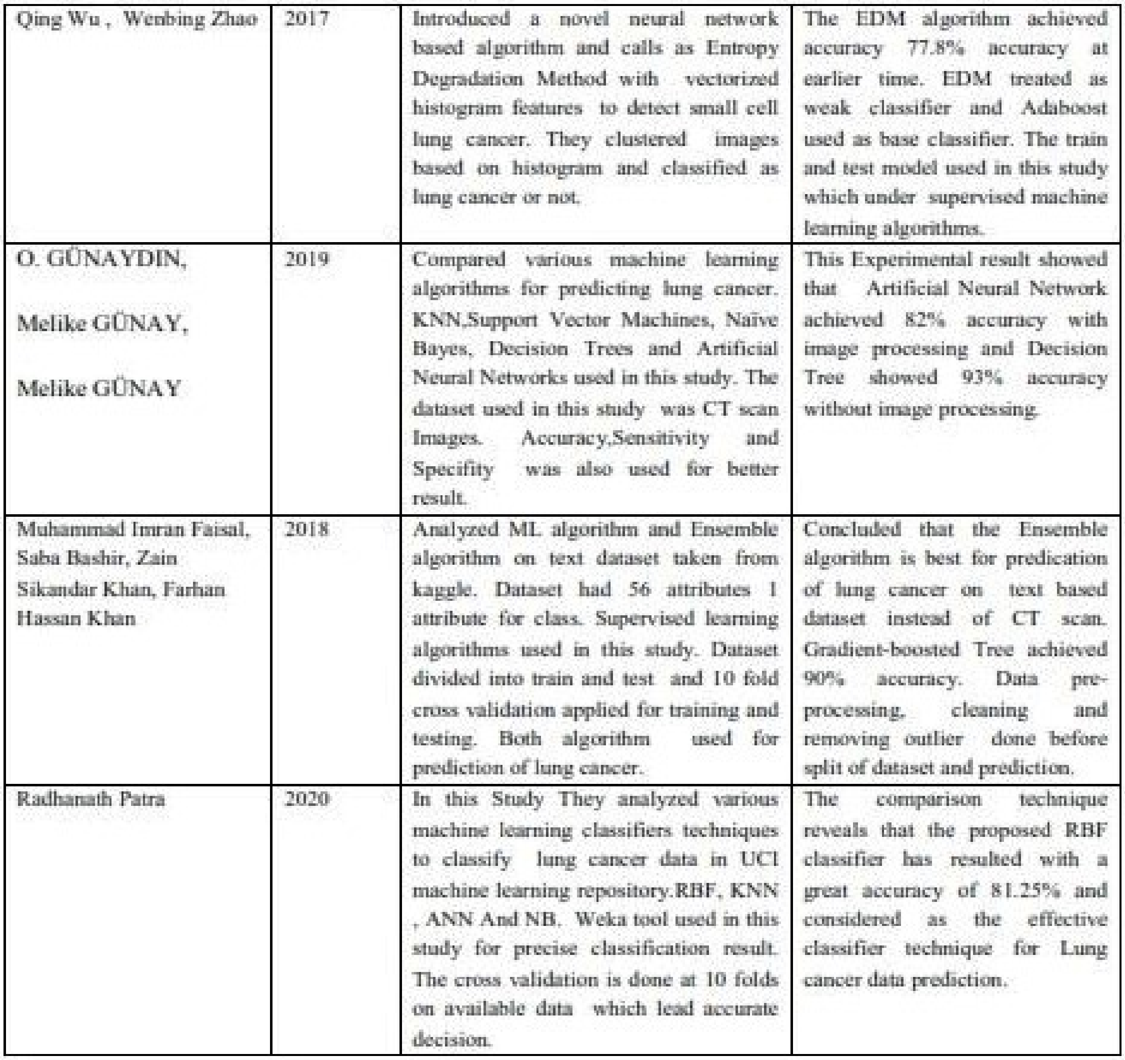
Radhanath Patra et.al (2020). In This Study They compared various machine learning algorithms to classify lung cancer.[8] Data is taken from UCI machine learning repository. The algorithms used in this study was RBF, KNN , ANN And NB. The tool used in this study was the Weka tool for precise classification result. 10 fold cross validation is done on available dataset to lead accurate decision. The compared technique reveals that the RBF had achieved 81.25% and considered as the effective classifier technique for Lung cancer data prediction.

Atharva Bankar et.al (2020) In this study they worked on text based dataset and Compared various Ensemble machine learning algorithms for predicting lung cancer.[9] The Ensemble algorithms used in this study was Decision Trees, Random Forest and XGBoost . Symptoms based dataset used in this study for predicting lung cancer at low cost. Accuracy, Sensitivity, Recall calculated with help of confusion matrix. The accuracy achieved on dataset was high. XGB and Random Forest has given accuracy 99.00%. It was analyzed that the Coughing of Blood, ,Genetic Risk, Clubbing of Finger Nails , Snoring and Passive Smoking are the main factors that are responsible for Lung cancer.

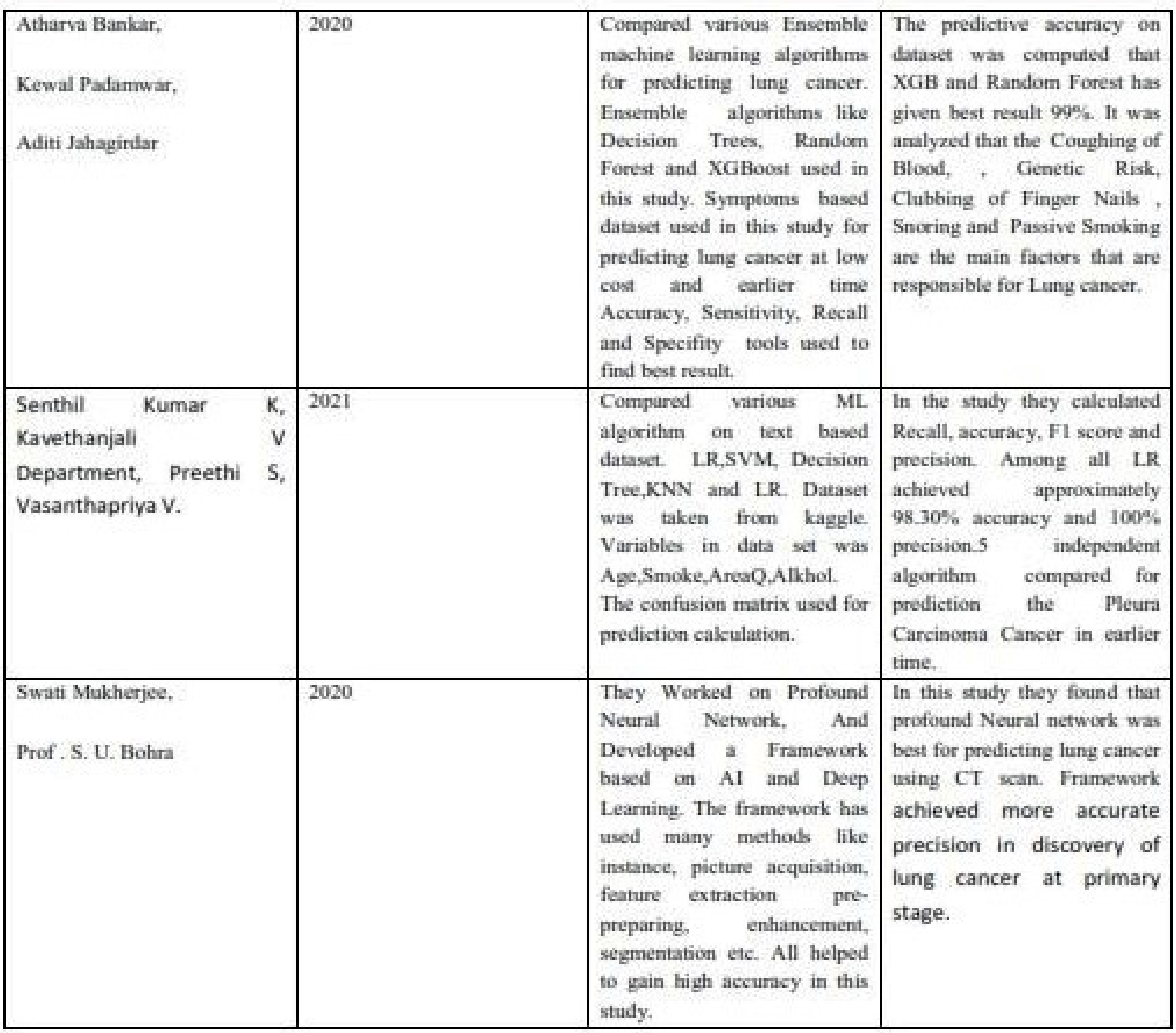
Senthil Kumar,et.al (2021). In this paper they compared various ML algorithms on text based dataset. The LR,SVM, Decision Tree,KNN and LR algorithms used in this paper.[10] Dataset was taken from kaggle. The dataset some variables like Age, Smoke, AreaQ,Alkhol Variables. They calculated Recall, accuracy, F1 score and precision for better result. Among all algorithms LR achieved approximately 98.30% accuracy and 100% precision. All independent algorithm compared for prediction the Pleura Carcinoma Cancer in earlier time.

Swati Mukherjee, et.al (2020). They Worked on CT scan Images and used Profound Neural Network algorithm, And Developed a Framework based on AI and DeepLearning. [11] The framework has used many methods like instance, picture acquisition, feature extraction pre-preparing, enhancement, segmentation etc that helped to gain high accuracy in this study. In this paper they found that profound Neural network was best for predicting lung cancer using CT scan. Framework achieved more accurate precision in discovery of lung cancer at primary stage.

Comparison Table



Qing Wu , Wenbing Zhao	2017	Introduced a novel neural network based algorithm and calls as Entropy Degradation Method with vectorized histogram features to detect small cell lung cancer. They clustered images based on histogram and classified as lung cancer or not.	The EDM algorithm achieved accuracy 77.8% accuracy at earlier time. EDM treated as weak classifier and Adaboost used as base classifier. The train and test model used in this study which under supervised machine learning algorithms.
O. GÜNAYDIN, Melike GÜNAY, Melike GÜNAY	2019	Compared various machine learning algorithms for predicting lung cancer. KNN,Support Vector Machines, Naïve Bayes, Decision Trees and Artificial Neural Networks used in this study. The dataset used in this study was CT scan Images. Accuracy,Sensitivity and Specifity was also used for better result.	This Experimental result showed that Artificial Neural Network achieved 82% accuracy with image processing and Decision Tree showed 93% accuracy without image processing.
Muhammad Imran Faisal, Saba Bashir, Zain Sikandar Khan, Farhan Hassan Khan	2018	Analyzed ML algorithm and Ensemble algorithm on text dataset taken from kaggle. Dataset had 56 attributes 1 attribute for class. Supervised learning algorithms used in this study. Dataset divided into train and test and 10 fold cross validation applied for training and testing. Both algorithm used for prediction of lung cancer.	Concluded that the Ensemble algorithm is best for predication of lung cancer on text based dataset instead of CT scan. Gradient-boosted Tree achieved 90% accuracy. Data pre-processing, cleaning and removing outlier done before split of dataset and prediction.
Radhanath Patra	2020	In this Study They analyzed various machine learning classifiers techniques to classify lung cancer data in UCI machine learning repository.RBF, KNN , ANN And NB. Weka tool used in this study for precise classification result. The cross validation is done at 10 folds on available data which lead accurate decision.	The comparison technique reveals that the proposed RBF classifier has resulted with a great accuracy of 81.25% and considered as the effective classifier technique for Lung cancer data prediction.



Atharva Bankar, Kewal Padamwar, Aditi Jahagirdar	2020	Compared various Ensemble machine learning algorithms for predicting lung cancer. Ensemble algorithms like Decision Trees, Random Forest and XGBoost used in this study. Symptoms based dataset used in this study for predicting lung cancer at low cost and earlier time Accuracy, Sensitivity, Recall and Specifity tools used to find best result.	The predictive accuracy on dataset was computed that XGB and Random Forest has given best result 99%. It was analyzed that the Coughing of Blood, , Genetic Risk, Clubbing of Finger Nails , Snoring and Passive Smoking are the main factors that are responsible for Lung cancer.
Senthil Kumar K, Kavethanjali V Department, Preethi S, Vasanthapriya V.	2021	Compared various ML algorithm on text based dataset. LR,SVM, Decision Tree,KNN and LR. Dataset was taken from kaggle. Variables in data set was Age,Smoke,AreaQ,Alkhol. The confusion matrix used for prediction calculation.	In the study they calculated Recall, accuracy, F1 score and precision. Among all LR achieved approximately 98.30% accuracy and 100% precision.5 independent algorithm compared for prediction the Pleura Carcinoma Cancer in earlier time.
Swati Mukherjee, Prof . S. U. Bohra	2020	They Worked on Profound Neural Network, And Developed a Framework based on AI and Deep Learning. The framework has used many methods like instance, picture acquisition, feature extraction pre-preparing, enhancement, segmentation etc. All helped to gain high accuracy in this study.	In this study they found that profound Neural network was best for predicting lung cancer using CT scan. Framework achieved more accurate precision in discovery of lung cancer at primary stage.

III CONCLUSION

In this work, it is concluded that various techniques are proposed for Lung Cancer prediction. The Lung cancer disease prediction technique have two type of dataset one is Image of CT scans and second one in Text or symptomatic. The various machine learning techniques are reviewed in this paper and concluded that the ensemble algorithms are given best result. In future hybrid ensemble machine learning algorithm will be designed on both image and symptomatic dataset for the Lung cancer prediction for better accuracy.

REFERENCES

[1] M. Selma, A. Mohamed, H. M. Yassine and B. Issam, "How to have a structured database for lung cancer segmentation using deep learning technologies," 2021 International Conference on Networking and Advanced Systems (ICNAS), 2021, pp. 1-5, doi: 10.1109/ICNAS53565.2021.9628946.

[2] D. Reddy, E. N. Hemanth Kumar, D. Reddy and M. P, "Integrated Machine Learning Model for Prediction of Lung Cancer Stages from Textual data using Ensemble Method," 2019 1st International Conference on Advances in Information Technology (ICAIT), 2019, pp. 353-357,

[3] R. P.R., R. A. S. Nair and V. G., "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1-4.

[4] Choubey, Ravi & Gautam, Pratima. (2021). Analysis of various machine learning Algorithms for Heart Disease Prediction. PIMT 13. 4.

[5] Q. Wu and W. Zhao, "Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm," *2017 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*, 2017, pp. 88-91.

[6] O. Günaydin, M. Günay and Ö. Şengel, "Comparison of Lung Cancer Detection Algorithms," 2019 Scientific Meeting on Electrical-Electronics&Biomedical Engineering and Computer Science (EBBT), 2019,pp.1-4..

[7] M. I. Faisal, S. Bashir, Z. S. Khan and F. Hassan Khan, "An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer," 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), 2018, pp. 1-4.

[8] Patra R. (2020) Prediction of Lung Cancer Using Machine Learning Classifier. In: Chaubey N., Parikh S., Amin K. (eds) Computing Science, Communication and Security. COMS2 2020. Communications in Computer and Information Science, vol 1235. Springer, Singapore.

[9] A. Bankar, K. Padamwar and A. Jahagirdar, "Symptom Analysis using a Machine Learning approach for Early Stage Lung Cancer," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020, pp. 246-250.

[10] Senthil . K. K, K. V, P. S and V. V, "Lung – Pleura Carcinoma Detection Using Machine Learning," 2021 3rd International Conference on Signal Processing and Communication (ICPSC), 2021, pp. 294-298.

[11] S. Mukherjee and S. U. Bohra, "Lung Cancer Disease Diagnosis Using Machine Learning Approach," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020, pp. 207-211.