International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6920
***Article Received**: 28 March 2023 **Revised**: 10 May 2023 **Accepted**: 29 May 2023*

Lung Cancer Detection and Classification using Machine Learning Algorithms

Ashwani Mishra[1], Sanjeev Gangwar[2]
[1]Department of Computer Science & Engineering
U.N.S.I.E.T VBS Purvanchal University Jaunpur, Uttar Pradesh, India.
E-mail: erashwani7072@gmail.com
[2]Department of Computer Application
U.N.S.I.E.T VBS Purvanchal University Jaunpur, Uttar Pradesh, India.
E-mail: gangwar.sanjeev@gmail.com

Abstract— Lung cancer is a clump of cells in the lung that are multiplying uncontrollably and improperly. Lung cancer is the deadliest disease, and its cure should be the primary focus of all scientific research. Although it cannot be prevented, we can lessen the danger. Thus, a patient's chance of life depends on the early identification of lung cancer. Several machine learning methods, such as Support Vector Machine, Logistic Regression, Artificial Neural Networks, and Naive Bayes, have been used for the investigation and prognosis of lung cancer. In this paper, Lung cancer prediction is finished by gathering the dataset from the survey and applying machine learning methods such as Support Vector Machine, Nave Bayes, K-Nearest Neighbors, Decision Tree, and Random Forest. With this result, it is revealed that Decision Tree attained the maximum accuracy of 100% as compared to the others.

Keywords- Machine Learning, Support Vector Machine, Naïve Bayes, Random Forest, Decision Tree.

I. INTRODUCTION

Cancer is abnormal division and uncontrolled group of cell, it also affects near by cells or tissues. Lung is a lower respiratory organ in our body through which we can inhale oxygen inside lungs and exhale carbon dioxide outside lungs. Lung cancer is type of cancer in which air sacs of bronchioles become reduce in size and a bundle of cells become collected. The most common cancer that leads to death globally is lung cancer. Some people have bad habits of smoking which is the greatest risk of lung cancer while second hand smoke also affects some people.

Asbestos and radon are some other agents which causes lung cancer in many people. Lung cancer causes many health problems such as shortening of breath cause due to decrease surface area of air sacs(alveoli), Coughing with blood because cells abnormally divide and become damage, Weight loss due to improper break down of glucose which provide energy to cell.

Lung cancer is the 2nd most common cancer in men & 5th most common cancer in both men and women together. Basically lung cancer is two types:

i. Small Cell Lung Cancer(SCLC)
ii. Non-Small Cell Lung Cancer(NSCLC)

A. Small Cell Lung Cancer(SCLC)

Small cell lung cancer is a type in which tobacco smoking is the main cause of the lung cancer, it contribute 10-15% of all lung cancer cases. SCLC has ability to grow and spread faster than other type of lung cancer. Small cell lung cancer is divided into the two categories:

i. Small cell Carcinoma
ii. Combined small cell carcinoma

Small cell Carcinoma is a very dangerous cancer because it also affect other body part such as cerbix, prostate and gastrointestinal tract along with lungs .

Combined small cell carcinoma occurs along with other types of lung cancer such as squamous cell carcinoma or adenocarcinoma.

B. Non-Small Cell Lung Cancer(NSCLC)

Non-small cell lung cancer occurs 85% of all lung cancer, it occurs in smokers as well as non-smokers also. It do not spread and grow faster as compare to small cell lung caner. NSCLC id divided into three categories based upon the type of cells in the tumor involved:

i. Adenocarcinoma
ii. Squamous cell carcinoma
iii. Large cell lung carcinoma

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6920
***Article Received**: 28 March 2023 **Revised**: 10 May 2023 **Accepted**: 29 May 2023*

Adenocarcinoma occurs in the out of portions of our lungs, it occurs mainly effects younger people slowly grow in the glands of alveoli. Squamous cell carcinoma occurs on those flat cells that line the inside of our airways, slowly grow in squamous cells in the bronchi usually in center of the lungs. Large cell lung carcinoma is least common among all types of NSCLC, It is fast growing and start spreading in large cells found anywhere in the lungs, but mostly in outer part. The Non-small cell lung cancer graphical representation of its type shown in below figure 1.

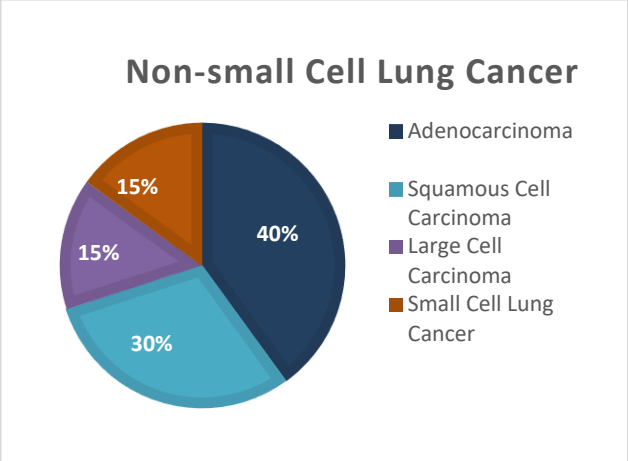

Figure 1. Types of Non-small Cell Lung Cancer

People who are suffering lung cancer may have such symptoms:

i. Coughing
ii. Chest Pain
iii. Wheezing
iv. Shortness of Breath
v. Weight loss without nonintentional

Lung cancer can be diagnose with the help of CT(Computed Tomography) Scan, X-Ray, MRI(Magnetic resonance imaging), PET(Positron Emission Tomography) Scans, Sputum cytology and bronchoscopy. Lung cancer can be treated by following ways such as surgery, Radiation therapy, chemo therapy, Targeted drug therapy, Stereotactic body Radio therapy etc.

II. LITERATURE REVIEW

Ibraim Goni et al [4]; defined "Lung Cancer Detection Using Convolutional Neural Network ", In this paper CNN(Convolutional Neural Network) and MLP model is develop to predict the lung cancer using the Lung Image database consortium(LIDC) and Image database resource initiative(IDRI) dataset which contain the one thousand image of lungs with different size. The CNN model obtain the 94.13% accuracy and MLP model achieved only 79.67%.

C. Anil Kumar et al [5]; proposed a Lung Cancer Prediction from Text Datasets Using Machine Learning. In this paper support vector machine(SVM) algorithm are use to build a model for predict lung cancer with Synthetic minority oversampling technique(SMOTE) and without SMOTE. To evaluate the model dataset collected from the University of California, Irvine, repository. The SVM model obtain accuracy with SMOTE 81% and without SMOTE 79%.

Abhishek Gupta et al [6]; proposed " A Study on Prediction of Lung Cancer Using Machine Learning Algorithms ", In this paper three machine learning algorithms are used(K-NN, Random Forest and Support Vector Machine). For making the model they collect the dataset from the Kaggle. They use Gabor, Sobel and Gaussians filters on collected biomedical image. They find the best model accuracy on Random Forest algorithms. The Random forest model obtain 84.2% , Support Vector Machine model obtain 82.1% and K-NN model obtain only 48.7%.

Elinor Nemlander et al [7]; defined " Lung Cancer prediction using machine learning on data from a symptom e-questionnaire for never smokers, formers smokers and current smokers", In this paper, they work on 504 patients using the e-questionnaire, which covers symptoms of lung cancer, and find 310 patients affected by lung cancer, while 194 patients are not affected. Don't smoke. In comparison to 36 predictors with an accuracy of 63% among former smokers and 26 predictors with an accuracy of 77% among current smokers, 17 predictors contributed to the prediction of lung cancer, correctly classifying 82% of the patients. They develop assessment tools that can help clinicians assess a patient's risk of having lung cancer.

Madhushree A et al[8]; describe a lung cancer detection using machine learning. They use Convolutional Neural Network(CNN), Support Vector Machine(SVM) and KNN machine learning algorithm. Develop model on Google collab using the python programming. To create a model they use the Kaggle Data Science Bowl(KDSB17) dataset which contain 2101 CT scans of Patient chest. KNN model obtain 84%, SVM model 88% and CNN model obtain 92% accuracy using this algorithms.

Vasupalli Saranya et al [9]; proposed " Lung Cancer Prediction Using Machine Learning", In this Paper Non parametric, supervised learning classifier are used to predict the lung cancer on the basis of the symptoms, and collect the dataset from the Kaggle which contain attributes like Age, Gender, Air Pollution, Alcohol, Swallowing Difficulty etc. They discover that KNN (K-nearest neighbor) is one of the algorithms with a 96% accuracy.

Smait Raut et al [10]; defined a lung cancer detection using machine learning approach. They collect the CT(computed Tomography) scans image from the website and DI-COM

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6920
***Article Received**: 28 March 2023 **Revised**: 10 May 2023 **Accepted**: 29 May 2023*

software. Using the MATLAB process on CT Scans image and analyze properties which differentiate between the cancerous image and normal image. They develop a system which is automatically detect the cancer cell by using machine learning algorithm and digital image processing.

Vemula Suvarchala et al [11]; proposed " Lung Cancer Prediction Using Machine Learning Methodologies", In this paper for prediction the lung cancer use the RBF(Radial Base Functions) classifier. For making the model, collect the data from the UC Irvine Machine Learning Repository, which consists of 32 instances with 57 characteristics. The RBF classification model has an accuracy of 81.25% for lung cancer.

Bushara A. R. et al[12]; develop a deep learning based lung cancer classification of CT images using augmented convolutional neural networks. In this paper DICOM image format and the CNN algorithm are both utilised to predict lung cancer. They get the CT scan image from the Image Dataset Resource Initiative of the Lung Imaging Dataset Consortium (LIDC-IDRI). They utilised the sigmoid activation function and RelU to build the model, and the CNN techniques yielded a 95% accuracy rate.

Aswathy S. U. et al[13]; defined a deep learning based BoVW-CRNN model for lung tumor detection in nano-segmented CT images. In this paper, the lung tumour is predicted using the Bag of Visual Words (BoVW) and Convolutional Recurrent Neural Network (CRNN) techniques. They gather information from the Lung Image Database Consortium (LIDC) and use nanoimages as input for enhanced images produced by the Gabor filter in order to create the model. Nanotechnology-based detection method with 98.5% accuracy for the CRNN classifier model and 96.5% accuracy for the BoVW classifier model.

Sumathi C et al[14]; proposed "Medical Imaging with Artificial Intelligence for Lung Disease Analysis: A Comprehensive Review". In this article, the model is built using the CNN, VGG16, VGG19, and ResNet 50 algorithms, and two datasets from the Kaggle repository are utilised. These datasets comprise normal, pneumonia, and covid19 pictures. When using the first dataset, the accuracy of the CNN model is 90%, the VGG16 model is 91%, the VGG19 model is 90%, and the ResNet 50 model is 91%.

III. DATASET

The dataset has been collected through a survey by form. This dataset includes attributes such as Age, Gender, Smoking, Chronic Disease, Coughing, Shortness of breath, Swallowing Difficulty, Chest pain, Alcohol consumption, Lung Cancer. These features are numerical attributes which can be easily used for prediction. The last attributes if this dataset is Lung cancer which predict the Lung Cancer result with two text values Yes and No. All related attributes, we are taken by the consultant doctor.

IV. PROPOSED METHODOLOGY

A. Support Vector Machine

SVM is one of the most well-liked algorithms for supervised learning, and it may be applied to both classification and regression issues. Even though, it is widely applied in Machine Learning Classification problems. The extreme case of the data will be seen since the support vector generates a decision boundary between these two sets of data and selects them. It will categories it based on the support vectors. The objective of the SVM algorithm is to find the hyperplane that best separates the data into two classes, while also maximizing the margin between the two classes. The objective function can be defined as:

$$\min\ 1/2\ w\text{^}2 \quad (1)$$

Subject to

$$y_i\ (wx + b) - 1\ \geq 0, i = 1 \ldots . n \quad (2)$$

where w is the weight vector, b is the bias term, x is the feature vector, y_i is the class label of the sample, and n is the number of samples.

B. Decision Tree

Decision Tree is a Supervised learning approach that can be used for both classification and Regression issues, however generally it is favored for addressing Classification problems. It is a tree-structured classification algorithm, where core nodes reflect the properties of a dataset, branches represent the decision rules and each leaf node represents the output. The Decision Node and Leaf Node are the two nodes present in a decision tree. Leaf nodes are the results of choices and do not have any more branches, while decision nodes are used to create decisions and have numerous branches. It is known as a decision tree because, like a tree, it begins with the root node and then extends to form more branches and a tree-like structure. Both numerical data and categorical data (YES/NO) may be included in a decision tree. Entropy of the parent is first computed, and then the information gain is calculated by deducting the weighted sum of the entropies of the offspring from that of the parent.

Information Gain= Entropy- [(Weighted Avg) * Entropy(each feature)] (3)

The one with largest information gain is considered as the root node and the process proceeds in till the classification is done. Each node in the decision tree specifies a specific symptom from the collection S = {S1,S2,S3,S4.....Sj}, where S denotes

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6920
***Article Received**: 28 March 2023 **Revised**: 10 May 2023 **Accepted**: 29 May 2023*

conditional characteristics. Vi, j stands for each branch's values, or the h-th range for the i-th symptom and leaves that provide a choice. D = {D1,D2,D3,D4,.....Dk} and W dk = {0,1} are their binary equivalents.

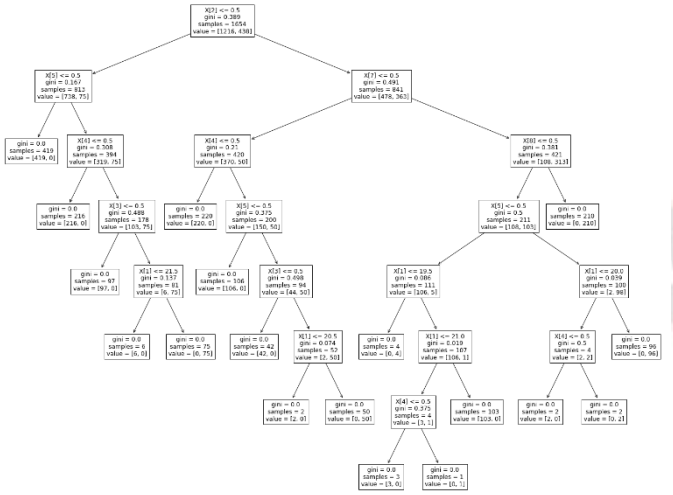

Figure 2. Proposed Decision Tree

When we apply decision tree algorithms to data, we get 100% accuracy.

C. K-Nearest Neighbour

The K-Nearest Neighbors (KNN) technique is a form of supervised machine learning algorithm that may be used to tackle challenges involving classification and regression-predictive. KNN predicts the values of new datapoints using "feature similarity," which further indicates modeling that the new data point will be given a value depending on how closely it resembles the points in the training set. KNN maintains all of the examples that are accessible and categories additional cases using a similarity metric. The parameter K in KNN denotes the number of nearest neighbors to be taken into account for picking a winner by majority vote. square root of n, where n refers to the total amount of data points. If n is even, we must add or remove 1 to make the value odd, which aids in better selection. Because KNN is a "lazy learner," we may apply it when the dataset is labelled, noise-free, and small.

Steps involved in K-Nearest Neighbors:

1. Choose the ‘k’ neighbour's number.
2. Determine the Euclidean distance between ‘k’ neighbours.
3. Pick the ‘k’ closest neighbours based on the Euclidean distance estimate.
4. Count how many data points there are in each category among these k neighbours.
5. The category for which the number of neighbours is highest should get the additional data points.

When we apply K-Nearest Neighbors algorithms to data, we get 95.4% accuracy.

D. Naive Bayes

The Naive Bayes algorithm, based on Bayes' theorem, is a supervised learning technique for classification problems. It is generally applied in text classification with a massive training set. One of the most straightforward and efficient classification algorithms is the naive bayes classifier, which aids in the development of rapid machine learning models capable of making accurate predictions. As a probabilistic classifier, it bases its predictions on the likelihood that a given item will occur. Let X is contains features X= {X1, X2, X3……., Xn} and Y is Output contain {Yes / No}. To get the probability, we utilized the following equation.

$$P(Y|X) = P(X|Y).P(Y) / P(X) \quad (4)$$

Where ,

P(Y|X) is a Posterior probability

P(X|Y) is a Likelihood probability

P(X) is a Marginal Probability

P(Y) is a Prior Probability

Naive Bayes algorithms provide an accuracy of 87.02% when we applied to data.

E. Random Forest

Random Forest is a popular machine learning algorithm that is used for various tasks, including classification and regression. In Random Forest, multiple decision trees are grown using bootstrapped samples of the training data. At each split, a random subset of the features is selected to determine the best split.

Steps involved in Random Forest Algorithms:

1. First, take ‘m’ number of random samples from ‘n’ number of different datasets.
2. Create a decision tree for each training set of data.
3. Every decision tree will have a result.
4. Finally, decide which prediction result received the most votes will be the final prediction result.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6920
Article Received: 28 March 2023 Revised: 10 May 2023 Accepted: 29 May 2023

V. RESULT & DISCUSSION

In this Research work, we have used various machine learning classification algorithms like Support Vector Machine, K-Nearest Neighbour, Naïve Bayes, Decision Tree and Random Forest.

Table II: Classification Report of Proposed Model

Proposed Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
K-Nearest Neighbour	96	98	97	94.9
Naïve Bayes	91	97	94	90.6
Decision Tree	100	100	100	100
Random Forest	98	100	99	98.4
Support Vector Machine	91	99	95	91.8

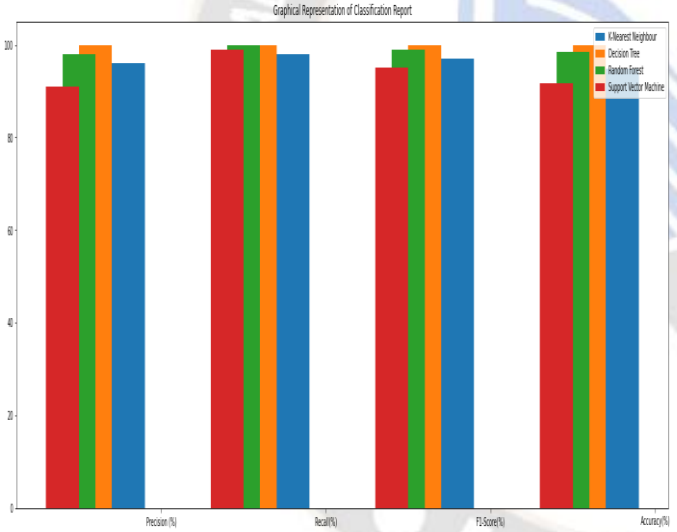

Figure 3. Graphical Representation of Classification Algorithms

The above Table II and Figure 3 make it very evident that, when compared to other classifier algorithms, the decision tree classifier outperformed them all by achieving 100% precision, 100% accuracy, 100% F1-score, and 100% recall.

VI. CONCLUSION

The disease lung cancer has a high mortality rate. Machine learning algorithms help to predict the lung cancer of the patient easily. There are many algorithms available, like linear regression, SVM, random forest, K-nearest neighbor, logistic regression, decision trees, and naive bayes. The decision tree is one of the algorithms with a good accuracy of 100% for predicting correctly. In the future, the models that have been generated during this research might be incorporated into an app that helps forecast lung cancer at an early stage.

REFERENCES

[1] S. 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, doi: 10.1109/icspc51351.2021.9451769.

[2] 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, doi: 10.1109/icecct.2019.8869001.

[3] M. A. Alzubaidi, m. Otoom and h. Jaradat, "comprehensive and comparative global and local feature extraction framework for lung cancer detection using ct scan images," in ieee access, vol. 9, pp. 158140-158154, 2021, doi: 10.1109/access.2021.3129597.

[4] Ibraim goni., et al. "lung cancer detection using convolutional neural network". Acta scientific computer sciences 4.9 (2022): 06-08.

[5] C. Anil kumar, s. Harish, prabha ravi, murthy svn, b. P. Pradeep kumar, v. Mohanavel, nouf m. Alyami, s. Shanmuga priya, amare kebede asfaw, "lung cancer prediction from text datasets using machine learning", biomed research international, vol. 2022, article id 6254177, 10 pages, 2022. Https://doi.org/10.1155/2022/6254177.

[6] Gupta, a., zuha, z., ahmad, i., & ansari, z. (2022). A study on prediction of lung cancer using machine learning algorithms. Research square platform llc. Https://doi.org/10.21203/rs.3.rs-1912967/v1

[7] Nemlander e, rosenblad a, abedi e, ekman s, hasselström j, et al. (2022) lung cancer prediction using machine learning on data from a symptom e-questionnaire for never smokers, formers smokers and current smokers. Plosone17(10):e0276703 https://doi.org/10.1371/journal.pone.0276703.

[8] Madhushree a, harshitha nayaka ys, chandrika m, madangowda hs, & mrs. Pallavi j. (2022). Lung cancer detection using machine learning. In international journal of advanced research in science, communication and technology (pp. 484–489). Naksh solutions. Https://doi.org/10.48175/ijarsct-5854.

[9] Vaishnavi. D, arya. K. S, devi abirami. T, m. N. Kavitha, 2019, lung cancer detection using machine learning, international journal of engineering research & technology (ijert) rticct – 2019 (volume 7 – issue 01).

[10] Raut, s., patil, s., & shelke, g. (2021). Lung cancer detection using machine learning approach. International journal of advance scientific research and engineering trends (ijasret).

[11] Vemula suvarchala, p., & madala, s. R. (2021). Lung cancer prediction using machine learning methodologies. Nveo-natural volatiles & essential oils journal| nveo, 1265-1272.

[12] Ar, b. (2022). A deep learning-based lung cancer classification of ct images using augmented convolutional

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6920
Article Received: 28 March 2023 Revised: 10 May 2023 Accepted: 29 May 2023

neural networks. Elcvia electronic letters on computer vision and image analysis, 21(1).

[13] Su, a., pp, f. R., abraham, a., & stephen, d. (2023). Deep learning-based bovw–crnn model for lung tumor detection in nano-segmented ct images. Electronics, 12(1), 14.

[14] Sumathi, c. (2022). Medical imaging with artificial intelligence for lung disease analysis: a comprehensive review. Journal of pharmaceutical negative results, 13(4), 1756-1768.