Lung Cancer Detection using Machine Learning

Bharathy S
Assistant Professor , CSE
Sri Sai Ram Engg., College
Chennai, Tamilnadu
Bharathy.cse@sairam.edu.in

Pavithra R
Student, cse
Sri Sai Ram Engg., College
Tamilnadu, Chennai
pavithraramkumar2001@gmail.com

Akshaya.B
Student, IT
Sri sairam Engg,college
Tamilnadu, Chennai
akshaya_1407@ieee.org

Abstract — Lung disease is one of the most common disease that is affected in our early stage to improve the rate of patients survival. For the radiologist the diagnosis of cancer is the most challenging part. An intelligent computer aided system is very much helpful for radiologist. Various studies for detection of lung cancer with the ML techniques. To predict the lung cancer mostly multi-stage classification is used. The classification system used for data enhancement and segmentation has been done. The segmentation method uses Threshold and marker-controlled watershed and binary classifier for classification method Lung cancer detection has higher degree of accuracy. The dataset is trained with various algorithms like Support Vector Machine (SVM), K-Nearest Neighbour, Decision Tree, Logistic Regression, Naïve Bayes and Random Forest using these algorithms higher accuracy is proven. An enhanced performance level of 88.5% accuracy has been produced with the Random forest algorithm.

Keywords:— Lung disease, Random Forest algorithm, K-Nearest Neighbour(KNN), Support Vector Machine (SVM), Decision Tree algorithm, Logistic Regression, Naïve Bayes algorithm.

I.INTRODUCTION

Lung disease causes several risk factors such as age, sex, diabetes, high blood pressure, cholesterol, abnormal pulse rate and other factors. Many techniques in data mining and neural networks have been taken to bring out the ailment of lung disease. The algorithms used for this disease is classified based on various methods like K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Naïve Bayes(NB), Random forest, Logistic Regression, Support Vector Machine(SVM). The nature of lung disease is very complex and hence, the disease must be handled in a proper way. The opinion of lung cancer in medical science and data mining are used for discovering various metabolic syndromes. Data mining with classification and clustering plays an important role in the prediction of lung disease and data investigation. Many problems damages the lung and even causes premature death. The people who are having the highest risk of getting cancer because of smoking. Lung cancer is divided into two types, namely non-small cell lung cancers(NSCLC) and small cell lung cancers(SCLC). It also kills more people whose who have affected by breast cancer, colon, prostate or ovarian. Lung cancer when their age increases to 60. Lung cancer is the growth of unusual cells in our lungs. These lung tissues divide rapidly and form tumors. There are some of the new techniques to analyse the lung cancer such as Thalassemia, Constrictive pericarditis, the person’s resting blood pressure(trestbps), fasting blood sugar(FBS), Electrical activity of our hearth at rest(Restecg), Maximum heart rate achieved(Thalach) and Coronary Calcium scan(CA) these are the data sets that needs to be trained and tested. where 75% of the data is used for training and remaining 25% is used for testing.

II. LITERATURE SURVEY

2.1 LUNG CANCER INCIDENCE PREDICTION USING MACHINE LEARNING ALGORITHMS.

The cancer is called malignant tumor caused by an irregular division of tissue or organ in cells. Although many types of cancer in both male and female. It is based on Statistical analysis of its data and neural-based models. The goal of the Back propagation algorithm is used for multi-layer perceptron to update the weight of the neurons with the gradient decent algorithm. Generally Initial weight are assigned randomly and it starts the input in its feeding to net and calculating its total potential by their corresponding weights following the hidden layer. The output produced by its Activation function of each neuron and the calculations are repeated till output layer. At that layer the output is compared with target and error. It is used in real-life applications, optimized problems and prediction. The Long Short term Memory Network(LSTM) is effective of recurrent network and used for the classification method and prediction analysis. The major components are cell, input gate, output gate and forget gate. Forget gate is used for the irrelevant data and input gate removal and accepts the forget gate. Output of LSTM uses the sigmoid activation function. It uses weights, it remembers the previous errors and minimization of network. The Support Vector Regression is kind of support vector machine to accept the real value in the binary format. Prediction problem is used effectively. It creates the sub-class of training data known as
support vectors and minimizes the observed data and predicted data distance to improve its performance.

2.2 MULTISTAGE LUNG CANCER PREDICTION AND DETECTION USING MULTICLASS SVM CLASSIFIER.

Lung cancer using Image Enhancement technique of better quality impression pictures. Images used for this purpose masking is subjected to a selective median filter. More dependable for preparing and identification of more dependable to consolidate the versatile edge to its own commotion recognition that is used. The Image Segmentation and Detection of cancerin CT images to gain its own better recognition of the image. The watershed lines depends to compare the edges of the marker and it is not influenced to its lower-construct edgesto problem that needs to be solved in neighbourhood minima. The Feature Extraction and Detection involves in huge arrangement of information that needs to be arranged in decreades . Feature Extraction of the normality or variation of normal picture. Feature Extraction arrange a large combination of pixels. Classification of cancer nodule of utilized SVM classifier of the machine learning algorithm. SVM uses the effective tool for hyper plane classifier that expands the edges. Cancer stage classification of the affected lung area and the total affected area.

2.3 SEX AND SMOKING STATUS EFFECTS ON THE EARLY DETECTION OF LUNG CANCER IN HIGH RISK SMOKERS USING AN ELECTRONIC NOSE.

Respiratory diseases like asthma, chronic pulmonary disease are identified by breath odor. This is because of the equilibrium of the air and pulmonary blood gas in the breath analysis of diagnostic disease in the lung cancer. An electronic nose (e-nose) is VOC in that array of the sensors with overlapping the digital VOCs. They also detect with the chemical reaction and can generate an electrical impulse. These are the sensors coated with reactive compound depends mainly on chemical constituents, an electrical sensor causing measurable resistance change. Data that it needs to be obtained in the form of pattern recognition technique to obtain the specific odor. E-nose which is capable of measuring a non-invasive breath samples in the real time analysis of chemicals. The e-nose system which contains 32 polymer sensors with some unique pattern of electrical resistance. e-nose for lung cancer patients used to facilitate the lung cancer in advanced stage, effects due to smoking comparison of high risk current. To study its majority and design in a cross sectional case control for the lung cancer patients for detected cancer. It has been ranged from 45-79 years of male or female from the history of past years. The “High risk Smoker” and “Lung cancer” in terms of sex. It is based on the LDCT lung cancer detection.

2.4 AUTOMATIC DETECTION OF ABNORMALITIES IN LUNG RADIOGRAPHS CAUSED BY PLANOCELLULAR LUNG CANCER.

Automatic algorithm for early planocellular detection using the lung Xray images, the fact that the lung cancer in this stage is detected too late, early diagnosis uses the radiography to analyze the diagnostic tools. It determines lung cancer using the Xray images, it involves the comparison between the extracted planocellular lung cancer structure and it has been analysed with lung Xray image by calculating with its coefficients and finding its maximum coefficient which indicates the suspected cancer affected area of the lung image. The result compromises the proposed algorithm that the lung cancer is predicted earlier. In this method the lung cancer is detected later that the planocellular has been extracted in the early method of using these analysis in scope of detecting the lung cancer.

III. PROPOSED METHODOLOGY

The proposed system will add some of the datasets for the lung cancer detection such as age, sex, Constrictive pericarditis (CP), person’s resting blood pressure (trestbps), Cholesterol, Fasting blood sugar (FBS), Records the electrical activity of your heart at rest (Restecg), person’s maximum heart rate achieved (Thalach), peak weaves (exang), the J point (the point of infection at the junction of the S wave and ST segment) becomes depressed during exercise therefore slopes sharply upwards (SLOPE), Coronary calcium scan (CA), Thalassemia (thal) are the datasets that needs to be trained and tested, from these datasets the target of affected and not affected by lung cancer. It uses the Random forest algorithm, K-nearest neighbour (KNN), Support Vector Machine (SVM), Decision tree algorithm, Logistic Regression, Naïve bayes algorithm to find out the highest accuracy and it compares with the target be be achieved will be the predicted result.
IV. DESIGN MODULES

1. PREPROCESSING TECHNIQUES

The datasets obtained from the Kaggle.com which contains of the 303 datasets. We are importing numpy library function from the pandas the store the data in the csv file format, the matplotlib and seaborn library function is used for the virtualization. Seaborn function is used to plot the numerical values in the form of graphs, it also involves the estimation of the percentage, mean value and standard deviation then it scaling to unit variance, it finds the average value of people affected and not affected. About 70% values lie in its between -1 to 1. By using the pandas function we are finding the target.

2. CLASSIFICATION OF DATA

Using the sklearn.model_selection we are importing the train_test_split. This technique of the train_test_split used for the evaluation of machine learning algorithm performance. The process of this is to take the datasets and to datasets needs to be cleaned. It involves taking the datasets and dividing it into two subsets. The datasets has to be trained and tested, the datasets splits up of 303 datasets to (203) datasets needs to be trained and (100) datasets needs to be tested, it needs of classification and regression problem.

3. MACHINE LEARNING ALGORITHMS

3.1 K-NEAREST NEIGHBOUR ALGORITHM

K nearest neighbors is the simplest form of the machine learning algorithm which is purely based on Supervised learning technique, this algorithm assumes many similarity between the no of cases and data available for cases and new data cases. The KNN algorithm uses feature similarity to find the values of new data values. It is also called as the lazy learner algorithm. First it selects the K nearest neighbour, it calculates the the Euclidean distance K Number of neighbours. Among these K neighbour the number of data points are categorized. Assign its data points to which the neighbour has maximum. It gives the range of 20 neighbour and prints the result in the form of graph.

3.2 RANDOM FOREST ALGORITHM

Random forest algorithm is also known as Supervised machine learning algorithm, and it has been briefly used in the Classification and Regression problems. The different samples of decision trees takes the major vote for classification and regression. The data sets that contains the continuous variables of the random forest algorithm has the regression and categorical variables of classification. It takes place in the ensemble technique. Ensemble means combining its multiple models, and it’s a collection of models it is used to make predictions to an individual model.
3.3 DECISION TREE ALGORITHM

The decision tree algorithm usually belongs to family of supervised learning algorithms. The decision tree algorithm is used for solving the Regression Problem and classification problems. It trains the model and it is used to predict the class or value that the target variable used for learning simple decision for their priored data. Predicting a class label to start the root of the tree. It has two nodes Decision node and leaf node, it is used to make the decisions and have a multiple branches of the leaf nodes are the output of those decisions. It uses a graphical representation for getting a possible solution to that problem of the conditions.

3.4 LOGISTIC REGRESSION

Logistic regression is known as a supervised learning algorithm that can be used to predict its target variable on a dependent category, it has a large set of data in the logistic regression, it can be either yes or no, true or false etc, it gives its value between 0 and 1 of probabilistic values. Logistic regression is very similar to linear regression, it fits a S shaped logistic function with two maximum values 0 or 1. Logistic function has the curve indicates the cells are cancerous. Logistic regression is known as significant machine because it provide probabilities of new data of continuous and discrete datas.

3.5 NAÏVE BAYES ALGORITHM

Naïve bayes algorithm is known as supervised learning algorithm based on its bayes theorem of solving its classification problem, it is used in classification of the im

age text which has its high-dimensional datasets, it is one of the simplest classification algorithm, building fast machine learning models, it has a probabilistic classifier it is used to predict its probability of an object on the basis. It is comprised of Naïve and bayes, it has certain feature of independent of other features based on its color, shape and taste recognized.

IV RESULT

The project mainly focuses on detecting of lung cancer using machine learning algorithm to find out the highest accuracy level using logistic regression, random forest, naïve bayes, k-nearest neighbour and decision tree with the datasets of age, sex, cp, trestbps, chol, fbs, restecg, thalach, ca using these datasets accuracy is detected. The higher accuracy of the algorithm is

V CONCLUSION

Lung cancer causes the cancer-related worldwide. Those 60% diagnosed with the lung cancer die after diagnosis for all the patients with lung cancer. Although the molecular pathology has lung tumor achieved with targeted treatments. Lung cancer diagnosis is essential for its selection of appropriate curative of non-invasive procedures. Thorax computerized tomography (CT) and its positron emission tomography are used for its non-invasive techniques. Machine learning techniques were used to process it in the raw data and provide a new novel in its lung cancer detection. However, the disease must be controlled in every stage and measures must be adopted. The future course that in the machine learning algorithms the effective way to predict the lung cancer must be diagnosed in its similar ways to analyse. Many feature selection methods are to be involved in the process selection method to predict the lung cancer.
REFERENCES

[1] A. Rehman, M. Kashif, I. Abunadi and N. Ayesh, "Lung Cancer Detection and Classification from Chest CT Scans Using Machine Learning Techniques," 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), 2021, pp. 101-104, doi: 10.1109/CAIDA51941.2021.9425209.

[2] R. D. Karkhiseyan, R. G, V, V. B. C and K. M, "A Review of Lung Cancer Detection using Image Processing," 2021 Smart Technologies, Communication and Robotics (STCR), 2021, pp. 1-4, doi: 10.1109/STCR51658.2021.9588835.

[3] Kubra Tunca, Boran Scekeroglu, Cagri Ozkan (2020,May) "Lung cancer prediction using machine learning algorithms". Turkey.

[4] S. S. Raoof, M. A. Isbar and S. A. Fathima, "Lung Cancer Prediction using Machine Learning: A Comprehensive Approach," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMA), 2020, pp. 108-115, doi: 10.1109/ICIMA48430.2020.9074947.

[5] Q. Firdaus, R. Sitt, T. Harsono and A. Anwar, "Lung Cancer Detection Based On CT-Scan Images With Detection Features Using Gray Level Co-Occurrence Matrix (GLCM) and Support Vector Machine (SVM) Methods," 2020 International Electronics Symposium (IES),2020 pp.643-648, doi:10.1109/IES0839.2020.9231663.

[6] N. S. Nadlami and S. Borkar, "Detection of Lung Cancer in CT Images using Image Processing," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 863-866, doi: 10.1109/ICOEI2019.8862577.

[7] Ö. Günaydın, 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, doi:10.1109/EBBT.2019.8741826.

[8] R. Y. Blalacoo, H. P. Jany, R. K. Gaitonde and V. Raut, "A novel approach for detection of Lung Cancer using Digital Image Processing and Convolution Neural Networks," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019, pp. 577-583, doi: 10.1109/ICACCS.2019.8728348.

[9] W. Rahane, H. Dalvi, V. Matur, A. Kalane and S. Jondhale, "Lung Cancer Detection Using Image Processing and Machine Learning HealthCare," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), 2018, pp. 1-5, doi: 10.1109/ICCTCT.2018.8551008

[10] R. Tekade and K. Rajeswari, "Lung Cancer Detection and Classification Using Deep Learning," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697352

[11] M. Vas and A. Dessai, "Lung cancer detection system using lung CT image processing," 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), 2017, pp. 1-5, doi: 10.1109/ICCUBEA.2017.8463851.

[12] S. Babade, S. Chaudhari, S. Jadhav and A. Nichal, "A review paper on computer aided system for lung cancer detection," 2017 International Conference on Big Data, IoT and Data Science (BID), 2017, pp. 97-102, doi: 10.1109/BID.2017.8336580.

[13] S. Kalaiyani, P. Chatterjee, S. Juval and R. Gupta, "Lung cancer detection using digital image processing and artificial neural networks," 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA), 2017, pp. 100-103, doi: 10.1109/ICECA.2017.8212773.

[14] Dhaware BU, and Pise AC. (2016) “Lung Cancer Detection Using Bayesain Classifier and FCM Segmentation.” IEEE, International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT): 170–174.

[15] M. B. A. Miah and M. A. Yousof, "Detection of lung cancer from CT image using image processing and neural network," 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEIIC), 2015, pp. 1-6, doi: 10.1109/ICEEIIC.2015.7307530.

[16] M. B. A. Miah and M. A. Yousof, "Detection of lung cancer from CT image using image processing and neural network," 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEIIC), 2015, pp. 1-6, doi: 10.1109/ICEEIIC.2015.7307530.

[17] A. Athura and R. S. D. Wahidabahu, "Lung tumor detection and diagnosis in CT scan images," 2013 International Conference on Communication and Signal Processing, 2013, pp. 1108-1112, doi: 10.1109/iccsp.2013.6577228.

[18] A. Chaudhury and S. S. Singh, "Lung Cancer Detection on CT Images by Using Image Processing," 2012 International Conference on Computing Sciences, 2012, pp. 142-146, doi: 10.1109/ICCSS.2012.43.

[19] W. Wang and S. Wu, "A Study on Lung Cancer Detection by Image Processing," 2006 International Conference on Communications, Circuits and Systems, 2006, pp. 371-374, doi: 10.1109/ICCCAS.2006.284656.