International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Machine Learning Techniques for Prediction of Lung Cancer

Nikita Banerjee, Subhalaxmi Das

Abstract*: Lung cancer has been one of the deadliest diseases in today's decades. It has become one of the causes of death in both man and woman. There are various reasons for which lung cancer occurs but classification of tumor and predicting it in the right stage is the most important part. This paper focused on the numerous approaches has been derived for lung cancer detection from different literature survey to advance the ability of detection of cancer. Digital image processing and data mining both are equally important because for prediction either image dataset or statistical dataset is used so for pre-processing the image dataset digital image processing is applied for statistical dataset data mining is applied. After pre-processing, segmentation and feature extraction we apply various machine learning algorithm for the prediction of lung cancer. So first we have provided a sketch of Machine learning and then various fields like in image data or statistical data where machine learning has been used for classification. Once the classification is done confusion matrix is generated for calculating accuracy, sensitivity, precision, these method is used to measure the rate of accuracy of the proposed model.

Keywords: Lung Cancer, Machine learning and its technique, Digital image processing

I. INTRODUCTION

The rapid growth of machine learning is very interesting for many people due to its numerous applications in various areas like it can be used for fraud detection, computer vision, bioinformatics, medical image diagnosis etc. This is used for prediction of cancer based on the medical reports like CT scan, X-Ray, and MRI etc, and has been proven that due to various machine learning technique it has become easier for the doctor to predict disease at right stage. Cancer is a leading cause of death globally and by 2018 it has been estimated as 9.8 million deaths and this estimation has been provided by world health organization, and the most common cancer is lung cancer, and death rate due to lung cancer is more as compared to other all type of cancer [1].

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Nikita Banerjee*, Department of Computer Science and Engineering, Collage of Engineering and Technology, Bhubaneswar, India. E-mail: nikitabanerjee1994@gmail.com
Subhalaxmi Das, Department of Computer Science and Engineering, Collage of Engineering and Technology, Bhubaneswar, India. E-mail: sdascse@cet.edu.in

Lung cancer is one of the leading causes of cancer death in both men and women [2]. There are various reason for lung cancer like smoking, explorer to radon gas etc but it is not necessary that the person who smoke will only suffer from lung cancer, it can also occur due to secondhand smoking. The treatment therapy monitoring and the lung nodule analysis by using the computed tomography (CT) medical images that are having useful strategies to diagnosis the lung cancer early and also to monitor the severity [3].
This paper consist of various machine learning techniques used for the prediction of cancer in both image data that is CT scan report through which we can predict the location of tumor or the size of tumor and CSV file which contain the data like age, gender smoking rate etc.
Paper has been dived into five sections. Section 1 consist of Enabling Terminology , section 2 Machine learning, Section 3 machine learning algorithm used for prediction, Section 4 and 5 consist of comparison Section 6 consist of discussion followed by conclusion and future scope.

II. ENABLING TERMINOLOGY

Pre-processing means cleaning the data so that it can be noise free and it would yield more accuracy. As cancer dataset can be an image data or a numerical data which will be in CSV (Comma separated values) format, and both the dataset has different process for pre-processing for image data we can used digital image processing and for clinical data we can use data mining technique. And after pre processing of data we apply machine learning for classification of the class and calculate accuracy.

A. Image Pre-Processing Using Digital Image Processing

Digital image processing is the technique where we can manipulate or perform some action in order to extract some useful information from the image. It starts from image pre-processing where we enhance the image by using various technique like histogram process, log transformation, etc then followed by image restoration is applied on the enhanced image by adding some noise like Gaussian noise, salt and pepper noise and based on the noise individual noise we add filter to remove the noise filter like mean filter, median filter etc, noise is added in image to get more clear picture. Once the noise is removed color conversion is adapted to convert the image from red, green, blue (RGB) to grey level or from RGB to HSV (hue, saturation, value).After the completion of image conversation image segmentation is enforced, the work of image segmentation is to segment the image into constituent parts, there are various techniques for image segmentation like edge detection, point detection, region based detection etc.

Image segmentation is very important in digital image processing because it keeps only that part which is needed. After image segmentation is executed it is proceed by feature extraction so feature extraction can be defined as the process by which we can reduce the dimensionality by which a set of the raw data is reduced to more manageable group for process there are various process of feature extraction like based on region, based on texture etc once the feature are extracted we can classify the data using machine learning technique.

B. Data Pre-Processing Using Data Mining

Data mining is a process of extracting meaningful data from a raw dataset. The different stages of data mining are data cleaning where we can insert some attribute to the missing values, or remove the duplicate value, then data integration which can be define as heterogeneous data from multiple source combined in a common source that is data warehouse. After data integration data we transform the data into appropriate form required by mining process this step is called data transformation, it includes binning, aggregation, normalization, clustering. Next it is followed by data reduction where we reduce the dataset into smaller volume with respect to maintain the integrity of original dataset, we can reduce the high dimensional data using principle component analysis and by numerosity reduction using histogram, sampling and clustering. The last step for pre-processing is data discretization which a part of data reduction and it is important for numerical data, it involve following methods binning, histogram, entropy based and clustering. After pre-processing feature extraction or feature selection can be done by using any learning algorithm like K nearest neighbor or K mean then on the basis of feature Machine learning algorithm can be applied.

III. MACHINE LEARNING

Machine learning is a subset of Artificial Intelligence the main difference between artificial intelligence and machine learning is that artificial intelligence allows the computer to think like human where as machine learning is a process where system learns from its past experience. In machine learning a set of data is trained using different machine learning methods which is called as training dataset. Once the training is complete then the machine is feed with new dataset which is called testing data set. Now machine can predict the object based on its training data. Machine learning is categorized into three categories supervised, unsupervised and reinforcement learning. Fig1 demonstrates the basic machine learning architecture.

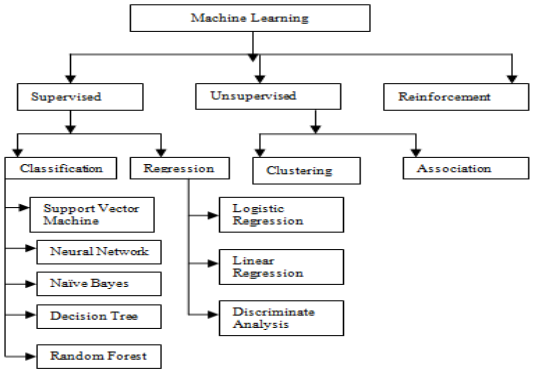


Fig1. Architecture of Machine Learning

Learning is categorized into three categories:

1) Supervised learning: Supervised learning is a learning in which we train the dataset which are already marked with correct answer (labeled data) after training we provide a new data that is test data set so that supervised approach can analysis the testing data set and based on the training data set it can provide correct result. Supervised learning is classified into two categories Classification (when the output variables are categorical (or discrete)) and Regression (when the output variable is continuous (or numerical)).

2) Unsupervised Learning: Unsupervised learning is a training of machine where machine learn by itself that is no information is given to the machine or we can say the data are not labeled so it allow the algorithm to act without any guidelines so it learn by the taken the similarities, patter and difference between two classes form a group for example a student learning itself based on its experience. Under supervised learning clustering in one of the technique, clustering can be defined as forming groups based on their types, and second is association which describe how thing are related to each other.

3) Reinforcement learning: It is based on reward and environment, its objective is to give best solution on the bases of maximum reward. The environment first send a state to the agent based on its knowledge to take the action, once the action take place environment send a pair of next state and reward back to the agent. The agent updates its knowledge with the reward. This continues till it collect maximum reward or we can say the algorithm is properly trained. Here the environment refers to object and agent refers to reinforcement learning algorithm. This concept is used in game theory.

IV. MACHINE LEARNING ALGORITHM USED FOR PREDICTION

Only pre-processing the data is not sufficiency because it need to be classify whether the patient is suffering from cancer or not,

what is its survival rate, if the patient is suffering from tumor then tumor is in which stage so that based on all this question it is very much important to have accurate prediction so that based on the prediction treatment can be given to the patient. For predicting the dataset various kind of machine learning techniques are used like support vector machine, neural network, decision tree etc, all the algorithm have different approaches of prediction and it is also possible that one algorithm can give better accuracy over another algorithm. Some of the famous machine learning algorithms used for predicting lung cancer is:

A. Support Vector Machine

Support Vector machine is a supervised learning method which fall under classification technique. Here object or the labeled data which are plot on n- dimensional space and then they are separated by hyper plane which is also called as decision boundary, and the objects are called as support vectors.

For a data are arranged in linear manner then in that case it can be separated by using hyper plane and if the data are arranged nonlinearly then for partitioning two classes we have to use kernel function. The type of kernel function are linear, nonlinear, polynomial, radial basis function and sigmoid.
Diagram given in fig 2 [4] shows the work of SVM by separating two classes that is Class A is circle and Class B is triangle by the use of hyper plane

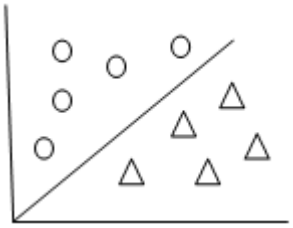


Fig2. Support Vector Machine

There are four scenarios for selecting a best hyper plane:
Scenario 1: Suppose we have three hyper planes A, B and C and two support vectors that are circle and triangle as shown in fig 3 [4]. Now we have to identify the correct hyper plane to classify the support vectors.

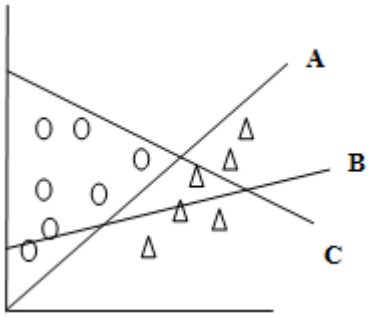


Fig3. Scenario 1 of SVM

So select the hyper plane which separate both the support vector properly, in this case we can see that hyper plane A separate both the classes properly.
Scenario 2: When we have three hyper planes A, B and C as shown in fig 4 [4] and all are divided properly.

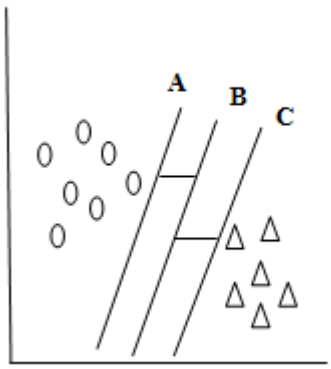


Fig4. Scenario 2 of SVM

The nearest distance to the hyper plane is called margin so based on the margin distance we will decide. So it can be seen that B segregate both the class properly.
Scenario 3: In presence of outlier (the object which get diverts from its group) as shown in fig 5 [4]

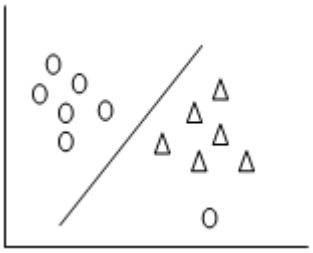


Fig5. Scenario 3 of SVM

In this case we can see that hyper plane A separate both the classes proper but one circle is present in the triangle class so that particular circle is outlier and it can be neglected.
Scenario 4: When the Object are non linearly arranged as shown in fig 6 [4].

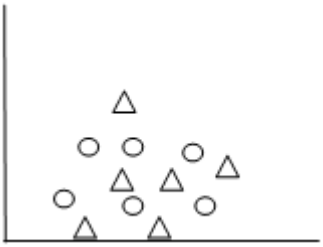


Fig6. Scenario 5 of SVM

In this case we cannot separate the two objects with a hyper plane, So for this type of problem we use kernel which takes as input low dimensional (Not separate able) feature space and give output high dimensional feature space and it is usually plot in 3D space as demonstrated in fig 7 [4].

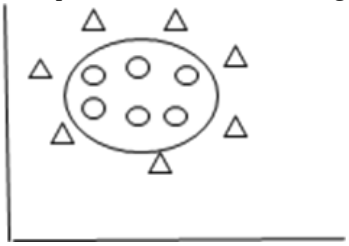


Fig7. Scenario 5 after using kernel

Based on the scenario application on SVM for the prediction of lungs cancer are:

Pradeep K R et al, [5] has used SVM with linear kernel function is used for classification of lung cancer data on two classes namely “Less survivability” as well as “More survivability” with labels as “Less” and “More” respectively. Based on classification precision rate prediction with an online decision support platform can aid doctors to provide patient-specific and evidence-based treatment that would benefit the patient.
Olusayo D. FENWA et al, [6] has used SVM with kernel function linear and RBF (Radial Basis Function) for classification of images into two classes namely “Idiopathic Pulmonary Diseases” and “Chronic Obstructive Pulmonary Disease”. The labels for these classes are using “1” and “2” for “Normal” and “Abnormal” respectively, based on classification confusion matrix is created to show classification and misclassification.

Suren Makaju et al, [7] have applied SVM to detect whether the tumour is malignant or benign. After extracting the feature, those features were used to train SVM. Based on training classification is done and after prediction on the predicted output scatter plot is built.

S.Sivakumar et al, [8] basically three types of SVM kernel function that is linear, polynomial and RBF has been used for prediction. Kernel function is used when the data set are arranged in non linear way. As the image were non linear so three type of kernel has been applied and based on the output generated by different kernel comparison was made.

Şaban Öztürk et al, [9] Various feature extraction has been used like GLCM, LBP, LBGLCM, GLRLM, SFTA using this technique feature matrix is generated on the feature matrix SVM has be applied for the predicting the performance of the feature extraction algorithm.

B. Neural Network

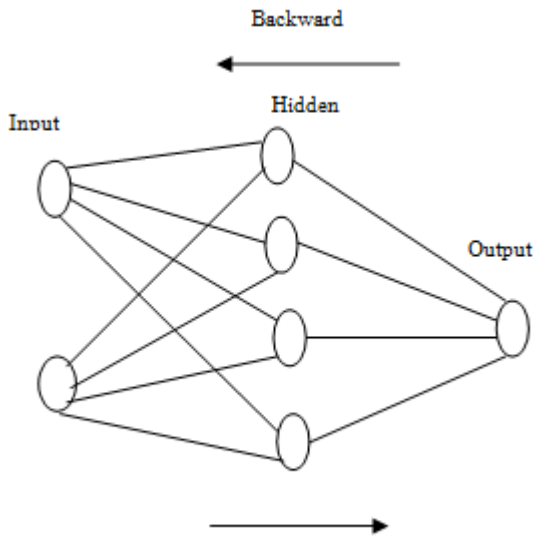
Neural network has been adopted from biological neurons and it is efficient for modeling large and complex problem. Advantage of neural network includes their high tolerance of noise data and ability to classify pattern even if they are not trained. Neural network is not only used for solving classification problem but also used for solve regression problem a neural network has three layer input layer, hidden layer, output layer.

Input layer receives the input pattern and forward it to one or more hidden layer, purpose of hidden layer is to perform computations on the weighted inputs and produce net input which is then applied with activation functions to produce the actual output. The hidden layer is connected to output layer, output layer receives input from hidden layer and in result it predicts the actual value. There is various type of neural network like feed forward neural network, recurrent neural network, convolution neural network etc.

Feed forward neural network is a simple neural network where the information moves from input layer to output layer in one direction and there is no loop in this network. As there is no formation of cycle it becomes a difficult to rectify if any error has occur in its previous state. To avoid this drawback Backpropagation algorithm is used. Backpropagation can be used as both forward pass and backward pass. Architecture of artificial neural network is demonstrated in fig 8.

Convolution network is an advanced neural network. CNNs belong to feed forward neural networks where a signal flows through the network without forming cycles or loops, in a typical CNN model, the main functional layers include convolution layer, pooling layer, fully connected layer.

Convolution layer consist of sets of filter which are applied to the input images. The objective of this layer is to extract high level feature as well as low level feature from the input in short it is used for dimension reduction. Pooling layer reduce the number of parameter of the convoluted features. There are two type of pooling max pooling which perform noise suppressant and average pooling perform dimension reduction as a noise suppressing mechanism. Fully connected input layer (flatter) convert the matrix into vector and feed it into a fully connected layer. The first fully connected layer take the input from the feature analysis and applies weight to predict the connected label and then the fully connected output layer gives the final probability for each label.



8. Architecture of ral Network

Use of neural network in Lung cancer prediction:

Akshay Jadhav et al, [10] ANN has been trained and tested using learning database to perform pattern classification then the artificial neural network classifies the uploaded images as cancerous or non-cancerous using Backpropagation Algorithm. Not only it classify whether the tumour is cancerous or not but also it display the range of the tumour.

Olusayo D. FENWA et al, [6] the artificial neural network extracts features such as texture and roughness from the images and performs accurate classification sequence, together with the train sets. Here Backpropagation algorithm has been used for training multilayer artificial neural network.

Jinsa Kuruvilla et al, [11] feed forward neural and back propagation with various network training function has been used for classification. Basically Backpropagation algorithm is trained with various networks training function like gradient descent Backpropagation(traingd), gradient descent with variable learning rate(traingda), Gradient descent with momentum (traingdm), Gradient descent with variable learning rate and momentum(traingdx),

Resilient back propagation (trainrp), Conjugate Gradient Algorithms (traincgf, traincgp, traincgb, trainscg), QuasiNewton BFGS (trainbfg), One Step Secant Algorithm (trainoss),Levenberg–Marquardt (trainlm) and Automated Regulariza-tion (trainbr).

Xin-Yu Jin et al, [12] The CNN model is designed as an 8-layer model. The 1st layer is the input layer, 2nd, 4th, 6th layers are designed as convolution layers, 3rd, 5th, 7th layers are used for pooling and the 8th layer is the softmax classification layer. Softmax is the activation function used for multi classification and it is used to determine the probability of multiple classes at once and softmax is the last layer of neural network so it is important that number of node in softmax should be same as output layer.

G. Kasinathan, S. Jayakumar and A.H. Gandomi et al, [13] here gradient value is calculated with respect to the parameters of CNN model that are used for gradient based optimization.

And author has also proposed Enhanced CNN with AlexNet. AlexNet is a Convolution neural network designed by Alex Krizhevsky. AlexNet contained eight layers, the first five are convolution layers, some of them followed by max pooling layers, and the last three were fully connected layers. It used the non-saturating ReLU activation function, which showed improved training performance over other activation function.

D. Palani et al, [3] here CNN has been used to classify the image as cancerous or non cancerous based on the feature extracted from hybrid association rule and decision tree.

C. Decision Tree

Decision tree is another supervised learning algorithm which comes under classification technique. Decision tree is a tree like structure, where each internal node denote a test sets branch represent the outcomes of test sets and leaf node contain the outcome of class label. It has three measuring parameters that is information gain, gain ratio and gini index. Information gain can be calculated based on the entropy (the measure of the amount of uncertainty in the data set), hence information gain can be defined as the difference between original information required (entropy of attribute before splitting) and new requirement (entropy of attribute after splitting). Based on the highest entropy that is before splitting the attribute is taken as the root node and after measuring the information gain attribute with highest information gain is selected as splitting attribute. Information gain comes under ID3 (Iterative Dichotomies 3) algorithm.

Gain ratio is the modification of information gain which is used to reduce the bias it comes under C4.5.Gini Index measure the how often the randomly chosen attributes are incorrectly identify it comes under CART(classification and regression tree).

Use of decision tree in lung cancer prediction:

D. Palani et al, [3] Entropy has been used to make decision whether the tumor is normal or abnormal. The information gain value is calculated for the process decision making by using all the input image features such as transition regions, range, morphological region, pixels as items sets. These all features with the highest normalized information gain value have been selected for making decision. The feature image with highest information gain becomes the root node, then based on that particular feature it will select its child node and it will continue till the training image set is not empty.

Haofan Yang et al, [14] decision tree has been applied on clinical dataset which consist of attributes like gender, age, reformed smoker, pack-year, race, length of smoking. Raw data has been inputted one by one and one attribute is selected at a time for descriptive value dataset. If the selected attribute contain valid dataset then it is stored in valid value dataset otherwise it is removed. This process continues till all descriptive value is used. So here decision tree has been used for pre-processing of clinical dataset.

Pradeep K R et al, [5] first attribute like gender, age, smoker, stage of tumor, location of tumor, whether the patient is diabetic or not collected timing (survival time in year) from hospital and NCCTG lung cancer data. Basically decision tree has been to predict the survivability of the patient using entropy and information gain (C4.5) and based on the survival rate suitable treatment can be given. C4.5 use IF-Then rule. Here it has been used as

IF Sex=MALE AND no years = (more than)>1 years AND prediction = More survivability
THEN Diagnosis=Further treatment recommended

C.M. Lynch et al, [15] decision tree is created in tree like structure, the technique is to divide the dataset into smaller part while simultaneously building a decision tree associated with these data that eventually ends at a single leaf or end node where the data subset cannot be viably split further. Here the final designation of the subset (in this case a survival time) is decided. For regression decision trees there are more classifications than in a typical classification decision tree that makes the outcomes near continuous (numeric vs. discrete outputs).In this case repeated cross validation is implemented to select a maximum tree depth of 10, a "minsplit" (minimum number of observations in a node to attempt a split) as 200, and a complexity parameter of 0.1 (described as: "any split that does not decrease the overall lack of fit by a factor of this parameter is not attempted" as parameters for the training task. Default values were used; a grid search in caret did not produce any discernibly different results in the output tree. The decision tree automatically pruned to a very short three-level depth and could not be coerced to be much more complex or better scoring despite parameter changes, reflecting the simplistic nature of this technique.

D. Random Forest

Random forest is one the ensemble method. Ensemble method is a technique where it combine the series of K learned model (or base classifier) that focused on creating an improved prediction model. It use decision tree classifier in a randomize way. A Training dataset of h tuple is given and for each iteration j a training set, H_i of h tuple is sampled with replacement from H(it work like a bootstrap).Which doing this some tuple may occur more than one tin and it may also happen some of the tuple does not occur at all. Once the bootstrap dataset is formed from the original dataset it used to make the decision tree. Let g be the number of attribute to determine the split at each node, where g is smaller than the available attribute and to build a decision tree classifier D_i randomly select node from g attribute as a candidate for splitting at the node. Cart is used to grow the tree and over fitting is not a issue in random forest. Its accuracy depends on the strength of the individual classifier.

Use of random forest in lung cancer prediction:

Y. Sumathipala et al, [16] as random forest is one of the ensemble methods to estimate the performance of the performance of the model on new data, each dataset are trained thousand times on bootstrapped cohort's samples from training dataset. The accuracy of each of the1000 models was measured as the AUC of the Out-of-Bag (OOB) samples.

C.M. Lynch et al, [15] one of the main issues with decision tree is over fitting, to avoid this situation random forest has been used. For the Random Forest, the number of trees was set to 500, manually selected as much for processing time as for the observation that performance had stagnated below this number. The "mtry" variable was selected via repeated cross-fold validations (using caret) for integer values between 1 and 10.

Minimum node size was set to 50, which was the default, after experimentation with the number failed to have any impact (other than over and under-fitting with extreme values below approximately 5 and above approximately 1000 [10% of the dataset], which were therefore options that were ignored).
Hawkins et al, [17] as due to feature extraction large amount of feature are high dimensional due to which over fitting issue has arise to avoid this situation random forest has been used.

E. Naïve Bayes

Naïve bayes classification is also called as Bayesian classifier is a probabilistic model which is based on bayes' theorem. It can predict class membership probabilities such as the probability that the given tuple belong to a particular class. It main task is to find the maximum probability from a given set of tuple and this maximum probability is called as maximum posteriori hypothesis and this can be calculated by using bayes theorem.
So the formula for bayes theorems is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

Where P(A|B) is the probability of target given prediction, P(B|A) is likelihood, P(A) is prior probability of class, P(B) is prior probability of prediction.
Conversely, Naive Bayesian is a simple probabilistic classifier based on Bayesian theorem with the (naive) independence assumption. Based on that rule, using the joint probabilities of sample observations and classes, the algorithm attempts to estimate the conditional probabilities of classes given an observation. Despite its simplicity, the Naive Bayes classifier is known to be a robust method, which shows on average good performance in terms of classification accuracy, also when the independence assumption does not hold. [2]

Use of naive bayes in lung cancer prediction:

Pradeep K R et al, [5] here naive bayes has been used to check the patient survival rate so that based on that treatment can be given to the patients. The query result (QR), QR, can be represented as:
QR = max (P (Treatment | Evidence))
Evidence: {Survival = 'Less survivability', Predictor Variable 1…Predictor Variable n}
Evidence: {Survival = 'More survivability', Predictor Variable 1…. Predictor Variable n}
QR = P (Survival='Less survivability' | Evidence)
QR = P (Survival='More survivability' | Evidence)
Evidence: {Treatment='Recommended ', Rule Variable1, Rule Variable n}
Y. Hemalatha et al, [18] on the statistical data naive bayes has been used to decide whether the person is suffering from cancer or not. As naive bayes is a probabilistic model so first calculate the probability of occurrence (yes) or not then calculate the probability of not occurrence (no) based on that it can be predicted whether it is cancer or not.

F. Logistic Regression

Logistic regression is used when the dependent variable are in binary in nature, it is widely used in classification problem by using probability. The value of the dependent variable ranges from 0 to 1 and it can be represented by an equation that is probability of event occurrence divided by probability of not event occurrence.
Logistic regression does not require linear relationship because it applies a non linear log transformation to predict odd ratio. At the same time it is important that the independent variable should not be correlated with each other.

Use of logistic regression in lung cancer prediction:

Y. Sumathipala et al, [16] to estimate the performance of the performance of the model on new data, each dataset are trained thousand times on bootstrapped cohort's samples from training dataset. The accuracy of each of the1000 models was measured as the AUC of the Out-of-Bag (OOB) samples.
Animesh Hazra et al, [19] logistic regression has been used to predict the death rate due to lung cancer. As the dataset has only two attribute either yes or no for which logistic regression can work more efficiently than other algorithms

G. Linear Regression

Linear regression is the most widely used regression technique as it is a regression method so its depended variable is continuous and independent variables can be continuous or discrete and the nature of the regression line or hyper plane is linear.
In a linear regression it show a relation between dependent variable and one or more independent variable using best fit straight line. It is used to predict the value of target variable based on predicted variable. Linear regression is sensitive to outlier.

Use of linear regression in lung cancer prediction:

C.M. Lynch et al, [15] The goal of the method is to fit a straight line to a set of data points using a series of coefficients multiplied to each input, like a weighting function, and an intercept. The weights are decided within the linear regression function in a way to minimize the mean error. These weight coefficients multiplied by the respective inputs, plus an intercept, give a general function for the outcome, patient survival time. In this way linear regression is easy to understand and quick to implement, even on larger datasets.

H. Linear Discriminate Analysis

discriminate analysis (LDA) is used to overcome the drawback of logistic regression as we can see that logistic regression is limited to only two class classification problems. If there are more than two classes on that case we use LDA. LDA is a statistical property used for dimensional reduction (project the high dimension space to low dimension space) technique in supervised learning classifier. It has two criteria that are maximize the mean between two classes and minimize the variance within each class. In medical field it helps to predict the patient disease based on various conditions given. It is also used in face reorganization and customer identification.

Use of LDA in lung cancer prediction:

Taruna Aggarwal et al, [20] LDA are used for classification of module and normal anatomy structure by analyzing geometrical feature distribution.

Geometrical features include area, perimeter, roundness, solidity, eccentricity, equivalent diameter, centroid and convex-area.

Şaban Öztürk et al, [9] LDA has been used for classification of histopathology images tumour detection using texture based feature extraction.

V. COMPARISONS

Based on the literature survey comparison has been shown in table 1. Comparison has been made on the basis of pre-processing, segmentation, feature extraction, classification method and accuracy.

Table 1 Survey of Machine Learning Algorithm

Serial No	Paper Name	Author	Pre-processing method	Segmentation Technique	Feature Extraction Technique	Classification Algorithm	Accuracy
[1]	An IoT Based Predictive Modeling for Predicting Lung Cancer Using Fuzzy Cluster Based Segmentation and Classification(2018)	D. Palani et al,	1. RGB to HSV 2.Transation region identification 3. Morphological thinning and cleaning Operation	Fuzzy C mean clustering	Local variance based, Transition regions based	Hybrid temporal association rule with decision tree classification with Convolution neural network	99.54% generated using hybrid model
[2]	Data mining in lung cancer pathologic staging diagnosis: Correlation between clinical and pathology information (2015)	Haofan Yang et al,	Duplicate data is deleted and saved in sql database	No segmentation	Rule extraction using support and confidence	Decision tree and association rule mining	Accuracy calculated at each rule extraction
[3]	Automated 3-D lung tumor detection and classification by an active contour model and CNN classifier (2019)	Gopi Kasinathan et al,	Gaussian filter	Active Contour Model	Fuzzy threshold method	Enhanced convolution neural network	Based on each nodule accuracy is calculated
[4]	Detection of Lung Cancer Using Backpropagation Neural Networks and Genetic Algorithm (2016)	Akshay Jadhav et al,	RGB to gray scale	No segmentation	Genetic Algorithm	Backpropagation Artificial Neural Network	Stages of cancer is shown
[5]	Lung Cancer Survivability Prediction based on Performance Using Classification Techniques of Support Vector Machines, C4.5 and Naive Bayes Algorithms for Healthcare Analytics (2018)	Pradeep K R et al,	As it was no image data so no preprocessing technique is used	As it was no image data so segmentation is not required	As it was no image data so feature extraction is not done	SVM, C4.5, Naïve bayes	C4.5 gives more accuracy
[6]	Classification Of Cancer Of the lung using SVM and ANN (2015)	Olusayo D. FENWA et al,	Gray scale conversion	No segmentation	Texture feature	ANN, SVM	ANN gives more accuracy than SVM
[7]	Novel Approach for Lung Image Segmentation through Enhanced Fuzzy C-Means Algorithm(2017)	C.Rangasw-amy et al,	Not mentioned	Enhanced Fuzzy C Mean clustering	Not done	Not done	-
[8]	Lung Cancer Detection using CT Scan Images (2018)	Suren Makaju et al,	Median filter, Gaussian filter, salt and pepper noise, speckle noise	Watershed segmentation	Region based	Support vector machine	86.6% accuracy generated by SVM

[9]	Lung Nodule Detection Using Fuzzy Clustering and Support Vector Machines(2013)	S.Sivakumar et al,	Not mentioned	Fuzzy C-Means (FCM), Fuzzy-Possibilistic C-Means	Texture feature	Support vector machine kernel has been used that are linear, polynomial, RBF	RBF gives more accuracy
[10]	Lung cancer classification using neural networks for CT images(2014)	Jinsa Kuruvilla et al,	Gray scale to binary image	Morphological Operation	Mean, standard deviation, skewness and kurtosis	ANN	Accuracy is calculated based on different type of ANN
[11]	Machine learning to predict lung nodule biopsy method using CT image features: A pilot study(2019)	Y. Sumathipala et al,	Not mentioned	Otsu's threshold and mathematical morphology operations	Texture based, Semantic feature extraction	Logistic regression, Random forest	Logistic regression gives more accuracy.
[12]	Prediction of lung cancer patient survival via supervised machine learning classification techniques(2017)	C.M. Lynch et al,	Not mentioned	Not mentioned	Not mentioned	Linear regression, Decision trees, Gradient boosting machines, Support vector machines	linear regression provide more accuracy
[13]	Feature Extraction and LDA based Classification of Lung Nodules in Chest CT scan Images(2015)	Taruna Aggarwal et al,	Median filter	Optimal threshold technique	Region based, GLCM	LDA	It provides 84% accuracy.
[14]	Pulmonary nodule detection based on CT images using Convolution neural network(2016)	Xin-Yu Jin et al,	Circular filter	Optimal threshold technique	ROI extraction	CNN	It provides 84.6% accuracy.
[15]	Application of Feature Extraction and Classification Methods for Histopathological Image using GLCM, LBP, LBGLCM, GLRLM And SFTA(2018)	Şaban Öztürk et al,	2D Gaussian smoothing filter, 2D median filter	Egde detection	GLCM, LBP, LBGLCM, GLRLM and SFTA	SVM, KNN, LDA and Boosted Tree algorithms.	SVM and Boosted Tree algorithms produced the highest success.
[16]	Predicting Malignant Nodules from Screening CT Scans(2016)	Hawkins et al,	Noise has been removed	Ensemble Segmentation approach	High dimension reduction	Random Forest	More than 90%
[17]	Predicting Lung Cancer Survivability using SVM and Logistic Regression Algorithms(2017)	Animesh Hazra et al,	Fill the missing value, normalization	Splitting dataset after feature selection	Feature selection using Pearson correlation coefficient (PCC)	SVM and Logistic regression	Accuracy of logistic regression is more than svm
[18]	Prediction of Lung Cancer Symptoms Using Naïve Bayes and J48 Classification Techniques(2019)	Y. Hemalatha et al,	Fill the missing value, delete duplicate data		Calculate mean and standard deviation	Naive Bayes, J48	Naive bayes gives more accuracy

VI. DISCUSSION

This paper consist of overview of proposed work by different researcher from 2013 to 2019 using machine learning algorithm either on digital image or through statistical approach. Most of the researcher has taken data set from the Lung Image Database Consortium image collection (LIDC-IDRI), The Cancer Genome Atlas (TCGA), SEER Database.

- LIDC- It consists of diagnostic and lung cancer screening thoracic computed tomography (CT) scans with marked-up annotated lesions [21]. It is collaborated by seven academic center and eight medical image companies to create the dataset. It consists of 1018 cases and the nodules are with diameter larger than 3 mm.

- TCGA- The Cancer Genome Atlas (TCGA), a program, molecularly characterized over 20,000 primary cancers and matched normal samples spanning 33 cancer types. This joint effort between the National Cancer Institute and the National Human Genome Research Institute began in 2006, bringing together researchers from diverse disciplines and multiple institutions [22].
- SEER Database- The surveillance, Epidemiology, and End Result abbreviation (SEER) program provide information on cancer statistics like gender, age smoking rate, tumor size, treatment record etc [23].

VII. CONCLUSION

The main focus of this paper is to show various machine learning algorithm used for the prediction of lung cancer at early stage. Image conversation, Gaussian filter, missing value handling has applied for image pre-processing. Survey has been carried out using two type of dataset first one is image dataset and another one is statistical dataset. In image dataset after pre-processing segmentation has been applied various method used for segmentation are watershed segmentation, active contour method, edge detection, fuzzy c mean cluster etc. For extracting the information from the dataset feature extraction has been done using gray level co-occurrence matrix, ROI extraction, region based. Algorithm like ANN, SVM, Logistic Regression, CNN, Decision Tree, Random Forest, Linear Discriminate Analysis, and Linear Regression has been used for classification stage. Based on the classification it can be predicted that neural network and support vector machine is generating more accuracy. In future other machine learning techniques along with the mentioned technique can be used for building a model which would yield more accuracy for the prediction of not only lung cancer but other cancer also.

REFERENCES

1. Cancer fact-sheet(www.who.int/en/news-room/fact-sheets/detail/cancer)
2. Krishnaiah, V., G. Narsimha, and Dr N. Subhash Chandra. "Diagnosis of lung cancer prediction system using data mining classification techniques." *International Journal of Computer Science and Information Technologies* 4.1 (2013): 39-45.
3. Palani, D., and K. Venkatalakshmi. "An IoT based predictive modelling for predicting lung cancer using fuzzy cluster based segmentation and classification." *Journal of medical systems* 43.2 (2019): 21.
4. SVM (https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/)
5. Pradeep, K. R., and N. C. Naveen. "Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4. 5 and Naive Bayes algorithms for healthcare analytics." *Procedia computer science* 132 (2018): 412-420.
6. Fenwa, Olusayo D., Funmilola A. Ajala, and A. Adigun. "Classification of cancer of the lungs using SVM and ANN." *Int. J. Comput. Technol.* 15.1 (2016): 6418-6426.
7. Makaju, Suren, et al. "Lung cancer detection using CT scan images." *Procedia Computer Science* 125 (2018): 107-114.
8. Sivakumar, S., and C. Chandrasekar. "Lung nodule detection using fuzzy clustering and support vector machines." *International Journal of Engineering and Technology* 5.1 (2013): 179-185.
9. Öztürk, Şaban, and Bayram Akdemir. "Application of feature extraction and classification methods for histopathological image using GLCM, LBP, LBGLCM, GLRLM and SFTA." *Procedia computer science* 132 (2018): 40-46.
10. D'Cruz, J., A. Jadhav, A. Dighe, V. Chavan, and J. Chaudhari. "Detection of lung cancer using backpropagation neural networks and genetic algorithm." *Comput Technol Appl* 6, no. 5 (2016): 823-827.
11. Kuruvilla, Jinsa, and K. Gunavathi. "Lung cancer classification using neural networks for CT images." *Computer methods and programs in biomedicine* 113.1 (2014): 202-209.
12. Jin, Xin-Yu, Yu-Chen Zhang, and Qi-Liang Jin. "Pulmonary nodule detection based on CT images using convolution neural network." *2016 9th International symposium on computational intelligence and design (ISCID)*. Vol. 1. IEEE, 2016.
13. Kasinathan, Gopi, et al. "Automated 3-D lung tumor detection and classification by an active contour model and CNN classifier." *Expert Systems with Applications* 134 (2019): 112-119.
14. Yang, Haofan, and Yi-Ping Phoebe Chen. "Data mining in lung cancer pathologic staging diagnosis: Correlation between clinical and pathology information." *Expert Systems with Applications* 42.15-16 (2015): 6168-6176.
15. Lynch, Chip M., et al. "Prediction of lung cancer patient survival via supervised machine learning classification techniques." *International journal of medical informatics* 108 (2017): 1-8.
16. Sumathipala, Yohan, et al. "Machine learning to predict lung nodule biopsy method using CT image features: A pilot study." *Computerized Medical Imaging and Graphics* 71 (2019): 1-8.
17. Hawkins, Samuel, et al. "Predicting malignant nodules from screening CT scans." *Journal of Thoracic Oncology* 11.12 (2016): 2120-2128.
18. Y. Hemalatha, Mr. G. AnanthNath.,Prediction of lung cancer symptoms using Naïve Bayes and J48 classification technique., IJSRD - International Journal for Scientific Research & Development| Vol. 7, Issue 01, 2019 | ISSN (online): 2321-0613
19. Animesh Hazra, Nanigopal Bera, Avijit Mandal., Predicting Lung Cancer Survivability using SVM and Logistic Regression Algorithms., *International Journal of Computer Applications (0975 – 8887) Volume 174 – No.2, September 2017*
20. Taruna Aggarwal, Asna Furqan, Kunal Kalra ., Feature Extraction and LDA based Classification of Lung Nodules in Chest CT scan Images., 978-1-4799-8792-4/15/31.00c 2015 IEEE
21. LIDC-IDRI(https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI)
22. TCGA(https://www.cancer.gov/about-cancer)
23. SEER Incidence Data(https://seer.cancer.gov/about/)

AUTHORS PROFILE

Nikita Banerjee received B.Tech in Computer Science and Engineering from Gandhi Institute for Technology, BPUT, Odiaha, India and perusing M.Tech in Computer Science and Engineering from College of Engineering and Technology, Odisha, India. Her special field of interest includes Machine Learning, Digital image process, Data Mining. Her major strength lies in Algorithm, Machine learning, Internet of web Technology.

Subhalaxmi Das received B.Tech in Computer Science and Engineering from Biju Patnaik University of Technology, Odisha, India and M.Tech in computer Science and Engineering from KIIT University, Odisha, India. She is currently working as lecturer in Dept of Computer Science and Engineering, College of Engineering and Techology, a constituent college of Biju Patnaik University of Technology, Odisha, India. She has 10 years of teaching experience. She has published some innovative ideas in International journal and Conference in the area of Spatial Data Mining. Her mojor strength lies in Algorithms, Data mining, Database Systems, Soft Computing.