Lung Cancer Detection using Decision Tree Algorithm

Prashantha G R
Associate Professor
Dept of CSE
JIT,Davanagere

Usha K
Assistant Professor
Dept of CSE
JIT,Davanagere

Zabi U R Rahaman
Dept of CSE
JIT,Davanagere

Syed Shahid
Dept of CSE
JIT,Davanagere

Shashank S R
Dept of CSE
JIT,Davanagere

***Abstract* —The most common cause of lung cancer death in the world is lung cancer, often known as "Lung Cancer." Since a result, rapid recognition, prediction, and diagnosis of lung cancer is crucial, as it helps to increase and automates the potential therapeutic procedure. Because of their effectiveness, pattern recognition techniques have been applied to improve the progression and management of hazardous disorders. Lung cancer has been studied and predicted using machine learning approaches such as KNN, Naive Bayes, Decision Tree, Svm Classifiers (SVM), Logistic Regression, and Neural Network (ANN). The existing circumstances that cause lung cancer, as well as the use of machine learning approaches, are analyzed in this paper, with a concentration on their strengths and limits.**

Keywords:- Diagnosis, prediction, effectiveness, analyze

I. INTRODUCTION

Machine learning is used in a variety of fields, including medical research, robotics, autonomous driving, Supervised learning has preexisting inputs and outputs where Unsupervised learning only takes in input values, whereas supervised learning uses input and output values that have already been established.Predicted judgments for unknown feature vectors are the outputs. This shows that the most essential characteristic for Machine Learning is prediction accuracy (ML). One use of machine learning is cancer. Over 14 million people are diagnosed with cancer each year around the world, with 266 thousand of these being breast cancer diagnoses. At least 140,000 persons would be affected by a 1% increase in forecast accuracy. The goal was to create a machine learning algorithm that was quantitatively superior to previous algorithms for breast cancer and comparable applications. Several databases are available.

All of these datasets are available at the University of California-Irvine Machine Learning Repository. Learning curves, like prediction accuracy, are graphs that show how accuracy improves as more training data is provided. Faster learning reduces the amount of training data needed to get the same prediction accuracy.

Lung cancer is one of the most dangerous diseases on the planet. Every year, lung cancer kills more people than any other cancer, including breast, brain, and prostate cancer. Lung cancer is the leading cause of cancer-related death among people aged 45 to 70.

Lung cancer claims the lives of more people each year than breast, colon, and prostate cancer combined, accounting for over a quarter of all cancer-related deaths.
Many existing technologies are used to detect lung cancer in advanced stages, including computed tomography (CT), chest radiography (x-ray), magnetic resonance imaging (MRI), and sputum cytology (most of these are expensive and time consuming). As a result, a new method for detecting lung cancer in its early stages is critical.

The primary side effects are torment in chest or rib, hack can be constant, dry, with mucus or with blood, respiratory contaminations, brevity of inhale, wheezing, entire body weakness or loss of craving, swollen lymph hubs.
Tumor cells are those cells that create, in spite of the fact that when the body doesn't require them, and. other than as typical old cells, they don’t pass. The malignancy cells present in lung causes lung development illness. Malignant tumors: 14 % of all new malignant growth analyze are primary lung disease. All the blood in the body goes from the heart through the lungs, so the malignant growth can undoubtedly spread to various parts of the body. Benign tumors are a less basic reason for respiratory infection. One model is hematoma. These can pack encompassing tissue, however they are typically asymptomatic.

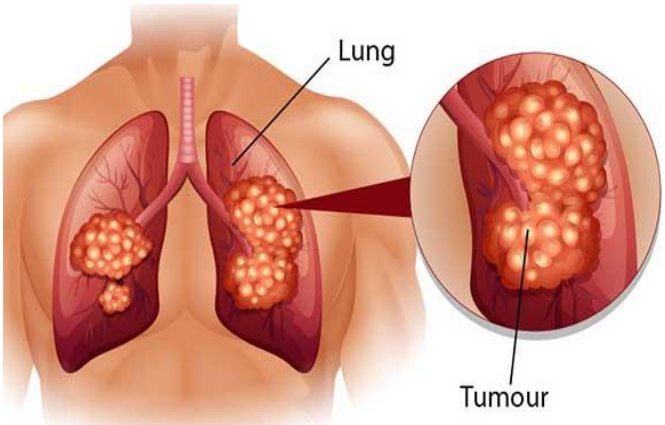

Fig1. Lung affected by tumors

At the point when air enters the nose or mouth, it goes down the trachea, additionally called the windpipe. After this, it arrives at an area called the carina. At the carina, the windpipe parts into two, making two principle stem bronchi. One prompts the left lung and the other to the right lung. From that point, similar to branches on a tree, the channel like bronchi split again into littler bronchi and afterward much littler bronchioles.
This ever-decreasing pipe work eventually terminates in the alveoli. Lung malignant growth spreads when cells sever or from a tumor and travel through the circulatory system the lymphatics to far off locales of the body and develop. These cells disclosure is basic issue for medical specialists.

II. RELATED WORK

The presented engineering, which consists of different channels of 2-D Convolutional networks, for which the production are consolidated using a deeply committed fusion method to induce the greatest classification, was explained by Arnaud A. A. Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J. van Riel, and Mathilde [1], but the morphological variety of knobs is continually more pronounced.

Clustering is explained in depth in the work "Unsupervised Profound Implanting for Clustering Analysis"[2].
Clustering is a critical component of many data-driven application domains, and it's traditionally been seen as a set of separate skills and computations. For clustering, centred learning representations have been used in a moderately limited amount of work. However, misclassification of any image does not provide the expected consequence.

Mario Buty1, Ziyue Xu1, and Mingchen Gao[3] published a paper on computed tomography imaging. It is a common method for identifying and evaluating lung cancer.
Expert qualitative assessments on various parameters characterising a nodule's appearance and form are frequently used in clinical practise to assess the malignancy of lung nodules, although these criteria are primarily subjective and arbitrarily defined.

Our method achieves a Dice-Sorensen Coefficient (DSC) of 63:44 percent without human aid, which is greater than the number (60:46 percent) achieved without intensive supervision. However, it provides less precision in this procedure.

This study proposed an autonomous lung cancer detection system that improves accuracy and yield while also reducing diagnosis time.

The amount of data in MR pictures is too great for human interpretation and analysis. There are four steps to the diagnosing technique. To categorise the Normal and Abnormal, a Probabilistic Neural Network was used.

Smoothing the picture with a Gaussian channel is the most important stage. In both the vertical and horizontal headings, this is followed by determining the image's slope by dealing with the smoothed picture using a convolution action with the Gaussian auxiliary. By recognising strong edges and saving the important weak edges, this technique alleviates concerns associated to edge discontinuities.

Emmanuel Adetiba and Oludayo .O. Olugbara proposed a model about Artifical neural network for lung cancer prediction.

Artificial neural network:
The data structures and utility of neural networks are designed to mimic associative memory. Neural nets learn by building models with known "info" and "results," generating probability weighted correlations between the two, and storing them in the data structure of the net. Thinking models "train" such systems to do tasks without the need for human interaction in the majority of circumstances.

task-explicit rules are being changed. For example, in image recognition, they may figure out how to distinguish photographs with malignant growth by looking at model pictures that have been correctly labelled as "disease" or "no illness," and then applying the findings to recognise tumours in diverse pictures.

As delineated in Figure, a artificial neuron has a great deal of neurotransmitters related with the data sources and each information has a related weight. A sign at input is increased by the weight, the weighted inputs are included, and a linear combination of the weighted inputs is gotten. A bias, which isn't connected with any data, is added to linear combination and a weighted sum is obtained as

$$z = w_{0+}w_1x_{1+}.....+w_nx_n.$$

Rivansyah Suhendra approach a model about Support Vector Machine
It is one of the most productive conventional technique utilized for grouping 'n' number of features. The classification is finished by finding a hyperplane. SVM passes a linear distinguishable hyperplane through a dataset so as to arrange information into two gatherings. Hyperplane is utilized as a separator for any measurement.

Best hyperplane is the one which amplifies the edge. The edge is the separation between the hyperplane and few close points. These nearby points control the hyperplane. This is greatest edge classifier. Maximal edge classifiers helps in expanding the edge of hyperplane. This is best since it sums up the mistake.
Parameshwar R. Hegde proposed a model about binary and multiclass classification for comparison of ML algorithms. Binary classification

In keeping with a classification rule, category is the duty of categorising the pieces of a supports the following into two groups (understanding which bundle each one belongs to). The following are examples of scenarios in which a decision needs to be made on whether someone has an abstract feature, a designated trademark, or a binary classification:

- Medical testing to choose whether a patient has certain disease or not – the arrangement property is the proximity of the ailment.
- A "leave or come behind short" test technique for instance picking if a specific has or has not been met – a go/no go gathering.
- Here, the two classes considered were malignancy and Non disease. Highlights were thought about in three blends; first the color, texture, both color and texture joined. Support vector machine gave the most elevated precision when similar classifiers were contrasted with texture highlights. Later when we joined both the features and tried SVM got the most noteworthy precision pace of 82.26%.

In this proposed strategy, distinguishing proof of the lung malignancy with the assistance of CT pictures is finished. The strategy has a few phases where in it starts with image acquisition, trailed by image preprocessing, segmentation, feature extraction and combination, classification and calculating accuracy using confusion matrix. The examination indicated the precision execution was around 87.8%. The general outcome demonstrated that the classifier effectively arranged the image into two classes. The best classifier was SVM with the exactness of up to 99.92%

III. PROPOSED METHODOLOGY

A. Data Preprocessing

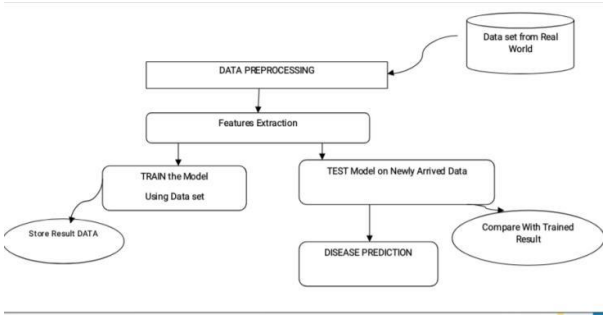

Fig2. Data Preprocessing

Preparation raw data to be used for a machine learning algorithms is the method of data preparation. It's the most crucial part of creating a machine learning model.

When embarking on a machine learning project, we don't always have access to adequate, well-prepared data. It is also required to clean and prepare the data before partaking in any data-related activity. As a result, we employ a pre - processing stage technique.

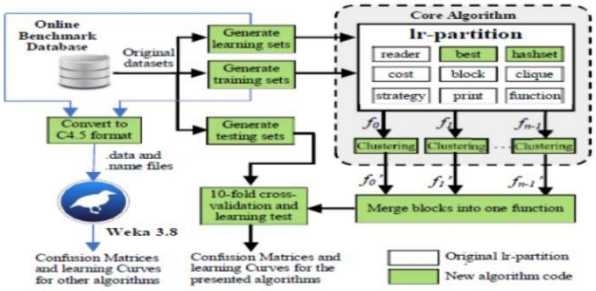

Fig3. Architecture of Lung Cancer Detection

Fig shows the detailed Step By Step flow of watershed segmentation. At the beginning the preprocessed image is given as input to the watershed segmentation. At beginning the flow starts with two parts. In first part the preprocessing of image is done as shown in figure.

In second Stage the preprocessed image is transformed into L*a*b Color space from RGB for which k-mean Clustering is obtained to find Concavity point splitting which is ended with Morphological processing. At the end output of both preprocessing and marker extraction is used to find watershed transform of an image.

There are different operational module in this application.
1. Preprocessing the data: The entire data is used to construct training sets for the primary algorithm and performance testing test sets.
2. The learning and training sets will be put into the original lr-partition technique, which will construct numerous blocks (without completely specified outputs).
3. Clustering: Every one of these blocks will be subjected to the new clustering method, which provide output values for the undisclosed input combinations.
4. Assemble and test: All of the aforementioned criteria will be integrated into a single large, properly documented function. The large function will be validated using test sets collected during preprocessing.
5. For comparison, the very same samples were run though Weka employing C4.5 and the original lrpartition method (without clustering).

In this system, we have proposed a model for an early prediction of cancer disease by using efficient machine learning techniques. The set of tasks that can be carried out in our proposed work is analyzed, designed, implemented and experimented using one of the tools such as Python by modifying the existing machine learning algorithms. These are a collection of data mining machine learning algorithms. The procedure may be used to data sets directly or invoked from our own programming language code. Techniques for data pre-processing, categorization, regression, association rules, and visualisation are all included in the Tools section. MySQL database server is used to meet storage needs. A cancer prediction system based on architectural machine learning techniques was utilised, merging the prediction system with ML technology. In this model, we employed a classification process known as a decision tree.The user must address queries concerning genetic and non-genetic

factors after logging into the cancer forecasting model. The estimation algorithm then calculates a risk value to each question based on the user comments.

In this system, we have proposed a model for an early prediction of cancer disease by using efficient machine learning techniques. The set of tasks that can be carried out in our proposed work is analyzed, designed, implemented and experimented using machine learning algorithms. These have collection of machine learning algorithms for data mining tasks.

A cancer prediction system based on architectural machine learning techniques was utilised, merging the prediction system with ML technology. In this model, we employed a classification process known as a decision tree. The user must respond to comments associated with genetic and non-genetic information after logging into the cancer prediction system. The prediction system generates a risk value to each question depending on the responses of the users. With the projections of the risk level, the estimation algorithm can determine the risk range.

Texture feature: One of the most commonly used techniques to extract textural data of Images is Gray Level Co-occurrence Matrix (GLCM). The GLCM technique gives sensible surface data of a picture that can be acquired just from two pixels. It utilizes statistical methods to look at the surface by thinking about the spatial connection between pixels.

A sequencing figure illustrates the relationships amongst processes and the order in which they occur. It's called a Message Sequence Chart. The interactions across items are portrayed in chronological order using a sequence diagram. It depicts the type of situations objects and classes, and even the messages that should be communicated between them in order for the scenario's functionality to be carried out. Sequence diagrams are commonly associated with use case representations in the system under development's Logical View. Other names for sequence diagrams are event drawings and event scenarios.

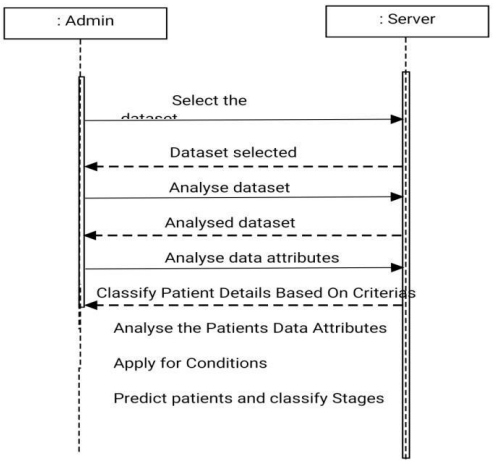

Fig4. Sequence Diagram

D . Machine Learning Algorithms

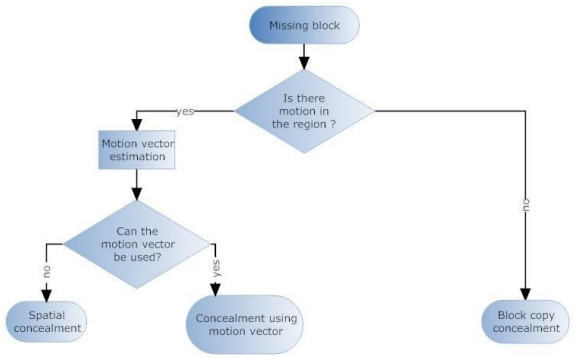

Fig5. Block Diagram of Decision Tree

Algorithm of a Decision Tree:
Decision trees use a lot of approaches to evaluate whether to split a junction into number of sub. With each generation of sub-nodes, the homogeneity of the developed sub-nodes improves. To put this another way, when the target variable changes, the node's purity improves. Based on all important parameters, the decision tree partitions the networks into sub-nodes, then picks the split that generates the most heterogeneous sub-nodes.

The model structure of the fundamental decision tree is the core software environment. The IMAD has a sample length of 1200 and a total quantity of training data of 100. The fuzzy coefficient is 1.35, and the beginning frequency of gathering intelligent Supplemental medical data is what it 's known in the medical field. The primary feature stoichiometric ratio is 0.24, but the weight parameter of the decision tree allocation is 0.25. This investigation applied the following data sets: (1)Mortality, child nutrition, immunizations, and infectious and weakens the immune system diseases are among the approximately 1000 indicators being used by World Health Organization (WHO). (2)CDC WONDER: data on US public health, including environmental data, murder statistics, and demographic data. (3)MIMIC Critical Care Database: The MIMIC Critical Care Database collects approximately 2000 patient ’s demographic and clinical data.

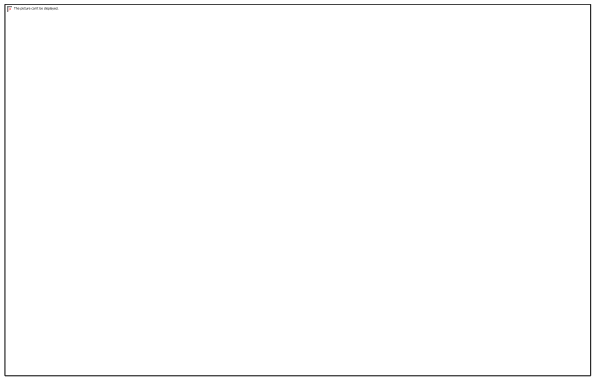

Fig6. The data sequence's original waveform

The IMAD data series was regressed and analysed using the autocorrelation feature decomposition. The IMAD data set was examined using, The model incorporated fuzzy feature reconstruction utilising the wavelet transform reconstruction technique, as well as regression analysis and a sophisticated diagnosing feature value decision tree model. A quadruple exemplifies the statistical distribution of feature weights for intelligent medical auxiliary diagnostic data, where and represent the entity set of feature weights for intelligent medical emergency backup diagnostic data (i.e., node and), is the collaborative statistics of feature weights for intelligent medical assistance diagnostic data, and represents the time delay of feature weight classification for intelligent medical auxiliary diagnostic data. The first setup. The symptomatic characteristics of I IMAD are studied using the rough set characteristic reconstructing technique.

The function parts of the innovative medical auxiliary medical knowledge that reflect pathological unique properties are extracted based from the results and the above evaluation, and the better decision tree model is used to obtain quantitative measurements of the intelligent healthcare pathological character traits, assisting the clinical diagnosis. The following are the detailed steps: (1)Input: a fuzzy histogram improved decision tree model and a finite set combination approach for intelligent healthcare supplementary diagnosing data attribute feature quantity; output: decision tree reliability evaluation function (3)For intelligent medical-assisted diagnosis, create a decision tree distributing characteristic training dataset. (4)Input: distribute a feature training set; output: kernel function for IMAD and selection in the fuzzy horizontal distributions node of the decision tree (5)Input: investigate the pathogenic traits of Decision Making.

Regression trees are used when the final result of the data is discrete or categorical, such as the relative importance of children in the classroom, a person's destruction or self preservation, loan approval, and so on; however, regression algorithms are used when the final result of the data is continuous, also including prices, an user's age, length of stay in a hotel, and so on.

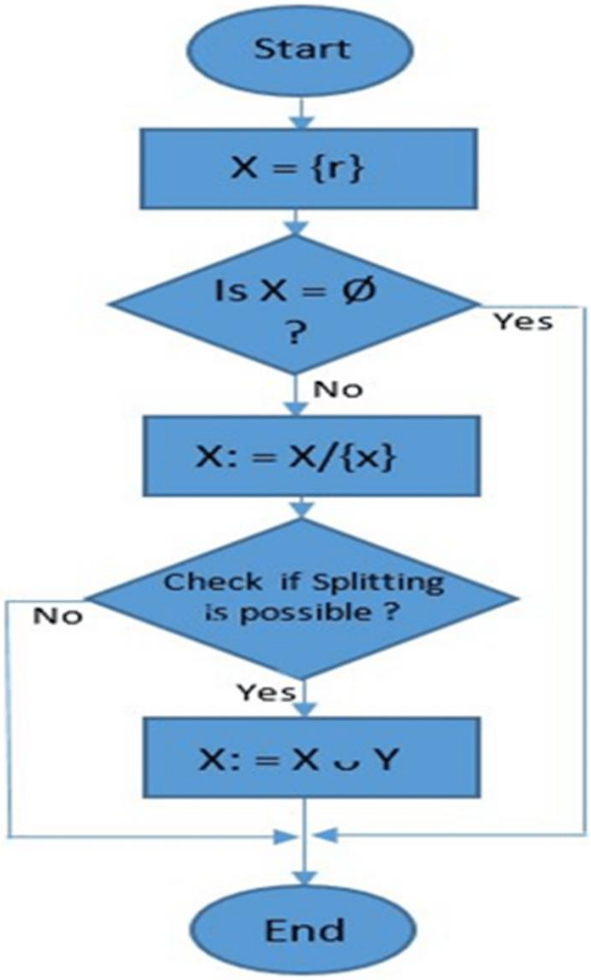

Fig7. Flow Diagram

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$
$$= (100 + 45) / (100 + 45 + 9 + 6) = 0.90$$

Recall: Recall gives us an idea about when it's actually yes, how often does it predict yes.

$$\text{Recall} = TP / (TP + FN)$$
$$= 100 / (100 + 6) = 0.94$$

DecisionTree :

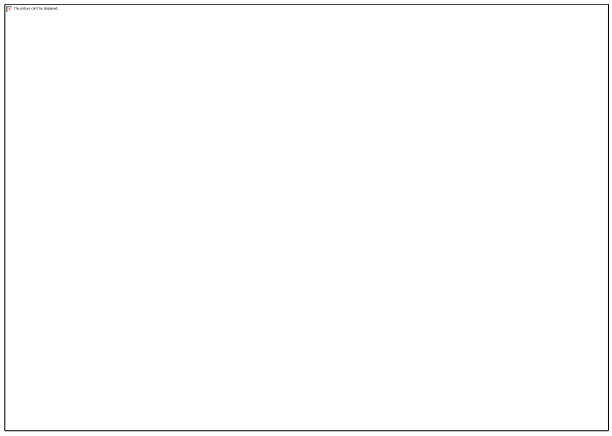

Fig8. Hyperplane of Decision Tree

$$\sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i x^t{}_j x_j \, K(\ x_i, y_i\ )$$

$$\text{subject to } \sum_{i-1}^{N} \alpha_i y_j = 0$$

As a participant's increases with age, so do their characteristics of the patients, according to the graph. The classification performance rate of the proposed methodology is higher than that of literature [4], [5], and [6], with literature [6] having the lowest classified accuracy rate, so while literature [4] and [5] have surprisingly high classification rates, with the best classification rate proceeding onward to be stable within about 80%. This shows that the method used in this study has a high classification impact for the vast majority of patient demographics and is suitable for IMAD.

Precision: Precision tells us about when it predicts yes, how often is it correct.

Precision = TP / (TP + FP)

=100/ (100+9) = 0.91

F measure = (2 * Recall * Precision)/ (Recall + Precision)

= (2 * 0.94 * 0.91) / (0.94 + 0.91) = 0.92

IV. EXPECTED OUTCOMES:

In this essay, we looked now at basics of learning algorithms and how it may be used to predict and diagnosis cancer. The majority of active studies has been on establishing statistical models using supervised machine learning techniques and classification algorithms in order to anticipate plausible medical outcomes. Based on their findings, it's evident that incorporating multifaceted massive datasets with a variety of feature selection and classification methodologies might yield in potential cancer forecasting tools. In comparison to the original lr-partition, the Seleukos methodology accelerated the degree of difficulty for the pancreatic cancer dataset by 14 percent over the period in the early stage, and it also proved to be four percentage points better. One C4.5 C4.5,Several of the industry in terms most well-known machine learning algorithms is still in its infancy. In 10-fold merge, Seleukos surpasses all other classes for the collections equilibrium, monk2, and nursery. Customers just need to enter a database file & three characteristics into the order to properly assess now that the approach has really been automated.

V. RESULTS:

Lung disease is the most widely recognized fatal malignancy in grown-ups around the world. The main objective of this project is to develop a software model using one of the machine learning classification algorithm called Decision tree

Decision is same as supervised learning which contains both preexistiing inputs and outputs, where the data is continuously split according to certain parameter .

By using decision tree we are going to predict the lung cancer based on 2 criteria one is based on stages of lung cancer and another one is based on age, we are representing these criterias in the form of graph, as mentioned below

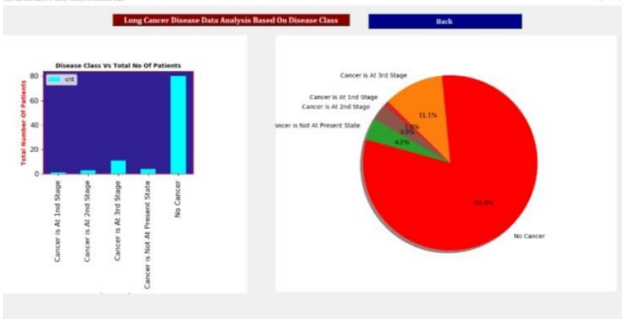

Fig9. Lung Cancer Prediction based on Diasease class

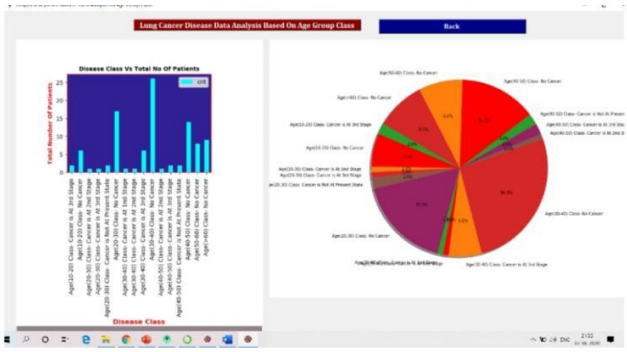

Fig10. Lung Cancer Prediction based on Age Group class

VI. CONCLUSION

Machine learning is widely used in various fields By using machine learning we are going to predict the lung cancer based on criterias By using decision tree (supervised learning) we are predicting based on age and the stages of lung cancer. cancer prediction system combining the prediction system with ML technology was used.

REFERENCES

[1] A. Kanakatte, N. Mani, B. Srinivasan, and J. Gubbi, "Pulmonary Tumor Volume Detection from Positron Emission Tomography Images," 2008 International

Conference on BioMedical Engineering and Informatics, 2008.
[2] Y. Lee, T. Hara, H. Fujita, S. Itoh, and T. Ishigaki, "Automated detection of pulmonary nodules in helical CT images based on an improved template-matching technique," IEEE Transactions on Medical Imaging, vol. 20, no. 7, pp. 595–604, 2001.
[3] S. H. Hawkins, J. N. Korecki, Y. Balagurunathan, Y. Gu, V. Kumar, S. Basu, L. O. Hall, D. B. Goldgof, R. A. Gatenby, and R. J. Gillies, "Predicting Outcomes of Nonsmall Cell Lung Cancer Using CT Image Features," IEEE Access, vol. 2, pp. 1418–1426, 2014.
[4] F. Taher, N. Werghi, and H. Al-Ahmad, "Bayesian classification and artificial neural network methods for lung cancer early diagnosis," 2012 19th IEEE International Conference on Electronics, Circuits, and Systems (ICECS 2012), 2012.
[5] P. Naresh, and D. R. Shettar, "Early Detection of Lung Cancer Using Neural Network Techniques," Prashant Naresh Int. Journal of Engineering Researchand Applications, ISSN : 2248-9622, Vol. 4, Issue 8(Version 4), August 2014, pp.78-83.
[6] S. Sivakumar and C. Chandrasekar, "Lung nodule detection using fuzzy clustering and support vector machines," International Journal of Engineering and Technology, vol. 5, no. 1, pp. 179-185, 2013.
[7] D. Song, T. A. Zhukov, O. Markov, W. Qian, and M. S. Tockman, "Prognosis of stage I lung cancer patients through quantitative analysis of centrosomal features," 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), 2012.
[8] S. V. Anand, "Segmentation coupled textural feature classification for lung tumor prediction," 2010 International Conference On Communication Control And Computing Technologies, 2010.
[9] A. Hashemi, A. H. Pilevar, and R. Rafeh, "Mass Detection in Lung CT Images Using Region Growing Segmentation and Decision Making Based on Fuzzy Inference System and Artificial Neural Network," International Journal of Image, Graphics and Signal Processing, vol. 5, no. 6, pp. 16–24, 2013.
[10] A. S. Deshpande, D. D. Lokhande, R. P. Mundhe, J. M. Ghatole, "Lung cancer detection with fusion of CT and MRI images using image processing.," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), volume 4 issue 3, March 2015.
[11] S. Akram, M. Y. Javed, U. Qamar, A. Khanum, and A. Hassan, "Artificial Neural Network based Classification of Lungs Nodule using Hybrid Features from Computerized Tomographic Images," Applied Mathematics & Information Sciences, vol. 9, no. 1, pp. 183–195, 2015.
[12] S. Kalaivani, P. Chatterjee, S. Juyal, 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.
[13] J. Alam, S. Alam, and A. Hossan, "Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifie," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), 2018.
[14] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, "Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," Am. J. Roentgenol., vol. 174, pp. 71– 74, 2000.
[15] V. Vapnik, "An overview of statistical learning theory," IEEE Transactions on Neural Networks, vol. 10, no. 5, pp. 988–999, 1999.
[16] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," Machine Learning: ECML-98 Lecture Notes in Computer Science, pp. 4–15, 1998.
[17] J. Quinlan, "Decision trees and decision-making," IEEE Transactions on Systems, Man, and Cybernetics, vol. 20, no. 2, pp. 339–346, 1990.
[18] M. Egmont-Petersen, D. D. Ridder, and H. Handels, "Image processing with neural networks—a review," Pattern Recognition, vol. 35, no. 10, pp. 2279–2301, 2002.