A Detailed Literature Survey and In-depth Analysis of Existing Methods for the Detection of Lung Cancer

Akshata Saptasagar, Sandhya Waghere, Rahul Badgujar, Atharva Misal, Omkar Raskar
Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India
akshata.saptasagar@gmail.com, sandhya.shinde@pccoepune.org, rahulbadgujar1508@gmail.com,
asmisal26@gmail.com, omkarmraskar@gmail.com

***Abstract*—There are various existing models which focus on the diagnosis and determination of stages of various cancer and other related diseases. This paper focuses on the diagnosis and stage determination of Lung cancer by compiling various MI models. This paper proposes a model that will not only diagnose the presence of disease but will also help the medical faculty in knowing the particular stage of the disease. Also, advanced analysis is provided by the models which give a brief overview of the disease and highlights the stakeholders about the curability of this disease. The model uses various algorithms and compiles some of the best-suited algorithms which will provide results with higher accuracy and precision.**

Keywords– Diagnosis, Stage Determination, Lung Cancer, Support Vector Machines, Decision trees, Enhanced classifiers, advanced analysis.

I. INTRODUCTION

Constant use of Tobacco, alcohol and other toxic products has become a risk factor for Cancer as well as other diseases. Also, unhealthy diet processes, least to no physical activities, and air pollution have become alarming aspects of many diseases. The disease in which there is an uncontrolled growth of cells is called cancer. This growth of cells when takes place in the lungs is termed lung cancer. Lung cancer can also happen when cell growth spreads from other organs to the lungs. Recent studies proved that lung cancer is the second most cancer worldwide. By 2020, Lung cancer accounted for 2.2 million cases worldwide. Occupational exposures or previously occurring lung diseases or indoor air pollution can also cause Lung cancer. Lung cancer also known as Lung malignancy is generally hard to diagnose as it can be considered as mere flu in its early stages. But this situation can be changed by detecting lung cancer at early stages with the help of various ML models. Around 236,740 lung cancer were detected in 2022 according to the American Cancer Society resulting in nearly 130K deaths.

Lung Cancer is mostly divided into two phases namely SCLC and NSCLC. SCLC is Small Cell Lung Cancer which is deadly serious and is found in the inner layers of the walls of bronchitis. NSCLC is Non-Small Cell Lung Cancer which is less serious as compared to SCLC and also is most commonly found across the world. In 2020, 2.2 million people were diagnosed with Lung Malignance out of which 1.7 million people lost their lives to cancer. This rate can be reduced by the detection of lung Malignancy at its early stages. Disease diagnosis can be implemented using the various Machine Learning models.

One such is the Support Vector Machine technique which involves classifying the dataset into several classes. the support vector classifiers help the model to divide the datasets into multiple classes providing high accuracies for small datasets. Decision trees are one of the models which help to segregate the data into various subsets and the number of decision trees can be coupled into the Random Forest model generating high accuracies for complex datasets. Enhanced Classifiers integrate the analysis of images of various models and based on voting classifiers the results can outsource which can then be further used to treat the disease.

A. Principle of Diagnosis and Stage Determination of Diseases

Diagnosis involves specialists making decisions about a person or group of individuals. It considers the whole lifetime of an individual including the individual's past life, experiences, style of living, attitude, and interests. Diagnosis requires knowledge, skill and the ability to synthesize and evaluate large portions of data. Diagnosis is a team effort which requires effective communication.

Determination involves detecting the anatomic extent of a particular disease. It helps in measuring to which extent cancer has spread across the human body. This can be done with the help of going through a process of scanning as well as biopsies and other related tests. The determination can be done in the form of numbered stagings in the TNM format. Stages involve the numbering severity of cancer cells on the count of 0 to 4. TNM includes a detailed analysis of tumours, nodes and metastasis.

II. RESEARCH AREA

In the present paper, we have proposed a model which will diagnose the disease presence. The model will further determine on which stage the current disease relies and using advanced analysis, the model will help to detect if the disease is curable or not. Also, this model can be used to state the treatment specifically to stages which will automatically be put into action once the disease is found to be at a certain stage.

The data set containing images of CT-scanned images will be processed by an extraction algorithm which will also help to extract certain features. Further using classification algorithms the abnormal images will be classified based on various stages. The advanced analysis will result in ranging the disease on the scale of curability. Also, the various algorithms used to build the model are used in such a manner that the results produced will be highly accurate and precise.

III. RELATED WORK

A. Analysis of CT Scan Images to Predict Lung Cancer Stages Using Image Processing Techniques

[1]Lung cancer is a malignant lung tumour that is characterised by unchecked cell proliferation. Due to many reasons detecting lung cancer earlier is essential. CT images are used widely as they have been proven more effective than X-rays for diagnosing the stage of lung cancer as early as possible. The main advantage of selecting digital image processing is the superiority of image data and optical ability over other techniques for lung cancer staging. Evaluating image cells and obtaining information is easier due to picture processing. The statistical parametric approach and analysing of the image based on grey level co-occurrence matrix are used to eliminate other features. GLCM gives a pairing of types that are related by second order. The GLCM has more characteristics than the statistical approach. Whereas the statistical approach has less complexity than GLCM.

Firstly we obtain the images for preprocessing. It includes various techniques as mentioned below. The features are then extracted with the help of the GLCM technique and statistical approach for determining the values of the features. At last, classifiers are used to partition cancer into two stages, one is limited and the other is extensive. Then the classifier performance is calculated

1) A. Image acquisition
    a) CT Scanned images of the patients are obtained. These images are to be preprocessed for the reduction of noise.
2) Pre-processing
    a) Smoothing: To eliminate the noise from the pictures smoothing technique is used,
    b) Enhancement: To make the picture quality better, enhancement is used. For getting desired results through enhancement, the Gabor filter is used. It is proven better than auto enhancement.
    c) Segmentation: To partition the image into several segments, Segmentation is used. A global threshold and Otsu's technique are used.
    d) Morphological Opening: In this technique, erosion and dilation are used.
3) Feature extraction: Using the grey level for analysis of the texture of the matrix Co-occurrence is done. The texture of the picture is quantified with the help of metrics known as image texture. This helps to that know the colour intensities are spread spatially in a picture. Then the GLCM is used to obtain the quantities of the grey-level picture elements.
4) Classifiers:
    a) SVM: It's used for classification as well as regression, but the most widely used is the classification problem.

Two methods were used to obtain the features which was the statistical parametric approach and the other was GLCM. The second approach has fewer features but is better than other methods as the dimensions in it exceed to the tenth power equivalent to total features and increase with it.

This study helped to preprocess the image and improve the precision of detecting the stage of lung cancer. By using SVM the statistical approach yields an accuracy of 78.95%.

In near future, the improvement of this model can be done to increase its efficiency and accuracy by giving it more information and fine-tuning.

B. Automated Detection of Lung Cancer Using CT Scan Images

[2]Hoque et al research's study developed an automated method for cancer diagnosis using grayscale CT scan pictures. Photos of lung cancer were added as the input. Using medical image processing techniques for pre-and post-processing, output images containing the area alone were created. Enhancing, filtering operation, and segmentation make up the preprocessing. Feature identification and extraction make up the post-processing.

Image acquisition is the initial stage of this approach. Preprocessing and postprocessing are the two steps that make up this. Filtering and segmenting are done during the preprocessing stage of picture enhancement. The post-processing stage involves feature extraction and identification.

1) A. Image Acquisition : The proposed strategy has been put into practice using a lung cancer dataset [2]. The photos are saved in image format in MATLAB of size 450x450 and are shown as RGB grayscale images with entries ranging from 0 to 1. For processing, they are converted to HSV (hue saturation value) format.
2) Pre-processing :
    a) Enhancing the image comes first. Contrast stretching is used, which has been shown to perform better on grayscale images[2].
    b) The second stage, known as filter operation, is used to enhance the edges and sharpen the image. In this

case, a median filter is utilised, which performed better than a mean filter or a Gaussian filter[2].

c) Segmentation is the third and last preprocessing step. In essence, it divides the image into sections based on their shared traits. The threshold-based Otsu technique has a wide range of applications, including binary image alteration and picture segmentation for further processing, such as feature analysis.

3) Post-Processing :

a) Feature Extraction: The region-propos MATLAB method returns a number of characteristics for each form in the picture. The proposed model makes use of an area, circularity (roundness and diameter), and solidity.

b) Identification: The identification process's objective is to calculate the ROI. The characteristics extracted in the feature extraction step are used to detect the ROI. The ROI has been correctly detected when a region is encircled by the value length of each feature.

The proposed method's outcome analysis is shown in table 1 below. The dataset includes the complete number of photos, including photographs that are True and Positive photos, True and Negative photos, False and Positive photos, and False and Negative photos, as well as the accuracy among these datasets. It illustrates that there are, respectively, 24, 23, and 24 TP values for batches 1, 2, and 3, as well as 1, 1, and 1 TN values for the same. It shows that the recommended method effectively recognises three sets of 26 photographs from all three batches 1, respectively which together include 78 images. For each of the three batches, the accuracy is 96.15%, 92.30%, and 96.15%, respectively. As a result, this study's accuracy was 95% which is pathologically approved.

In this research, they suggest using a feature selection algorithm and for a classifier using the Support Vector Machine(SVM) to identify lung cancer in CT scan pictures. The major objective of this study is to accurately and as accurately as possible diagnose lung cancer. When compared to other studies, this research simplifies processes that improve accuracy while also saving time. There is still a chance to use grayscale photographs because the accuracy rate can still be increased.

C. Early Stage Lung Cancer Prediction Using Various Machine Learning Techniques

[3]Lung cancer is chiefly activated due to cigarette smoking. 2.09 million cases are diagnosed to date. This diagnosis of lung cancer mostly includes a variety of complex processes. World Health Organization also stated that by 2018, most deaths occurred regarding Lung Cancer. There are two forms of Lung Malignancy and they are SCLC and NSCLC. SCLC is deadly serious and is mostly found in the bronchial wall's inner layers. SCLC spreads rapidly as compared to NSCLC. NSCLC is less dangerous as compared to SCLC in the early stages. Catalog Classification is used to classify whether a person possesses lung cancer or not. Various ML algorithms like SVM, KNN, Clustering can be used to solve such problems. Machine learning mostly deals with statistical models eliminating the need for instructions which depend on patterns and inferences

ICD9 demonstrative codes can help to differentiate a particular chronic disease from rest of the chronic diseases. But here several visits are needed by the patient to make the algorithm familiar. Fuzzy-C-Means or Fuzzy-Possibilistic-C-Means can be used but the accuracy only notes at 80.36%. Also, it was discovered that to achieve better accuracy, one has to make use of hybrid models. Neural Networks like ANN and CNN were used to diagnose cancer but the output acquired from these models provided output with less accuracy compared to that of ML models[3]. And the accuracy of these network models can be increased by coupling them with advanced neural technology but implementing those needed high knowledge and eventually, the concluded model resulted in greater complexities.

The proposed work in the paper consists of the model that estimated the performance of various other models. The first step in building the model comprises Data pre-processing which includes Importing the required libraries and importing the necessary datasets. Successive steps to importing datasets include Redressing Missing data followed by Converting data into Categorical data where all the independent categorical features are converted into numerical features. Feature selection was done to extract the necessary features eliminating the rest which might have led to future computational complexities. further data gets split up into train, test and validation sets. The algorithms used for comparison include the following:

1) Support Vector Machine: It uses a classifier named Support Vector Classifier. This classifier assists in model training using training dataset and predicts the output into several classes.
2) Random Forest: It involves creating multiple decision trees and the output is predicted by combining the votes of n-decision trees.
3) K-Nearest Neighbor: KNN uses similarity measures to calculate the difference between actual and observed data points. Critical estimations involve k vectors to be odd but this model uses 2 classes issues to wear off the tie between the classes.
4) Neural Networks: Using an ANN network, various forms of data are fed to the input layer which then processes the data and transfers the processed output to hidden layer. Further, activation functions are applied in it which helps in re-tuning weight within input and hidden layer and transferring the result to the output layer.
5) Voting Classifier: Hard voting is implemented which generally takes majority votes by the predicted class labels. This model is generally a hybrid model as it uses SVM, Random Forest and KNN to train the classifier.

The results involve parametric accuracies of all 5 models with 0.8 and 0.2 as training and testing datasets respectively.

Accuracy of SVM rates for 95%. Random Forest gives 97.5% accuracy whereas KNN gives 97% accuracy. Neural networks account for 95.99% accuracy. The voting classifier gives the highest accuracy of all which is 99.5%.

The model helps in predicting early-stage cancer in people. The model gave a brief description of famous ML classification models and also compared the accuracies of SVM, KNN, Random Forest, Neural Networks and Voting Classifiers. Thus, the voting classifier is considered to be best suited for predicting cancer at its early stages. Also, the scope can be extended by using Logistic regression models or Extra trees classifiers or by using Boosting methods.

D. Application of Machine Learning Methods for Determining the Stage of Cancer

[4]Cancer is among the main global causes of death based on research by WHO, accounting for 20% of fatalities in the Europe region. The models created using machine learning techniques can aid in the process of identifying a patient based on their physical symptoms. Predicting the stage of cancer's particular symptoms is the aim of the study.

The term "staging" is used to define the tumour's location, size, kind, lymphadenopathy distribution, and the existence of metastases. The TNM method is being used to better the information flow between doctors and to aid patients in understanding their illnesses.

To denote the size of the tumour, the TNM methodology is employed for the 'T'. When it is not possible to assess the adjacent lymph nodes, a value of X is applied. When there is no malignancy in the adjacent lymphoid tissue, the value is 0. To illustrate the magnitude, place, and amount of neighbouring lymph nodes where cancer has been disseminated, N is linked from one to three. When malignancy has spread towards other parts of the body, the letter M is used to denote this.

When the illness has not migrated to other bodily regions, the zero value is employed. The primary research issue is the prediction of a patient's oncological stage of the disease utilising the TNM method. 2) Techniques: A lot of data is processed and analyzed using machine learning techniques. Three techniques were used to generate regression models: Decision trees, Support Vector Machines, Ensemble Algorithms, and 3) Measures To analyze the models, four measures are used. A) MSE B) R-Squared C) MAE D) RMSE 4) Experimental procedure: The initial training sample contains information on 18329 patients with breast cancer, whereas the testing set contains 6112 entries.

There are three sets of datasets taken into account. The first contains 18329 training and 6112 test samples. The second contains 22732 training and 5150 test samples and the third contains 16506 training and 4126 test samples. Ten-fold cross-validation is performed on a set of training data to improve the accuracy of the model and avoid retraining. Nine FineTree, MediumTree, and CoarseTree models are produced using the Decision Tree approach. Six models are built using SVM and six using Ensemble Algorithms.

The minimal leaf size is the fundamental variable that is utilized to build various tree models. For the FineTree Model, MediumTree Model, and CoarseTree Model, respectively, the model parameters are 4, 12, and 36. The results of the measures used to analyze the models have very few variations. The simulation outcomes showed that the Decision Tree-based models had the highest R-squared 0.99 values. The Decision Tree approach was shown to require the least train time, under 2 secs, but the Support Vector Machine method requires training times ranging from 14. 691 to 177. 3 sec. Models FineTree, MediumTree, and CoarseTree are the quickest.

IV. PROPOSED WORK

Disease diagnosis and determination-based models mostly include one algorithm and either focus on the diagnosis of diseases or the determination of stages of the particular disease. Models where more than one algorithm is used result in hampering the accuracy. Also due to the complex integration of algorithms, the model's efficiency is circumscribed. The proposed model will be using more than one algorithm but will do the work of both diagnosing and determining the stages of lung cancer. The complexities of algorithms will be curbed and thus resulting output will be incisive. Treatment switching will be easily noted down for stages and thus this will help in switching treatments after the disease increments or decrements from the present stage.

V. CONCLUSIONS AND FUTURE SCOPE

Several ML models will be used in combination to generate precise results. Also, the complexities of models are taken into consideration. The produced results will be detailed as well as highly accurate. The model building can be used for various other disease detection mechanisms by only changing the feature sets. This will help in providing a complete overview of the disease for a particular human being along with treatments and medicinal aids attached to certain stages.

REFERENCES

[1] M. Islam, A. H. Mahamud and R. Rab, "Analysis of CT Scan Images to Predict Lung Cancer Stages Using Image Processing Techniques," 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2019, pp. 0961-0967, doi: 10.1109/IEMCON.2019.8936175.

[2] A. Hoque, A. K. M. A. Farabi, F. Ahmed and M. Z. Islam, "Automated Detection of Lung Cancer Using CT Scan Images," 2020 IEEE Region 10 Symposium (TENSYMP), 2020, pp. 1030-1033, doi: 10.1109/TENSYMP50017.2020.9230861.

[3] C. Thallam, A. Peruboyina, S. S. T. Raju and N. Sampath, "Early Stage Lung Cancer Prediction Using Various Machine Learning Techniques," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1285-1292, doi: 10.1109/ICECA49313.2020.9297576.

[4] M. Todorova, "Application of Machine Learning Methods for Determining the Stage of Cancer," 2020 International Conference Automatics and Informatics (ICAI), 2020, pp. 1-4, doi: 10.1109/ICAI50593.2020.9311355.

[5] Juan Cui, Fan Li, Guoqing Wang, Xuedong Fang, J David Puett, and Ying Xu. Gene-expression signatures can distinguish gastric cancer grades and stages. PloS one, 6(3):e17819, 2011.

[6] P. Basak and A. Nath, "Detection of different stages of lungs cancer in ct-scan image using image processing techniques", International Journal of Innovative Research in Computer and Communication Engineering, vol. 5, pp. 9708-9719, 2017.

[7] Va Dominic, Dr. Deepa Gupta, Sangita Khare, and Aggarwal, Ab, "Investigation of chronic disease correlation using data mining techniques", in 2015 2nd International Conference on Recent Advances in Engineering and Computational Sciences, RAECS 2015, 2015.

[8] James D. Brierley BSc, MB, FRCP, FRCR, FRCPC, Mary K. Gospodarowicz MD, FRCPC, FRCR (Hon) Christian Wittekind MD, "TNM Classification of Malignant Tumours Eighth Edition", Union for International Cancer Control (UICC), Pages:1-241, 2017.