LUNG CANCER PREDICTION USING EXTENDED KNN ALGORITHM

[1] Ajitha E, [2] Dr. Diwan B
[1]Assistant Professor, Department of Computer Science and Engineering, St. Joseph's Institute of Technology, Chennai-119
[2]Associate Professor, Department of Computer Science and Engineering, St. Joseph's College of Engineering, Chennai-119
[1]ajithamano2129@gmail.com, [2] diwandiwan@gmail.com

Abstract: *The paper proposes an Extended version of the KNN Algorithm that is used for Lung Cancer Prediction based on the CT - Images given as the input during the Training phase. The 2-D image undergoes a Modified Gabor Filtration technique wherein the features of the image are extracted for Edge Detection. This further undergoes Feature Extraction followed by Binarization which is fed as Production data to the Machine Learning model. Based on the Extended KNN Algorithm, the model evaluates the testing data and corresponding predictions are made. The model predicts the Cancer Stage based on the input CT - Image which is passed to the doctor for further medication.*

Keywords: **Extended KNN – Modified Gabor Filtration – Edge Detection – Feature Extraction – Binarization**

1. INTRODUCTION

Lung Cancer is considered to be the deadly disease that has got the highest mortality rate of 46.7 per 100,000 persons and 31.9 per 100,000 persons amongst the Male and Female population respectively. Following Cancer Statistics projected by WHO, Lung cancer, kills 1.76 million people every year. In India, with a mortality rate of 5/100,000, it is responsible for 63,475 deaths annually. The major reason for this fatal death due to lung cancer is the delayed diagnosis and rapid growth of Cancerous cells. These cells can be carried away from the lungs through blood, or lymph fluid that surrounds lung tissue. Lymph flows through lymphatic vessels, which drain into lymph nodes located in the lungs and the center of the chest. Lung cancer often spreads toward the center of the chest because the natural flow of lymph out of the lungs is toward the center of the chest.

The growth of lung nodules more than diameter 3mm is considered to be cancerous cells. The CT – Scan image is given as input which is further analyzed by the radiologist. The prolonged diagnostic method aggregates the growth of tumor cells. These lung nodules exhibit large variation in density which is not visible in radiograph at the initial stage. These nodules can further develop anywhere within the lung field namely ribs and structures beneath the digraph which becomes fatal if not treated immediately. To overcome these issues and to provide an efficient way to detect Lung Cancer cells, we have proposed a Supervised Machine Learning Model which takes CT-Scan images as input and performs image processing techniques like Image Enhancement, Gabor Filtration, Edge Detection, Region Growing, Analysis of Three- Layer Segmentation, Feature Extraction or Binarization during the Training Phase. The model predicts the lung cancer stage during the Testing Phase. The prediction algorithm named Extended KNN is implemented for better accuracy and precision.

2. LITERATURE SURVEY

In Automatic detection of lung nodules based on Computer-Aided Detection as cited by Disha Sharma et.al. (2011) follows the Training and validation of algorithms were not accurate as the Cancerous cells are manifested by itself as a non-calcified pulmonary nodule that can be detected accurately by reading the lung Computed Tomography (CT) images. But due to the unavailability of CT-San Images, the existing system continued with CAD which is composed of two CAD sub-

procedures is presented as CAD which is to analyze the internal parenchymal nodules and CADJP is to identify the nodules attached to the pleura surface.

The goal of a CAD system is to identify regions of interest in the image that can reveal specific abnormalities and alert physicians to these regions. In Lung nodule classification using deep features in CT images," Database that contains CT images of the thoracic region for 1010 patients along with the annotation data of suspicious nodules (for both benign and malignant nodules) for a size greater than 3 mm from up to four radiologists collected over a long period. The data that holds diagnostic data for a small number of cases (157 patients) e from biopsy, surgical resection, progression, or reviewing the radiological images to show 2 years of nodule size has been used for the analysis here.

The existing system, as cited by Arvind Kumar Tiwari et.al. (2016) is based on X-Ray images for the detection of cancerous cells by using various textural features for disjoint segments of the lung and focused on a subset of lung disorders. Further, they used Computer-Aided Devices (CAD) for a complex system, for the detection of various lung disorders. The aim is to get more accurate results using various stages of image processing like image enhancement, image segmentation, texture analysis etc.

In image enhancement, images are compared with Gabor filter, auto enhancement, and fast Fourier transform techniques. Since only a restricted range of waveform data that can be transformed, which leads to disintegration of wavelets. To avoid this, we and need to apply a window weighting function to the waveform such that the spectrum leakage can be avoided. Henceforth, the spectrum leakage can be avoided. Image processing used in the existing system extracts a small number of features, which leads to the loss of data. Once the position of the nodule is determined, the nodule is then extracted from the entire image. Then, the features such as total area, average area, maximum area, and average eccentricity, average equivalent diameter, standard equivalent diameter, weightedX, weightedY, number of nodes, and number of nodes per slice were calculated. Thus, in this approach, there is a loss of efficiency which in turn leads to lesser productivity of the model.

3. PROPOSED SYSTEM

The proposed system is based on detecting the Lung Nodules and their growth size represented in CT-Scan Images. The model is used for faster and accurate prediction of Lung Cancer such that better medication is made possible for the patients. The dataset of CT – Scan images is in Dicom format which is divided into two Training dataset and Testing Dataset wherein 75% of Data is used in Training Phase and 25% of Data is used for Testing and Cross-Validation. The data is fed into the Image Processing Stages wherein the image undergoes Segmentation, Modified Gabor Filtration, and Feature Extraction.

The existing model is based on the application of the Gabor filter to 2D X-ray lung images for the analysis and they are related directly to Gabor wavelets. Image enhancement is the initial and important stage in lung cancer detection; it takes X-ray lung images as the input from the database. In the Gabor filter, the number of rotations and dilations are present and these are time taking. This results in a blurred image which is not appropriate for the next stages in image processing.

To overcome this drawback a Modified Gabor Filter (MGF) approach has been implemented. The Machine learning Model is Trained using an Extended KNN Algorithm. Henceforth the model can predict whether or not the Lung Nodule is a cancerous cell. The production data is fed into the Supervised Machine Learning model which makes a prediction based on the imagery input. The prediction made is displayed to the Radiologist.

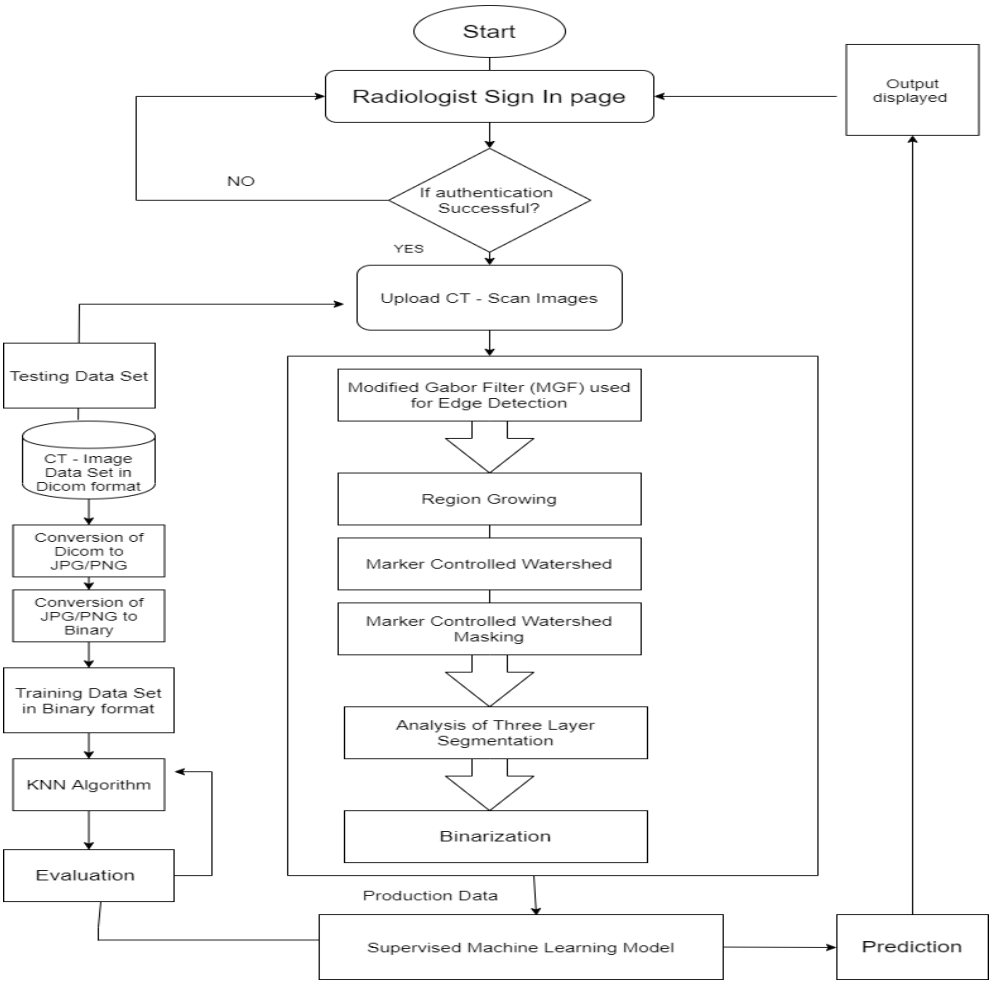

Figure 1: Proposed Architecture Diagram of lung Cancer Prediction using Extended KNN Algorithm

4. METHODOLOGY

The Dataset are collected from Kaggle where 2D Slices of CT- Scan images of 115 patients are collected which are in DICOM format. These 2D images are reconstructed into 3D structures of lungs, further, the image undergoes Denoising which enables the image to be eligible for Image Processing Techniques. The following diagrams depict an overview of the Image Processing stages. The Initial stage is Image Acquisition wherein the 2D Slices of the Image are arranged to reconstruct it into a 3D structure such that the image is eligible to undergo Modified Gabor Filtration. Modified Gabor Filtration (MDF) is performed in the Image Segmentation phase.

The spatial aspect ratio is considered at the Kernel Stage which leads to a reduction in image distortion at the beginning itself. This helps in obtaining significantly clearer images that are used for Segmentation. The thickness of the image slices is calculated and the ones which are suspected to have cancerous lung nodules are converted into Hounsfield units (HU) to isolate other regions. We have also plotted the Frequency graph for Hounsfield Values across their corresponding Frequencies. The image slices are displayed following their thickness and are most likely to be cancerous cells.

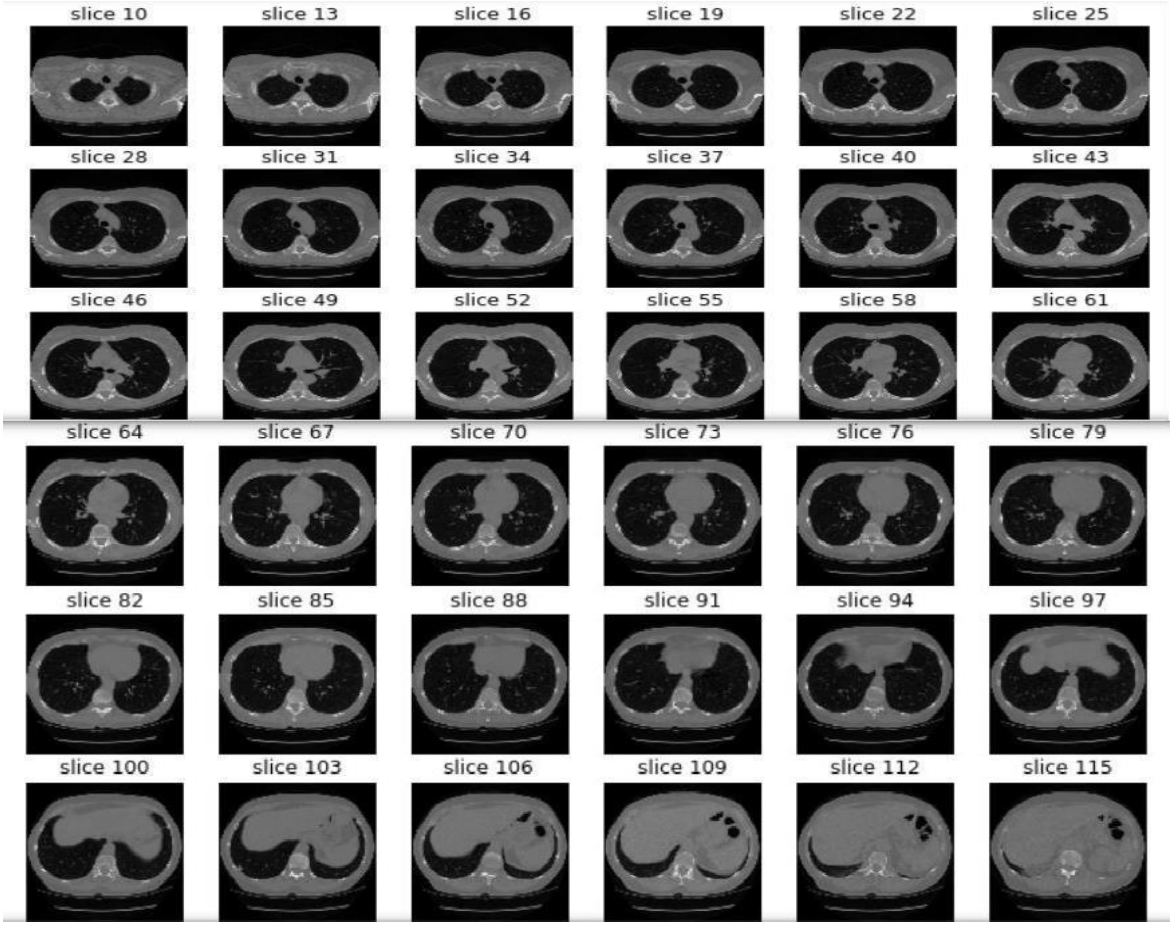

Figure 2: Samples of Image Slices

Masking of the image is performed by highlighting Region of Interest (ROI) by outlining the Edges of the Lung Image focusing the Lung Nodules of size greater than 3mm.

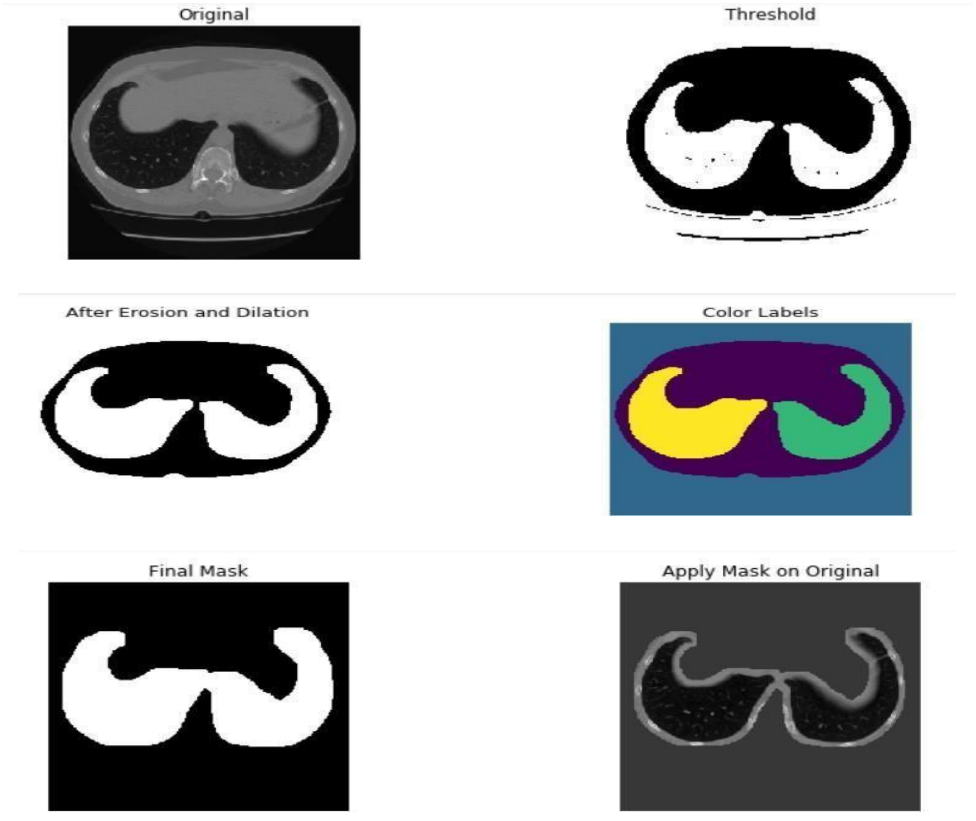

Figure 3: Stages of Image Masking

The Images undergo a Feature Extraction process wherein the images are further filtered to remove the repetitive features which may cause excessive time and space complexity during the Training phase. The extracted features are further plotted as a Feature Importance graph which plays a vital role in identifying the best features that can give an accurate prediction. This process of selecting the best features is termed Feature Selection which improves the efficiency of the machine learning model.

Feature Selection is the process of reducing the input variables by identifying the desirable feature which is intended to give precise output. The paper proposes Statistical – based feature selection wherein the relation between each input variable and the target variable is evaluated. Based on the statistical data produced the strongest relationship with the target variable is taken as the best match and the corresponding feature is considered for the Training phase.

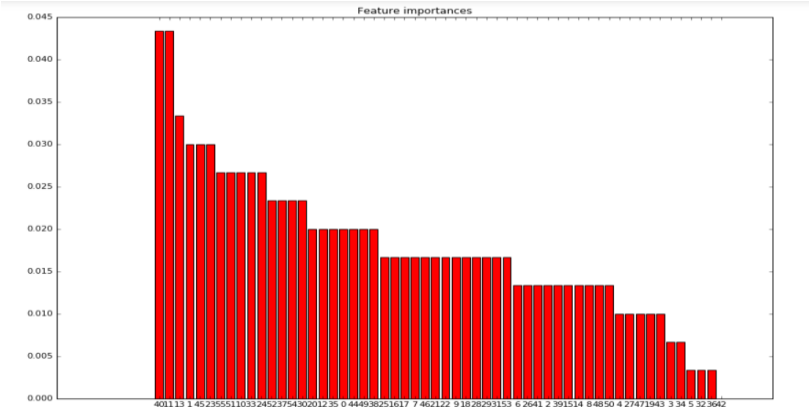

Figure 4: Feature Importance Graph

This method proves to be effective as it removes the non – informative and redundant predictors from evaluating, henceforth their sizes and the growth rate are analyzed. The Supervised Machine Learning model is trained with the labeled dataset of 115 patients and further undergoes the Prediction phase.

5. EXTENDED K NEAREST NEIGHBOR ALGORITHM

The existing K Nearest Neighbor (KNN) Algorithm is a non-parametric method used for Classification and Regression problems. The input consists of the K closest training data points in the feature space such that the output depends on the similarity of features amongst the K closest neighbors.

The proposed system aims to overcome the shortcomings of the existing algorithm and improve its efficiency by 25% by Rescaling Features in K- Nearest Neighbors Algorithm. KNN is a Distance-Based algorithm where KNN classifies data based on proximity to K-Neighbors. Then, often we find that the features of the data we used are not at the same scale/units. This unit difference causes Distance-Based algorithms such as KNN to not perform optimally, so it is necessary to rescaling features that have different units to have the same scale/units.

- Hyper Parameters are used to control the learning process during the Training Phase, by doing so the efficiency of the algorithm can be optimized.
- Introducing Hyper Parameters to the Training phase allows the algorithm to take different learning approaches at various steps of its Training Phase.
- Extended KNN is implemented using the sci-kit-learn library in Python, wherein the Extended KNN class takes Nine Hyper Parameters as its Arguments which is the value of K in the algorithm and the distance metric is set to Manhattan.
- Out of these Nine Hyper Parameters passed into the Extended KNN class the best set of values of Hyper Parameters are chosen. This process is called Hyper Parameter Tuning.
- The strategy used for Hyper Parameter Tuning is GridSearchCV. In this approach, the

machine learning model is evaluated over a range of Hyper Parameter values from which the algorithm searches the Best Set of Hyper Parameter Values. These Hyper Parameter values further undergo Cross-Validation to avoid Overfitting or Underfitting problem.

6. PERFORMANCE ANALYSIS

On implementing the Extended KNN Algorithm the accuracy, precision and overall efficiency of the Machine Learning increases to 94.55% as the existing KNN gave an accuracy of 70.29%. The Machine Learning Model undergoes K-Fold Cross Validation which statically estimates the performance of the model. K-Fold Cross Validation is used to protect the model against Overfitting which eventually reduces the performance of the predictive model.

K-Fold Cross Validation Method guarantees that the score of the model doesn't depend on the Training or Testing Dataset. The dataset is divided to K number if Subsets and the Holdout method is repeated K – times. Henceforth, K-Fold cross-validation is an improvement of Holdout validation.

7. CONCLUSION AND FUTURE SCOPE

Lung cancer is the deadliest disease with the highest mortality rate of 11.57 lakh new cancer patients are registered and 7.84 million cancer patients are dying every year. The proposed model proves to substantially reduce the delay in diagnosis and remove the ambiguity of Lung cancer stages as it essentially proves to give an accuracy of 94.55%. This also avoids medical errors caused amongst the medical staff during the diagnosis which itself increases the fatality of the disease.

The system can be further extended by helping doctors get better diagnostics, and hence, detect diseases faster. They can be extended to varied imagery inputs likewise MRI, PET, CT- FFT Scan images which in turn allows reliability and ease of use. This contributes to the betterment of patients whose health plays a vital role in empowering the Human Resource of a Nation. Medical Image Processing enables accurate detection of Cancerous cells which is, in turn, a boon for doctors who can provide faster treatment. The model is beneficial in the faster detection of cancerous cells such that patients can be treated accordingly.

REFERENCES

[1] Permata T S and Bustamam A 2015 Clustering protein-protein interaction network of TP53 tumor suppressor protein using Markov clustering algorithm International Conference on Advanced Computer Science and Information Systems (ICACSIS) pp 221–226.

[2] Kajal N et al 2015 Early Detection of Lung Cancer Using Image Processing Technique: Review International Journal of Advent Research in Computer and Electronics (IJARCE) 2(2), E-ISSN: 2348-5523.

[3] Febr Mokhled S. AL-TARAWNEH, "Lung Cancer Detection Using Image Processing Techniques", Leonardo Electronic Journal of Practices and Technologies, June2012.

[4] Nguyen, H. T., et al, "Water snakes: energy-driven watershed segmentation", IEEE transactions on patter analysis and machine intelligence, volume 25, number 3, pp.330-342, March 2003.

[5] Suzuki K., "false-positive reduction in the computer-aided diagnostic scheme for detecting nodules in chest radiographs", academic radiology, volume 13, number 10, pp.10-15, February 2005.

[6] American Cancer Society, "Cancer Statistics, 2005", CA: A Cancer Journal for Clinicians (2002).

[7] D. Lin and C. Yan, " Lung nodules identification rules and extraction with fuzzy neural network" IEEE neural information processing, vol.4 Feb 2005.

[8] D. Lin and C. Yan, "Lung nodules identification rules extraction with neural fuzzy network", IEEE, Neural Information Processing, vol. 4,(2002).

[9] B. Zhao, G. Gamsu, M. S. Ginsberg, L. Jiang, and L. H. Schwartz, "Automatic detection of small lung nodules on CT utilizing a local density maximum algorithm", Journal of applied clinical medical physics, vol. 4, (2003).

[10] A. El-Baz, A. A. Farag, PH.D., R. Falk, M.D. and R. L. Rocco, M.D., "detection, visualization, and identification of lung abnormalities in chest spiral CT scans: phase I", Information Conference on Biomedical Engineering, Egypt (2002).

[11] Referred to https://www.kaggle.com/ for procuring dataset of CT- Scan Medical image of over 1500 patients.

[12] American Cancer Society, "Cancer Statistics, 2005", CA: A Cancer Journal for Clinicians, 55: 10-30, 2005.