IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

SNNO: 2347–3657
Volume 8,Issue2June ,2020

Prediction Lung Cancer– In Machine Learning Perspective

Mrs. Mukka Shirisha, Mr. Mohd Sirajuddin, Mrs. Kshetravati N Sangami

Abstract—

Recent years have shown an increasing mortality rate from lung cancer; therefore, it is essential to determine whether or not the tumor has transformed into cancer. If an accurate prediction can be made at an early stage, not only will many lives be saved, but doctors will be able to begin treatment sooner. Checking the tumor's size, location, etc., with the help of computed tomography is crucial for ensuring the tumor's health. In this paper, we propose a framework for early cancer prediction that has the potential to save a large number of lives. Our primary research interests lie in the areas of computer science known as Digital Image Processing (DIP) and Machine Learning (ML). The preprocessing stage of digital image processing has gained a lot of notoriety in recent years. The next step involves putting the pre-processed image through a segmentation phase, passing the resulting image on to a feature extraction phase, and then using machine learning classification algorithms like SVM (Support Vector Machines), Random Forest, and ANN (Artificial Neural Network) to train the features extracted from the image. The prognosis for the tumor's malignancy or benign nature is based on the classification results obtained.

I. INTRODUCTION

Cancer, chikungunya, cholera, and other diseases are on the rise alongside the world's growing population. Among these, cancer is rapidly rising to prominence. Usual killer in terms of mortality rates. The human body is made up of trillions of cells, so cancer can develop at virtually any point. Human cells normally replicate themselves to meet the body's fluctuating demands for new cells. Eventually, all of our cells will die from old age or damage, and then new ones will form to take their place. However, when cancer cells develop, this regulated process fails. An increasingly abnormal cellular state is characterized by the survival of old or damaged cells when they should die and the proliferation of new cells in inappropriate locations. These surplus cells have the potential to proliferate indefinitely, leading to abnormal growths known as tumors. This tumor quickly metastasizes to other organs.

Malignant tumors are those that have the potential to spread to other parts of the body, whereas benign (non-cancerous) tumors do not. (Cancerous) is the development of cell which has power to spread \sin other section of body this spreading of infection is called \metastasis. There is numerous sort of cancer like Lung cancer, \leukemia, and colon cancer etc. Lung cancer has seen a sharp rise in prevalence since the 1800s. Lung cancer may be caused by a number of different factors, the most common of which are smoking, radon gas exposure, passive smoking, and asbestos exposure. Small-cell lung cancer (SCLC)

And non-small-cell lung cancer (NSCLC) is the two primary subtypes of lung cancer (NSCLC). In general, the growth and spread of non-small cell lung cancer are slower than those of small cell lung cancer. Small cell lung cancer (SCLC) is nearly always caused by smoking and is characterized by rapid tumor growth and the formation of large, metastatic tumors. Most cases begin in the bronchi, which are located in the middle of the chest. The cumulative number of cigarettes smoked is associated with an increased risk of dying from lung cancer. [1]

There are a variety of warning signs that might point to lung cancer, and these include:

Symptoms include dyspnoea (effort-related shortness of breath), haemoptysis (bloody coughing), and persistent coughing or a shift in the patient's usual coughing pattern.

1,2,3 Assistant Professor
1,2,3 Department of CSE
1,2,3 Global Institute of Engineering and Technology Moinabad, Ranga Reddy District, Telangana State.

Pattern, dysphonic (hoarse voice), clubbing of the fingernails, chest discomfort or abdominal pain, cachexia (weight loss, weariness, and lack of appetite), and wheezing (uncommon), the disorder is called dysphasia (difficulty swallowing), Shoulder, chest, and arm pain; Bronchitis or pneumonia;

• Health deterioration and sudden weight loss [1]

Medical imagings tests (such as a chest X-ray, CT scan, or MRI) are used to detect lung cancer and help doctors pinpoint the precise position of any tumors so they can provide the most effective therapy. In order to preserve the most lives possible, it is crucial that the condition be diagnosed as soon as possible. There is a lot of background noise in medical pictures, making accurate prediction almost impossible. Therefore, a machine learning algorithm will be used to the pre-processed medical picture in order to detect lung cancer.

II. ENABLING TERMINOLOGY

A. Segmentation

Lung image segmentation seeks to exclude the windpipe, tubular branches, alveoli, and other background structures from a preprocessed picture in order to isolate the lungs' parenchyma and determine their size. Muscle groups in the picture to improve accuracy and simplify feature extraction. Segmentation may be accomplished in a variety of ways. Applying edge detection to the preprocessed picture is the first stage in segmenting a lung CT scan image, followed by applying a threshold to the edge to exclude intensity below the threshold while keeping intensity above the threshold in consideration. Next, morphological methods such as morphological closure and morphological opening will be used. The lung volume will be calculated in step (4) by using morphological segmentation. [2]

B. Machine Learning Context to Lung Cancer

Tumors are classified as benign or malignant using machine learning. Here are the relevant algorithms for predicting lung cancer:

Since image data are high dimensional, a support vector machine (SVM) may aid in cancer prediction by partitioning the dataset into two groups via the use of kernel functions. In order to classify the pictures, a hyper plane is used to display the images in a three-dimensional space using a kernel function such as a polynomial kernel, Gaussian kernel, radial basis function, etc. If we take a CT scan of the lungs, for instance, the resulting image will be pre-processed before being trained using RBF kernel. During training, the images will be labelled as 1 and 2 to represent normal lung tissue and tumor, respectively. [3]

2 Artificial Neural Networks (ANNs) ANNs are a concept borrowed from ANNs found in biology. This is a multi-layer feed-forward neural network. Layers of input, concealed, and output make up the network. First, the image is fed into the input layer, where the activation value is determined; then, at the output layer, the activation function is calculated and aggregated to get O(x); finally, the difference between O(x) and the desired output, known as error, is calculated using the Backpropagation algorithm, in which a weight is assigned to the error and the error is propagated backwards until it reaches an optimal value. [4]

3) Random Forest - In other words, a random set of choices is a decision tree. Because it employs bagging theory, it can process huge arrays of variables without losing any of them.

III. SUMMARY OF LITERATURE SURVEY

Researchers have suggested a number of methods for predicting cancer, including Palani et al [5].'s IoT-based predictive modeling using fuzzy logic Tumor datasets were classified using a C-mean clustering-based segmentation method, an incremental classification algorithm based on association rule mining, and a decision tree for classification; then, using the output from the incremental classification model and additional features, a convolutional neural network was used to predict whether a tumor was benign or malignant.

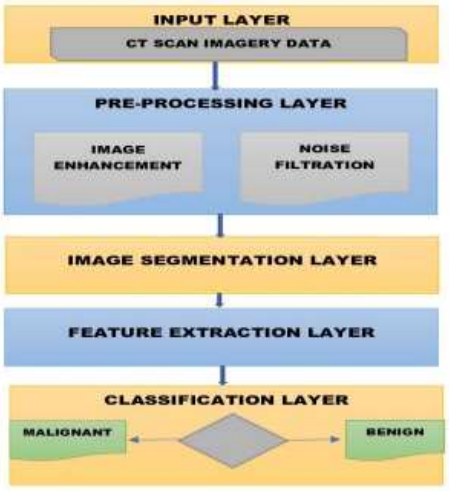
Lynch, et al., [6] The performance of these machine learning algorithms is evaluated using the root mean square error to forecast an individual's chance of survival.

To prevent over fitting, we preprocess the parameters by giving a default value, and then we train each model using 10-fold cross validation. Texture-based feature extraction is used to extract features like contrast and brightness from the picture collection, as suggested by FENWA et al., [3]. Both the artificial neural network and the support vector machine algorithms from the class of machine learning tools are used, and their results are compared in terms of accuracy. In order to determine which feature extraction method yields the most accurate predictions when applied to a certain machine learning algorithm, stark et al. [7] created a model in which a total of five different feature extraction methods were utilized in separate categorization algorithms.

IV. PROPOSED FRAMEWORK FOR CANCER PREDICTION

A new model has been presented based on a review of the existing literature; it includes a pre-processing block, a segmentation block, a feature extraction block, and a finalization block. A stumbling hurdle for categorizing. A CT scan report is mostly utilized in cancer prognosis. However, CT scan reports are often full with noise that is invisible to the naked eye, thus different digital image processing techniques are crucial for obtaining a noise-free picture. The goal of digital image processing is to get meaning from digital images by means of analysis and editing. In the first phase of digital image processing, known as "image pre-processing," numerous tools, such as histogram equalization and spatial filters, are used to improve the quality of the picture. After the picture has been preprocessed, it may be restored using methods such as applying noise (such as salt and pepper noise or Gaussian noise) and filters (such as the median filter or the mean filter). If the picture is not already grayscale, then colour conversions are performed. The innovative structure suggested is seen in Fig.

Once the picture has been segmented into its constituent pixels, a technique known as feature extraction may be carried out. As a subset of dimensionality reduction, feature extraction condenses large amounts of raw data into digestible chunks of picture data, such as regions and textures, for further processing. After the feature has been extracted, it is next classified using a variety of machine learning methods.



There gradient-based watershed segmentation to be one of several approaches to watershed segmentation. Grayscale is preprocessed using the gradient magnitude.

Picture; it has high pixel value along the object edge and low \spiel value in another left section. As a result, we get a final picture that has been segmented and from which features may be extracted.

C. Feature Extraction Layer

Segmentation's results are sent into a feature extraction pipeline. As a result of our feature extraction process, we have two distinct sets of information: region-based and texture-based features. image and based on texture we have extracted features like mean is used to find average intensity, standard deviation is used to measure average contrast, smoothness is used to measurably separate similar textures, and area in context to image means pixel of the image, perimeter in context to image mean vector containing the distance around the boundary of each region in the image, and centric means the centre of mass of the region and is in 1 X 2 vector form.

V. PERFORMANCE EVALUATION

Metrics for assessing model performance are used for analysis. Metrics are selected for use with machine learning projects based on their specific requirements. Categories of model assessments these variables/parameters/etc. are:

A. Confusion Matrix

The confusion matrix is a matrix that describes in great detail any misclassifications or incorrect classifications that have occurred. It has four possible outcomes: true positive (accurately forecast the positive class), false positive (incorrectly predict the positive class), true negative (accurately anticipate the negative class), and false negative (incorrectly predict the negative class).

B .Division By Class Accuracy

Our forecast accuracy may then be gauged based on the results. Prediction accuracy and total number of predictions produced are two metrics that may be used to assess it.

$$P_{acc} = \frac{T_{neg} + T_{pos}}{T_{pos} + F_{pos} + F_{neg} + T_{neg}}$$

C. Recall

It measures the proportion of actual positive that are correctly identified.

SNNO: 2347–3657
Volume 8,Issue2June ,2020

$$P_r = \frac{T_{pos}}{T_{pos} + F_{neg}}$$

D. Precision
It measure the proposition of positive identification is actually correct.

$$P_{prec} = \frac{T_{pos}}{T_{pos} + F_{pos}}$$

E. F1 score
F1 score is the average of both precision and recall.

$$P_{F1} = 2 * \frac{Recall * Precision}{Recall + Precision}$$

VI. RESULT AND DISCUSSION

Artificial neural networks, Random forests, and Support vector machines are employed in the suggested model to determine if a tumor is malignant or benign at its earliest stages. Machine. The accuracy of artificial neural networks has improved in recent years, and this is true for both region-based and texture-based features. When the suggested model is compared to the accuracy, it is clear that the latter has been improved upon while the former has witnessed a decrease in recall. Mat lab R2017a was used for digital image processing, whereas Jupiter notebook was used for machine learning classification. Below is a side-by-side comparison of the two options.

TABLE I. REGION BASED FEATURE

	Accuracy	Precision	Recall	F1 score
Random Forest	79%	100%	50%	67%
SVM	86%	100%	67%	80%
ANN	92%	100%	69%	81%

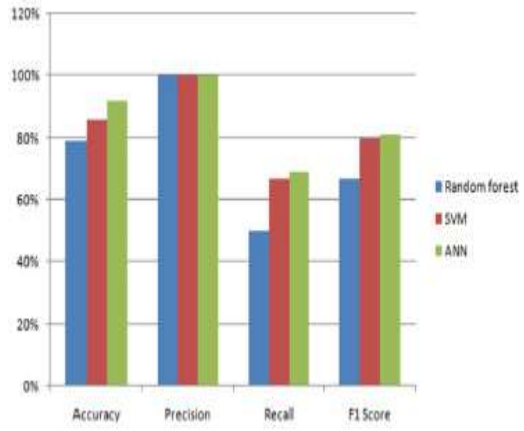


Fig 2. Performance measure based on region based Extraction

TABLE II. TEXTURE BASED FEATURE

	Accuracy	Precision	Recall	F1 Score
Random Forest	70%	89%	47%	62%
SVM	80%	90%	57%	69%
ANN	96%	100%	69%	81%

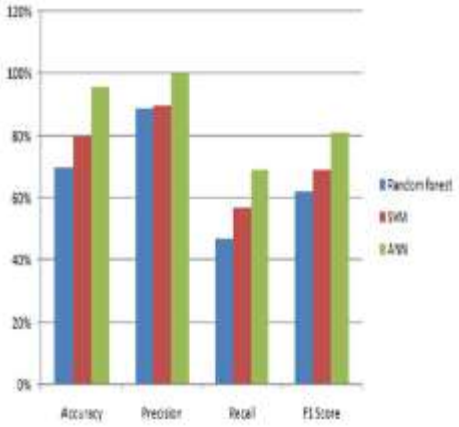


Fig 3. Performance measure based on texture based Extraction

VII. CONLUSION AND FUTURE SCOPE

The suggested approach provides a high-level summary of early lung cancer prediction. Following the first malignancy/benignity tumor prediction, we produce a confusion matrix. We compute the accuracy, recall, precision, and F1 score for each

SNNO: 2347–3657
Volume 8,Issue2June ,2020

machine learning method using the confusion matrix as our basis.

The results show that our suggested model is able to differentiate between benign and malignant tumors and that artificial neural networks are more accurate than conventional methods in identifying malignant tumors based on their textures and regional distributions.

Soon, deep learning will be superior to machine learning for tasks such as image categorization, object identification, and feature extraction. The bigger the number of hidden layers in a CNN network, the more accuracy it is believed to provide.

REFERENCES

[1] 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.

[2] Zhang, Junkie, et al. "Pulmonary nodule detection in medical images: a survey." Biomedical Signal Processing and Control 43 (2018): 138- 147.

[3] Fenway, Lousy D., Funmilola A. Ajala, and A. Adigun. "Classification of cancer of the lungs using SVM and ANN." Int. J. Compute. Technol. 15.1 (2016): 6418-6426.

[4] Adour, Marisa, and Michael Mayo. "A survey of neural network-based cancer prediction models from microarray data." Artificial intelligence in medicine (2019).

[5] Palani, D., and K. Venkatalakshmi. "An IoT based predictive modeling for predicting lung cancer using fuzzy cluster based segmentation and classification." Journal of medical systems 43.2 (2019): 21.

[6] 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.

[7] Öztürk, Şaban, and Bayram Academic. "Application of feature extraction and classification methods for histopathological image using GLCM, LBP, LBGLCM, GLRLM and SFTA." Procedia computer science 132 (2018): 40-46.

[8] 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.

[9] 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.