A R DIGITECH
International Journal Of Engineering, Education And Technology (ARDIJEET)
www.ardigitech.in ISSN 2320-883X,VOLUME 07 ISSUE 04 01/11/2019

Machine Learning Approach Via an Ensemble of Classifiers for Computer Aided Lung Nodule Diagnosis

[1]*Archna Paliwal*
[1]*Department of Computer Science & Engineering RDEC,Ghaziabad*
archana.p@gmail.com

Abstract: The objective is to foretell methods for lung cancer prediction that are based on machine learning. Using supervised machine learning to analyse datasets(SMLT) to analyse the full dataset, validate the data, clean and prepare it, and visualise the results, as well as to conduct uni-, bi-, and multivariate analyses, missing value treatments, and variable identification. Using supervised classification machine learning techniques, we aim to provide a machine learning-based approach for accurate lung cancer prediction.In addition, we will evaluate the user interface of a GUI for lung cancer prediction by characteristics and compare and contrast the performance of several machine learning algorithms using the dataset provided by the transportation traffic department.

Keywords: Dataset, Machine Learning Classification method, Python, Prediction of accuracy result

Introduction

A devastating disease, cancer affects many people's lives. During vital functions, the lungs' primary function is to draw oxygen into the bloodstream and exhale carbon dioxide. Lung cancer develops when cells and tissues proliferate uncontrollably. Cancer, which is the first malignancy, which is the leading cause of cancer-related mortality in men and the second leading cause of cancer-related mortality in women. Nearly one million elderly people die each year because of cancer worldwide [1].Tumours may only be classified as either benign or malignant. Cancer comes in many forms, including colon cancer, leukaemia, melanoma, and many more [2]. Since the early eighteenth century, the incidence of cancer has significantly increased. Many other things may cause carcinoma, including smoking, secondhand smoke, exposure to gases like radon, asbestos, and many more.There are two subtypes of lung cancer, small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC).At partner diplomas in higher degrees, the radiologists may use computerised tomography (CT) and opportunity scanning techniques to find the harmful nodules [3]. Their origin is in the bronchi, which are located in the chest's midsection. Malignant neoplasm symptoms includesymptoms such as difficulty breathing when moving, lethargy, speech impediment, dysphasia, blood in the cough, lack of appetite, and pain in the shoulder, chest, or arm [4].Considering the symptoms, the crucial task of detecting cancer in its early stages may be quite challenging.carcinoma has the highest death rate of any cancer kind because its symptoms are most severe in the latter stages. Doctors rely on correct designations for different types of carcinoma to help them determine and choose the best therapy [5]. While physician recommendations remain the most important part of any designation process, current data suggests that various AI class methodologies might assist physicians in improving their procedures. Misuse class tactics are a common way to lessen the likelihood of errors caused by inexperienced physicians [6].

One use of artificial intelligence is machine learning (ML), which allows computers to learn and improve themselves via experience rather than

code. ML classifiers have gained notoriety for their ability to identify lung and breast malignancies. There are three main types of machine learning algorithms: supervised, unsupervised, and reinforcement learning. In order to improve the accuracy of our model for identifying cancer in CT scans, we used an ensemble classifier. This classifier comprises five separate ML supervised algorithms, such as decision tree, KNN, SVM, RF, MLP, logistic regression, etc. [7]. Medical facilities and clinics may use ML to reduce the diversity of study measures. This paper's primary objective is to classify carcinoma detection as either benign or malignant. There are four stages to the suggested method. The first is noise filtering as part of the pre-processing of CT test images. The second is segmentation using 'Otsu' thresholding. The third is feature extraction, which extracts many characteristics such as area, perimeter, centroid, and so on. Alternatives for categorising and using implemented mathematical assessment on the data are generated using a predictive model, such as carcinoma prediction [8].

In order to boost performance and accuracy metrics, ensemble classifiers propose merging model choices. Compared to using a single model, ensemble approaches often provide more precise findings. One way in which ensembles might increase performance is by lowering the variance of the prediction mistakes made by the causal models [9].

Plans call for ML methods to be used in the identification of tumours inside the body. To identify tumour cells and retrieve sensitive values, ML models are used after feature extraction [10]. In order to find or reduce the likelihood of screening for carcinoma malignancy, pre-analysis will be useful. Smoking, alcohol use, obesity, and other risk factors showed a statistically significant impact at the pre-analysis level [11]. The vast array of categorization concerns essentially encompasses the diagnostic and prognostic challenges associated with cancer [9].

Proposed System

To aid in future decision-making, machine learning supervised classification algorithms will be used to analyse datasets and identify patterns; these patterns will then be used to forecast which patients are most likely to be impacted. Gathering facts The data set used to forecast network assaults is divided into two parts: the training set and the test set. The Training set and the Test set are often divided in a 7:3 ratio. Applying the data model developed using the ensemble learning model to the training set allows one to make predictions about the test set depending on how well the model performed in the tests.

Compares well-known machine learning techniques, which boosts accuracy.

Machine learning methods' potential for attribute prediction-based cancer detection in real-world settings is the subject of these publications.

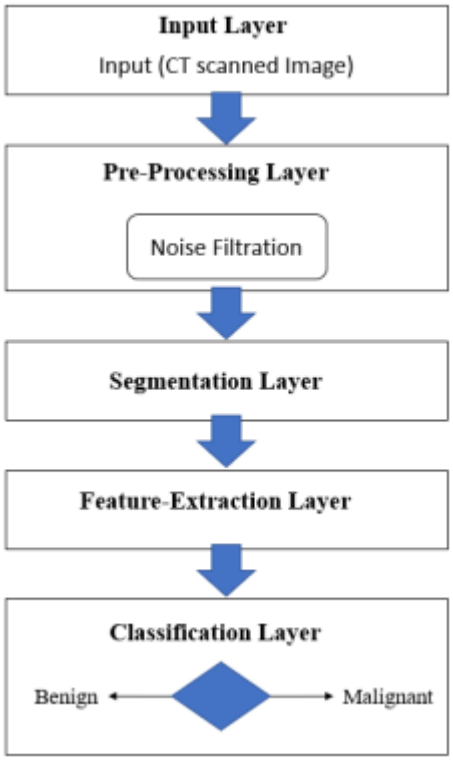

Figure 1. Process of Predicting cancer from images

In order to improve the picture quality from every angle, pre-processing is often used. Eliminating unwanted noise and preparing the picture for further processing are additional examples of pre-processing tasks.removing undesired parts using a variety of filters and methods . Because there is noise in the dataset photographs and certain areas are unnecessary. After removing unwanted edges from CT scan pictures, we scaled them using OpenCV and performed the Gaussian Blur noise filtering technique to remove noise from the input image.

When you divide a picture into smaller, more manageable pieces, you're engaging in image segmentation. To segment the picture, we use the Otsu thresholding approach, which binarizes the image and provides us with the threshold value suitable for binarization. We next segment the picture to get the right area by using that threshold value. Based on the intensity of the specified location, we were able to create a threshold value. Then, because the margins of the border might provide more noise to the picture, we eliminate them. The next step is to extract features from the acquired area once we've generated labels for it.

Once the picture has been segmented, it is sent to an extraction layer. From the labelled areas, several region-based parameters are retrieved, including area, perimeter, centroid, mean intensity, solidity, and eccentricity. A picture's area is equal to its pixel count.

An inbuilt function computes the perimeter by calculating the distance between each adjoining pair of pixels around the border of the region. The centroid, in this context, is the centre of mass of each region, represented by a 1x2 vector. The mean is the average intensity. Solidity is the ratio of pixels in the region to pixels of the convex hull. Eccentricity is the ratio of focal distance over the major axis length. A value of 0 indicates a circle and a value of 1 indicates a straight line. Classification data is derived from these retrieved characteristics.

Following the feature extraction step, we can see how the classification approach uses specialised ensemble methods to determine whether the tumour is benign or malignant. It is simple to determine whether the tumour is malignant after using the category technique, and the outcome provides accurate prediction [15]. The following are examples of basic ensemble classifiers built using various ML models.

Using the provided input data, SVM [12] performs prediction, regression, and classification. It divides the input dataset into its constituent parts using a boundary known as a hyper-plane, and then uses those parts to classify the dataset. By using SVM, linear and non-linear regions may be distinguished.

To distinguish between impacted and unaffected regions of a picture, one may use a linear separation classifier [14]. By expressing the non-linear type, we will isolate the impacted element or location in non-linear separation.

One well-known modelling approach used to assess epidemiologic datasets is Logistic Regression [13] (LR). After learning the LR model's coefficients, the LR method produces reliable predictions after first determining the use of logistic features. One name for LR is a binary classifier, and it uses the sigmoid function, which is mostly dependent on guesswork, to determine the classification category. An MLP is a kind of feed-forward neural network that takes in very rapid data and produces very quick outputs. In order to train the community, MLP employs backpropagation. The neural network uses backpropagation methods to generate predictions. Backpropagating the faults allows us to fix the edge weights for a specific instance. Input, hidden, and output layers make up the neural network model. In most cases, the representation of the dataset's feature count is given by the number of input layer neurons. According to the need, there may be n hidden layers with n neurons each, and the number of neurons on the output layers is typically equal to the number of classes or labels in the dataset. An example of a classification statistics method is K-Nearest Neighbour [18].

The highlighted house's 'K' closest training samples serve as associate input. The class of the associates around the associate item determines its predicted class.

Conclusion

Here, we present a new method for CT-scan lung cancer detection that uses an ensemble classifier and compares its performance to that of an RF classifier. We used five different machine learning models—SVM, LR, MLP, Decision-tree, and KNN—in Ensemble-Classifier. The suggestedtechnique provides in-depth understanding of how to forecast lung cancer using a CT scan. We prepare the dataset for training and testing by extracting features based on regions, and then we divide it in half. We use the confusion matrix to determine whether a malignancy is malignant or benign, and then we compile a report detailing our classification process, which includes metrics like recall, accuracy, precision, and F1-score. Additionally, we created ROC curves for both models. We may conclude that the constructed ensemble classifier can distinguish between malignant and benign cancers, since the accuracy of the Ensemble-Classifier is 85%. The recall value shows that the model has properly recognised the highest number of malignant tumour instances, and the matrices generated by the Ensemble Classifier are quite close to the ones obtained by random forest.

References

[1] Torre L A, Siegel R L, Jemal A. Lung Cancer Statistics[J]. Advances in Experimental Medicine & Biology, 2015, 893:1-19.

[2] Chen W, Zheng R, Baade P D, et al. Cancer statistics in China, 2015[J]. CA:A Cancer Journal for Clinicians, 2016, 66(2):115-132.

[3] Chen W, Zheng R, Zeng H, et al. Epidemiology of lung cancer in C hina[J]. Thoracic cancer, 2015, 6(2): 209-215.

[4] None. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening[J]. New England Journal of Medicine, 2011, 365(5):395-409.

[5] de CarvalhoFilho A O, Silva A C, de Paiva A C, et al. 3D shape analysis to reduce false positives for lung nodule detection systems[J]. Medical & biological engineering & computing, 2017, 55(8): 1199-1213.

[6] Yuan S, Ying W, Dazhe Z. Computer-Aided Lung Nodule Recognition by SVM Classifier Based on Combination of Random Under sampling and SMOTE[J]. Computational & Mathematical Methods in Medicine, 2015, 2015:1-13.

[7] Robert D. Ambrosini, Peng Wang, Walter G. O'Dell. Volume change determination of metastatic lung tumors in CT images using 3-D template matching[J]. Proceedings of SPIE - The International Society for Optical Engineering, 2009, 7260.

[8] Ashwin S, Ramesh J, Kumar S A, et al. Efficient and reliable lung nodule detection using a neural networkbased computer aided diagnosis system[C]//2012 International Conference on Emerging Trends in Electrical Engineering and Energy Management (ICETEEEM). IEEE, 2012: 135-142.

[9] Tan M, Deklerck R, Jansen B, et al. A novel computer-aided lung nodule detection system for CT images[J]. Medical Physics, 2011, 38(10): 5630-5645.

[10] Lavanya K, Durai M A S, Iyengar N. Fuzzy rule based inference system for detection and diagnosis of lung cancer[J]. International Journal of Latest Trends in Computing, 2011, 2(1): 165-171.

[11] Kaya A, Can A B. A weighted rule based method for predicting malignancyof pulmonary nodules by nodule characteristics[J]. Journal of Biomedical Informatics, 2015, 56: 69-79.

[12] Kim B C, Yoon J S, Choi J S, et al. Multi-scale gradual integration CNN for false positive reduction in pulmonary nodule detection[J]. Neural Networks, 2019, 115: 1-10.

[13] Lin-Lin W, Chun-Lei X. Hybrid method of image segmentation usingwatershed transform and improved FCM[J]. Computer Engineering & Applications, 2010, 46(14):189-191.

[14] Armato S G, Giger M L, Moran C, et al. Computerized Detection ofPulmonary Nodules on CT Scans[J].Radiographics, 1999, 19(5):

[15] Okada K, Comaniciu D, Krishnan A. Robust anisotropic Gaussian fittingfor volumetric characterization of Pulmonary nodules in multislice CT[J]. IEEE Transactions on Medical Imaging, 2005, 24(3): p.409-423.

[16] Kostis W J, Reeves A P, Yankelevitz D F, et al. Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images[J]. IEEE Transactions on Medical Imaging, 2003,2(10):1259-1274.

[17] Xie H, Yang D, Sun N, et al. Automated pulmonary nodule detection inCT images using deep convolutional neural networks[J]. Pattern Recognition, 2019, 85: 109-119.

[18] Ding J, Li A, Hu Z, et al. Accurate Pulmonary Nodule Detection in Computed Tomography Images Using Deep Convolutional Neural Networks[C]International Conference on Medical Image Computing and ComputerAssistedIntervention. Berlin: Springer, 2017:559- 567.

[19] Dou Q, Chen H, Jin Y, et al. Automated Pulmonary Nodule Detection via 3D ConvNets with Online Sample Filtering and Hybrid-Loss ResidualLearning[C] International Conference on Medical Image Computing and ComputerAssisted Intervention. Berlin: Springer, 2017: 630- 638.

[20] Lv X Q, Wu L, Gu Y, et al. Detection of low dose CT pulmonary nodules based on 3D convolution neural network[J]. Optics and Precision Engineering, 2018, 26(5): 1211-1218