LUNG CANCER DETECTING USING MACHINE LEARNING

Jayesh Koli
computer engineering department Sandip institute of technology and research centre Nashik, India
jayeshmkoli13@gmail.com

Prof. Pramod.G. Patil
Computer engineering departmentSandip institute of technology andresearch centre Nashik, India pgpatil11@sitrc.og

Krutika karad
Computer engineering departmentSandip institute of technology and research centre Nashik, India
krutikarad1821@gmail.com

Atharv Borse
Computer engineer department Sandip institute of technology and research centre Nashik, India
atharvaborse30@gmail.com

Pranav Wagh
Computer engineer department Sandip institute of technology and research centre Nashik, India
pranavwagh7666@gmail.com

Abstract - The Main Objective of this research paper is to find out the early stage of lung cancer and explore the accuracy levels of various machine learning algorithms. After a systematic literature study, we found out that some classifiers have low accuracy and some arehigher accuracy but difficult to reached nearer of 100%. Low accuracy and high implementation cost due to improper dealing with DICOM images. For medical image processing many different types of images are used but Computer Tomography (CT) scans are generally preferred because of less noise. Deep learning is proven to be the best method for medical image processing, lung nodule detection and classification, feature extraction and lung cancer stage prediction. In the first stage of this system used image processing techniques to extract lung regions. The segmentation is done using K Means. The features are extracted from the segmented images and the classification are done using various machine learning algorithm. The performances of the proposed approaches are evaluated based on their accuracy, sensitivity, specificity and classification time.

Key Words*: *Structural Co-occurrence Matrix (SCM), Classifier, Data Set, ROC curve, Malignant nodule, Benign nodule.

1. INTRODUCTION

The cause of lung cancer stays obscure and prevention become impossible hence the early detection of lung cancer is the onlyone way to cure. Size of tumour and how fast it spread determine the stage of cancer [1]. Lung cancer spreading widely all over the world. Death and health issue in many countries with a 5-year survival rate of only 10–16% [2][3]. In some cases, the nodules are not clear and required a trained eye and considerable amount of time to detect. Additionally, most pulmonary nodules are not cancerous as they can also be due to non-cancerous growths, scar tissue, or infections [4]. Even though many researchers use machine learning frameworks. The problem with these methods is that, in order to evaluate the best performance, many parameters need to be hand-crafted which is making it difficult to reproduce the better results [5]. Classification is an important part of computation that sort images into groups according to their similarities [6][7]. In the structure of cancer cell, where most of the cells are overlapped with each other. Hence early detection of

cancer is more challenging task [8][9]. After an extensive study, we found that ensemble classifier was performed well when compared with the other machine learning algorithms [10]. The existing CAD system used for early detection of lung cancer with the help of CT images has been unsatisfactory because of its low sensitivity and high False Positive Rates (FPR).

2.LITERATURE REVIEW:

In paper [11] Pankaj Nanglia, Sumit Kumar et all proposed a unique hybrid algorithm called as Kernel Attribute Selected Classifier in which they integrate SVM with Feed-Forward Back Propagation Neural Network, which helps in reducing the computation complexity of the classification. For the classification they proposed three block mechanisms, pre-process the dataset is the first block. Extract the feature via SURF technique followed by optimization using genetic algorithm is the second block and the third block is classification via FFBPNN.

The overall accuracy of the proposed algorithm is 98.08%.In paper [12] Chao Zhang,Xing Sun, Kang Dang et all perform a sensitivity analysis using the multicenter data set. They chosen two categories Diameter and Pathological result. Diameter were divided into three sub groups.0-10mm,10-20mm,20-

30mm. In 0-10mm group sensitivity 85.7% (95% Cl,70.8%-100.0%) and specificity 91.1% (95% Cl, 86.8%-95.2%) were found. In 10-20mm group sensitivity 85.7% (95% Cl,77.1%-94.3%) and specificity 90.1% (95% Cl, 84.8%-95.4%) . found. The algorithm had provided the highest accuracy of 85.7% for adenocarcinoma and 65.0% for Squamous cell carcinoma.

In paper [13] Nidhi S. Nadkarni and Prof. Sangam Borkar focuses their study mainly on the classification of lung images as normal and abnormal In their proposed method median filter was used to eliminate impulse noise from the images. Mathematical morphological operation enables accurate lung segmentation and detect tumour region. Three geometrical features i.e. Area, perimeter, eccentricity was extracted from segmented region and fed to the SVM classifier for classification.

In paper [14] Ruchita Tekade, Prof. DR. K. Rajeswari studied the concept of lung nodule detection and malignancy level prediction using lung CT scan images. This experiment has conducted using LIDC_IDRI, LUNA16 and Data Science Bowl 2017 datasets on CUDA enabled GPU Tesla K20. The Artificial Neural Network used to analyze the dataset, U-NET architecture for segmentation of lung nodule

from lung CT scan images and 3D multigraph VGG like architecture for classifying lung nodule and predict malignancy level. Combining these two approaches have given the better results. This approach given the accuracy as 95.66% and loss 0.09 and dice coefficient of 90% and for predicting log loss is 38%. In paper [15] Moffy Vas, Amita Dessai, studied mainly on the classification of lung images cancerous and non-cancerous.

In paper [15] Moffy Vas, Amita Dessai, studied mainly on the classification of lung images cancerous and non-cancerous. In their proposed method pre-processing was done, in which unwanted portion of the lung CT scan was removed. They used median filter to eliminate salt and pepper noise. Mathematical morphological operation enables accurate lung segmentation and detect tumour region. Seven extracted features i.e. energy, correlation, variance, homogeneity, difference entropy, information measure of correlation and contrast respectively was extracted from segmented region and fed to the feed forward neural network with

back propagation algorithm for classification. The algorithm looks for the least of the error function in the weight space gradient descent method. The weights are shuffled to minimise the error function. The training accuracy was 96% and testing accuracy was 92%. The sensitivity was 88.7% and specificity was 97.1%. In paper [16] Radhika P R, Rakhi.A.S.Nair, mainly focused on prediction and classification of medical imaging data. They used UCI Machine Learning Repository and data.world. dataset. Used various machine learning algorithm for comparative study and found that support vector machine gives higher accuracy 99.2%. Decision Tree provide 90%, Naïve Bayes provide 87.87% and Logistic Regression provide 66.7%. In paper [17] Vaishnavi. D1, Arya. K. S2, Devi Abirami. T3 , M. N. Kavitha4, studied on lung cancer detection algorithm. In pre-processing they used Dual-tree complex wavelet transform (DTCWT)in which the wavelet is discretely sampled. GLCM is second order statistical method for texture analysis which provide a tabulation of how different combination of Gray level

co-occur in an image. It measures the variation in intensity at the pixel of interest. They used Probability Neural Network (PNN) classifier evaluated in term of training performance and classification accuracy. It gives fast and accurate classification.

In paper [18] K.Mohanambal , Y.Nirosha et al studied structural co-occurrence matrix (SCM) to extract the feature from the images and based on these features categorized them into malignant or benign. The SVM classifier is used to classify the lungnodule according to their malignancy level (1 to 5).co-occur in an image. It measures the variation in intensity at the pixel of interest. They used Probability Neural Network (PNN) classifier evaluated in term of training performance and classification accuracy. It gives fast and accurate classification.

In paper [18] K.Mohanambal , Y.Nirosha et al studied structural co-occurrence matrix (SCM) to extract the feature from the images and based on these features categorized them into malignant or benign. The SVM classifier is used to classify the lungnodule according to their malignancy level (1 to 5).

SYSTEM MODEL
DATA EXPLORATION:

Three datasets are used in this research containing labelled nodules positions image.

1. TCIA Dataset:

The cancer imaging archive (TCIA) host collection of de-identified medical images, primarily in DICOM format. Collectionsare organized according to disease and image modality (such as MRI or CT). CT images data used to support the findings of this study have been depositedinthe(doi.org/10.7937/K9/TCIA.2015.A6V7JIWX).

1. Lung Image Database Consortium Image Collection (LIDC-IDRI) consists of lung CT scans of 1018 patients (124GB) in DICOM format. Four experienced radiologists independently reviewed the lung CT scans and annotated the nodules in the dataset.

2. ALGORITHMS AND TECHNIQUES

1. The U-Net Convolutional Network is used for biomedical image segmentation. It takes an input image and an output mask of the region of interest. It first generates a vector of features typically in a convolutional neural network, and then use another up- convolutional neural network to predict the mask given by the vector of features [20][21][22]. This is a binary classification task using morphological and radiological features extracted from the images and masks. Logistic regression is particularly strong in binary

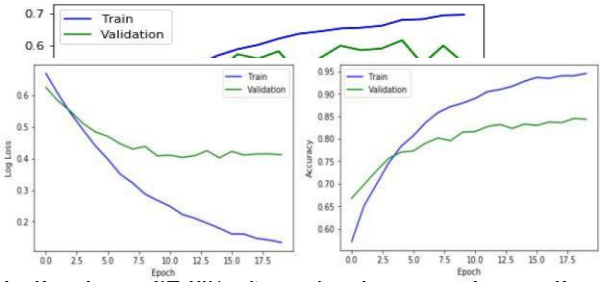
classification which provide top candidate model for completion of this task. Gaussian Naïve Bayes is suitable for the continuous numerical features. It takes the mean and variance for each feature in each class .

MODEL EVALUATION AND VALIDATION

Model 1: U-Net Convolutional Neural Network for nodule segmentation

RESULT DISCUSSION:

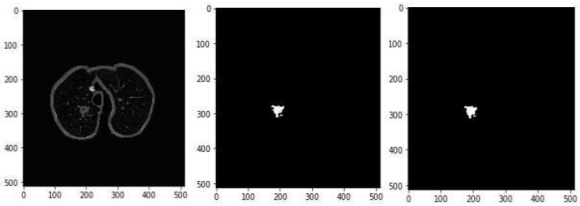
The data was split into 80% training and 20% validation set with a train test split function. Due to the long training time of 3 hours for 2 epochs, a cross validation was not performed. The U-net model converged in 10 epochs and give a dice coefficientof 0.678 which indicating a 67.8% overlap between the predicted nodule masks and ground truth nodule masks. However, there was 78% percentage of predicted masks that have at least 1 pixels of overlap with the ground truth masks. The objective of this research is to accurately detect the position of the nodules, the sensitivity and the number of false positives rate per scan [30][31][32]. There were a large number of FP per TP which is further reduced in the second model below.



*indicating a 67.8% of overlap between the predicted nodule masks andground truth nodule masks.*Model 2: Convolutional Neural Network for reducing false positives of detected nodules

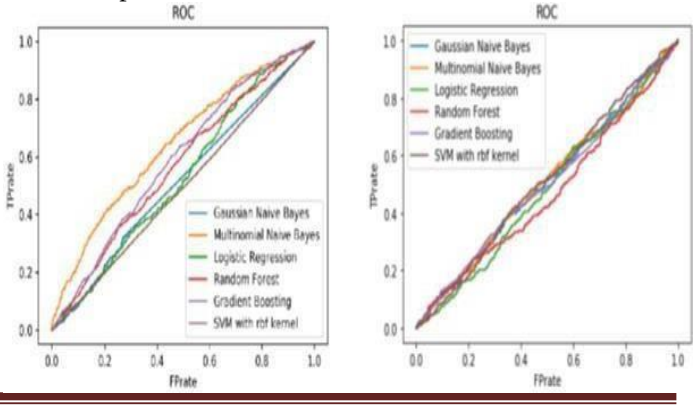
Model 3: Classification of cancer or non-cancer with handpicked features

The final features selected as predictors included Diameter, Spiculation, MeanHU, and Eccentricity. This



was determined through A/B testing to find the combination of features that performed the strongest on the best performing model. The classification of cancer with classifier using handpicked features performed stronger than the CNN at a logloss of approximately 0.55, an AUC of 0.64, and an average precision of 0.41. In comparison, these models trained with random labels achieved a logloss of 0.59, AUC of 0.50 and an average precision of 0.29 [33][34]. The probability of cancer in the dataset is 0.26, so the stratified random labels performed similarly to the proportion of classes while the true labels performed substantially better.

Multiple classifiers performed well, similarly after they were optimized with a grid search algorithm. This shows that these models performing similarly in its ability to exploit the information in the input features to make its predictions [35][36]. Furthermore,



transforming the training data into discretized categories by rounding resulted in less than a 0.05% increase in logloss, indicating the robustness of these models.

Model	Log Loss True Label	Log Loss Random Label	AUC_True Label	AUC_Random Label	Average Precision_TL	Average Precision_RL
Gaussian Naïve Bayes	0.5850	0.8037	0.6380	0.5053	0.4145	0.2929
Multinomial Naïve Bayes	0.5528	0.5920	0.6457	0.5050	0.4100	0.2093
Logistic Regression	0.5525	0.5939	0.6548	0.4823	0.4132	0.2655
Random Forest	0.5533	0.6038	0.6150	0.4681	0.3769	0.2624
Gradient Boosting	0.5672	0.5964	0.6173	0.5019	0.3274	0.2862
SVM-rbf kernel	0.5893	0.5931	0.5017	0.5108	0.2514	0.3787
Ensemble*	0.5519		0.6459		0.4133	

Table 2 Different Models are compared between True labels and Random labels

Model 4: Convolutional Neural Network for cancer or non-cancer prediction with detected nodules.
The CNN model reached a validation loss of 0.5646 and an AUC of 0.6231. This is similar but marginally worse than the bestperformance of the classifiers with handpicked features. This may be due to diameter being the strongest parameter to detect cancer. CNNs are designed to be size and scale invariant, but rather focus on the features.

CONCLUSION

CAD system for lung cancer includes the stages of pre-processing, nodule detection, nodule segmentation, feature extraction and classification of the nodule as benign or malignant. Once the nodules are detected and segmented the feature extraction process begins. The features necessary for classification are extracted using feature extraction techniques from the segmented nodule. Based on the features extracted, a classifier is used for classifying the nodule as benign or malignant. The performance

of both the CNN and classifiers were similar, with the classifiers performing slightly better. Compared to the performance of radiologists, the sensitivity of nodule detection was within the range of radiologists at 65% with the two stage neural networksvs 51-81.3% with radiologists. The false positive rate is much higher than the neural networks which is at 6.78 false positives.

REFERENCES:

[1] N.Camarlinghi, “Automatic detection of lung nodules in computed tomography images: Training and validation of algorithms using public research

[2] R. L. Siegel, K. D. Miller, and A. Jemal, ``Cancer statistics, 2016”, CA,Cancer J. Clin., vol. 66, no. 1, pp. 730, 2016.

[3] Detecting and classifying nodules in Lung CT scans, **http://modelheelephant.blogspot.com/2017/11/detecting-and-classifying-nodules-in.html,2017**.

[4] Diego Riquelme and Moulay A. Akhloufi ,”Deep Learning for Lung Cancer Nodules Detection and Classification in CT Scans”,www.mdpi.com,2020.

[5] Anita Chaudhary, Sonit Sukhraj Singh,”Lung Cancer Detection on CT Images by using Image Processing”,IEEE,2012.

[6] Gawade Prathamesh Pratap, R.P. Chauhan, “Detection of Lung Cancer Cells using Image Processing Techniques”, International Conference on Power Electronics, Intelligent Control and Energy Systems(ICPEICES),2016.

[7] Pooja R. Katre,Dr. Anuradha Thakare,“Detection of Lung Cancer Stages using Image Processing and Data Classification Techniques”,International Conference for Convergence in Technology,IEEE,2017

[8] Rituparna Sarma, Yogesh Kumar Gupta“A comparative study of new and existing segmentation techniques”,ICCRDA,2020.

[9] Eali Stephen Neal Joshua1*, Midhun Chakkravarthy1, Debnath Bhattacharyya2 ,” An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: A Systematic Study”,International Information and Engineering Technology Association (IIETA),2020.

[10] Pankaj Nanglia,Sumit Kumar, Aparna N. Mahajan, Paramjit Singh, Davinder Rathee,“A hybrid algorithm for lung cancer classification using SVM and Neural Networks”,The Korean Instituteof Communicationand Information Science(KICS),2020. Also available at www.elsevier.com/locate/icte.

[11] Chao Zhang,Xing Sun, Kang Dang et all "Toward an Expert Level of Lung Cancer Detection and Classification