Lung Cancer Detection and Classification using Machine Learning Algorithm

Meraj Begum Shaikh Ismail
Department of Computer Science & IT,Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad -431 004 (MS) India
merajfatema01@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 are higher 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.

Keywords— Structural Co-occurrence Matrix (SCM), Classifier, Data Set, ROC curve, Malignant nodule, Benign nodule.

I. INTRODUCTION

The cause of lung cancer stays obscure and prevention become impossible hence the early detection of lung cancer is the only one 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).

II. 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%) were found. In 20-30mm group sensitivity 78.9% (95% Cl,66.0%-91.8%) and specificity 91.3% (95% Cl, 83.2%-99.4%) were

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, extracting feature and classification purpose. They used 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 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 lung nodule according to their malignancy level (1 to 5).

III. SYSTEM MODEL

A. DATA EXPLORATION

Three datasets are used in this research containing labelled nodules positions for image segmentation and cancer/non-cancer labels for classification [19].

1. TCIA Dataset

The cancer imaging archive (TCIA) host collection of de-identified medical images, primarily in DICOM format. Collections are 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 deposited in the Lung CT-Diagnosis repository (doi.org/10.7937/K9/TCIA.2015.A6V7JIWX).

2. 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.

3. Kaggle data science bowl 2017 provides lung CT scans of 1595 patients (146GB) in DICOM format and having a set of labels, which denote that if the patient was diagnosed with lung cancer in future, even one year after the scan were taken.

B. ALGORITHMS AND TECHNIQUES

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. The features are continuous and numerical, but can be discretized into categories. The following classifiers were explored [23][24][25].

1. Logistic regression is particularly strong in binary classification which provide top candidate model for completion of this task.

2. Gaussian Naïve Bayes is suitable for the continuous numerical features. It takes the mean and variance for each feature in each class [26].

3. Multinomial Naïve Bayes required the categorical data. In this feature transformed into discrete steps. This may be more suited than Gaussian NB since some of the feature distributions representing a class is not normally distributed. For example, diameter with non-cancer is strongly skewed to the left [27].

4. Support Vector Machines draws a separation line that maximizes the points representing the classes in a multidimensional feature space. A kernel trick can be used to fit a more defined boundary [28].

5. Random Forest frequently used on kaggle for classification tasks. It creates many decisions tress with random samples and features and takes a vote on its output. This is used to prevent overfitting.

6. Gradient Boosting also frequently used on kaggle for classification tasks. It's similar to Random Forest but instead of random samples for each tree, it takes the samples with the highest error on the previous tree to train the successive trees.

7. Ensemble classifiers are created by averaging the output of several of the above models.

C. MODEL EVALUATION AND VALIDATION

Model 1: U-Net Convolutional Neural Network for nodule segmentation [29].

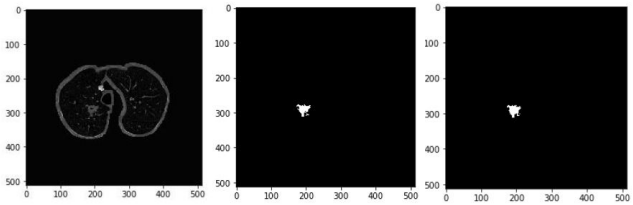

Figure 1 U-Net image segmentation. Processed CT image (left), ground truth label (center), predicted label (right)

CI. 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 coefficient of 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.

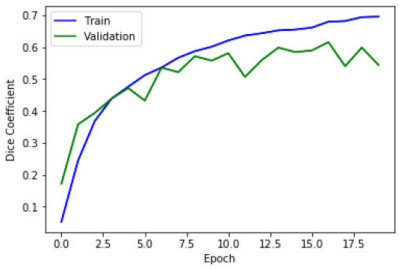

Figure 2 A dice coefficient of 0.678 was reached, indicating a 67.8% of overlap between the predicted nodule masks and ground truth nodule masks.

Model 2: Convolutional Neural Network for reducing false positives of detected nodules

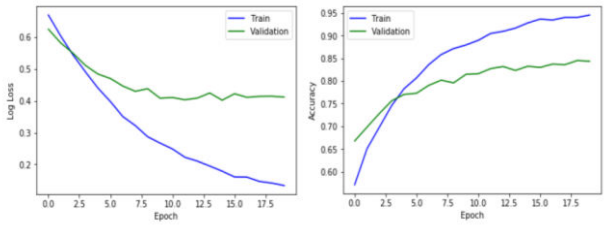

Figure 3 CNN converges to a validation accuracy of 84.4% at classifying a detected nodule as TP or FP

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.

	Sensitivity	**Average of** FPs per scan	FPs per TP
Before nodule classification	***0.75***	***0.060***	***11.1***
After nodule classification	***0.65***	***0.011***	***2.32***

Table 1 shows Sensitivity, TP and FP rates per scan

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.

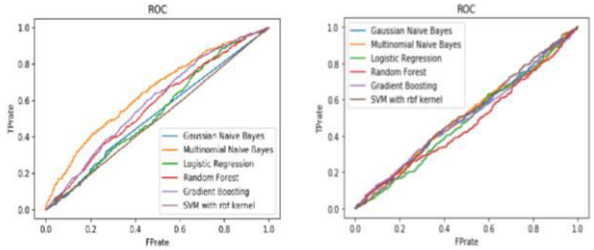

Figure 4 ROC plots True Positive Rate and False Positive Rate for true labels (left), and random labels (right).

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 best performance 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.

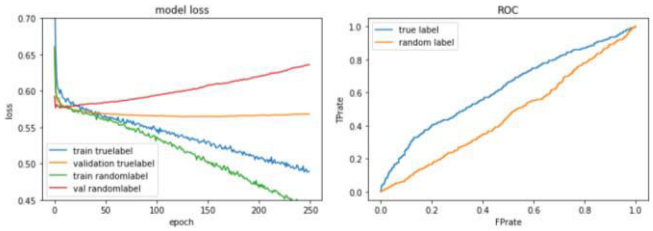

Figure 5 Model loss for training and validation is compared between true labels and random labels (left) ROC curve is substantially improved for true labels compared to random labels (right)

CII. 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 networks vs 51-81.3% with radiologists. The false positive rate is much higher than the neural networks which is at 6.78 false positives per case with the neural networks vs 0.33-1.39 false positives per case with radiologists. Despite the large number of false positives rate, by solely using the largest nodule detected for cancer prediction. The precision with the classifiers is substantially higher at 41% compared to 1-2% by radiologists.

REFERENCES

[1] Smita Raut1, Shraddha Patil2, Gopichand Shelke",Lung Cancer Detection using Machine Learning Approach", International Journal of Advance Scientific Research and Engineering Trends(IJASRET),2021.

[2] N.Camarlinghi, "Automatic detection of lung nodules in computed tomography images: Training and validation of algorithms using public research databases", Eur. Phys. J. Plus, vol. 128, no. 9, p. 110,Sep. 2013.

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

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

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

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

[7] 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.

[8] 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

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

[10] 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.

[11] 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 Institute of Communication and Information Science(KICS),2020. Also available at www.elsevier.com/locate/icte.

[12] Chao Zhang,Xing Sun, Kang Dang et all "Toward an Expert Level of Lung Cancer Detection and Classification Using a Deep Convolutional Neural Network",The Oncologist,2019. Also available at www.TheOncologist.com.

[13] Nidhi S. Nadkarni and Prof. Sangam Borkar, "Detection of Lung Cancer in CT Images using Image Processing", Proceeding of the Third International Conference on Trends and Informatics (ICOEI),IEEE,2019.

[14] Ruchita Tekade, Prof. DR. K. Rajeswari, "Lung Cancer Detection and Classification using Deep Learning",Fourth International Conference on Computing Communication Control and Automation(ICCUBEA),IEEE,2018

[15] Moffy Vas, Amita Dessai, "Lung Cancer detection system using lung CT image processing",IEEE,2017

[16] Radhika P R, Rakhi.A.S.Nair, "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms",IEEE,2018

[17] Vaishnavi. D1, Arya. K. S2, Devi Abirami. T3 , M. N. Kavitha4, "Lung Ancer Detection using Machine Learning", International Journal of Engineering Research & Technology (IJERT),2019.

[18] K.Mohanambal1, Y.Nirosha2, E.Oliviya Roshini3, S.Punitha4, M.Shamini5,"Lung Cancer Detection Using Machine Learning Techniques",IJAREEIE,2019

[19] Pragya Chaturvedi,Anuj jhamb,Meet Vanani,Varsha Nemade,"Prediction and Classification of Lung Cancer using Machine Learning Techniques",ASCI,2020.

[20] Prathyusha Chalasani, S.Rajesh,"Lung CT Image Classification using Deep Neural Networks for Lung Cancer Detection", International Journal of Engineering and Advanced Technology (IJEAT),2020.

[21] A Asuntha, Andy Srinivasan,"Deep learning for lung Cancer detection and classification", springer.com, 2020.

[22] GAP Singh,PK GUPTA,"Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans",Springer.com,2017.

[23] Kanchan Pradhan,Priyanka Chawla,"Medical internet of things using machine learning algorithms for lung cancer detection",Journal of Management Analytics,2020.

[24] Abdelhameed Ibrahim,Shaimaa Mohammad,Hesham Arafat Ali, "Breast cancer detection and classification using thermography: a review ",International Conference on Advanced Machine Learning Technologies and application, 2018.

[25] Preeti Katiyar, Krishna Singh,"A comparative study of lung cancer detection and classification approaches in CT images",International Conference on Signal processing and Integrated Networks(SPIN),IEEE,2020.

[26] Ozge Gunaydin, Melike Gunay, Oznur Sengel,"Comparison of lung cancer detection algorithms ", Scientific Meeting on Electricals, Electonics and Biomedical Engineering and Computer Science (EBBT),IEEE,2019.

[27] Mr. Sandeep,A.Dwivedi, Mr.R.P.Borse,,"Lung Cancer Detection and Classification by using Machine Learning & Multinomial Bayesian", (IOSR-JECE),2014.

[28] Wasudeo Rahane, Himali Dalvi, Yamini Magar,Anjali Kalane"Lung Cancer Detection using Image Processing and Machine Learning HealthCare",International Conference on Current Trends toward Converging Technologies,IEEE,2018.

[29] Wafaa Alakwaa, Mohammad Naseef,Amr Badr,"Lung Cancer Detection and Classification with 3D Convolutional Neural Network (3D-CNN)",(IJACSA),2017.

[30] Sanjukta Rani Jena, S Thomas George, D Narain Ponraj, "Lung cancer detection and classification with DGMM-RBCNN", Springer.com,2021.

[31] Jun Sang, Mohammad S Alam, Hong Xiang,"Automated detection and classification for early stage lung cancer on CT images using deep learning", Pattern Recognition and Tracking,2019.

[32] Saadaldeen Rashid Ahmed Ahmed,Isra Al-Barazanchi,Ammar Mhana, Haider Rasheed Abdulshaheed "Lung cancer classification using data mining and supervised learning algorithm on multi-dimensional data set", Periodical of Engineering and Natural Science (PEN),2019.

[33] Kun-Hsing Yu,T sung-Lu Michael Lee,Ming-Hsuan Yen,Sckou et all, "Reproducible Machine Learning Methods for Lung Cancer Detection using Computed Tomography Images: Algorithm Development and Validation", Journal of Medical Internet Research (JMIR),2020

[34] Anam Masood, Binsheng, po Yang, Pingli, Hauting Li et all, "Automated decision support system for lung cancer detection and classification via enhanced RFCN with multilayer fusion RPN ", Transaction on Industrial Informatics, IEEE,2020.

[35] Maxim D Podolsky, Anton A Barchuk, Vladimir I,Gusarova et all,"Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels",Asian Pasific Journal of Cancer Prevention,2016.

[36] Tanzila Saba,"Automated lung nodule detection and classification based on multiple classifiers voting ", Microscopy Research and technique,2019.