Detection and Classification of Lung Nodule in CT Scan Images Using CNN

Azhar Hussain¹, Javeria Jabeen², Muhammad Wajid³*, Muhammad Haroon⁴

¹Department of Computer Science, MNS University of Agriculture, Pakistan. Email: azharhussain1049@gmail.com
²Department of Computer Science, MNS University of Agriculture, Pakistan. Email: javeria.jabeen@mnsuam.edu.pk
³Department of Computer Science, NFC Institute of Engineering & Technology Multan, Pakistan. Email: wajid.gm495@gmail.com
Muhammad Haroon⁴
⁴Department of Computer Science, Xi’an University of Technology, Xi’an, Shaanxi, China. Email: 1201214002@stu.xaut.edu.cn

*Corresponding author: wajid.gm495@gmail.com

Abstract

The Identification of lung malignant cells at premature stage remains as remarkable research area for re-searchers. Lung nodules are account as a normal reason for death in people all over the world. Identification of lung nodules at the initial stage can build the pace of endurance of patients. In this article computed tomography (CT) scan images as input to classify lung malignant cells of non-small cell lung cancer and categories according to subtypes of cancer by using the image processing technique with a convolutional neural network (CNN). The segmentation of the CT scan is performed to simplify the representation for meaningful and easier analysis. The feature extraction technique is used which is independent from the size and rotation of the image. These extracted features will use as an input to convolutional neural network (CNN), it passes through different layer of neural network at the end output layer classifies the affected and non-affected areas of the lung nodule with respect to subtypes (adenocarcinoma, squamous cell carcinoma, and large cell carcinoma). The deep learning algorithms VGG 16 show accuracy 85.01% while another deep learning algorithms VGG 19 show best accuracy 98.74% to correctly classify the nodules and subtypes.
**Keywords:** Computed Tomography scan, Deep Learning, Lungs Cancer, Lungs Nodules, None-Small Cell Lungs Cancer.

1. **Introduction:**

According to American cancer society, over 150,000 lung cancer patients succumb. The motility rate is the highest in all natures of malignance. Lung tumor is one of the most perilous illnesses on the planet. More people are dying because of lung nodule related to other malignance diseases like breast, brain, and prostate cancer. The overall death in the world related to cancer among 25% is caused by cellular breakdown in lung which is highest related to other cancer diseases (www.cancer.org). We can decrease this ratio by detecting cancer at the early stage using image processing. Image processing is used in a different area of the medical field to improve performance and accuracy. Image processing is assumed to be as fundamental part in the identification of lung tumor. Lung tumor is categories into two main classes’ small cell and non-small cell lung tumor [1]. We can use image processing on a CT scan to detect the non-small lung nodules that is the most well-known kind of malignance in lung. It is nearly nine out of ten cases.

Particle pollution, such as that caused by exhaust smoke, raises the risk of lung cancer. Some genetic factor also may play a role for developing lung cancer. If you have a family history of lung cancer, you are at risk of developing the illness [2].

Over 150,000 lung cancer patients die each year, according to the American Cancer Society. The motility rate is highest in all natures of malignance [3]. Lung tumor is one of the most perilous illnesses on the planet. More peoples are dying because of lung nodule related to other malignance diseases like breast, brain, and prostate cancer. The overall death in the world related to cancer among 17.6% is caused by cellular breakdown in lung which is highest related to other cancer diseases [3].

We can decrease this ratio by detecting cancer at the early stage using image processing. Image processing is used in a different area of the medical field to improve performance and accuracy. Image processing assumed as fundamental part in the identification of lung tumor. Lung tumor is categories into two main classes’ small cell and non-small cell lung tumor.

We can use image processing on a CT scan to detect the non-small lung nodules that is the most well-known kind of malignance in lung.
Detection and Classification of Lung Nodule in CT Scan Images Using CNN

It is nearly nine out of ten cases. Non-small lung cancer growth rate is slow and does not show any symptoms at the initial level. Due to the slow growth rate its nodules size is very small. These nodules at the early stage show in very small size on CT scan images. In most cases, non-small lung cancer is detected after maturing due to small.

The patient’s survival rate has decreased. The survival rate is straightforwardly dependent upon the development rate and its time when it is identified. Due to the slow growth rate of non-small lung cancer, it is time taking and or difficult to diagnose its initial level. We can use the image process technique with deep learning neural network model to diagnose cancer. Using this model, we can improve cancer diagnosis accuracy which is helpful for the survival of cancer patients. In this way, we can decrease the death ratio in cancer patients using image processing with a convolutional neural network.

2. Related Work:

There are some prior papers which are reviewed and presented here. The author uses computer aided diagnosis (CAD) to validate the opinion made by radiologist on CT scan images. The CAD system is develop using convolutional neural network to validate the CT scan results. The suggested model can predict the malignancy based on the scans provided as input [1].

Al Mohammad et al. proposed a system develop a CAD base system that can improve our screening process in medical lab. It will reduce the rate of missed nodules and assist the radiologist in detection of lung nodule with the help of computer vision base system. In this paper we analyze a performance of CAD base system and factor that effect the performance in identification of lung tumor. Al-so investigate the how CAD base system effect the radiologist in decision making [4].
WGANS base over sampling technique compare the performance with commonly use data augmentation technique and CNN. These schemes is tested on publicly available lung image database consortium dataset. The experiment shows that WGAN over sampling technique assist the CNN to boost classification performance as compared to convolution data augmented method. [5].

A temporal subtraction (TS) picture is created by subtracting a prior image from the present image, which is then twisted to match the structures of the previous and current images. The TS method re-places existing abnormalities in a medical picture by removing normal structures and enhancing interval alterations such as new lesions. Many artefacts that remain on the TS picture, however, can be identified as false positives [6].

Lung Cancer is one of most common disease it can become the cases of death. It becomes difficult for doctor to identify the cancer at benign state and later which becomes the causes of death. Three different algorithms CNN, DNN, SAE are used for classification of lung nodules. The convolutional neural network among others shows the best performance for diagnosing lung cancer with the accuracy of 84.15% [7].

The model is on a dataset provided by LIDC-IDRI. In this model use different function like error back propagation, adaptive weighting scheme to improve accuracy of nodule detection. Our proposed model multi-view knowledge base collaborative (MV-KBC) deep learning shows the accuracy 91.6% to successfully classify the nodule part form other lung in CT scan images [8]. Classify lung nodule using image processing and pattern recognition technique which is based on indexes and SVM used classifier. The presented methodology shows mean accuracy of 98.11% [9].

The present approach of ANN combines the second, third and fourth steps and supervision is involved only in defining a structure. This approach is semi supervised in which after defining the structure system segment and classify the cancer and non-cancer area of lung on the base of ANN [10]. Median method is used while converting RBG to gray scale image. To Identify the body part from binary image region props method is used while gray level co-occurrence matrix uses to extract the feature and this feature is pass to machine learning algorithms like SVM [11].

The deep belief network (DBN) is tested for classification of malignant nodules without computing the texture feature. DBN contains undirected connections between its top two layers and
downward-directed between all its lower-level layers. The experiments show that the sensitivity rate is 73.40% and specificity rate is 82.20% [12].

A quick examination of the CT image histogram is performed to determine an appropriate threshold value for improved segmentation results with K-means clustering algorithms. To avoid a false positive rate different sets of 2D and 3D characteristics are collected from nodules. The system's sensitivity to tiny nodules is 83.33 percent [13].

The structural co-occurrence matrix (SCM) based method is used for classification of lung nodule as malignant and benign as well as their degree of malignance. The three algorithms multilayer perception, SVM and k-nearest is applied to classify the benign and malignant and nodule malignant levels from one to five [14].

A GLCM-based surface descriptor is used in this technique a DCNN to train the features of nodes and a Fourier-shape descriptor to represent the heterogeneity of nodules. Our Fuse-TSD method has an AUC of 96.65%, 94.45%, and 81.24 percent, respectively [15].

3. **Data Collection:**

The dataset used in research collected from Kaggle. Kaggle is one of largest community of data scientist that provide datasets and online computing resources. The CT scan image is generated in dcm format after that it is converted to png format. The model supports images in png or jpeg format. The dataset is divided into four folders which contains the image of three different types of lung cancer which include Adenocarcinoma, Larger cell carcinoma, Squamous cell carcinoma and one-fold-er contains the normal cell data.

Each of these types is further divided into train, test and validate folder for model learning purposes. The training folder represent the training set which contains the 70% of total CT scan images, test folder represents the test set which contains the 20% of total images and validation folder represent the validation set which contains the 10% of total images which is used to validate model after its training process.

4. **Data Annotation Method:**

The accuracy of data is essential for training of model. The model is the best fit if the model is trained on data which is accurately labeled. Semi-automatic approach adopted to enhance labeling accuracy and efficiency.
Physicians perform a three-level annotating procedure. Each scan was allocated to one resident physician for initial annotation, in first round of the label is assigned. In the second round the images were forwarded to attending doctors with several years of clinical experience. The attending physicians rated the annotations as accurate or wrong in the second round, and identified any nodules that were missed in the first round. Finally, the chief physician conducted a second evaluation and reached a conclusion.

![Semi-Annotation Labeling](image)

**Fig. 2: Semi-Annotation Labeling.**

The annotations is used to train an initial detector. Because of the varied labelling standards, the detector had a sensitivity of 0.85 and a FROC score of 0.59 on the test data. The first locations of the lung nodule are provided by this basic detector. After that, the remaining annotation steps are the same as the three-level annotation steps. This method creates a closed loop. The annotation is updated on a regular basis, and the performance of lung nodule annotation improves with time.

5. **Material and Methods:**

The proposed method consists of different phases. In the first phase is data collection and in the second phase the dataset is prepared using the semi-annotation method. The dataset consists of four types of lung cancer CT scans images which include adenocarcinoma, squamous cell carcinoma, large cell carcinoma and normal lung cell images. Dataset is split into training and testing and validation by 70%, 20% and 10% of total dataset. The training dataset is given as input to neural network algorithm Visual geometric group (VGG19).
Detection and Classification of Lung Nodule in CT Scan Images Using CNN

The neural network has been proved as one of powerful tool for classification task in medical. The connection between neural network layer and multi-level feature fusion is vital for neural network model. We have to use VGG 19 which consists of 19 layer which extremely efficient algorithms for image processing. By using this CNN algorithms on data to get feature from image approximately 40,000 feature is processed per patient CT scan image and generate the output-to-output frame.

![Deep Learning Network Diagram](image)

*Fig. 3: Model Training Structure*

The deep learning model passes through three different phases training, test and validation. The model is trained on 70% of total images, the CT scan images are given as input to deep learning algorithms. The algorithms use the number of hidden layers for recognizing the edge and relevant features that are used for learning of model while training process. After that the model is fed with testing images and at the end testing is performed.

Training: The training folder contains 70% of the total image for the training of prediction model.

Test: The folder contains 20% of total images for testing purpose of trained model.

Validate: The validation folder contains the 10% of total dataset to validate the train model is accurately predicting the input images.

The preprocessed image is cropped into different size 96x96, 128x128, 196x196 and 224x224 with batch size 38 and 64. Build model with different combination of image size and batch size. Analyzed the performance of model. For both indexes, the results for 96x96 are not as good as those for 128x128. Because the input size is limited, there isn't enough data and features for the
negatives to properly train the model. If input sizes were larger, and there was more redundant information and even noise added during the training, making the model more difficult to train.

After trying different combination of size and batch size the model shows the best performance on input size 224x224 with batch size 64 for classification of lung nodule and sub types with accuracy is 98.74 % and model precision and recall value 98.34 % and 96.57%. The model uses Rectified liner unit (ReLU) activation function it consider the value greater than 1 as 1 and less than 1 as 1. For normalization of input between neural network layers the normalization pixel is defined while building model and dropout value is defined, it is used when model is overfitting. This dropout value is used for removing the overfitting problems. The SoftMax activation function is used for output layers.

After the model training and testing process is complete the model needs to be evaluated. The validation image is from dataset is give as input to trained model. The 10% images of total dataset are feed as input to model, the accuracy is on validation is 86.46% for classification purpose.

6. Results and Discussion:

The non-small cell cancer and its sub types (adenocarcinoma, squamous cell carcinoma, larger cell carcinoma) is classified by using the deep learning neural network algorithms (VGG19). The purposed model uses the CT scan image dataset which consist of 1000 cancer related CT scan images. 215 images belong to normal cell, 338 CT scan images for Adenocarcinoma cancer, 260 CT scan images for Squamous cell carcinoma and 187 CT scan images for large cell carcinoma and dataset is split into train, test, validate, training folder contains 70% of total images, test contains 20% of total images and validation contains 10% of total dataset.

Evaluate the performance of the model we need to understand the basic terminologies.

True Positive (TP): Input CT scan image contains a cancer nodule, and it is classified as cancerous.

False Positive (FP): Input CT scan image without cancer nodule and it is classified as cancerous.

True Negative (TN): Input CT scan image without cancer nodule and it is classified as non-cancerous.

False Negative (FN): Input CT scan image contains a cancer nodule, and it is classified as non-cancerous.
Detection and Classification of Lung Nodule in CT Scan Images Using CNN

Accuracy: Accuracy is calculated by adding TP and NP and its sum is divided by sum of TP, TN, FP, FN.

\[
\text{Accuracy} = \frac{(\text{TP}+\text{NP})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})}
\]

Precision: Precision is calculated by adding FP and TP score values and divided its result by TP.

\[
\text{Precision} = \frac{(\text{FP}+\text{TP})}{\text{TP}}
\]

Area Under the Curve (AUC): AUC is calculated by multiplying true positive with difference of false positive and its result is added with difference of true positive which is multiply with difference of false positive at the end sum of both value is divided by 2

\[
\text{AUC} = (\text{TP} \times \text{dFP}) + (\text{dTP} \times \text{dFP})/2
\]

Recall: Recall is calculated by adding FN and TP score values and divided its result by TP.

\[
\text{Recall} = \frac{(\text{FN}+\text{TP})}{\text{TP}}
\]

F-1 Score: F-1 score is calculated as shown in formula multiply ‘2’ with Recall, and Precision. And divide this result with sum of recall and precision.

\[
\text{F-1 Score} = \frac{2 \times \text{Recall} \times \text{Precision}}{(\text{Recall} + \text{Precision})}
\]

The purposed model is trained by deep learning algorithms visual geometric group (VGG19) over the 1000 CT scan image. For the sake of high accuracy, the dataset is prepared by semi-automatic method. After the image dataset is preprocessed the CT scan image is given to neural network algorithms with a different combination of input image size and batch size. The neural network gives benchmark performance on image size of 224x224, and batch size is 64. Its accuracy is 98.74% during training and 86.46% while validation.
Figure no 4 shows the history of accuracy of model. The epoch is taken as X-axis and accuracy is taken as Y-axis, the graph shows the accuracy against the 20-epoch value throughout training and validation process.

The history of precision of while training and validation process of model is shown in figure 4. In figure show that the epoch is taken as X-axis and precision is taken as Y-axis, the graph is showing the precision against the 20 epochs throughout training and validation process model.
Detection and Classification of Lung Nodule in CT Scan Images Using CNN

Fig. 6: History of AUC

In figure 6 show that history of the area under the curve during training and validation. The epoch is taken as X-axis and AUC (Area under the curve) is taken as Y-axis, the graph shows the AUC against the 20 epochs throughout training and validation process of model.

Fig. 7: History of Loss.

In figure 7 shows loss rate. The epoch is taken as X-axis and Loss rate is taken as Y-axis, the graph shows the Loss against the 20 epochs throughout training and validation process model. The loss value shows how wrong label is predicted by model. The model incorrectly finds the pattern in input images due to this model assigning wrong label to input images.
Fig. 8: History of F1-Score

In figure 8 shows F1 Score. The epoch is taken as X-axis and F-1 score (which is already discussed in above section) is taken as Y-axis, the graph shows the F-1 score against the 20 epochs throughout training and validation process model.

Conclusion:

In this research we have implemented the identification of non-small cell lung cancer and its sub-types using deep learning algorithms. The dataset consists of three different types of lung cancer which include adenocarcinoma, squamous cell carcinoma, and large cell carcinoma and the fourth one is normal lung cell. The neural network algorithms VGG 19 to classify the non-small lung cancer with respect to its type.

This computer aided solution proves as more convenient for identification of lung nodule as earlier stage. As result, there is a decrease in deaths related to lung cancer. Our proposed solution classifies the non-small lung cancer with an accuracy of 98.74. This solution is used to assist the physician to correctly identify non-small lung cancer at its earlier stage.

Future Work:

In future work we should try other deep learning algorithms with different segmentation techniques. In our research we have used labeled CT scan image as input to deep learning algorithms.

Declaration:

We declare that this manuscript is original, has not been published before and currently is not being considered for publication elsewhere.
Detection and Classification of Lung Nodule in CT Scan Images
Using CNN

We know of no conflicts of interest associated with this publication, and there has been no significant financial support for this work that could have influenced its outcome.

The dataset used in our research is collected from Kaggle. Kaggle is one of largest community of data scientist that provide datasets and online computing resources. The dataset is not in dcm format, it is not .png format. The CT scan image is generated in dcm format after that it is converted to png format. The model support images in png or jpeg format. The dataset is divided into four folders which contains the image of three different types of lung cancer which include the Ade-no carcinoma, Larger cell carcinoma, Squamous cell carcinoma and one folder contains the normal cell data.

Credit Author Statement


References:


Detection and Classification of Lung Nodule in CT Scan Images Using CNN