IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10756

Deep Learning Algorithm for Classification and Prediction of Lung Cancer using CT Scan Images

V.M. Sanjay Kumar[1], R. Santhosh[1], Dr. S. Roselin Mary[2]

Student, B.E Computer Science Engineering, Anand Institute of Higher Technology, Chennai, India[1]

Professor, CSE, Anand Institute of Higher Technology, Chennai[2]

Abstract - Cancer is the second leading cause of death globally and among all the cancers, Lung cancer is considered to be the major reason for all cancer-related deaths. Detecting lung cancer at early stages can improve the survival rate, but it is a difficult task as lung cancer shows very few symptoms. To diagnose lung cancers at an earlier stage, Computed Tomography (CT) images are used. Nowadays medical decision-making is performed using only CT scan images and an automatic assessment of Computer-Aided Diagnosis (CAD) system. Computer Aided Diagnosis (CAD) system is a medical diagnosis tool that is very useful for today's medical imaging practicality. The primary aim of this work is to develop an advanced computer-aided diagnosis (CAD) system using deep learning algorithms that will efficiently extract data from CT scan images and provide precise and timely diagnosis of lung cancer. The work is divided into three phases - segmentation, feature extraction and classification. The CT scan images are segmented using (OTSU) Thresholding. This work focuses on utilizing the deep learning techniques, namely Convolutional Neural Network (CNN) for feature extraction and recurrent neural network (RNN-LSTM) for lung cancer classification and obtain high accuracy.

Keywords - Lung Cancer Diagnosis, CT Scan Images, Deep Learning, Convolutional Neural Network, Recurrent Neural Network and Long Short Term Memory.

I. INTRODUCTION

Cancer is a disease that is the second leading cause of death globally, and most of the cancerrelated deaths are caused due to lung cancer. Lung Cancer is a malignant tumor which is been characterized by uncontrollable growth of cancerous nodules in the lung tissue. Lung cancer is a lethal disease i.e. it is very harmful and sufficient to cause deaths. More than one million of deaths are occurring yearly due to lung cancer [5]. According to 2015 Global Cancer Statistics, lung cancer accounts for 13% of 14.1 million new cases and 19.5% of cancer related deaths [1]. The Globocan 2018 states that there are 67795 new **978-1-7281-4042-1/19/$31.00 ©2019 IEEE**
cases and 63475 deaths in India due to Lung cancer [17]. The World Health Organization report states that lung cancer is responsible for approximately 2.09 million new cases and 1.76 million deaths in 2018 [17].

Early detection of lung cancer will improve the chances of patient's survival. Therefore, there is an immense need for supporting the medical decisionmaking process so that lung cancer can be detected at an early stage [11]. However, to diagnose lung cancers at an earlier stage, the pulmonologist and radiologists started using chest Computed Tomography (CT) images [8, 12]. Nowadays CT scan images are proving very helpful in diagnosing the cancerous lung nodules.

Moreover, to support medical decision making, performing an automatic assessment of pulmonary nodules using the Computer-Aided Diagnosis (CAD) systems is also beneficial. Computer Aided Diagnosis (CAD) system is an efficient medical diagnosis tool and is very useful for today's medical imaging practicality. These CAD systems are mostly used by the physicians to provide an additional second opinion in order to obtain an accurate diagnosis.

Machine learning algorithms and Deep learning algorithms are the two emerging techniques that have recently attracted many researchers [14]. Deep learning techniques has also achieved tremendous success in computer vision. These techniques provide a uniform feature extraction-classification framework to users and free them from troublesome handcrafted feature extraction [10]. Deep learning techniques also provides the opportunity to increase the accuracy of the early detection of diseases [12].

The primary aim of this work is to develop an advanced computer-aided diagnosis (CAD) system using deep learning algorithms that will extract data from CT scan images and provide precise diagnosis of Lung cancer. In this work, the deep learning techniques, namely - convolutional neural network and recurrent neural network are utilized to propose a model for lung cancer diagnosis using CT scans and to obtain high accuracy.

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10756

II. LITERATURE SURVEY

In [1], they proposed a 3D multipath VGG-like network, which was evaluated on 3D cubes, extracted from Lung Image Database Consortium and Image Database Resource Initiative (LIDCIDRI), Lung Nodule Analysis 2016 (LUNA16) and Kaggle Data Science Bowl 2017 datasets. Predictions from U-Net and 3D multipath VGG like network were combined for final results. The lung nodules were classified and malignancy level is detected using this architecture with 95.60% of Accuracy and 0.387732 of logloss. The traditional method of directly feeding the segmented CT scans into 3D CNNs for classification proved to be inadequate. So in [2], they used a modified U-Net trained on LUNA16 data to detect nodule candidates. The regions of CT scans with segmented lungs where the most likely nodule candidates were located by the U-Net output was fed into 3D Convolutional Neural Networks (CNNs) which classified the CT scan as positive or negative for lung cancer. The performance of the proposed CAD system outperformed the current CAD systems, and provided more efficient training and detection and more generalizability to other cancers.

In [3], they invented an image-based computeraided detection (CADe) algorithm by using regions with CNN features (R-CNN) for detecting the lung abnormalities. They further evaluated the performance of image-based CADx using CNN and that of image-based CADe using R-CNN for detecting various kinds of lung abnormalities like lung nodules and diffuse lung diseases. In [4], the strength of deep learning technique for diagnosing lung cancer on medical image analysis problem was examined. They then proposed a new deep learning algorithm for learning high-level image representation for achieving high classification accuracy in medical image binary classification tasks. They evaluated the model on Kaggle Data Science Bowl 2017 (KDSB17) data set, and compared it with some related works proposed in the Kaggle competition. The novelty of [5], was to accurately model the appearance of the detected lung nodules by using a new developed 7th order MGRF model. A deep autoencoder (AE) classifier was fed by the two feature groups to distinguish between the malignant and benign nodules. The proposed framework demonstrated the promise to be a valuable tool for the detection of lung cancer evidenced by achieving classification accuracy of 91.20%. In [6], they proposed a U-net architecture for lung CT image segmentation. This proposed architecture consisted of a contracting path that extracted highlevel information and a symmetric expanding path that recovered the needed information. The results showed an accurate segmentation with 0.9502 Dice-Coefficient.

III. DATASET

The Lung Image Database Consortium dataset is used in this work. The Lung Image Database Consortium image collection (LIDC-IDRI) consists of lung cancer screening thoracic CT scans [15, 16]. There are 1018 helical thoracic CT scans of 1010 different patients. Total 244,527 DICOM images are present in the dataset.

IV. PROPOSED SYSTEM

In this work, the combination of CNN and LSTM architectures are discussed for the detection of lung cancer using the LIDC dataset. Fig 1 shows the proposed architecture.

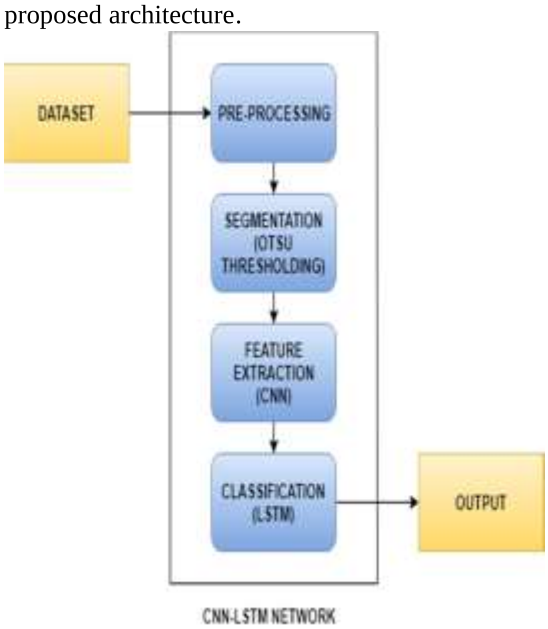


Fig. 1 Proposed Architecture

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10756

In the proposed CNN-LSTM system, CNN consists of convolution and max-pooling layers only. The max-pooling layers output is fed to the LSTM layer.

= CNN () (1)

In equation (1), is the input provided to the CNN network. is the output of the CNN network that is further fed to the next LSTM network. Feature vectors are the output of CNN network and are passed to the LSTM network. Finally the LSTM network provides the classification.

V. SYSTEM ANALYSIS

A. Methodology

The architecture of the proposed CNN-LSTM system is shown in Fig 2. The detailed explanation of all the layers, which are present in the proposed system is given below:

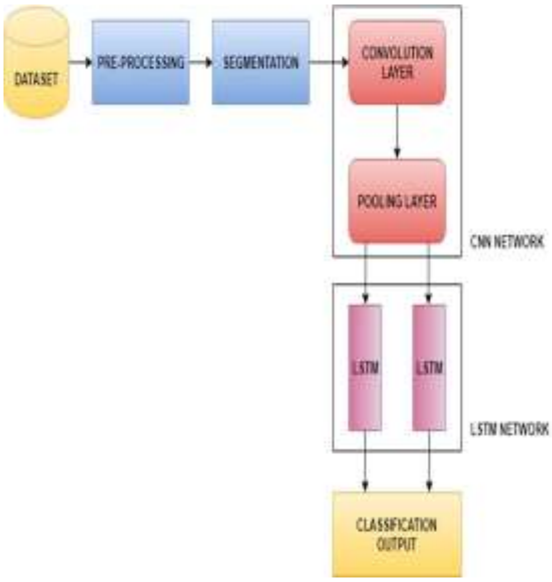


Fig. 2 Architecture of CNN-LSTM

1. Input Layer

The LIDC dataset is provided to the input layer. The input LIDC dataset consists of CT scan images of patients in DICOM format.

2. Pre-processing

The input LIDC dataset contains images in DICOM format. This format is not easy to understand, so there is a need of pre-processing. In pre-processing step, the DICOM images are converted to .png or .jpeg. Median filter is then applied on the images to remove noise. These images are further converted to grey scale images.

3. Segmentation

The simplest image segmentation method used in this work is the thresholding method. This method is based on a threshold value and is used to turn a grey-scale image into a binary image. The key idea is that, the thresholds are derived using OTSU method. Again bitwise addition is performed to convert the binary images to grey scale images.

4. CNN

The architectures of a CNN networks are mostly designed to extract information from 2D structure of an input image. The CNN architecture used in this work is GoogleNet. This architecture consisted of a 22 layer deep CNN. GoogleNet contains a combinations of 9 inception modules. Each module includes some pooling, convolutions at different scales

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10756

and concatenation operations. GoogleNet also uses 1x1 feature convolutions which works as a feature selector. An overview of inception module is given in Fig 3.

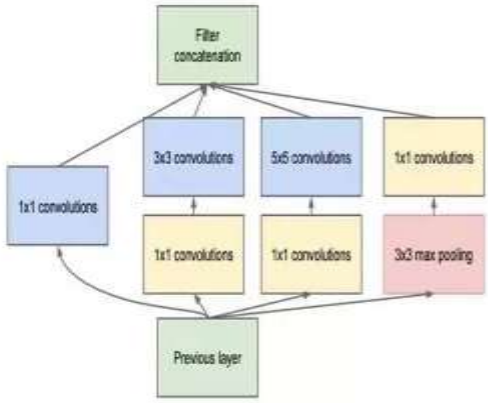


Fig. 3 Overview of Inception module

4.1 Convolutional Layer

The main purpose of convolution layer is to transform the segmented images into a set of feature maps. This is done by performing convolution operation i.e. defined as the sum of the dot product of two functions after one function is reversed and shifted. A kernel filter is used to filter the input segmented images and the resultant convolved output produces the feature maps.

$$F(i,j)=(I*K)(i,j) \quad (2)$$

Where, F(i,j) = set of feature maps, I(i,j) = input (segmented) image and K = kernel filter.

4.2 Max-Pooling Layer

In this layer, down sampling of the feature maps is done by taking the maximum value from a small window length map.

$$F=Max\{F(i,j)\} \quad (3)$$

Where, F(i,j) = set of feature maps produced by convolutional layer, F = reduced feature maps.

5. RNN (LSTM)

The last layer of the CNN -LSTM is the LSTM layer. The LSTM consists of four gates. The equations of the gates are given below.

= (,) (4) = (,) (5)

= (,) (6)

Where, = indicates the input gate, = indicates the forget gate, = indicates the output gate, = indicates activation function, = weights for the respective gates, = output of previous lstm block, = current input and = bias for the respective gates

The LSTM networks major purpose is to map the features extracted from the previous layer to their appropriate class. The classification of the extracted features is done using softmax activation function.

6. Output Layer

The final result is obtained in the form of label or class i.e. normal and cancer.

IJARCCE
ISSN (O) 2278-1021, ISSN (P) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 10, Issue 7, July 2021
DOI 10.17148/IJARCCE.2021.10756

B. Flow of the Proposed System

The proposed system is majorly divided into two phases: the training phase and the testing phase. Fig 4 shows the phases of the system.

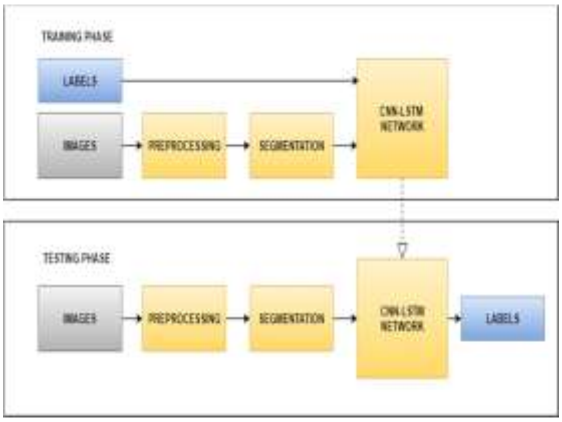
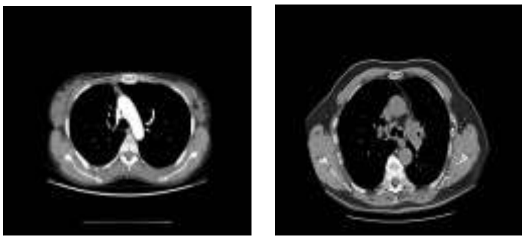


Fig. 4 Phases of Proposed System

C. Results

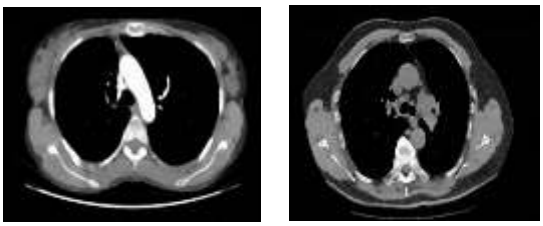
The implementation of proposed CNN-LSTM system is carried out in python language using PyCharm IDE and Windows XP/7 Operating system. The user is required to first register themself. After registration the user needs to login. Once the user logins successfully, input images should be provided. The LIDC dataset consists of CT scan images of patients. These CT scan images are provided to the input layer. Fig 5 shows input images.



(a) (b)

Fig. 5 Input images

In pre-processing step, noise from the images is removed using the median filter. These images are further converted to grey scale images. Fig 6 shows pre-processed images.



(a) (b)

Fig. 6 Pre-processed images

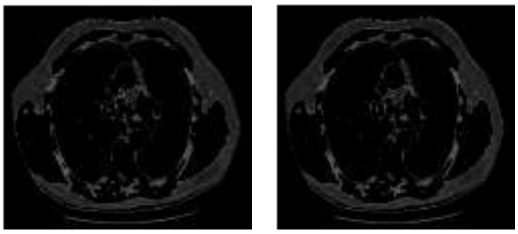
The images are then segmented. Fig 7 shows the segmented images.

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10756



(a) (b)

Fig. 7 Segmented images

The segmented images are passed to the CNNLSTM network for Classification. The network provides the classification by displaying the label. Fig 8 shows the label, Normal or cancer i.e. the final output.

Fig. 8a Classification Result of Input a

Fig. 8b Classification Result of Input b

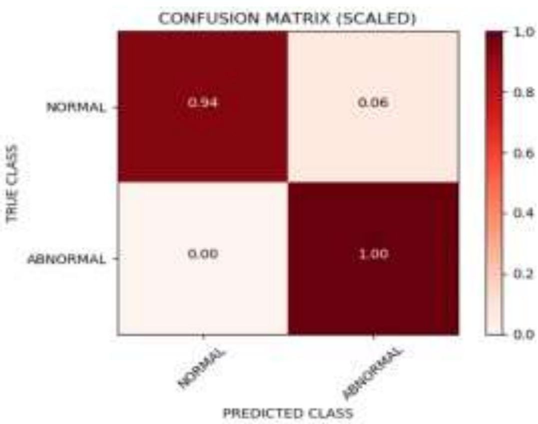


Fig. 9 Confusion Matrix

Performance measure such as accuracy is calculated using equation (7).

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10756

$$Accuracy = \frac{(\qquad)\qquad(\qquad)}{\qquad\qquad} \qquad (7)$$

From the matrix shown in Fig 9, accuracy calculated = 97%. Based on these accuracy measure, the performance of the system is evaluated.

D. Comparative Analysis

The proposed CNN-LSTM system is compared with other existing systems. Different deep learning architectures are been used in existing systems. The comparison is performed based on performance such as accuracy.

TABLE 1. COMPARATIVE STUDY

Sr. No	Architectures	Accuracy
1.	CNN- U Net architecture[2]	82%
2.	Deep AE[4]	92.2%
3.	R-CNN[3]	95.2%
4.	CNN-VGG architecture [1]	95.6%
5.	CNN-LSTM (proposed system)	97%

From table 1, it is clear that the proposed systems accuracy is the highest accuracy achieved so far

VI. CONCLUSION

The proposed system is a hybrid CNN-LSTM model used for Lung cancer detection. The process of detection of lung cancer starts with accepting CT Images. These CT images are further pre-processed and segmented. Finally the classification is performed using proposed CNN-LSTM algorithm. In this, CNN model performs the feature extraction and LSTM model performs prediction and classification. The proposed CNN-LSTM system is compared with other existing system based on the accuracy measure. The main aim of this work is to improve the accuracy of the prediction systems. This aim is achieved by the proposed system as it provides an accuracy of 97%, which is the highest accuracy achieved so far.

REFERENCES

[1] Ruchita Tekade “Lung Cancer Detection and Classification using Deep Learning” 2018 Fourth International Conference on Computing Communication Control and Automation
[2] Alakwaa, Wafaa, Mohammad Nassef, and Amr Badr "Lung Cancer Detection and Classification with 3D Convolutional Neural Network (3D-CNN)" Lung Cancer 8.8 (2017).
[3] Kido, Shoji, Yasusi Hirano, and Noriaki Hashimoto "Detection and classification of lung abnormalities by use of convolutional neural network (CNN) and regions with CNN features (R-CNN)" IEEE, 2018.
[4] Serj, Mehdi Fatan, et al. "A Deep Convolutional Neural Network for Lung Cancer Diagnosis arXiv:1804.0817 (2018). [5] Shaffie, Ahmed, et al. "A Novel AutoencoderBased Diagnostic System for Early Assessment of Lung Cancer" 25th (ICIP). IEEE, 2018.
[5] Skourt, Brahim Ait, Abdelhamid El Hassani, and Aicha Majda. "Lung CT Image Segmentation Using Deep Neural Networks" Procedia Computer Science 127 (2018).
[6] Swapna, G., Soman Kp, and R. Vinayakumar "Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals." Procedia Computer Science 132 (2018)
[7] Xie, Yutong, et al. "Knowledge-based Collaborative Deep Learning for Benign-Malignant Lung Nodule Classification on Chest CT." IEEE transactions on medical imaging (2018).
[8] Iyer, Akshay, Hima Vyshnavi AM,"Deep Convolution Network Based Prediction Model For Medical Diagnosis Of Lung Cancer-A Deep Pharmacogenomic Approach: deep diagnosis for lung cancer." IEEE, 2018.
[9] Tafti, Ahmad P., et al. "Diagnostic Classification of Lung CT Images Using Deep 3D Multi-Scale Convolutional Neural Network." 2018(ICHI). IEEE, 2018.
[10] Jin, Taolin, et al. "Learning Deep Spatial Lung Features by 3D Convolutional Neural Network for Early Cancer Detection." (DICTA), 2017 International Conference on. IEEE, 2017.
[11] da Nóbrega, Raul Victor Medeiros, et al. "Lung Nodule Classification via Deep Transfer Learning in CT Lung Images."31st International Symposium on. IEEE, 2018.

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10756

[12] Fan, Lei, et al. "Lung nodule detection based on 3D convolutional neural networks." (FADS), 2017 International Conference.

[13] Tang, Hao, Daniel R. Kim, and Xiaohui Xie. "Automated pulmonary nodule detection using 3D deep convolutional neural networks."15th International Symposium on IEEE, 2018.

[14] Paul, Rahul, et al. "Combining deep neural network and traditional image features to improve survival prediction accuracy for lung cancer patients from diagnostic CT." 2016 IEEE International Conference.

[15] The Lung Image Databas e Consortium LIDC and Image Database Resource Initiative IDRI

[16] Kaggle, "Data science bowl 2017."

[17] Statistics of Lung Cancer in India

[18] LUNA16, "Lung nodule analysis 2016.