INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 07 ISSUE: 06 | JUNE - 2023 SJIF RATING: 8.176 ISSN: 2582-3930

Recognition of Lung Cancer Using CNN Model

[1] **Ms. Sukhwani Divya M.**, [2] **Dr. Pawar U. B.**, [3] **Prof. Daund R. P.**

[1] PG Student, [2] Asso. Professor, [3]Asst. Professor,

[1]Department of Computer Engineering,

[1]SNDCOERC, Yeola, India

Abstract: Lung cancer is one of the deadliest types of cancer; thousands of people are infected with this type of cancer, and if it is not detected in the early stages of the disease, then the patient's chance of survival will be very small. For the above reasons and to help overcome this terrible early diagnosis with the help of artificial intelligence procedures are most needed. It is also one of the most common and contributing to death among all cancers. Cases of lung cancer are increasing rapidly. There are about 70,000 cases a year in India. In the past decade, cancer detection using deep learning models has been a hot topic, especially in medical image classification. It is worth noting that CNN models are more advanced in dealing with the diagnosis of diseases such as lung cancer due to the higher performance and capabilities of CNNs. This system presents an approach that uses a convolutional neural network (CNN) to classify tumors located in the lungs as malignant or benign. The accuracy obtained by CNN is 99%, which is more efficient compared to the accuracy achieved by traditional existing systems. This is done by applying a convolutional neural network technique to a dataset of lung cancer CT scans.

***Index Terms* - lung, computed tomography, convolutional neural network, detection, classification**

I. INTRODUCTION

This application focuses on the early detection of lung cancer to give patients the best chance of recovery and survival using a CNN model. Using a dataset of thousands of high-resolution lung scans, the model pinpoints when lesions in the lungs are cancerous. This will dramatically reduce the number of false positives that plague current detection technology, allow patients earlier access to life-saving interventions, and give radiologists more time to spend with their patients.

Medical imaging and related fields have taken center stage in a digitally connected world. Medical information is used by several pupils from biometrics to insurance. These are increasing the need for this information to be trustworthy and reliable across platforms. Efficient systems need to be developed to use this information, which is usually in the form of medical images, in the automatic diagnosis of diseases.

One of the leading causes of human death is cancer. The most common malignant tumor is lung cancer and its worldwide incidence is increasing by 2% per year. Lung cancer is linked to the use of tobacco products in 90% of cases. An asymptomatic person (smoker) undergoes a computed tomography (CT) scan, which is one of the best ways to diagnose lung cancer. The detection of CT-like nodules is not a simple task because they include low density and small size etc. in the area of complex anatomical contrasts similar to other structures. A number of computer-aided tumor detection and characterization techniques have been proposed in the literature. Two key categories for the development of these techniques are computer aided identification (CAD) and CADx diagnostics. The installation of the CADx device minimizes the number of unnecessary biopsies and reduces mental trauma for patients with benign tumors. Therefore, CADx acts as a second approach and helps experts to diagnose cancer at earlier stages of the disease. This diagnosis is accurate and effective.

In recent years, deep learning techniques have demonstrated the ability to automatically detect features from training images and manipulate contact, even hierarchy,between the characteristics of the deep neural structure. The new learning system can also solve feature computation, selection and integration problems without using complex processing steps and pattern recognition. Using CNN as an extractor and classifier, this paper proposes a technique to classify pulmonary nodules as benign or malignant, which can provide a specialist with a second opinion.

II. LITERATURE REVIEW

Convolutional Neural Network (ConvNet/CNN), Deep Learning or other machine learning algorithms have been used by various researchers to perform experiments on different types of lung cancer detection. Shrinivas Arukonda used the technique of convolutional deep neural networks. In this system, it is able to detect lung cancer in its earlier stages due to its survival rate.

The dataset is used from the Data Science Bowl 2017 Kaggle competition, LUNA16. in the first step they detected the annotated nodes. A 32 X 32 X 32 cube is around the nodes, with the node in the middle. A region of interest mask is applied to the lungs. Cubes are then formed around the predicted nodes and the prediction is made using a second 3D CNN. The accuracy achieved in the model was found to be 94.30%.

Various research studies with CADx systems have already been proposed to improve the accuracy of lung cancer detection. Hua et al. proposed a deep learning technique for lung cancer detection. The authors achieved a sensitivity of 73.40% and a specificity of 82.20% with the deep belief architecture (DBN). The authors also used a CNN architecture and obtained a sensitivity of 73.30% and a specificity of 78.70%. Kumar et al. proposed a deep learning system with an accuracy of 75.01% using stacked autoencoder (SAE).

1) In 2017, authors Pooja R. Katre and Anuradha Thakare described various image processing techniques for lung cancer detection. In their proposed approach to noise removal and enhancement, the Median Filter method is used. The best thing about a median filter is that it removes noise without blurring the image. Preserves region edges. Used to remove salt and pepper noise from an image. Gabor filter is used in the enhancement stage because it provides better results compared to fast Fourier and automatic enhancement. The purpose of this work is to detect the tumor at an early stage. CT scan images are accepted as input. After image processing, there is a feature extraction phase in which the area, perimeter and eccentricity of the image are calculated. The Support Vector Machine algorithm is used to classify the data. The above signs help to identify the size of the tumor and based on this the stage of the cancer is determined.

2) In December 2017, Suren Makaju, P.W.C Prasad, Abeer Alsadoon and A.K. Singh worked on a lung cancer CAD system with CT Scan Images as his primary focus. They believe that the best input data for this research are CT images. The proposed model uses algorithms to remove noise before image processing. It uses the same segmentation as the current system, i.e. the watershed algorithm, and supports well-defined feature extraction before classification by SVM. The author used images from the LIDC dataset and the system provides 92% accuracy and 50% specificity.

3) In May 2015 Md. Badrul Alam Miah and Mohammad Abu Yousuf designed a neural network-based CAD system for early detection and diagnosis. ANN and Fuzzy Clustering, IP, Curvelet Transform, Multinomial Bayesian Algorithm, Backpropagation, Gray Coefficient Weight Estimation and SVM are the basis of these observations. The goal is to create a fast and robust, more accurate system with rotation, scaling and translation feature extraction. A dataset of 300 images obtained from hospitals is used. The steps in the proposed system are image acquisition, processing, binarization, thresholding segmentation, feature extraction, and neural network classification. The steps in image processing are grayscale conversion, normalization, noise reduction, binarization and removal of unwanted part of the image. Feature extraction uses features like image center, height-to-width ratio, average distance between black pixels and center, etc. Two neural network outputs are used for classification. The system provides an accuracy of 96.67%, higher than all existing systems.

4) In March 2014, prof. Sanjeev N. Jain and Bhagyashri G. Patil Several methods to detect cancer cells from CT lung scans. The purpose of this paper is to find cancer cells and provide more accurate results using different segmentation techniques such as thresholding and watershed segmentation. In thresholding, a threshold value is set to distinguish the object of interest from the background. If the pixel value is greater than the threshold value, then it belongs to the object, otherwise it is in the background. Thus, the region of interest can be extracted using a thresholding approach. In watershed segmentation, the background and object are separated using different markers: internal markers associated with the object of interest and external markers associated with the background. It is a simple, intuitive and fast method. According to the research, the watershed segmentation approach has more accuracy (85.27%) than the thresholding approach (81.24%).

5) In 2012, Mokhled S. Al Tarawneh published a comparative article between different image processing techniques and the algorithms they use for a CAD system for lung cancer detection. The main goal of the article is to find out the properties for accurate comparisons between images with different processing techniques. The three steps are image enhancement, segmentation and feature extraction. The goal of image enhancement is to improve image quality and provide better input for classification. Gabor filter, automatic enhancement and fastfourier transform improve the enhancement speed. The Thresholding and Watershed methods are used for segmentation, of which the watershed provides better segmentation quality. Feature extraction uses binarization and masking. Binarization and masking when implemented together gives the optimal result.

6) In 2017, China's Lei Fan and his group of researchers used a deep learning algorithm to detect CAD lung cancer. In this paper, image processing is not applied to lung CT scans. The images are directly fed as input to the convolutional neural network, which consists of two convolutional layers, two max pooling layers, one fully connected layer and one output layer. Rectified linear units (ReLU) are applied between the convolutional and maximum pooling layers. The system provides an overall accuracy of 67.7%. It is concluded that support vector machines have lower classification accuracy than 3D convolutional neural network for the same number of input samples.

III. SYSTEM ARCHITECTURE

This architecture presents lung cancer detection based on chest CT images using CNN. In the first stage, regions of the lung are extracted from the CT image and individual slices are segmented in this region to obtain tumors. The segmented tumor regions are used to train the CNN architecture. Then, CNN is used to test patient images. The main goal of this study is to determine whether the tumor present in the patient's lungs is malignant or benign. Figure 1 shows the block diagram of the proposed system. As shown in the figure, the trained system will be able to detect the presence of cancer in a lung CT image.

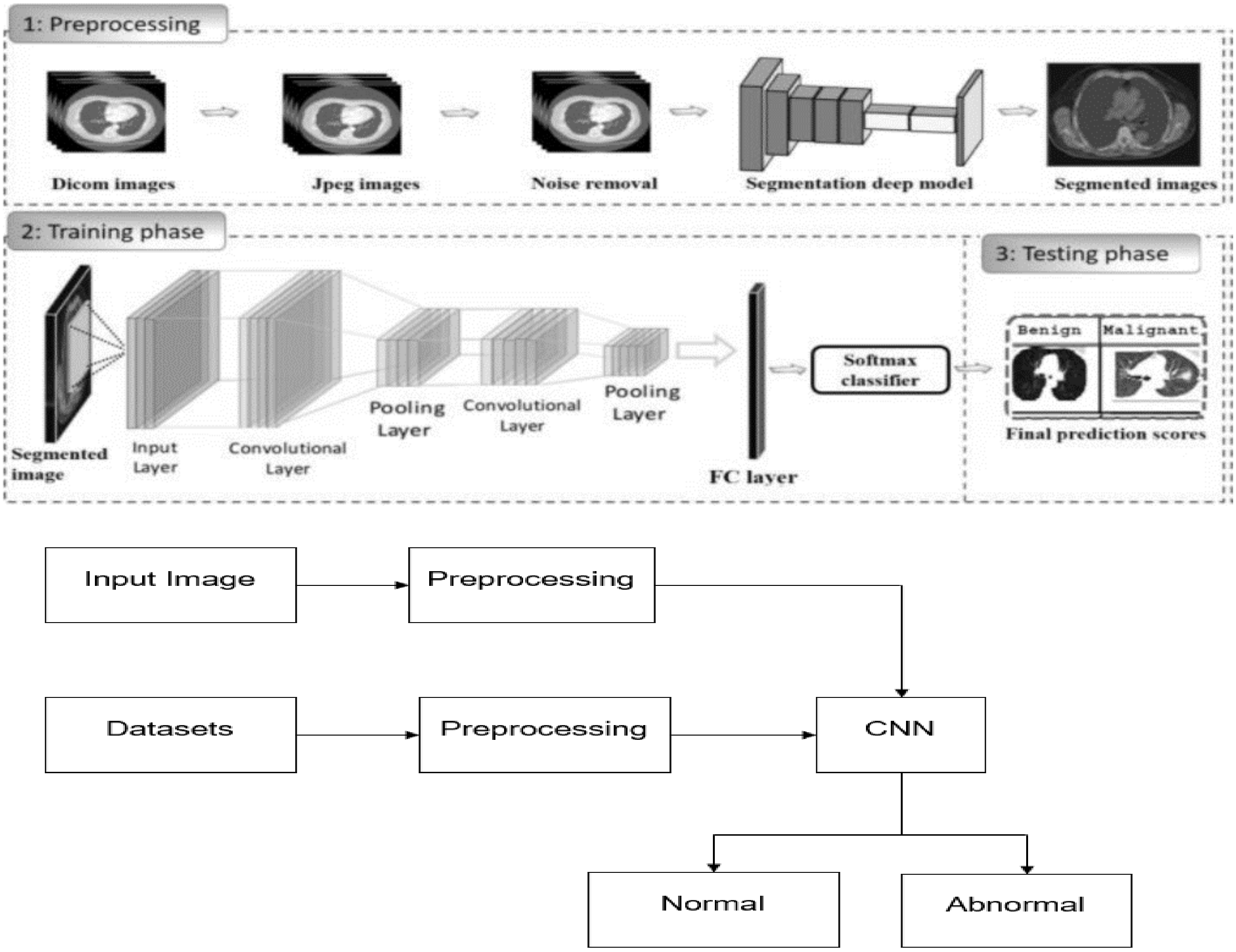


Figure. 1.1. System Overview

Data file: -

The database used is obtained from the Lung Image Database LUNA16, Data Science Bowl 2017. It is a classification database of lung nodules containing scans of a total of 1018 patients. Each patient's CT scan, in turn, contains approximately 150 to 550 images in dicom format. The database provides four classifications namely (i) Unknown, (ii) Benign, (iii) Malignant and (iv) Metastatic

PRE-PROCESSING

In the preprocessing stage, a median filter is used to restore the test image by minimizing the effects of degradation during acquisition. Various lung nodule preprocessing and segmentation techniques are discussed.

The median filter simply replaces the value of each pixel with the median value of its neighbors, including itself. Thus, pixel values that are very different from their neighbors will be eliminated.

Convolutional Neural Network

A convolutional layer creates a feature map by convolving different subregions of the image with a trained kernel. Furthermore, non-linear activation functions such as sigmoid, tange or rectified linear (ReLu) can also be used. Another method to reduce calculations is a pooling layer, where an image/feature map region is selected and the maximum of them is selected as a representative pixel. A 2x2 or 3x3 grid can therefore be reduced to a single scalar value. a traditional fully connected layer can also be used in conjunction with convolutional layers and are usually used towards the output stage.

Convolution layer 1:

The 3-D hdf5 data forms the input to the first convolutional layer. This layer with kernel size 50x50 with step 6. The output of this layer produces 78 features. The mass padding is set to a variation of 0.01 Gaussian and the bias is set to a constant zero. This output is then fed to the Rectified Linear (ReLu) layer to null all negative activations. The primary application of this layer is the detection of the lowest level features, such as whether there is a classification in an image region.

Convolution layer 2:

The output of the first convolutional layer is fed to the second, which has a kernel size of 3x3 and a step of 1. This layer pads the data with one zero closure. The weight padding is the same as convolutional layer 1 and the bias is set to a constant value of 1. This layer is also followed by the ReLu layer. This layer is intended to use the information predicted from the previous layer and detect the classification pattern - eg popcorn, diffuse, etc. So, it learns from the training phase which patterns are benign and which are malignant. In this way, CNN achieves two goals – it learns features hierarchically and eliminates the need for specific feature engineering.

Maximum pooling layer:

After convolution layer 2 comes the max-pooling layer, where the most sensitive node of the given kernel is extracted. The kernel size used in the proposed network is 13x13 with a step offset of 13. This is primarily intended to reduce the computational effort. Since each CT scan consists of 500 images, if we have a batch size of 50, the number of required calculations can be significantly large, leading to frequent memory overload. The maximum pooling layer is mainly used to alleviate memory and data bottlenecks by reducing image dimensions.

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 07 ISSUE: 06 | JUNE - 2023 SJIF RATING: 8.176 ISSN: 2582-3930

Dropout Layer:

A dropout layer is used in the network to prevent overfitting. This is done by turning off random neurons in the network. Our proposed network uses a waste layer with a drop ratio of 0.5. The intent of this layer is to improve the quality of classification on test data previously unseen by the network.

Fully connected layer:

A fully connected layer is used which provides two outputs. It uses a Gaussian mass padding of 0.5 and a constant bias padding of 0. The two output neurons from this layer give the classification of benign or malignant.

This layer is primarily intended to combine all features into a single top-level image and will eventually form the basis for the classification step.

IV. METHODOLOGY & ALGORITHM

Modules 1:

Importing the given image from the dataset We need to import our information index using the preprocessing function of keras ImageDataGenerator, next we will do size, scale, range, zoom range, flat flip. At this point, we import our image dataset from the organizer through the information generator job. Here we will set the training, test and approval, as well as set the target size, group size and class mode from this capacity that we have to prepare.

Module 2:

Prepare the module according to the given image data set To prepare our data set using the work with classifier and matching generator we also prepare the steps according to the age at that time the absolute number of ages, the approval information and the approval steps using this information we can prepare our data set

Module 3:

Working Interaction of Layers in a CNN Model An algorithm in Deep Learning that can take an image as input, assign biases and weights to multiple image characteristics, and classify one image from another is called a Convolutional Neural Network. Other classification algorithms require huge pre-processing unlike ConvNet which does not require much processing. ConvNets do not require filters that are hand-crafted, the algorithm can learn itself if properly trained unlike traditional methods. The visual cortex present in the human brain is an organization that is synonymous with and highly inspired by the operation and architecture of Conv Net. A receptive field is an area that consists of neurons that respond only in enclosed areas. Their union includes four layers with 1,024 data units, 256 units in the main hidden layer, eight units in the subsequent secret layer, and two yield units.

Input layer:

Image knowledge is contained in this layer of the convolutional neural network. Image information is solved using three-dimensional frames. The input layer is responsible for transforming the image dimension into a singular column. Suppose there is an image of measurement "28x28=784", this dimension needs to be converted to a singular column before inserting it into the input.

Convolution layer:

The convolution layer is also known as the "feature extractor layer" because the image highlights were removed in this layer. As a matter of first importance, the image part is connected to the Convo layer to perform the complexity operation as we saw earlier. Calculates the speck entry between the open fields and the channel. The result is a single whole number of the yield quantity. At this point, direct a similar info image with Step through the following responsive field and do the same once more. It repeats a similar interaction over and over again until it goes through the entire image.

Pooling layer:

Reduction of the spatial volume of the image after convolution is performed using a pooling layer. A pooling layer is used halfway between two convolutional layers. Without using a pooling layer, when the fully connected layer is applied after the convolutional layer, it could require a lot of computing power. Thus, the spatial volume of the input image can only be reduced by using maximum pooling. In a step of two, maximum pooling was applied in one depth slice. We can observe that the input is 4 x 4

shortened to 2x2.

Fully connected layer (FC):

A fully associated layer includes loads and slopes. A fully connected layer connects neurons in an individual layer to subsequent layers. Training is used to classify images between different classes.

Output layer:

The principle of one-hot encoding is followed, which includes the name in the output layer.

Module 4:

Django Framework Model Deployment and Yield Prediction In this module, the prepared deep learning model is transformed into a document of various leveled information design (.h5 record), which is then transferred to our Django system to provide a better user interface and yield prediction, whether given Benign cases, Malignant cases, Normal cases.

V. RESULTS & IMPLEMENTATION

We have used various python libraries such as 'dicom' (library to work with DICOM Files), 'scikit-image' (an image processing library), and 'OpenCV' (an open-source, real-time computer vision library) to preprocess the dataset before fitting into the model. We have then split the preprocessed dataset into 60%, 20%, and 20% as training, validation, and testing respectively. Using the TensorFlow Library and the Kera's API, we have trained the model by using the training dataset. We have trained the model for 30 epochs in a system having Intel i7-4790k 4th Generation Processor, 16 GB of RAM, and a platform of Windows 10 architecture.

Metrics Evaluation: After completion of the training process, we have found that the model yields 83.33% as the training accuracy. We further assess the model performance by examining various metrics such as Testing accuracy, Precision, Recall, Kappa-Score, and F-score. Accuracy is a measure of how many instances that are correctly classified out of the total number of instances. The proposed model provides us an excellent testing accuracy of 100%. The reason behind obtaining 100% accuracy might be due to the small dataset having a smaller number of instances. Working with a larger dataset might affect the accuracy of the model.

Accuracy is a measure of how many instances that are correctly classified out of the total number of instances. The proposed model provides us an excellent testing accuracy of 100%. The reason behind obtaining 100% accuracy might be due to the small dataset having a smaller number of instances. Working with a larger dataset might affect the accuracy of the model. Precision shows how many of the total selected instances are correctly classified, whereas recall shows how many correctly classified instances are selected. The computed precision and recall of our model are depicted in Table 1.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

TABLE I PRECISION AND RECALL

Classes	Precision	Recall
Benign	1	1
Malignant	1	1

Kappa score (equation 15), is the measure of agreement between two raters. In our case, predicted values and the truth values are the two raters. Using Observed Accuracy and Predicted Accuracy, we have computed the Kappa score of the model and obtained an attenuation of 1.

$$Kappa\ Score(K) = \frac{p_o - p_e}{1 - p_e}$$

The harmonic mean of Precision and Recall is known as F-score or F-measure. We obtained a F-score of 1 using equation 16, which signifies perfect Precision and Recall of the model.

$$F - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

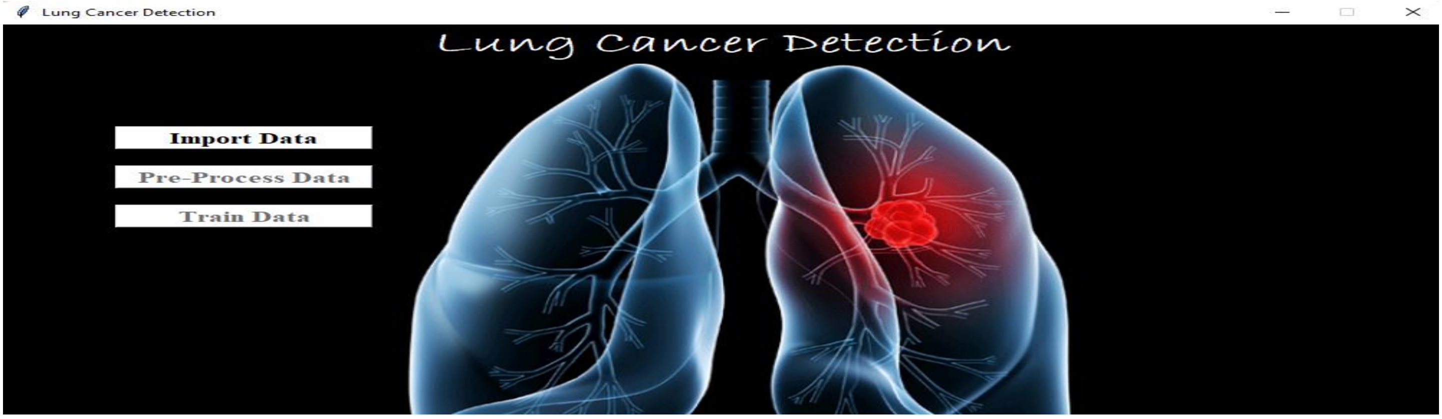
VI. IMPLEMENTATION STEPS

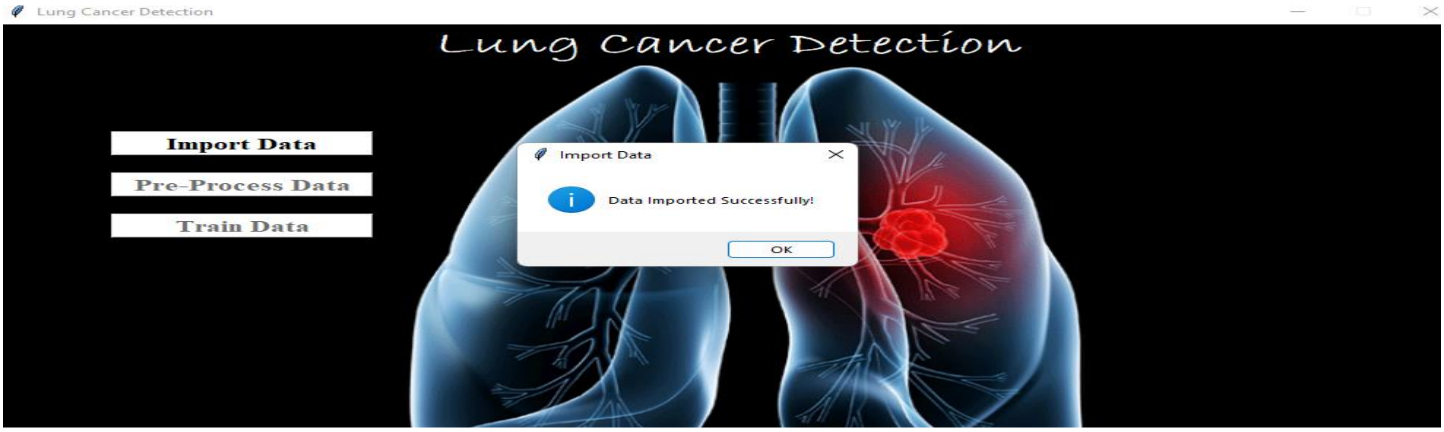
1.Install all the libraries mentioned in the requirements.txt file with the command pip install -r requirements.txt

2.Open your terminal/command prompt from your project directory and run the file python lcd_cnn.py by executing the command python lcd_cnn.py

```
PS C:\Lung-Cancer-Detection-main> python lcd_cnn.py
2023-05-24 18:54:43.040771: W tensorflow/stream_executor/platform/default/dso_loader.cc:64] Could not load dynamic library 'cuda
rt64_110.dll'; dlerror: cudart64_110.dll not found
2023-05-24 18:54:43.041699: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] Ignore above cudart dlerror if you do not have
a GPU set up on your machine.
WARNING:tensorflow:From C:\Python\lib\site-packages\tensorflow\python\compat\v2_compat.py:107: disable_resource_variables (from
tensorflow.python.ops.variable_scope) is deprecated and will be removed in a future version.
Instructions for updating:
non-resource variables are not supported in the long term
2023-05-24 18:54:58.665848: I tensorflow/core/platform/cpu_feature_guard.cc:193] This TensorFlow binary is optimized with oneAPI
 Deep Neural Network Library (oneDNN) to use the following CPU instructions in performance-critical operations:  AVX AVX2
To enable them in other operations, rebuild TensorFlow with the appropriate compiler flags.
2023-05-24 18:54:58.675629: W tensorflow/stream_executor/platform/default/dso_loader.cc:64] Could not load dynamic library 'nvcu
```

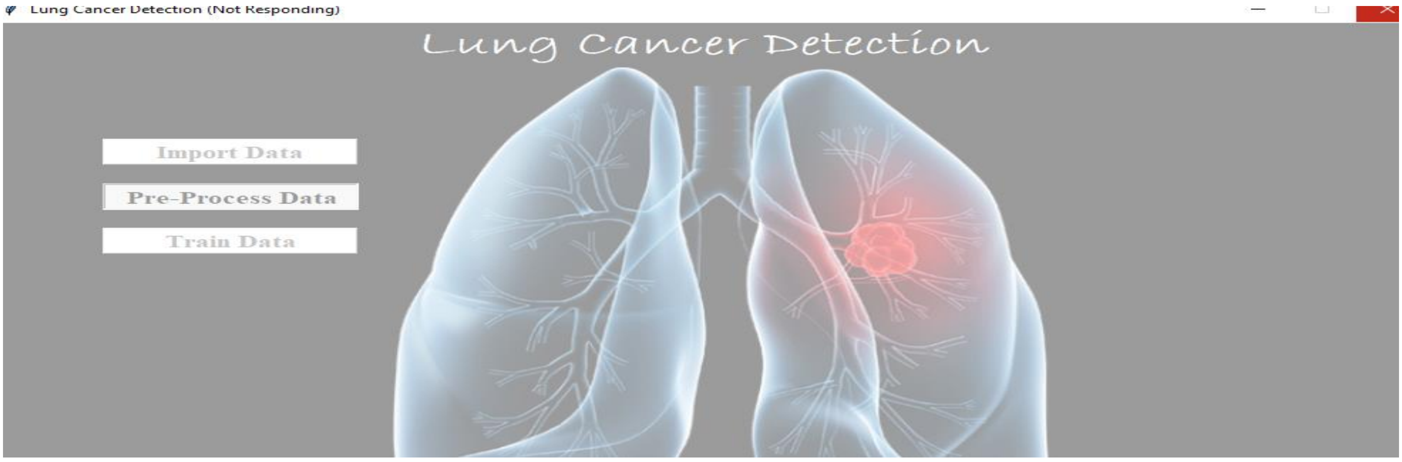
3. Import the data

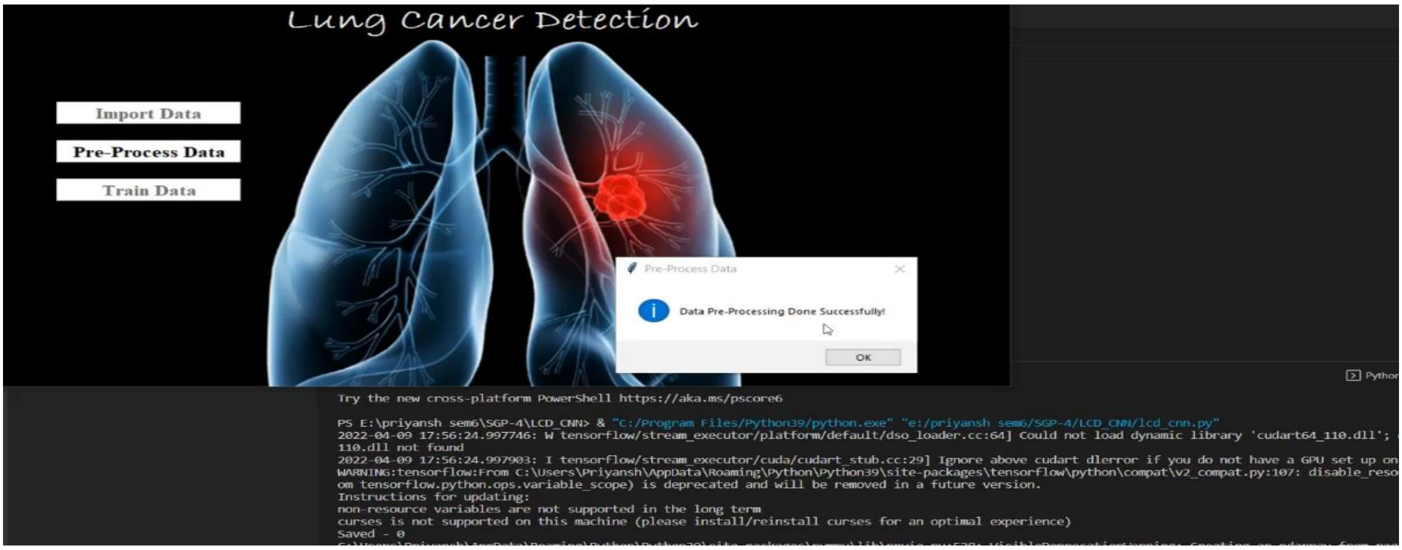




INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 07 ISSUE: 06 | JUNE - 2023 SJIF RATING: 8.176 ISSN: 2582-3930

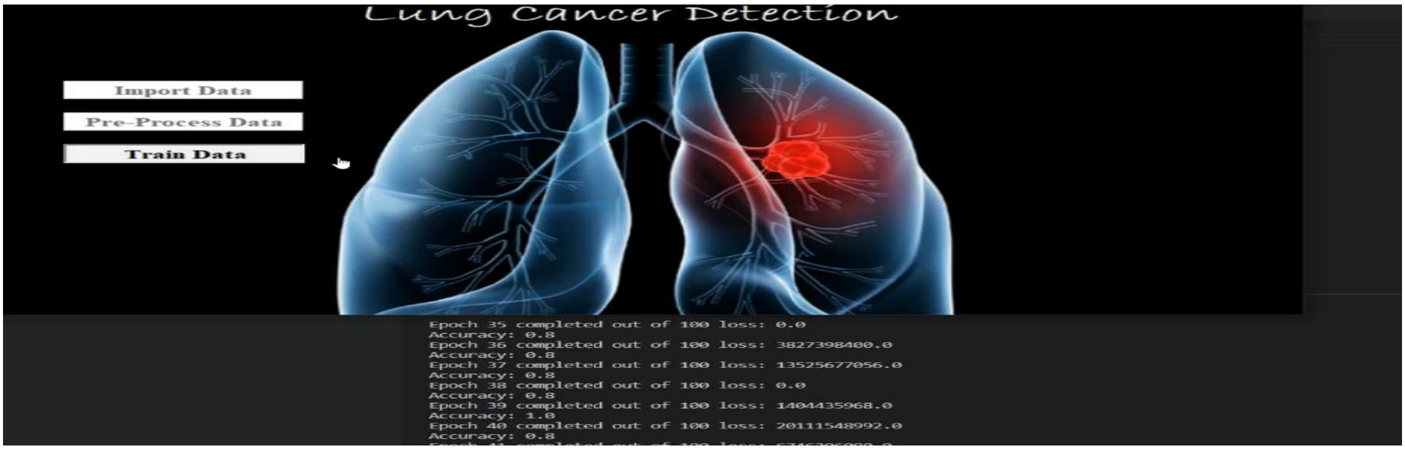
4.Preprocess the data



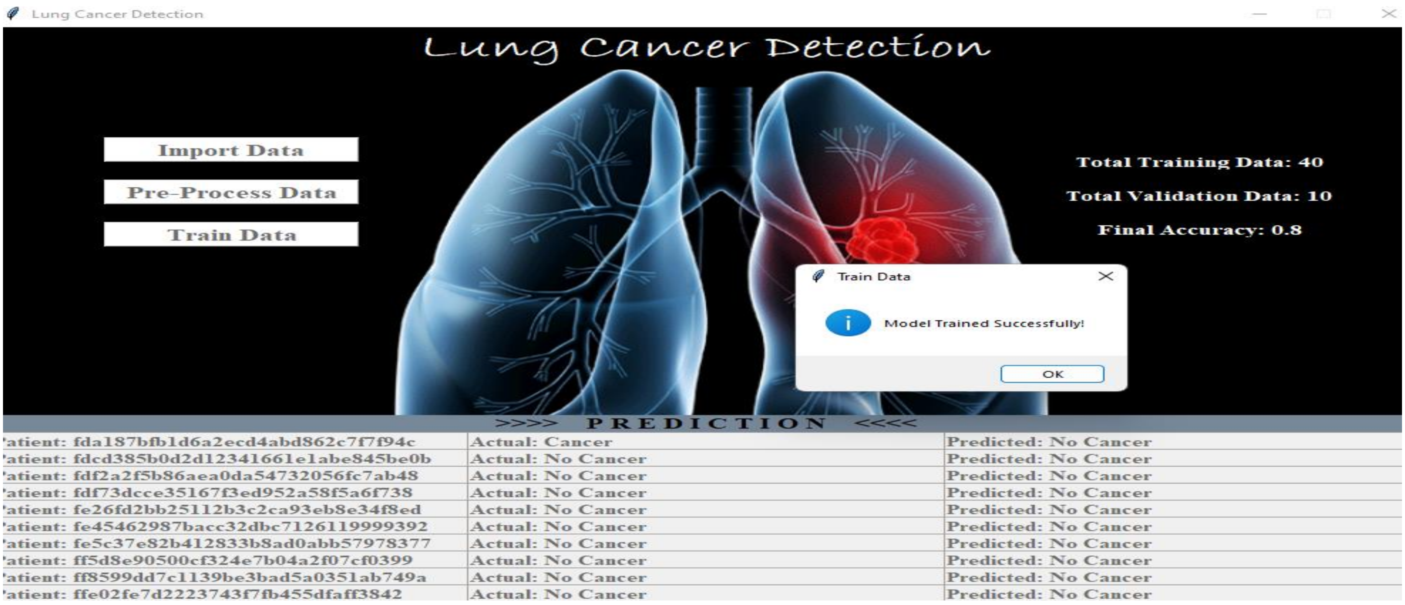


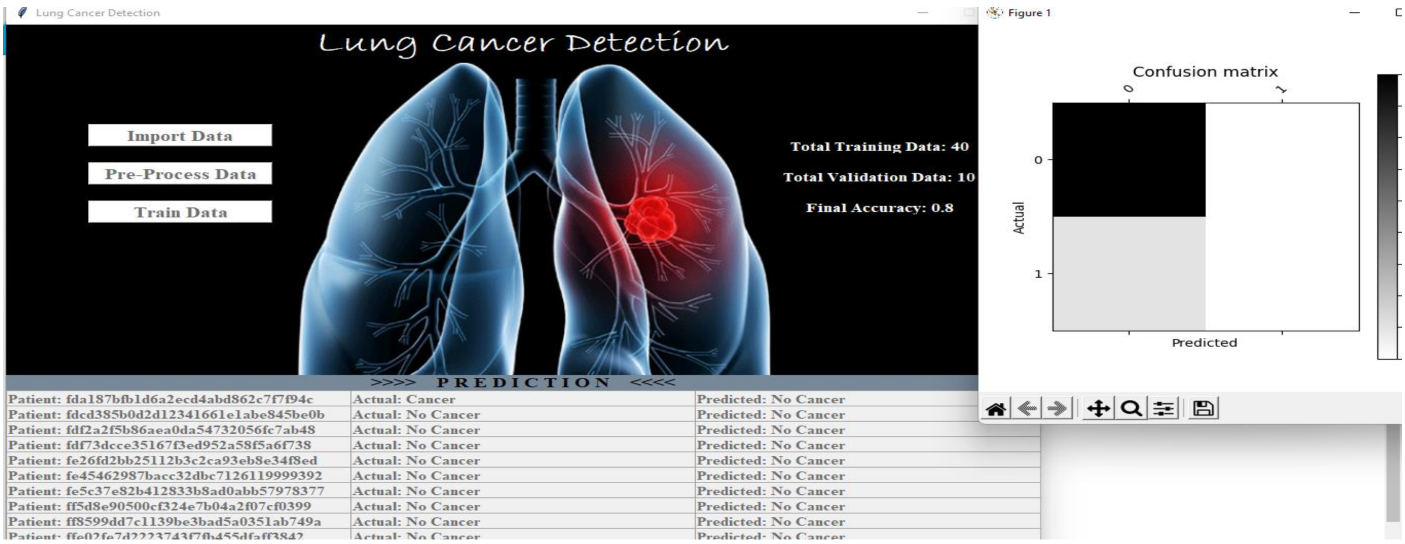
5.Train the data

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 07 ISSUE: 06 | JUNE - 2023 SJIF RATING: 8.176 ISSN: 2582-3930



6.Get the expected result





VII. INCORPORATED PACKAGES

- **Pandas** – This library helps to load the data frame in a 2D array format and has multiple functions to perform analysis tasks in one go.
- **NumPy** – NumPy arrays are very fast and can perform large computations in a very short time.
- **Matplotlib** – This library is used to draw visualizations.
- **Sklearn** – This module contains multiple libraries having pre-implemented functions to perform tasks from data pre-processing to model development and evaluation.
- **OpenCV** – This is an open-source library mainly focused on image processing and handling.
- **TensorFlow** – This is an open-source library that is used for Machine Learning and Artificial intelligence and provides a range of functions to achieve complex functionalities with single lines of code.

VIII. CONCLUSION & FUTURE SCOPE

A convolutional neural network-based system was implemented to detect malignant tissues present in the input lung CT image. An image of a lung with different shape, size of cancer tissues was fed as an input for training the system. The proposed system is able to detect the presence and absence of cancer cells with an accuracy of about 96%.

In addition to the deep convolutional network, the same dataset was classified by the Multilayer Perceptron Network Backpropagation algorithm using GLCM features. The results show only 93% accuracy.

In this proposed work, the obtained specificity is 100%, which shows that there is no false positive detection. Also, the accuracy, sensitivity and specificity of the proposed system is high compared to previously available conventional neural network-based systems.

In the near future, the system will be trained with large data sets to diagnose the type of cancer with its size and shape. The overall accuracy of the system can be improved by using a 3D convolutional neural network and also by improving the hidden neurons using a deep network.

REFERENCES

1. http://globocan.iarc.fr/Pages/fact_sheets_ cancer.aspx.

2.https://www.livemint.com/Politics/3eXX60XBig4bWZ25Kr1iQO/India-recordedabout39-million-cancer-cases-in-2016data.html

3. Using Deep Learning for Classification of Lung Tumors on Computed Tomography Images.

4. M.S. Al-Tarawneh, "Lung cancer detection using image processing techniques," Leonardo Electronic Journal of Practices and Technologies, vol. 20, pp. 147– 58, May 2012.

5. LUNA16, "Lung tumor analysis 2016." https:// luna16.grand-challenge.org/.

6. A Manikandarajan, S Sasikala, Detection and Segmentation of Lymph Nodes for Lung Cancer Diagnosis. National Conference on System Design and Information Processing – 2013.

7. M.S. Al-Tarawneh, "Lung cancer detection using image processing techniques," Leonardo Electronic Journal of Practices and Technologies, vol. 20, pp. 147– 58, May 2012

8. Albert Chon, Peter Lu, NiranjanBalachandar "Deep Convolutional Neural Networks for Lung Cancer Detection".

9. Wavelet Recurrent Neural Network for Lung Cancer Classification":3rd ICSTcomputer,2017.

10. A.Kavitha, Anusiyasaral and P.Senthil," Design Model of Retiming Multiplier For FIR Filter &its Verification", International Journal of Pure and Applied Mathematics, Vol116 No12, 2017, pp. 239-247

11. S Sasikala, M Ezhilarasi, Combination of Mammographic Texture Feature Descriptors for Improved Breast Cancer Diagnosis. Asian Journal of Information Technology, 2016.

12. K.Malarvizhi, R.Kiruba, "A Novel Method Of Supervision And Control Of First Order Level Process Using Internet Of Things ",Journal Of Advanced Research In Dynamical And Control Systems, Vol 9 No.6 2017, Pp.1876-1894.