

LUNG CANCER DETECTION USING MACHINE LEARNING ALGORITHM

Dr. R K khare¹

Anjali Singh, Akanksha, Chanchal Sahu^{2,3,4}

1 Associate Prof. Computer Science and Engineering, SSTC , Bhilai

2,3,4 Students, Computer Science and Engineering, SSTC , Bhilai

Abstract: Lung diseases are illnesses that damage the lungs and impede the breathing process. Lung cancer is one of the leading causes of death in individuals all over the world. Early detection can increase the chances of survival in humans. If the condition is detected in time, the average survival percentage for people with lung cancer rises from 14 to 49%. While computed tomography (CT) is significantly more effective than X-ray, a comprehensive examination involves a number of imaging treatments to complement each other. A deep neural community for identifying lung tumors from CT images is being created and tested. A deep neural community for identifying lung tumors from CT images is being created and tested. A densely coupled convolution neural network (DenseNet) and adaptive boosting set of rules dataset of 201 lung images is utilized for classification as ordinary or malignant, with 85 percent of the images used for teaching and 15 percent used for testing and classification. The experimental findings revealed that the proposed approach attained a precision of 90. 85 percent of the time.

Keywords: DenseNet, Image Processing, Deep Learning, Convolution Neural Networks (CNN).

I. Introduction:

Lung cancer is the leading cause of cancer death worldwide, accounting for 2.09 million new cases and 1.77 million deaths in 2018. In the early 2000s, four case-controlled studies from Japan revealed that combining chest radiography and sputum cytology in screening was effective for lowering lung cancer

mortality. Two randomized controlled studies conducted from 1980 to 1990 found that screening using chest radiographs was no longer effective in lowering lung cancer mortality. Although the usefulness of chest radiographs in lung cancer screening remains disputed, they are more cost-effective, easier to obtain, and offer a lower radiation dosage than low-dose computed tomography (CT). Another downside of chest CT is the high number of false positive (FP) outcomes. It has been claimed that 96% of nodules discovered by low-dose CT screening are FPs, resulting in needless follow-up and invasive investigations. Chest radiography is not as sensitive as chest CT, but it is more advanced in terms of specificity. Taking these factors into consideration, the development of a computer-aided diagnosis (CAD) version for chest radiographs might be beneficial by increasing sensitivity while maintaining low FP findings.

The most recent use of convolutional neural networks (CNN), a branch of deep learning (DL), has resulted in remarkable, modern-day advancements in radiography. DL-based completely designs have also demonstrated potential for nodule/mass detection on chest radiographs, indicating sensitivities within the range of 0. Furthermore, radiologist performance for recognising nodules improved with these CAD designs compared to without them. In medical practise, radiologists frequently struggle to detect nodules and distinguish between benign and malignant nodules. Normal anatomical systems frequently resemble nodules, which is why radiologists must pay close attention to nodule morphology and marginal residents. Because these flaws are caused by the circumstances

rather than the radiologist's abilities, even skilled radiologists might make mistakes. Detection and segmentation are the two basic strategies for identifying lesions using DL. The detection approach is a neighborhood-level classification, whereas the segmentation method is a pixel-level classification. The segmentation method can provide more distinct data than the detection method. In clinical practice, identifying a lesion's scale on the pixel-degree scale increases the chance of producing a proper analysis. Pixel-level classification also makes it easier to keep track of changes in lesion size and shape, because the shape may be utilized as a reference during detection. It also allows you to recall not only the long and short diameters, but also the location of the lesion while calculating the effect of treatment. However, to the best of our knowledge, no studies have been conducted on the use of the segmentation approach to detect pathologically proven lung malignancies on chest radiographs. The goal of this study is to teach and verify a DL-based model capable of identifying lung cancer on chest radiographs using the segmentation technique, as well as to evaluate the properties of this DL-based completely version to enhance sensitivity while keeping low FP outcomes.

The following points highlight this newsletter's contributions:

- This study advanced a deep learning-based model for detecting and segmenting lung cancer on chest radiographs.
- Our dataset is extremely fine since all of the nodules/loads were pathologically tested lung tumors that had been pixel-degree labelled by radiologists.
- The segmentation strategy has become more informative than the categorization or detection procedures, which are no longer the most successful for lung cancer detection most cancers however also for observe-up and treatment efficacy.

II. Literature review

In 2021, techniques that use gene expression information are steeply-priced but notably correct. In evaluation, there may be a radiometric technique that is cost-effective although its accuracy is not aggressive.

P. Aonpong et al. [1] advised a genotype-guided radiomics technique (GGR) that results high accuracy and occasional fee. This approach is accomplished sequentially: pre-processing, radiomics characteristic extraction and selection (enter functions), prediction. This prediction, called GGR, includes two steps that use models. The first model uses gene estimation and the second model predicts recurrence using the anticipated gene. This approach uses the general NSCLC radiographic data set, which incorporates CT pix and gene expression information. Experimental results of this technique show that the prediction accuracy can be substantially progressed from current radiometric approach and ResNet50 to 83.28% by means of the proposed GGR. Statistics show that the main motive of disorder and demise in lung most cancers is how it is diagnosed inside the early tiers.

To examine the mutagenic reputation of Epidermal Growth Factor Receptor (EGFR), F. Silva et al. [2] advocated MLP for the very last category of EGFR mutation repute. This EGFR evaluation consists of the nodule, the lung containing the principle nodule, and each lungs. This proposed method consists of two major levels. The first section is the feature gaining knowledge of task. The 2nd section is an end-to-stop class version based totally on transfer mastering techniques. This approach makes use of the LIDC-IDRI and NSCLC-Radio genomics datasets. Experimental outcomes show that it has the exceptional capability to predict.

In 2020, the main motive of lung cancer demise and disease is how it's miles diagnosed in the early tiers. H. Yu et al. [3] proposed the "Adaptive Hierarchical Heuristic Mathematical Model (AHHMM)" for the automatic prognosis of lung cancer. This approach includes five steps. The first step is to get the picture. The second degree is Pre-processing. The 0.33 step is Binarization. Next, Thresholding and segmentation. Finally, characteristic extraction and detection by means of a Deep Neural Network (DNN). Modified K-manner clustering was extensively utilized for pre-class pics. Experimental results show that this approach has an accuracy of 96.67% of the lung cancer dataset. Radiologists frequently use screening for a huge range of CT scans for correct examination. Automated

algorithmic solutions may assist, however the dating among algorithmic solutions and physicians is also a assignment.

To solve this hassle, a gadget known as low-dose CT scans has been proposed through O. Ozdemir et al. [4] This system at once analyzes the CT scans and provides calibrated ratings. These are device-based totally three-dimensional convolutional neural networks. This technique is executed sequentially: Pre-process, Computer-Aided Detection module (CADE) for segmentation, Computer-Aided Diagnosis module (CADx). CADx relies upon on the overall performance of the CADE, so its miles developed and configured concurrently. This technique makes use of LIDC-IDRI, LUNA-16, Kaggle datasets. This device is more reliable in the real international. The proposed version can diagnose lung most cancers with ninety six.Five% accuracy.

Q. Zhang et al. [5] stated that many may be saved if lung cancer is recognized within the early levels. Early detection of lung cancer nodules is a completely challenging, time-ingesting, and repetitive undertaking for radiologists. They proposed a machine called a “Multi-Scene Deep Learning Framework (MSDLF)” with the “vesselness filter” to boom the accuracy and decrease the fake advantageous criteria. The predominant purpose of this analysis is to decide large nodes (> three mm). A version designed by 4-channel CNN. The technique includes the subsequent steps: Preparation of statistics set, mending of lung contour and segmentation of parenchyma, the removal of the vessel, the records set standardization, design of the CNN Design, segmentation, and class, normalized round sampling. The LIDC-IDRI dataset is used in this technique.

A. Masood et al. [6] said that drawing lung nodules manually by radiologists is a time-consuming and tedious assignment. To help radiologists, the system a 3D Deep Convolutional Neural Network (3DDCNN) is supplied. Their system works higher than superior systems. The aggregate of deep gaining knowledge of and cloud computing is used to correctly locate lung nodules. They used the Multi-Region Proposal Network (mRPN) in their architecture. The technique

consists of training Datasets, data augmentation, pre-processing, proposed model architecture, education technique, cloud-based 3DDCNN CAD machine. The ANODE09, LUNA-sixteen, LIDC-IDRI.

III. Proposed Model:

Convolutional neural networks and deep learning

Deep learning-trained Convolutional Neural Networks (CNN) have grown to dominate pattern detection, identification, segmentation, and classification applications in both medical and non-medical domains. Indeed, if enough training data is available, CNNs have essentially surpassed the previous generation of Radiomic/texture analysis algorithms discussed above. In our own study, after collecting and curating sufficiently large training sets at the end of 2016, our CNN-based techniques began to outperform the prior state-of-the-art texturing and SVM-based methods. While a thorough explanation of such strategies is beyond the scope of this article, it is important to appreciate the fundamental distinctions and advantages over earlier methods.

Feature learning vs. feature selection

Unlike radiomic/texture analysis approaches, CNN techniques create features from scratch rather than picking from a palette of manufactured or pre-selected sets that rely on the algorithm developer's contextual knowledge.

Hierarchical features

The initial few layers of a CNN often consist of multiple layers of features, allowing the network to learn the correlations between features in a far more complicated manner than a single feature extraction step can. Consider the following illustration: a textural characteristic, such as the local entropy of the joint histogram, can be utilised to detect hypotheses that extend into the parenchyma. A CNN, on the other hand, can discover that spiculations surround the whole perimeter of the nodule and that this is a symptom of a cancerous nodule.

End-to-end learning

CNNs are often trained "end-to-end," which means that the entire network is trained to optimise the problem of interest, adjusting all network parameters until the peak classification rate is obtained. In contrast, each stage of our LungX texture-based solution had to be designed and optimised separately, with no certainty that the entire pipeline would be optimum.

Segmentation-free

Because segmentation is performed implicitly inside the algorithm, the CNN technique may run without the nodular segmentation step. We discovered substantial sensitivity of the prediction score to the segmentation stage in a later investigation of our LungX system.

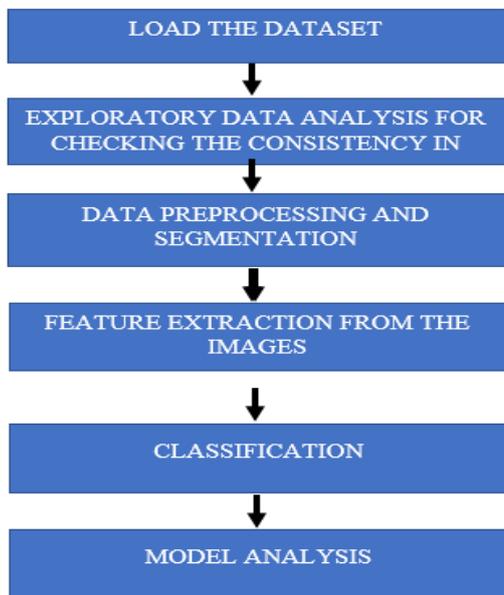


Fig 1 Flow chart of proposed system

IV. Operation of the Proposed Model

We utilised a dataset from the Kaggle datasets of CT Scan slices image datasets in this study. I first performed exploratory data analysis to ensure the consistency of the data. Then, using the graphs and plots from the analysis, we go to the following phase. That is data pre-processing, which includes segmentation, 3D visualisation, and volume rendering. Following this, we extracted attributes from the photos, including modality, image size, pixel spacing, the position of the CT Scan slice, and other image-related

data. Because the images were in dicom format, the procedure was considerably easier than with other formats. Following feature extraction came categorization and for that we had split the data into test and train with test size as 0.2 that is 20%. Then we did networking and keras tensor flow (CNN) model creation, and we received the model summary. And we used the matplotlib package to create a graph of accuracies and loss. Finally, characteristics were taken from the model that we created.

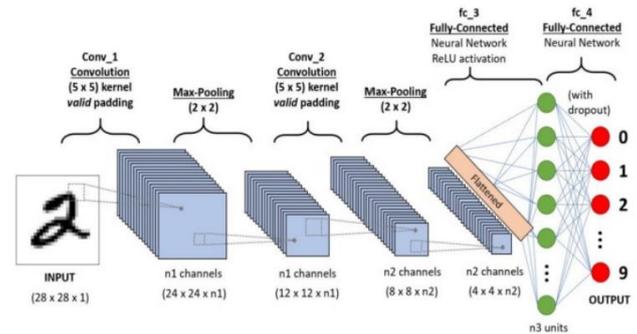


Fig2: CNN Architecture

The convolutional layer is the first parameter-containing layer that is passed on to the next Pooling layer. The linked layer is the most important layer component of Convolutional Neural Networks, with average(optimized) pooling and max pooling (CNNs). Which has been recognised for picture classification in computer vision.

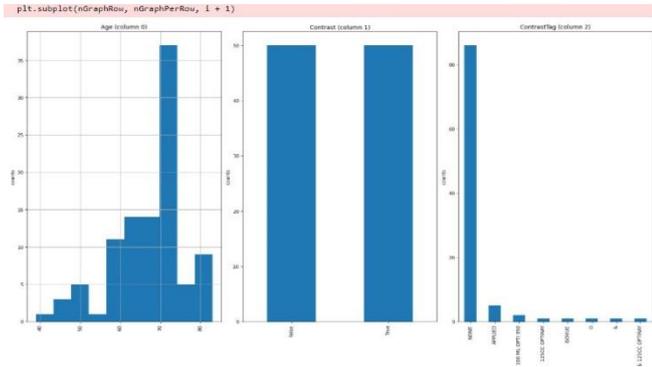
V. Result:

Step 1.EXPLORATORY DATA ANALYSIS

```

j>
  named: 0
  Age Contrast ContrastTag raw_input_path id tiff_name dicom_name
  0 60 True NONE .data\50_dicom_cases\Contrast00001(1).dcm 0 ID_0000_AGE_0060_CONTRAST_1_CT.tif ID_0000_AGE_0060_CONTRAST_1_CT.dcm
  1 69 True NONE .data\50_dicom_cases\Contrast00001(10).dcm 1 ID_0001_AGE_0069_CONTRAST_1_CT.tif ID_0001_AGE_0069_CONTRAST_1_CT.dcm
  2 74 True APPLIED .data\50_dicom_cases\Contrast00001(11).dcm 2 ID_0002_AGE_0074_CONTRAST_1_CT.tif ID_0002_AGE_0074_CONTRAST_1_CT.dcm
  3 75 True NONE .data\50_dicom_cases\Contrast00001(12).dcm 3 ID_0003_AGE_0075_CONTRAST_1_CT.tif ID_0003_AGE_0075_CONTRAST_1_CT.dcm
  4 56 True NONE .data\50_dicom_cases\Contrast00001(13).dcm 4 ID_0004_AGE_0056_CONTRAST_1_CT.tif ID_0004_AGE_0056_CONTRAST_1_CT.dcm
  
```

#COLUMN DISTRIBUTION PLOT



STEP 2: IMAGE VISUALISATION (DICOM AND TIFF FILE)

#PROCESSED TIFF DATA

```
tiff_data.head(5)
```

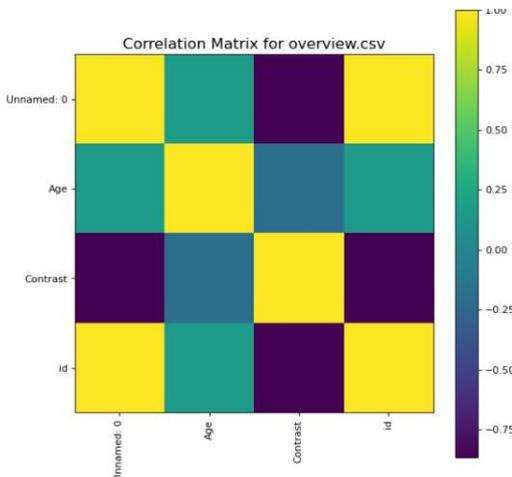
	path	file	ID	Age	Contrast	Modality
0	C:/Users/irvsk0/anaconda3/Lungancer/tiff_...	ID_0000_AGE_0060_CONTRAST_1_CT.tif	0000	60	True	CT
1	C:/Users/irvsk0/anaconda3/Lungancer/tiff_...	ID_0001_AGE_0069_CONTRAST_1_CT.tif	0001	69	True	CT
2	C:/Users/irvsk0/anaconda3/Lungancer/tiff_...	ID_0002_AGE_0074_CONTRAST_1_CT.tif	0002	74	True	CT
3	C:/Users/irvsk0/anaconda3/Lungancer/tiff_...	ID_0003_AGE_0075_CONTRAST_1_CT.tif	0003	75	True	CT
4	C:/Users/irvsk0/anaconda3/Lungancer/tiff_...	ID_0004_AGE_0056_CONTRAST_1_CT.tif	0004	56	True	CT

#PROCESSED DICOM DATA

```
dicom_data.head(5)
```

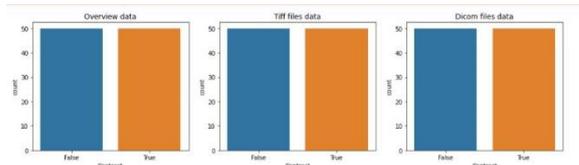
	path	file	ID	Age	Contrast	Modality
0	C:/Users/irvsk0/anaconda3/Lungancer/dicom_...	ID_0000_AGE_0060_CONTRAST_1_CT.dcm	0000	60	True	CT
1	C:/Users/irvsk0/anaconda3/Lungancer/dicom_...	ID_0001_AGE_0069_CONTRAST_1_CT.dcm	0001	69	True	CT
2	C:/Users/irvsk0/anaconda3/Lungancer/dicom_...	ID_0002_AGE_0074_CONTRAST_1_CT.dcm	0002	74	True	CT
3	C:/Users/irvsk0/anaconda3/Lungancer/dicom_...	ID_0003_AGE_0075_CONTRAST_1_CT.dcm	0003	75	True	CT
4	C:/Users/irvsk0/anaconda3/Lungancer/dicom_...	ID_0004_AGE_0056_CONTRAST_1_CT.dcm	0004	56	True	CT

#CORRELATION MATRIX



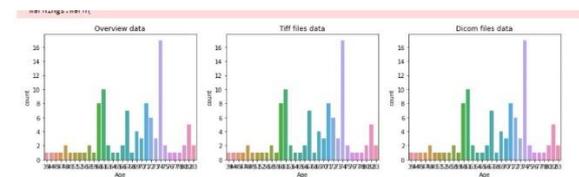
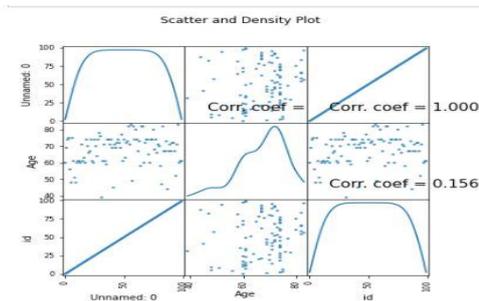
#COMPARISON PLOTTING

(CONTRAST)



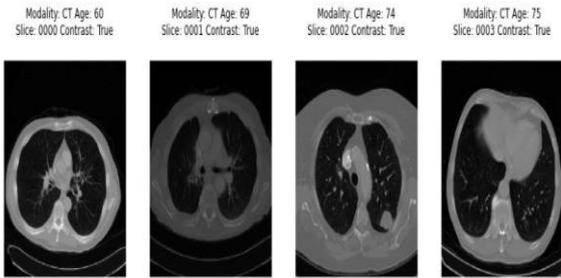
(AGE)

#SCATTERPLOT

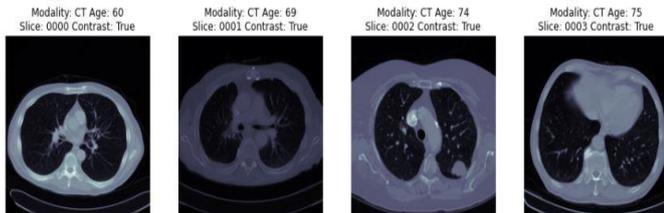


#SHOWING IMAGES

```
show_images(tiff_data,16,'TIFF')
```

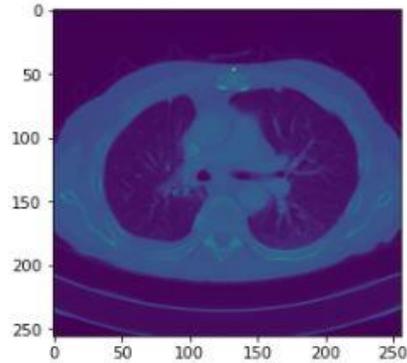


```
show_images(dicom_data,16,'DICOM')
```

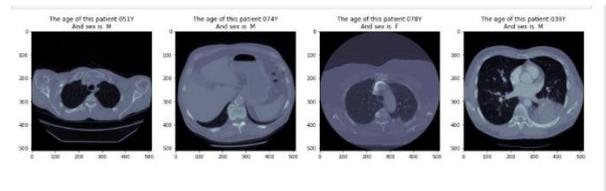


```
test_image = jimread(all_images_list[1])
plt.imshow(test_image[0])
```

<matplotlib.image.AxesImage at 0x1956fe9b9a0>

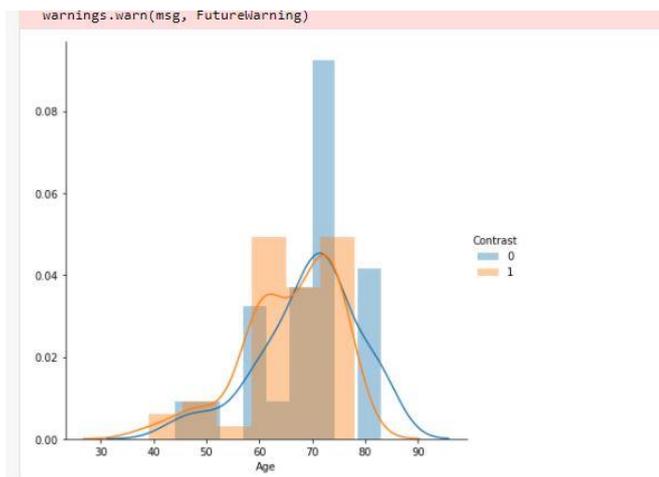


VISUALISING IMAGE WE ARE DEALING WITH-



STEP 2- PREPROCESSING AND SEGMENTATION

#PLOT (AGE VS CONTRAST)

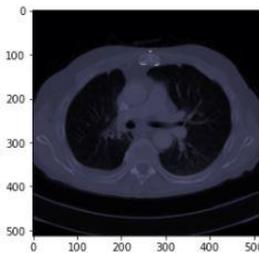


#TEST IMAGE

STEP 3- FEATURE EXTRACTION

File Location.....: C://Users//rvsk0//anaconda3//Lungcancer//dicom_dir
SOP Class.....: 1.2.840.10008.5.1.4.1.1.2 (CT Image Storage)

Patient's Name...: TCGA-17-Z011,
Patient ID.....: TCGA-17-Z011
Modality.....: CT
Study Date.....: 19820630
Image size.....: 512 x 512
Pixel Spacing...: [006.445312e-01, 006.445312e-01]
Slice location...: -220



STEP 4- CLASSIFICATION

#MODEL

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 256, 256, 32)	320
activation (Activation)	(None, 256, 256, 32)	0
batch_normalization (Batch Normalization)	(None, 256, 256, 32)	128
max_pooling2d (MaxPooling2D)	(None, 128, 128, 32)	0
dropout (Dropout)	(None, 128, 128, 32)	0
conv2d_1 (Conv2D)	(None, 128, 128, 64)	18496
activation_1 (Activation)	(None, 128, 128, 64)	0
batch_normalization_1 (Batch Normalization)	(None, 128, 128, 64)	256
max_pooling2d_1 (MaxPooling2D)	(None, 64, 64, 64)	0
dropout_1 (Dropout)	(None, 64, 64, 64)	0
conv2d_2 (Conv2D)	(None, 64, 64, 128)	73856
activation_2 (Activation)	(None, 64, 64, 128)	0
batch_normalization_2 (Batch Normalization)	(None, 64, 64, 128)	512
max_pooling2d_2 (MaxPooling2D)	(None, 32, 32, 128)	0
dropout_2 (Dropout)	(None, 32, 32, 128)	0
conv2d_3 (Conv2D)	(None, 32, 32, 256)	295168
activation_3 (Activation)	(None, 32, 32, 256)	0
batch_normalization_3 (Batch Normalization)	(None, 32, 32, 256)	1024
max_pooling2d_3 (MaxPooling2D)	(None, 16, 16, 256)	0
dropout_3 (Dropout)	(None, 16, 16, 256)	0
conv2d_4 (Conv2D)	(None, 16, 16, 512)	1180160

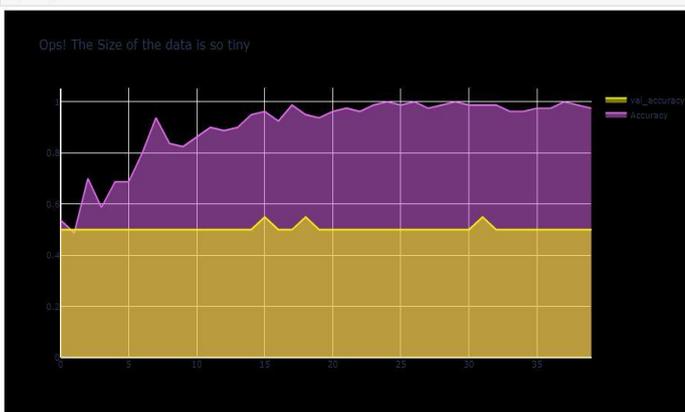
#MODEL SCORE

```

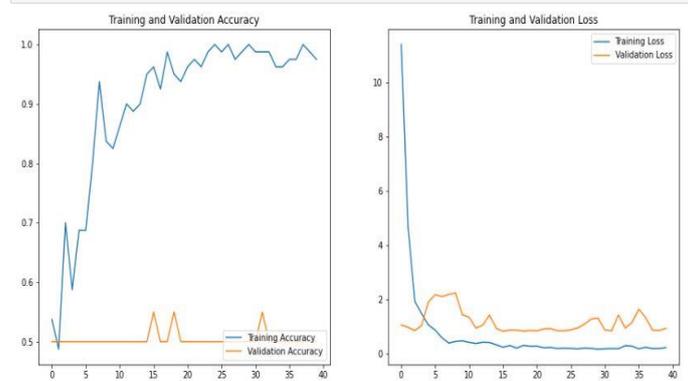
: score = model.evaluate(input_test, output_test, verbose=0)
: score
: [0.9387299418449402, 0.5]

```

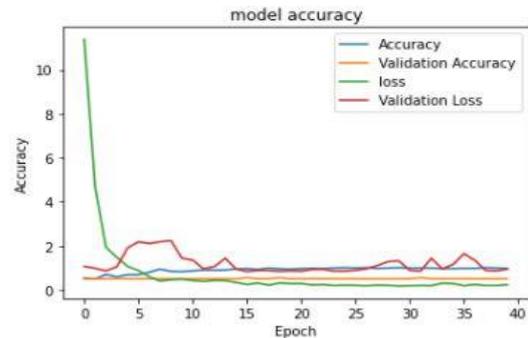
#PLOTTING ACCURACY AND VAL_ ACCURACY



#PLOTTING THE TRAINING ACCURACY AND VALIDATION ACCURACY, TRAINING LOSS AND VALIDATION LOSS



#VISUALISE TRAINING/VALIDATION ACCURACY AND LOSS



#FEATURE EXTRACTION

- Shields, and Otto Muzik. "Computerized Detection of Lung Tumors in PET/CT Images." 2006 International Conference of the IEEE Engineering in Medicine and Biology Society (2006): n. pag. Web.
- [15]. "Getting Started with the Keras Sequential Model." Guide to the Sequential Model - Keras Documentation. N.p., n.d. Web. 12 June 2017.
- [16]. Golan, Rotem, Christian Jacob, and Jorg Denzinger. "Lung Nodule Detection in CT Images Using Deep Convolutional Neural Networks." 2016 International Joint Conference on Neural Networks (IJCNN) (2016): n. pag. Web.
- [17]. Hawkins, Samuel H., John N. Korecki, Yoganand Balagurunathan, Yuhua Gu, Virendra Kumar, Satrajit Basu, Lawrence O. Hall, Dmitry B. Goldgof, Robert A. Gatenby, and Robert J. Gillies. "Predicting Outcomes of Non-small Cell Lung Cancer Using CT Image Features." *IEEE Access* 2 (2014): 1418-426. Web
- [18]. Neural Network Models (supervised)" 1.17. Neural Network Models (supervised)
- [19]. Scikitlearn 0.18.1 Documentation . Google, n.d. Web. 12 June 2017.
- [20]. Lynne Eldridge MD. (2013, March 22). Lung Cancer Survival Rates by Type and Stage [Online].
- [21]. Available: <http://lungcancer.about.com/od/whatislungcancer/a/lungcancersurvivalrates.htm>. [10.]
- [22]. Morphological Operators, CS/BIOEN 4640: Image Processing Basics, February 23, 2012.
- [23]. Image Processing -Laboratory 7: Morphological operations on binary images, Technical University of Cluj-Napoca, Computer Science Department
- [24]. Albert Chon, Peter Lu, Niranjan Balachandar "Deep Convolutional Neural Networks for Lung Cancer Detection".
- [25]. Wavelet Recurrent Neural Network for Lung Cancer Classification":3rd ICSTcomputer,2017.
- [26]. Khare, R.K., Sinha, G.R. and Kumar S., (2015), CAD for Lung Cancer Detection: A Review, *International Journal of Modern Trends in Engineering and Research (IJMTER)*. Vol. 2. Issue 7, pp 333-338.
- [27]. Khare, R. K., Sinha, G. R. and Kumar S., (2017), Fuzzy Based Contrast Enhancement method for Lung Cancer CT Images, *International Journal Of Engineering And Computer Science*. Vol. 6, Issue 5. pp 21201-21204.
- [28]. Khare, R. K., Sinha, G. R. and Kumar S., (2017), Mass Segmentation Techniques For Lung Cancer CT Images, *International Journal on Recent and Innovation Trends in Computing and Communication*, Vol. 5, Issue 11, pp 184-187.
- [29]. Khare, R. K., Sinha, G. R. and Kumar S., (*017), Cancer Detection Using Neuro Fuzzy
- [30]. Classified in CT Images, *International Journal on Future Revolution in Computer Science & Communication Engineering* . Vol. 3, Issue 12, pp 258-261