Cancer Detection using Lung CT Images

[1] Vaishnavi Rao, [2] Suhas M Suresh, [3] Shrinidhi P Shetty, [4] Omkar T P, [5] Kavitha Sooda
[1,2,3,4,5]Department of Computer Science & Engineering,
B.M.S. College of Engineering, Bangalore, India
Email: {[1]raovaishnavi98@gmail.com, [2]suhasagasthya@gmail.com, [3]nidhi.shetty.1106@gmail.com, [4]omkartp98@gmail.com,
[5]kavithas.cse@bmsce.ac.in}

***Abstract*— Higher death rate has been due to cancer based on past history and particularly due to lung cancer. Manual diagnosis, evaluation of CT scans and identifying the type of disease is time consuming and hence leading to slower identification of cancer. Identification of small cell lung cancer cells through automated analysis of lung CT images. Early and quick analysis of lung cancer counters the high mortality rate by detecting at an early stage which can be achieved through deep learning. The proposed method of diagnosis consists of image processing and cancer detection stages to identify the presence of cancerous cells in the CT images of the patient. The diagnosis performs three main image processing steps to distinguish presence of cancer nodules in the lung through CT images of the patient. These are carried out using neural networks – specifically the CovNets (CNNs). The performance of the proposed system shows satisfactory results and the proposed method gives 88.86 percent accuracy.**

Keywords—Computer Tomography, Lung Cancer, Nodules, Tuberculosis, Image Processing, Convolutional Neural Networks

I. INTRODUCTION

Cancer has been widely known to be the highest cause of deaths among the human population all over the world. Lung cancer is known to be second largest cause of death after heart attack. Thus, it is the most prominent cause of cancer-related deaths [1]. Essentially, lung cancer refers to uncontrolled growth in the lungs which leads to formation of lung nodules. Every cell undergoes division to form new cells. Failure of this activity hence leads to growth of the nodules, also termed as tumor. Nodules that can be removed and cause no serious damage are benign. Nodules that grow aggressively and into other body parts are malignant. Fig. 1 shows a cancerous CT scan.

Early detection of these malignant nodules are generally difficult; and this, coupled with poor environmental conditions due to high air pollution, unhealthy smoking habits and other harmful practices prove to be the main reason for discovery and treatment of cancer at a late stage, when nodules have grown into other areas and the patient's life is at stake [2].

The main methods used by doctors for discovering any lung related issues include sputum cytology (study of phlegm in the lungs to determine if the cells are cancerous), computerized tomography (CT scanning), X- rays, MRI scanning, etc. Drawbacks when it comes to the aforementioned methods is that they take up a lot of manual labor and time. This hence causes delay in detection of diseases, especially time – crucial diseases like lung cancer and tuberculosis.

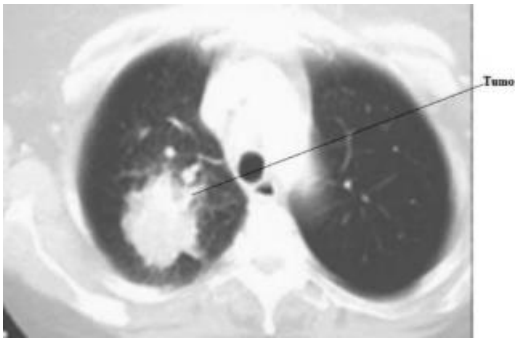

Fig 1: CT image of lung cancer patient

Interpreting and identifying diseases from medical images has been made easier through constantly improving technology. This includes aided diagnosis and detection by computers [5].

CT scans are essentially 3D imaging wherein several 2-dimensional x-ray images construct a single 3-dimensional image, thus inherently increasing accuracy.

Despite popular use in lung disease diagnostics, CT scans still require certain processing such that they can be easily readable by the human eye. Figure 1 displays CT scan of cancer nodules in the lungs [6]. Thus, computer aided diagnosis is helpful in quick identification and classification of the type of lung diseases a patient might have with more accuracy than any other implementation [7]. Furthermore, there's preference for CT scans as the noise in the images are lesser. This automatically makes it easier to find nodules and nodes, thus making the images sensitive [7].

Many machine aided methods employ a combination of deep learning or ML algorithms with DIP or preprocessing steps, with various implementations tried and tested for diagnostic purposes.

Neural networks aid in the implementation of machine learning (ML). In this, training examples are used and analysed by a computer to perform a task. A model structured to be an object recognition model would use a training set and find visual patterns in the images that consistently correlate with particular labels [9].

Neural networks are popularly organized to have nodes layered. These will be feed-forward, i.e., data moves in only one direction through the layers.

Convolutional neural networks are widely used for image processing, segmentation, detection, recognition, etc.

The aim of proposed system is to represent a fast and robust system for detecting lung cancer at an early stage and our proposed system provides a higher accuracy than many

other existing techniques. The system provides early detection of cancer at a low cost. In our project, we will be using CT scan images of different patients having lung cancer to analyze and classify the datasets using CNN algorithm which increases the possibilities of early detection. Since CNN's works by training a large dataset and the model built with large dataset, the accuracy of the system is supposed to be high [10]. The whole proposed system is planned to be implemented as a web-based application and also as an API that can be integrated with the existing CT scan systems at the radiology laboratories. Thereby facilitating faster and cost saving analysis of CT images and also provide a secondary means of verifying the results obtained from a radiology laboratory.

II. Literature Survey

The traditional flow of analysis in any computer aided CT image diagnostics begins with an initial image preprocessing. This stage enhances the CT images such that ease of diagnosis is accomplished. It follows many methods such as noise reduction, smoothing, filtering, etc. This stage is followed by segmentation of the lung region and calculation of region of nodules. The CT images are segmented so that only the region of interest is selected, i.e., the lungs and the areas of infection within the lungs. This stage helps minimize the regions of feature extraction and focus on where the nodules will most likely be present [11]. It involves processes such as morphological opening, closing, edge detection, region growing, etc. The next step is performing feature extraction to obtain initial diagnosis. Here, there are two sub – stages, namely shape based extraction and feature based extraction. This stage is significant for its identification of presence or absence of any abnormality or malaise in the segmented lung regions. Finally, passing through a machine learning algorithm that classifies the disease based on the initial diagnosis, thus a final diagnosis is provided as an output [12].

Also, The CT scan images are ideal due to the quality of the content present in them along with the images having lesser noise ratio compared to images obtained from other techniques. These images are comparatively more sensitive and easier detection of the size of the tumor is possible. CT scans are more sensitive than chest radiography for the detection of tumorous growths. Manual interpretation of a huge amount of CT images is time consuming for radiologists. CT scans allow visualization of the lungs within a few seconds. Typically, one CT scan with highly detailed anatomy contains over 400 slices, i.e., 400 images [13]. A single slice of CT scan contains the tumor and the cancer part in the lung area. This lung area is so small that it could be smaller than half of the lung region. Manual segmentation of the CT images is thus very difficult. Moreover, in the CT images, it is tough to distinguish very small tumors in denser area regions of the lungs. In addition to this, differences of intensity in CT scan images and misjudgment by doctors and radiologists on the structure of anatomy seen in the scans might cause trouble in marking the cancerous cell [14].

In recent years, the latest research of lung cancer detection include detection using machine learning and multinomial Bayesian and Bayesian classifier, detection using BPNN and SVM, size estimation of nodule using image segmentation and back propagation, etc. The future models for detection of diseases may include a cognitive approach that may consist of intelligent networks or a Bayesian model [15].

Considering this process of image processing that in involved of a CT Image and the drawbacks of the Xray over CT Images, an alternate system is proposed that uses the newer techniques and methodologies such as a Neural Network (CNN) which is faster in processing, less image processing that has to be done manually and can accommodate the huge datasets [16].

III. Methodologies

For the implementation of the proposed solution, the dataset that we have used is LUNA 16 (Lung Nodule Analysis) part of a Kaggle competition. The dataset consists of 9 subsets and has candidates and annotations csv files where candidates_v2.csv file which contains the filename, class to which the CT scan belongs and the x, y, z co-ordinates of the lung nodules. The annotations.csv contains filename and diameter of the cancerous cell in that particular image with its voxel co-ordinates. The high-level design of the proposed system can be seen in Fig 2, which includes two phases of image processing.

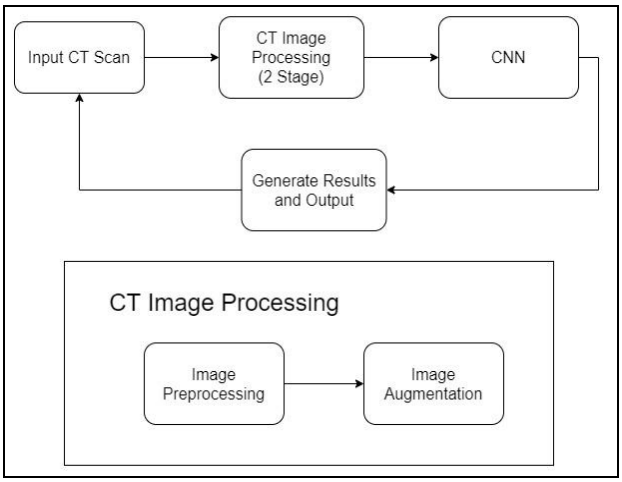

Fig 2: Proposed System Workflow

The main flow of analysis in any computer aided CT image diagnostics begins with an initial image preprocessing. This stage enhances the CT images such that ease of diagnosis is accomplished. It follows many methods such as noise reduction, smoothing, filtering, etc. we perform multiple operations in order to convert 3d to 2d image, RGB to greyscale conversion from the CT scans, creation of sub images from the CT scans, etc. This stage is followed by augmentation of the lung images.

The dataset we found was skewed, meaning that the dataset contained majority of negative images rather than equal amount. This would lead to a biased model and class-based accuracy impact. In order to fix this, we had to explore data augmentation as an option. Through which we could balance the dataset, by increasing the number of positive images by generating them using the existing images instead of downloading new ones. The data augmentation technique applied in our proposed system was rotation. We rotated each positive image by 90,180 and 270 degrees in order to generate three new images using one, which increased our dataset threefold and made the dataset balanced and unbiased.

Fig 3 shows the flow of stages in the proposed system that has been implemented.

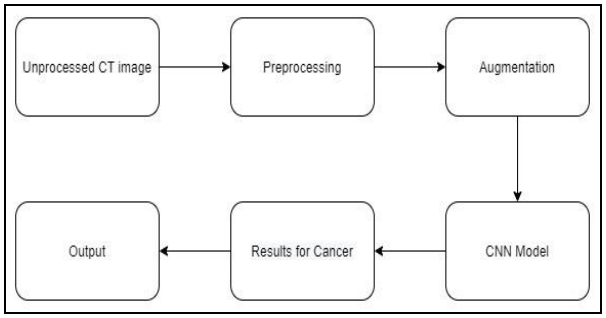

Fig 3: High Level design of the proposed system.

Finally, passing through CNN network that flattens the image, extracts features, identifies patterns and classifies the image as cancerous or not and provides an output of the same to the user.

IV. IMPLEMENTATION

The total workflow can be sub divided into stages of image processing, classification and result generation. Each stage further consists of sub processes that need to be executed.

Preprocessing assists in suppressing information irrelevant to the diagnostics process, thereby decreasing computational time and error rate. Therefore, improving quality of relevant features and simultaneously suppressing unwanted noise and fluctuations is the aim to be achieved.

Before proceeding with preprocessing methods, a dataset of a considerable size is acquired, consisting of authentic CT images of both normal lungs and lungs with disease (i.e., lungs with nodules). This dataset of raw images is then used as an input for the preprocessing stage.

The input raw images are processed in two phases – Image preprocessing and augmentation. After the processing is complete, the data extracted from the image is fed to the classifier to generate results. The main objective of image preprocessing is to enhance the image visual quality in order to diminish the irrelevant parts of the images. During preprocessing we perform multiple operations in order to convert 3d to 2d image, RGB to grayscale conversion from the CT scans. The preprocessed dataset is created in 3 parallel mode, that is three images are created at once from the huge dataset. This ensures speedy delivery of the preprocessed dataset required. Fig 4 shows the flow of preprocessing.

Initially the raw CT images are given as input to the CT Image class where the .mhd format images are read. The voxel coordinates are obtained from that image using the candidates_v2.csv file which contains the image name and corresponding x, y, z coordinates. The voxel coordinates are 3 dimensional coordinates where a possible lung nodule formation is suspected. From the voxel coordinates obtained, the whole three-dimensional CT image is cropped in and around the coordinates to create a sub image which eases the process of detecting the cancer by decreasing the area, volume and structure.

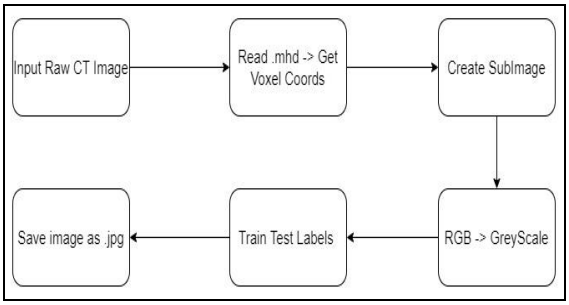

Figure 4: Image preprocessing workflow

Every time a sub image is created, the three-dimensional CT scan slice is converted to a two-dimensional image with the mapping of the data. The two dimensional image further eases the process of detecting the cancer since all three coordinate mapping are converted to two coordinate mappings. Then the normalizePlanes() function of CTScan Class is called whose primary objective is to convert the hounds units to grayscale units. The image is then saved for data processing and augmentation. The whole of the dataset is divided into test train labels and saved to a specific directory under the name train and test.

Based on initial investigations, the dataset was found to be skewed or biased towards negative image. This unbiasedness can lead to poor class specific accuracy. Since the model we are using is CNN, it is data dependent and larger dataset leads to better accuracy. In order to make the dataset unbiased and balance the percentage of positive to negative images given as input for training, image augmentation was done. Image augmentation is a method to artificially expand the dataset without acquiring new images but by applying various image transformation techniques on the existing images. Such techniques include random rotations, shifts, shear, color space variations, flips, adding random noise, resize, transformations, varying exposure and so on.

For our proposed system, we have rotated the existing image by 90, 180 and 270 degrees in order to generate three new images using one image. This increases the biased class by three times and balances the training data to be fed as input to the CNN model.

In image processing of the CT images for diagnosis, this is the final step. Its main aim is to isolate and choose desired segments or features from the images. When a large dataset is considered, and it has a high level of redundancy, it is efficient to have the input as a representation set of features. Transforming the input to obtain features defined in a set is known as feature extraction. Calculated features are represented in a vector format and act as the input to the classifier. These can either be texture features or geometry.

Geometry based features include perimeter, shape, size, are, etc. and energy, entropy, correlation, etc. texture-based features are considered. These features are then used as training features to develop classifier, and thence classify the input model of processed CT images [6].

The techniques for extracting shape are of two types: region and contour-based methods. The entire ROI is considered for region based. While contour-based strategies emphasize on information of contour of an object.

Regions are characterized by their geometric properties, and these form the basis for shape-based feature extraction or

geometry-based feature extraction. The basic geometric properties are measured in scalar, which include perimeter, area, eccentricity [8].

A neural network is employed for diagnosis. A multilayer feed forward neural network with supervised learning method is reliable and efficient for the purpose of detection of cancer nodules. Convolutional Neural Networks are essential for our model because of their significant ability to reduce the images into a form where processing can be performed easily, and without losing features that are critical for getting a good prediction [25].

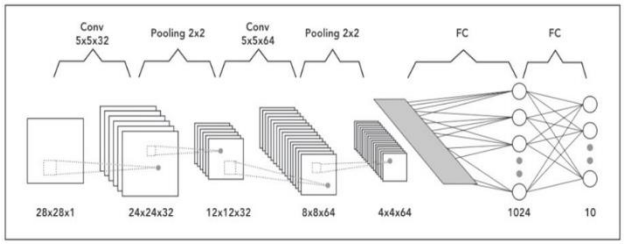

Fig 5: General CNN Architecture

The dataset chosen entails a large number of CT images, with dataset size going up to 60 gigabytes. Along with a large size, the dataset includes a variety of features. CNNs are highly scalable and can work with a huge amount of features, thus overcoming the limitations we face. Therefore, Convolutional Neural Networks are most suited for our model, and we have chosen this architecture to implement our model.

Computers interpret input images as pixels arranged in arrays based on resolution. Through this, three values are obtained, namely, height, width and dimension.

The first layer in the model is a convolution layer. Feature extraction for the input image occurs in this layer. It uses focused image squares to learn the features while maintaining the integrity of the pixels and their relationship with each other. The operation of convolution takes an input of two matrices: the input image and a filter matrix [25].

The two types of features to be extracted are differentiated as shape based and texture based features. These features, on applying a filter, will be obtained for further processing and subsampling through the remaining layers of the CNN.

Convolution can perform different operations based on different filters applied. Operations such as edge detection, Gaussian blur, sharpening of image, etc. can be performed. The same can be seen in Fig 6.

The input image will have the dimensions as 32x32x3 where the pixel values of width x height are 32, along with 3 values for the RGB values of the image.

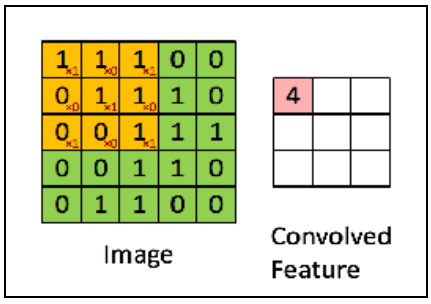

Fig 6: Convolution operation

Referring to fig 7, The convolution layer will compute the output by finding the dot product of weights of a small region of the image and the input volume associated with it. If 3 filters are used, the resultant dimensions are 32X32X3 [26]. This layer acts as a image feature extraction of ct images. It also extracts spatial and temporal dependencies through appropriate filters and activation functions. These features will be used by pooling layer for subsampling of other features.

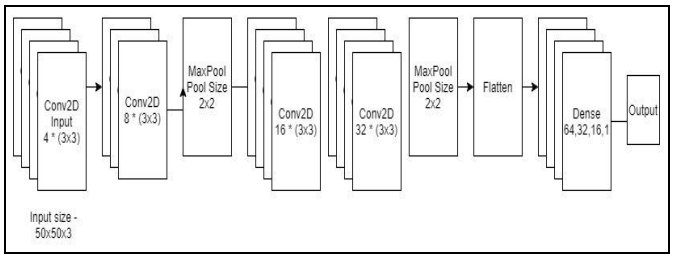

Fig 7: Architecture of Proposed CNN

Pooling Layer – Pooling reduces the spatial size of the convolved feature. It also decreases the computational power required to process the data. Pooling layer facilitates the extraction of rotational and positional independent features. The two types of pooling include max and average pooling. In our proposed system, max pooling layers of 2x2 are included which reduces the noise of the data and also reduces the dimensionality. The same can be seen in Fig 8 which shows two types of pooling,

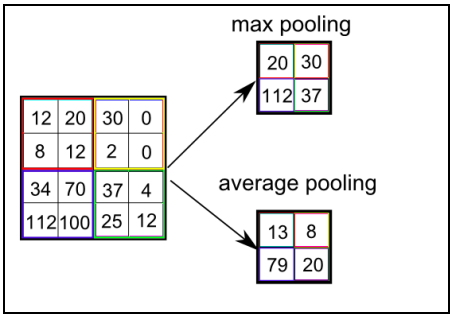

Fig 8: Pooling and its types

Flatten and Dense Layer – Responsible for converting the 2d image matrix data into 1d column vector. The column vectors are fed into the feed forward neural network with back propagation where the weights are adjusted and model is trained for best fit. Based on the proposed architecture consists of sandwiched maxpool layers between two convolutional layers which is similar to a VGG net. The reason for the above sandwich is that it increases the accuracy and the model is optimal for the given CT image input. A flatten layer converts the 2d matrix to single column matrix which is then given as input to the various dense layers with ReLU activation function. The last activation function is Sigmoid and gives a binary output. The model generated after the training is complete is saved, and used for the web interface.

A challenge faced was tuning the model to give acceptable accuracy and give proper output. In order to increase the accuracy, there were multiple trials in order to tune the parameters, we experimented on the layers of CNN increasing the Convolutional layers and adding maxpool layers in between them to reduce the dimensionality of the

convolved feature matrix. This increased the accuracy of the model by almost 10%.

Based on the output from the classifier the CT image that was given as input is classified as either normal or having cancer.

After necessary hyper parameter tuning such as including maxpool layers and increasing convolutional layers and adjusting the model to resemble a VGG net, the model that we developed had a train accuracy of 99.99987 percent with test accuracy of 88.8669 percent. The performance of the model was evaluated considering the four-evaluation metrics that is val_acc, val_loss, accuracy and loss.

V. Conclusion

CT images have been a boon to medical analysis since its inception. It helps to identify muscular and tissue related abnormalities. With the advancements in image processing and inferences using various algorithms, the applications and obtaining of results using CT scans have found exponential growth. In this paper, we aimed to do early detection of cancer, If the doctor had suspected other diseases other than cancer, this system would provide correct insights. Providing verification and assistance to the radiology laboratories over their conclusions of a CT scan.

VI. Future Enhancement

Future Enhancements for the proposed system can include several more parameters can be added to the algorithm to detect a greater number of diseases automatically in the CT scan image. This system can be expanded to non-lung areas to further increases the automated detection. The algorithm and image processing techniques and filters can also be specialized based on the disease under consideration to provide more efficient and accurate results.

REFERENCES

[1] Jin, X., Zhang, Y., & Jin, Q. (2016) "Pulmonary Nodule Detection Based on CT Images

[2] Using Convolution Neural Network." 2016 9Th International Symposium On Computational Intelligence And Design (ISCID), 2016

[3] Nunes É.D.O., Pérez M.G., Medical Image Segmentation by Multilevel Thresholding Based on Histogram Difference, presented at 17th International Conference on Systems, Signals and Image Processing, 2010.

[4] Rendon-Gonzalez, E., & Ponomaryov, V., "Automatic Lung nodule segmentation and classification in CT images based on SVM." 2016 9Th International Kharkiv Symposium On Physics And Engineering Of Microwaves, Millimeter And Submillimeter Waves (MSMW), 2016.

[5] R. Mason, F. Murray, J. Nadel, "Mason, R. Murray & Nadel's Textbook of Respiratory Medicine 4th Edition", Elsevier, 2005.

[6] R. Mason, F. Murray, J. Nadel, "Mason, R. Murray & Nadel's Textbook of Respiratory Medicine 4th Edition", Elsevier, 2005.

[7] Suren Makaju, P.W.C. Prasad, Abeer Alsadoona, A. K. Singh, A. Elchouemic, "Lung Cancer Detection using CT Scan Images", 2018

[8] Roy, T., Sirohi, N., & Patle, A., "Classification of lung image and nodule detection using fuzzy inference system." International Conference On Computing, Communication & Automation, 2015

[9] Shapiro L.G., Stockman G.C., Computer Vision: Theory and Applications, Prentice Hall, 2001.

[10] Jaspinder Kaur, Nidhi Garg, Daljeet Kaur, "Segmentation and Feature Extraction of Lung Region for the Early Detection of Lung Tumor", 2012

[11] Md. Badrul Alam Miah & Md. Abu Yousuf, "Detection of Lung Cancer from CT Image Using Image Processing and Neural Network", 2015

[12] Ganesh Sable, Harsha Bodhey, "Lung Segmentation and Tumor Identification from CT Scan Images Using SVM", 2012

[13] Shubhangi Khobragade, Aditya Tiwari, C.Y. Pati and Vikram Narke, "Automatic Detection of Major Lung Diseases Using Chest Radiographs and Classification by Feed-forward Artificial Neural Network"

[14] K. Veropolous, C. Campbell, G. Learnmonth, B. Knight, J. Simpson, "The Automated Identification of Tubercle Bacilli using Image Processing and Neural Computing Techniques", 1998.

[15] Kavitha Sooda, T R Gopalakrishnan Nair, "The future networks—a cognitive approach", Cognitive Informatics, Computer Modelling, and Cognitive Science, 161-176, 2020. doi.org/10.1016/B978-0-12-819443-0.00009-X

[16] P. Stark, "Use of imaging in the staging of non-small cell lung cancer", Up To Date, 2008.

[17] C. Jacobs, E. M. van Rikxoort, T. Twellmann, E. T. Scholten, P. A. de Jong, J. M. Kuhnigk, M. Oudkerk, H. J. de Koning, M. Prokop,C. Schaefer-Prokop, and B. van Ginneken, "Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images", Medical Image Analysis, 2014.

[18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", Nature, vol. 521, pp. 436444, 2015.

[19] M. Firmino, A. H. Morais, R. M. Mendona, M. R. Dantas, H. R. Hekis, and R. A. Valentim, "Computer-aided detection system for lung cancer in computed tomography scans: Review and future prospects", Biomedical Engineering Online, vol. 13, p. 41, 2014.

[20] Sanjukta Rani Jena, Dr. Thomas George, Dr. Narain Ponraj, "Feature extraction and classification techniques for the detection of lung cancer: a detailed survey", 2019

[21] Junji Shiraishi, Hiroyuki Abe, Feng Li, Roger Engelmann, Heber MacMahon, Kunio Doi, "Computer-aided Diagnosis for the Detection and Classification of Lung Cancers on Chest Radiographs: ROC Analysis of Radiologists' Performance", 2006

[22] Suzuki, K., Kusumoto, M., Watanabe, S. I., Tsuchiya, R., & Asamura, H. (2006) "Radiologic classification of small adenocarcinoma of the lung: radiologic-pathologic correlation and its prognostic impact," The Annals of Thoracic Surgery. 81(2): 413-419.

[23] B. Gupta and S. Tiwari, " Lung Cancer Detection using Curvelet Transform and Neural Network," International Journal of Computer Applications, 2016.

[24] Venkateshwarlu K., Image Enhancement using Fuzzy Inference System, in Computer Science & Engineering, Master thesis, 2010.

[25] https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148

[26] http://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

[27] https://cs231n.github.io/convolutional-networks/

Vaishnavi Rao is pursuing B.E in Computer Science and Engineering, B.M.S. College of Engineering. Her interests include Deep Learning, Artificial Intelligence and Data Analytics. She is currently working as a Software Development Test Engineer at Akamai Technologies.

Suhas M Suresh is pursuing B.E in Computer Science and Engineering, B.M.S. College of Engineering. His interests include Distributed Systems, Private and Public Cloud and System Architecture. He currently works as Operations Engineer Intern at Endurance International Group, Bangalore

Shrinidhi P Shetty is pursuing B.E in Computer Science and Engineering, B.M.S. College of Engineering. Her interests include Application Security, C programming and UI Design. She currently works as Intern at Zscaler Softtech India Pvt.Ltd

Omkar T P is pursuing B.E in Computer Science and Engineering, B.M.S. College of Engineering. His interests include Artificial Intelligence, Networking and UI Development. He currently works as Intern at Hewlett Packard Enterprise R&D.

Kavitha Sooda holds Ph.D in Computer Science and Engineering. She has eighteen years of teaching experience and completed her Post-Doctoral work on Higher Education System from IISc, Bengaluru. Her interest includes routing techniques, QoS application, Cognitive Networks, Evolutionary Algorithms and Higher Education System. Currently she works as Associate Professor at B.M.S. College of Engineering, Bengaluru