Lung Cancer Detection using CT Scan Images

Syed Abudhagir Umar, Chenigaram Abhigna, Cheemalapati Venkata Vignesh, Teki Rohith Raj

***Abstract:** Lung cancer is a fatal disease that takes numerous lives every year around the world. However, detecting this disease in its initial stages can help save the lives of the people. CT imaging is the best technique used for imaging in the field of medical sciences. It is used by doctors but it is hard for medical examiners to decipher and recognize cancer through the computer-assisted tomography scan images. Hence, Computer-aided diagnosiswill be very supportive for the medical examiners to identify and recognize the cancerous nodules in cells precisely. The primary agenda of the project is to assess diverse computer based methods, explore present finest method, deduce its limitations and setbacks. Then, proposing a latest model with upgrades and advancements to the present leading model. Techniques appliedfor diagnosis of lung cancerare organized based on theprecision. Numerous methods were surveyedon everystride and the complete limitations and setbacks were identified. A lot of techniques had low precision and few had high precision. But none of those were satisfying. Therefore, our target is to increase the precision of themodel.*

***Keywords:** Computer-Aided Diagnostic, CT Scan Images, Cancer, Image Processing, Lung Cancer Diagnosis.*

I. INTRODUCTION

Lung cancer is considered quite possibly primitive and deadliest sickness in the modern world. On the other hand, premature identification and medical assistance of the sickness will help to save numerous lives. Although the CT scan imaging technique is the best medical field around the world, it is Hard for clinicians to recognize and spot the cancerous cells from the CT scan pictures. As a result, computer-assisted diagnostics will a service to clinicians in properly identifying malignant cells. Lung cancer gets more dangerous as the stages of the cancer cells increase in the lungs. This makes it the highest harmful and widely disseminated cancer in the world. The agenda of the study is to find cancer cells in the lungs at an early stage.

Manuscript received on 02 August 2022 | Revised Manuscript received on 07 August 2022 | Manuscript Accepted on 15 October 2022 | Manuscript published on 30 October 2022.
* Correspondence Author

Dr. Syed Abudhagir Umar, Department of Electronics and Communication Engineering, B V Raju Institute of Technology, Vishnupur (Telangana), India. E-mail: syedabudhagir.u@bvrit.ac.in

Chenigaram Abhigna*, Department of Electronics and Communication Engineering, B V Raju Institute of Technology, Vishnupur (Telangana), India. E-mail: 18211a0445@bvrit.ac.in

Cheemalapati Venkata Vignesh, Department of Electronics and Communication Engineering, B V Raju Institute of Technology, Vishnupur (Telangana), India. E-mail: 18211a0444@bvrit.ac.in

Teki Rohith Raj, Department of Electronics and Communication Engineering, B V Raju Institute of Technology, Vishnupur (Telangana), India. Email: 18211a04m5@bvrit.ac.in

The CAD system, which is an interdisciplinary approach based on Image Processing and Machine Learning approaches, is used to detect lung cancer. Because of the formation of cancer cells, where the majority of cells are overlaid, anticipating lung cancer is the most difficult difficulty. Image processing techniques are drawing attention in various medical fields for detection and therapy purposes. The most significant component in determining the abnormality concerns in the targeted image is the time factor. The importance of image quality and accuracy in identifying diseases quickly cannot be overstated. The enhancement stage determines how image quality is assessed and progressed. We can readily identify the tumors in an image with the help of image processing methods such as image enhancement, image segmentation, and feature extraction. The development of a Computer based Detection set-up for the quick identification of cancerous nodules. utilizing an integrative method built on machine learning techniques and image processing methods.

II. RELATED WORK

[1] provided a method which gives classification between nodules and regular biological architecture of a lung. In this method, they extracted various features. The classifier and optimal thresholder for segmentation used is Linear Discriminant Analysis. The accuracy is 84%, sensitivity; 97.14% and specificity;53.33%

[2] used CNN in their Computer Aided Detection in order to identify the lung cancer. It is found the precision is 85%, sensitivity is 83% and specificity is 87%. The primary benefit of this method is that they used circular filter in Region of interest (ROI) extraction phase, this usage reduced the expense of preparing and acknowledgment strides. However, even though the implementation cost is decreased, The model didn't have satisfactory precision.

[3] used K mean unsupervised learning algorithm for clustering or segmentation. This gathers the pixel dataset based on Particular Attributes. The Back propagation network was used for classification in the particular method The Model's precision is 91%. Image pre processing Technique used in this model is also used in later models for removal of noises.

[5] created a watershed division based framework. Gabor Filter utilized to work on nature of the picture. It contrasts the exactness and brain fluffy model and area developing technique. The framework has a precision of 90.1% which is relatively higher than the model with division utilizing brain fluffy model and locale developing strategy.

This model purposes marker controlled watershed division is used to remove the over segmentation problem. This is the main advantage of the system

[4] applied fuzzy interference system to identify cancerous nodule in their model. Gray transformation is used for image contrast enhancement and Prior to Segmentation Image binarization is implemented. The resulting image is then separated using an active compiler method. Cancer classification was executed by applying a non-invasive technique. Factors such as location, description, entropy, correlation, maximum axis length, and minimum axis length were taken to train the separator. the precision of the model is 93.99%. The disadvantage is that it does not catergorize cancer as dangerous or not dangerous.

[6] explored different study classification methods to classify the at hand information in the UCI machine learning area into cancerous and non-cancerous. The input data is prepossessed. After binarisation, commonly used tool in Weka tool is chosen to divide the dataset. By compassion it is found that, RBF classifier has shown a wonderfull precision of 81%. RBF classifier is said to be effective classifier method for Lung cancer data prediction.

[7] here, preexisting lung cancer patients' data is collected and sent for image processing. Matlab is used to study the data set that includes diagnostic image information. This is accomplished by the segmentation and classification of medical images. The patients' Computed Tomography (CT) lung pictures are divided as Normal and Abnormal respectively. To concentrate on the tumor part, the aberrant photos are segmented. Features taken from the photos are used for classification. To produce superior classification performance, the feature extraction stage is prioritized. This information then acts as an input to machine learning algorithms to identify a pattern. This gave some solid insights into what combination of parameters is most probable to result in an abnormality. The ultimate objective is to discover efficient and widespread categorization techniques using well-known ml techniques like FPCM and better versions.

[8] this particular model is taken to identify a cancer nodule in a CT scan of the lungs with help of a water separation method. Then the image is further categorized as dangerous or mid using SVM. This system has achieved a decent accuracy of 92%. Significant improvements are observed in this model when compared to previous ones. But this model lacks in the detection of the various stages of lung cancer. As per future scope developments, this model wasn't satisfactory. However few techniques like good pre-screening and decryption process can be employed to increase precision.

[9] This work suggested a deep neural network based on Google Net. This network apparently has Maximum dropout ratio, which is proven to decrease processing time There is overfitting during learning time, which has been taken care of using a dropout layer. More than 55% of the neurons are at fully connected level and they are found to be much greater than existing model. There were tests performed using three pre-trained CNN architectonic in the LIDC. This suggested model which used a deep neural network has shown an increased precision.

III. IMPLEMENTATION DETAILS

We start by importing NumPy, matplotlib, os, OpenCV, pickle, and sklearn libraries. we now access the data set and turn the CT scan images into grayscale images. The dimensions of these images are 512 pixels. This is done because a grayscale image needs less information for each pixel, so it will be easy to train the data. The CT scan images are obtained through benchmark.

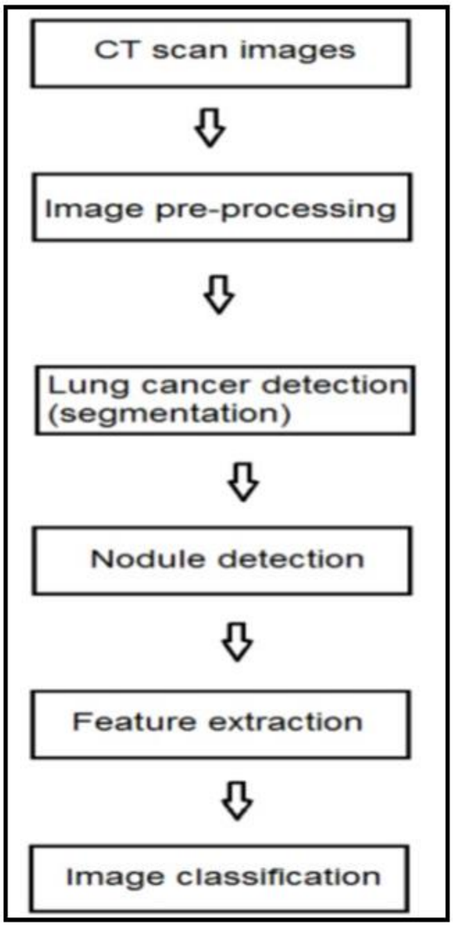

Fig. 1. Flow chart for the steps followed for implementation

A. Proposed Model

Dataset: CT scan Images for Lung Cancer Detection (512 pixels)

The dataset collection consists of two different folders which are the folder of YES images and the folder of NO Images. Both YES and NO folders contain different CT scan images taken from various patients.

The 'YES' folder has CT scan pictures with lung cancer and the 'NO' folder contains CT scan pictures with no lung cancer. There are 155 CT scan images of positive cases and 98 CT scan images with negative cases. That is, with no lung cancer. All the images are 512 pixels respectively. After performing data augmentation on this existing data the dataset increased to a total of 2065 images with 1085 images labeled as 'YES' and 980 images labeled as 'NO'.

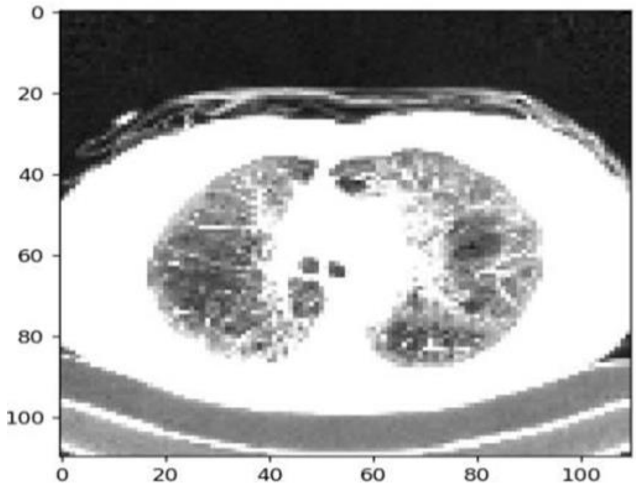

Fig. 2. Lung CT scan Image

In Our model, we import NumPy, matplotlib, os, OpenCV, pickle, and sklearn libraries. Then access the data set and turn the CT scan images into grayscale images. The dimensions of these images are 512 pixels. This is done because a grayscale image needs less information for each pixel, so it will be easy to train thedata.

We use Convolutional Neural Networks (CNN) to add layers to the model which will give very accurate results. We import Tensorflow The layers which we import are Dense, Dropout, Activation, Flatten, Conv2D, and max-pooling. We take four Conv2D layers, three layers of max-pooling, three Dense layers, one Flatten layer, and one Dropout layer respectively. Overall 12 layers.

We now run the model with 25 Epochs to get high accuracy and less loss. We take a CT scan image and resize and reshape it to test it. Apply deep auto encoders to get the output.

B. Steps to Build a Classification Model

Imported all necessary modules:

The layers which we import are Dense, Dropout, Activation, Flatten, Conv2D, and max-pooling. A dense is the layer of neurons. Here, every neuron sends input to the next layer of neurons.

The Activation layer is a non-linear layer in CNN that has an activation function that gives an activation map. Flatten layer is used to transform all the resulting 2D arrays into a single long linear vector.

Conv2D generates a convolution kernel which is convolved with layer input to construct a tensor of outputs. Maxpooling2D helps with ED spatial data. We now run the model with 25 Epochs to get high accuracy and less loss. Now, we take a CT scan image and resize and reshape it and we will test it. We take four Conv2D layers, three layers of max-pooling, three Dense layers, one Flatten layer, and one Dropout layer respectively

The First layer is the convolution layer which consists of 32 filters with the size of the filter (3,3) and a default step is 1. By detecting variations in image intensity levels, The filters are used to detect spatial patterns such as edges in an image. The size of the filter (3,3) is the dimensions of the matrix of the filter. A stride in the filter is used to modify the amount of movement over the image. As the stride size is 1, the filter will move one pixel, or unit, at a time.

The Second layer is max pooling with pool size as 2, Max pooling is the process in which, from the region of the feature map of the filter the maximum element is taken. A feature map with the most important features of the previous feature map will be the output.

The Third layer is again the Conv2D layer which has 64 filters and kernel size of 3 with activation function as ReLU, with padding controls the size of the output feature map. A convolution kernel with layers input produces a tensor of outputs.

The layer four with a pool size of 2 is max pooling.

The layer five is the Conv2Dlayer which has 128 filters and withsizeofthekernel3withReLUasthe activation function.

The layer six with a pool size of 2 is max pooling.

The Seventh layer is the Conv2Dlayer which has128 filters and with a size of the kernel 3 with ReLU as activation function.

The Eighth layer is the Flatten. Flatten layer is used to transform all the resulting 2-Dimensional arrays into a single long linear vector.

The Ninth layer is the dropout layer which is used to drop the unwanted neurons to the next layer, 0.25 specifies the probability where of nullifying the neurons to prevent overfitting.

The last three layers are the Dense fully connected layers with an unalike number of neurons. And for the first three dense layers, the activation function is defined as the ReLU. All the neurons from the previous layers send input to the neurons in the Dense layer, allowing the model to simply identify the relationship between the values of the data in which the model is working for the last dense layer for the prediction we used softmax as the activation function, where the probability of each class is returned and is summed to one, the class having the highest probability is returned for output.

Generally, the softmax activation function is applied to help with multiclassification but existing model has sigmoid as the activation function.

C. Testing the Dataset

When the image which does not have cancer is taken then the result is shown as 0 and we get False as the output. When an image that does have cancer is taken then the result is shown as 1 and True

IV. RESULTS

The final results are indicated in the form of TRUE or FALSE. After running tests on the dataset It undergoes classification. Further on the layers which we imported come to play. Finally the testing on the data set occurs which will gives us the final results. If the particular Image has cancerous nodules i.e is malignant the results show TRUE. If there are no cancerous nodules in the CT scan image and is benign the result is FALSE.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

When collate to the preceding model, The addition of extra CNN layers help with the increases precision of our model. The comparison table shows the differences between previous model and the one we proposed**.**

The accuracy of this model is 90.6%.

A. Model Accuracy:

Model Accuracy shows us how precisely a model is predicted. It is a very popular method to check how good our model is based on training data. The graph shows a steady increase in the accuracy of the model as epochs increases. This indicates that our model is very accurate.

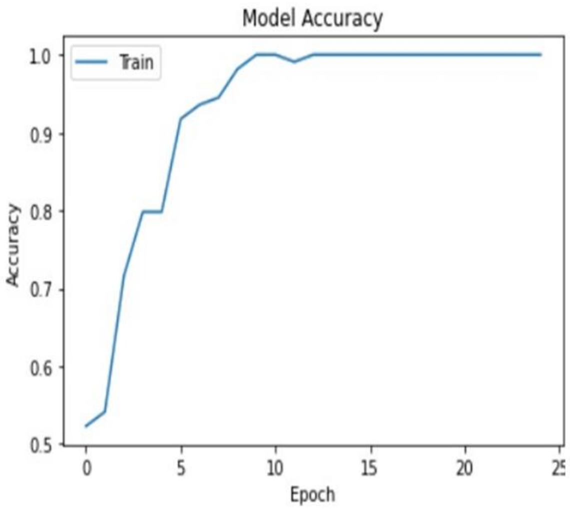

Fig. 3. Model Accuracy Graph

B. Model Loss

Model Loss indicates how inaccurate a model's forecast in on a single case. Model loss and Model's precision are indirectly proportional. Lesser the loss is better the precision will be of the model. The graph shows that the loss of the model is decreasing as the epochs are increasing. This indicates that model is perfect as loss is zero.

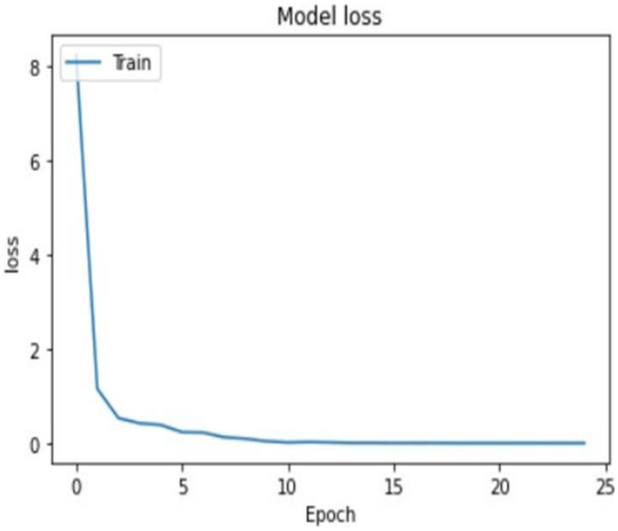

Fig. 4. Model Loss Graph

C. Validation Accuracy

Validation Accuracy shows how precise the model is on new data. Validation accuracy is same as model accuracy but here we test it on the new data. The graph shows a steady increase in the accuracy of the model as epochs increases. This indicates that our model is very accurate

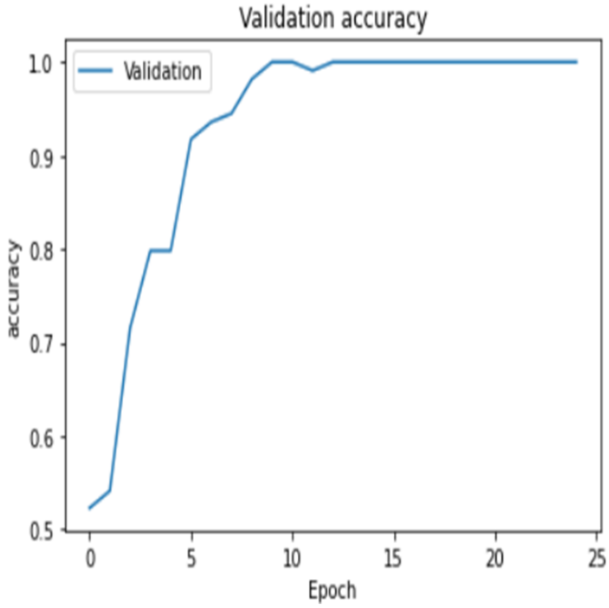

Fig. 5. Validation Accuracy Graph

D. Validation Loss

Validation Loss shows how good the model fits new dataset. Validation loss is same as model loss but here we test it on the new data. The graph shows that the loss of the model is decreasing as the epochs are increasing. This indicates that model is perfect as loss is zero.

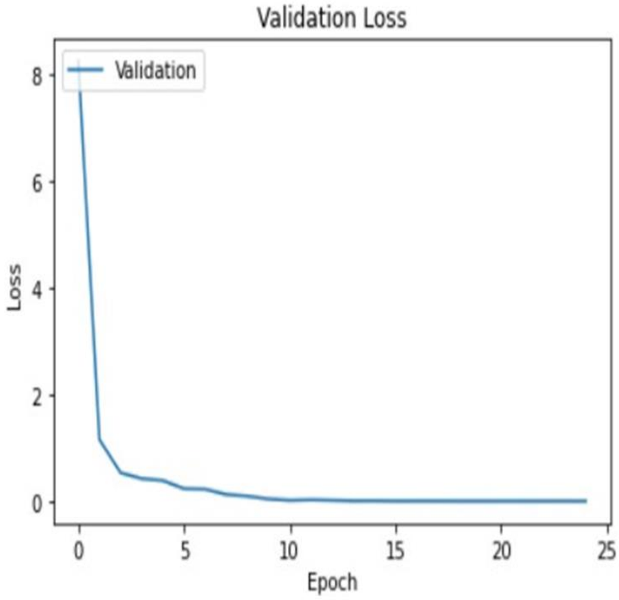

Fig. 6. Validation Loss Graph

E. Recall and Precision

Precision = True Positives / (True Positives + False Positives)

The recall is also known as sensitivity.

Recall = True Positives / (True Positives + False Negatives)

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

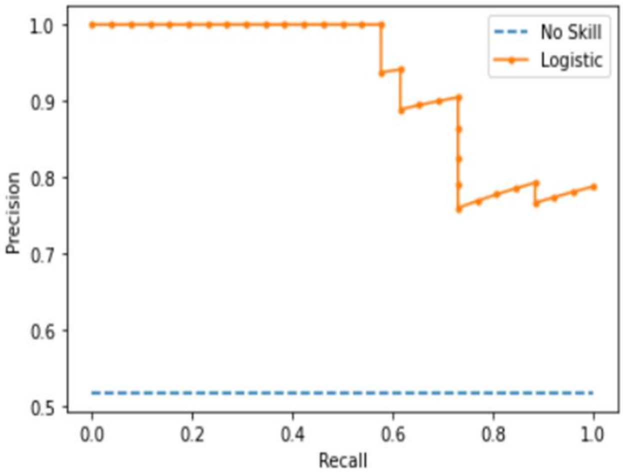

Fig. 7. Recall and Precision Graph

V. CONCLUSION

We Successfully Developed Machine Learning Model using Convolutional Neural Networks to detect Cancerous Nodules in a CT scan image and increased the accuracy by adding extra layers of CNN. Multiple layers of CNN are used in our model which improved the results of the model. We elaborately explored the implementation of a Computer-Aided Diagnosis (CAD) system using CT Scan images theoretically and practically. The techniques used were evinced to be efficient in the early identification of Lung cancer and also distinguishing if the cancer is dangerous or benign. We report on the development of an early cancer detection model using CADCT image analysis.

FUTURESCOPE

In future, in-depth research should be continued in the mentioned and various other models. Also proposed models that are developed using CNN must be validated. It is necessary to substantiate the existing systems to speed the early detection process of lung cancer and for more efficient real time applications.

REFERENCES

1. Aggarwal, T., Furqan, A., & Kalra, K. (2015) "Feature extraction and LDA basedclassification of lung nodules in chest CT scan images." 2015 International Conference On Advances In Computing, Communications And Informatics (ICACCI), DOI:10.1109/ICACCI.2015.7275773. [Crossref]
2. Jin, X., Zhang, Y., & Jin, Q. (2016) "Pulmonary Nodule Detection Based on CT Images Using Convolution Neural Network." 2016 9Th International Symposium On Computational Intelligence And Design (ISCID). DOI: 10.1109/ISCID.2016.1053. [Crossref]
3. Sangamithraa, P., & Govindaraju, S. (2016) "Lung tumour detection and classification using EK-Mean clustering." 2016 International Conference On Wireless Communications, Signal Processing And Networking (Wispnet). DOI: 10.1109/WiSPNET.2016.7566533.B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished. [Crossref]
4. Roy, T., Sirohi, N., & Patle, A. (2015) "Classification of lung image and nodule detection using fuzzy inference system." International Conference On Computing, Communication & Automation. DOI:10.1109/CCAA.2015.7148560. [Crossref]
5. Ignatious, S., & Joseph, R. (2015) "Computer aided lung cancer detection system." 2015 Global Conference On Communication Technologies (GCCT), DOI:10.1109/GCCT.2015.7342723. [Crossref]
6. Radhanath Patra, (July 2020). Prediction of Lung Cancer Using Machine LearningClassifier,CCIS,volume1235,19
7. ApoorvaMahale (2017) Lung Cancer Detection using Data Analytics and Machine Learning- https://cdas.cancer.gov/approved-projects/1462/
8. Suren Makaju, P.W.C. Prasad, Abeer Alsadoon, (December 2017) A. K. Singh, A. Elchouemi Lung Cancer Detection using CT Scan Images 6th International Conference on Smart Computing and Communications, ICSCC 2017,Kurukshetra,India.
9. Tulasi Krishna Sajja, Retz Mahima Devarapalli, Hemantha Kumar Kalluri (October 2019) Lung Cancer Detection Based on CT Scan Images by Using Deep Transfer Learning IIETAhttps://doi.org/10.18280/ts.360406. [Crossref]

AUTHORS PROFILE

Dr. Syed Abudhagir Umar, Associate Professor, ECE Dept. B.E(Electronics and Communication Engineering) degree from Sethu Institute of Technology, Tamil Nadu, India. M.E(Electronics and Control Engineering) degree from Sathyabama Institute of Science and Technology, Tamil Nadu, India. Ph.D(Grid Computing - Department of Electronics and Communication Engineering) from Anna University, Tamil Nadu, India. Sabbatical Research Fellow (Deep Learning - Department of Computer Science Engineering) from Bennett University, Greater Noida, India. He has a total of 16 years of experience which contains 11 years of teaching experience, 4 years of research experience and 1 year of industry experience. Currently working at B V Raju Institute of Technology as Associate Professor in the department of Electronics and Communication.

Chenigaram Abhigna, Final year student pursuing Engineering focused in Electronics and Communications from B V Raju Institute of Technology Enrolled in Special Lab Focused on Artificial Intelligence and Machine Learning. Worked on Several projects such as Stock Price Prediction and Lung Cancer Using CT scan Images. Worked as a Digital Marketing Intern previously. Currently Interning in Hewlett Packard Enterprise as Technical Solution Consultant. Participated in Nation Wide Model United Nation(MUN) Conferences. An Honorary Mention Awardee in College MUN (BVRITMUN 2020). A Goal oriented, Disciplined individual with excellent communication skills. Strengths are Active Participation, Creativity, Leadership and Social Skills. A skilled artist with deep interest in Arts, Crafts and Literature.

Cheemalapati Venkata Vignesh, BTech(ECE) degree from B V Raju Institute of Technology (BVRIT). Lung Cancer Detection using CT Scan Images and Stock Prize Prediction using machine learning are the projects which he has worked upon in the Machine Learning special lab provided by the college. Consistently strive to be in challenging and exciting learning positions which will help to become a lifelong learner. Done internships in the fields of technology and marketing which provides new perspectives to tackle any problems. Strengths are communication, adaptability, discipline, preparedness and friendly with any new environment. An avid reader of books and strive to be a researcher.

Teki Rohith Raj,who has graduated BTech in electronics and communication stream from BVRIT.I have done projects in Lung Cancer Detection using CT scan images and text recognition from special lab, which is part of the academic year. I have genuine interest in furthering my knowledge in an interdisciplinary manner. I have done internships in BSNL and 360 digimg and gained practical knowledge from them. Multitasking, listening, and critical thinking are my strengths. I have thorough understanding of fundamentals and keen interest in studies. I have taken several initiatives and have undertaken my projects and events in the college. I have commitment desire of a good student, meticulous with the right blend of combined with analytical ability, supported by excellent communications skills, brings out the intellectual independence.