International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 04 | Apr 2022 www.irjet.net p-ISSN: 2395-0072

Automatic Pulmonary Nodule Detection in CT Scans using Xception, Resnet50 and Advanced Convolutional Neural Networks models.

Dr. S. V. G. Reddy[1] , V Bhuvaneshwari[2], Aniket Kumar Tikariha[2] , Yagna Sriram Amballa[2] , Balumuri Sesha Sahith Raj[2]

[1]Associate Professor, Department of Computer Science and Engineering, GITAM University, Visakhapatnam, Andhra Pradesh, 530045, India.
[2]Student, Department of Computer Science and Engineering, GITAM University, Visakhapatnam, Andhra Pradesh, 530045, India.

Abstract - *Lung cancer is one of the most dreadliest diseases that is more probable to grow rapidly with the spread of metastasis. Metastasis is the formation of additional secondary malignant growths away from the primary cancer location. The ability to recognize and diagnose the malignant nodules and categorize them as benign, malignant, or indeterminate(normal) on chest computed-tomography (CT) scans is extremely crucial for early lung cancer diagnosis and treatment. For that purpose, with the increasing advancement of technology numerous machine learning and deep learning techniques have come into existence to diagnose lung cancer where the machines are taught to predict outcomes. By using such means to precisely detect the cancerous pulmonary lung nodules can aid in the timely manifestation of lung cancer. However, it's not an easy task to develop a reliable lesion detection approach due to irregularity in the patterns of lung lesions, it's shape, size and the complex nature of the surrounding conditions. In our proposed computer-aided design system we perform cancerous nodule detection by using advanced CNN model and pre-trained CNN models like Resnet50 and Xception. In our advanced CNN models, we integrated several approaches for improved image pre-processing and employed methods such as SMOTE and class weighted approach to account for the dataset's imbalance. By adjusting the imbalances in our dataset, we were able to considerably enhance our model's accuracy. For this project we use the lung cancer screening thoracic computed tomography (CT) images from the IQ-OTHNCCD lung cancer dataset which is collected from kaggle. The dataset contains 1190 images totally. These 1190 images are the CT scan slices of 110 cases. Each case approximately having 10 slices. These images are categorized into 3 classes: normal, benign, and malignant. Among them, there are 40 malignant instances, 15 benign cases, and 55 normal cases. In this project we try to build our own Convolutional Neural Networks to classify the images into one of the three classes and we also employ the pre-built architectures, namely RESNET50 and XCEPTION, trained on the image net dataset and compare their performances on certain metrics.*

***Key Words*: Lung cancer, Lesion detection, Deep Learning, Advanced Convolutional neural network(CNN), *Computed-Tomography (CT) scan*, Resnet50, Xception.**

1. INTRODUCTION

Lung cancer is the type of cancer that begin in the lungs. Our bodies are made up of trillions of cells. Each cell has its own life cycle. Healthy cells in our bodies die at some time throughout their lifespan and are replaced by new ones. When this process does not go as planned, i.e., when cells do not die when they are old or injured, but instead continue to multiply abnormally, resulting in an overabundance of cells, tumors form. These tumors are classed as normal tumors when they do not pose a threat to a person's life. Malignant tumors are cancerous tumors that cause harm to our bodies[1]. When detected early on, these tumors are considered benign since they can be treated well. However, these tumors have a significant possibility of metastasizing and becoming malignant over time when left undiagnosed. Lung cancer is consistently cited as the leading cause of cancer death, accounting for over 18 lakh deaths. According to the GLOBOCAN- 2020 assessment on cancer occurrences among people and fatalities, approximately about 193 lakh new cancer cases were diagnosed worldwide, with around 100lakh cancerdeaths[2].

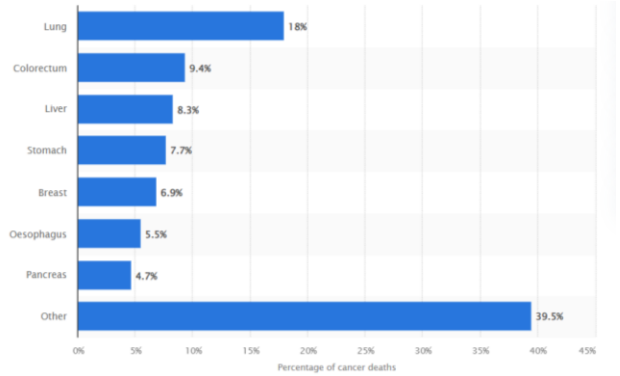

Fig -1: Global cancer mortality rate in 2020, by type of cancer (Source: Statista)

1.1 Types of Lung Cancer

There are two major categories of lung cancer. They are Non-small cell lung cancer and Small cell lung cancer. Among the two, NSCLC is the most frequently encountered one when compared with the SCLC. The two variations of cancer develop and are treated in different ways on the basis of their severity. About 80% to 85% of the total lung cancers are non-small cell lung cancers. Adenocarcinoma, squamous cellcarcinoma, and large cell are few sub-types of NSCLC[3]. NSCLC is substantially less severe than small cell lung cancer and is much easier to treat. SCLC accounts for about 10% to 15% of all cancers. When opposed to NSCLC, SCLC spreads and affects other organs more swiftly. In most situations, it is extremely difficult to diagnose this type of cancer since it has spread to other places of the body, making diagnosis and treatment extremely tough. Because this cancer develops quickly, chemotherapy and radiation treatment are effective. However, this type of cancer has a tendency to re-attract people even after treatment.

1.2 Nodule

A nodule is a "spot on the lung" which usually is an abnormal growth that forms in a lung that is seen on an CT scan. We may have one nodule on the lung or several nodules. This little round or oval solid overgrowth of tissue is surrounded by normal lung tissue. These nodules can either be benign or malignant[4]. Nodules are very common. Not all nodules are malignant. About 95% of lung nodules are benign which form due to respiratory problems or other illnesses which do not require treatment. As pulmonary nodules tend to have a diverse complicated features such as lesion size, shapes, and classification characteristics, it becomes difficult for CAD systems to accurately identify the lesions and diagnose lung cancer[5]. Thus, it is extremely essential to incorporate the right detection mechanisms to detect the pulmonary lung nodules and to improve the accuracy in finding lesions that are tiny in size and are usually left undetected.

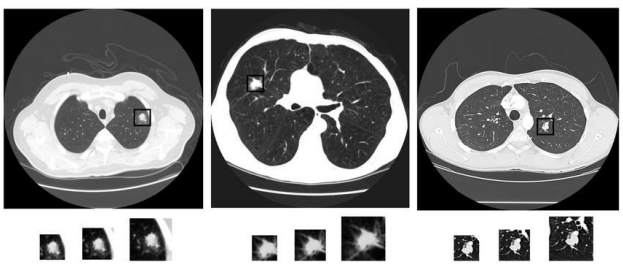

Fig -2: Different categories of lung tumors-

Benign, Primary malignant and metastatic Malignant.

1.3 General lung cancer statistics

Lung cancer is indeed one of the leading causes of death worldwide. Lung cancer accounts for almost two-thirds of all cancer-related fatalities. In 2018, there were 181 lakh new cancer diagnoses and 95 lakh cancer-related deaths globally[6]. Predictions have been made that number of cancer cases by 2040 may reach roughly 295 lakh, with 164 lakh cancer-related deaths. Lung cancer is one of the most serious lesions that may have a dramatic effect on a person's health in a short period of time due to its ability to readily migrate from one region of the body to another without exhibiting any severe symptoms in the early stages. Each year, lung cancer claims more lives than breast, colon, and prostate cancer combined. It is anticipated to be one of the leading causes of death in the American population. In the United States, it was anticipated that 1,806,590 new cases of cancer would be discovered, with 606,520 people at risk of death from the disease. According to a National Institutes of Health study of lung cancer occurrences and deaths from 2013 to 2017, the annual rate of new cancer cases is 442.4 per 1 lakh men and women, while the annual rate of cancer death is 158.3 per 1 lakh men and women. According to the National Institutes of Health, about 16,850 toddlers and adolescents whose ages falls under 0 and 19 years will be diagnosed with cancer in 2020, with 1,730 of them dying of the disease[7].

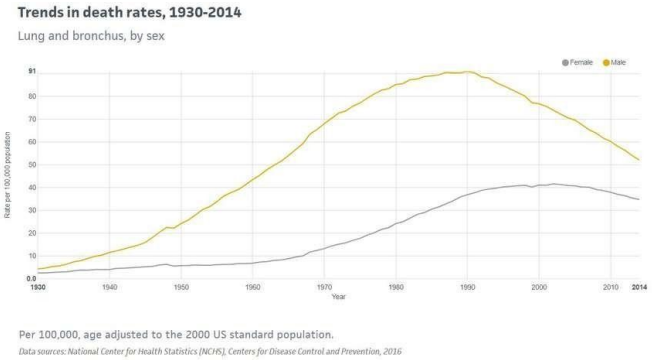

Fig -3: Death Rate Trend of Lung Cancer in the US.

1.4 Overview of the scenario

Only 15% of lung tumors are detected in their early stages. Various methods are being followed to treat this disease like chemotherapy etc. However, lung cancer patients with various clinical stages have drastically varying prognoses. Patients in stage IA groups who have a survival rate of 5 years are more than 90%, while patients in stage IV who have a survival rate 5-years is fewer than 10%. However, the survival probability further falls to 3.5% when cancer tends to spread to different other organs. Thus, faster diagnosis of lung cancer is a critical step to provide improved chances of survival. Early

detection of this cancer depends on how accurately the malignant nodules present in the lungs are detected in CT scans.

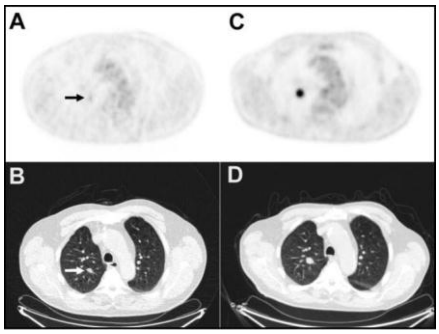

Fig -4: CT scans identified a lung nodule which, over 11 months progressed and was confirmed as lung cancer later.

1.5 Convolutional Neural Network

In our project we will be performing medical image analysis on the images of lung CT scans, analyse them and perform lung cancer detection. The deep learning model that has more significance when it comes to dealing with medical image analysis is the Convolutional Neural Network[8]. In our research, we develop our own convolutional neural networks and use a few approaches to improve the accuracy of our model and also employ the prebuilt architectures, namely RESNET50 and XCEPTION, trained on the image net dataset.

1.6 Glimpse of working methodology of our proposed CNN model

The working methodology of our proposed CNN models begins with taking the CT scans of various cases from the IQ-OTH/NCCD lung cancer dataset. Then we segment the lung parenchyma and slice the radiographs using image pre-processing methods to obtain the sliced image of the lesion which is fed to the convolutional neural network as an input image. The input image then goes through the different layers of the CNNs.

Input -> Droupout -> Convolution -> RELU -> Pooling -> Convolution -> Softmax-> Dense -> Pooling -> RELU -> Fully connected layer

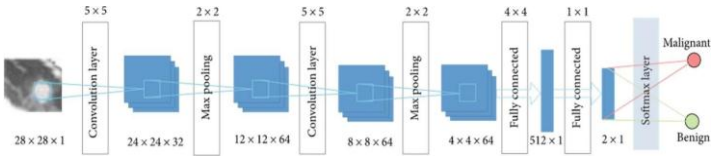

Fig -5: The architecture of the proposed CNN model to classify sliced CT scan input images as malignant or benign.

The CNN system extracts the potential features while the model is being trained on a set of a variety of images from the IQ-OTHNCCD lung cancer dataset[9]. Later when a new test input is given to the proposed CNN model, it compares learned features with the input data and classifies the inputted sliced tumor to be either normal or benign or malignant.

2. LITERATURE REVIEW

In [10], the researchers propose a model to help in the identification of lung nodules, that transfers and enhances a CNN with many resolutions for lung nodule candidates categorization through knowledge transfer. Small nodules with poor resolution and big nodules with high resolution can both be identified using this approach. The methodology used in the research includes various steps namely rough candidate nodule finding and judgment. The results of the experiments performed by the authors during the course of this research suggest that it is possible to overcome the challenge posed by the wide range of sizes and forms of lung nodules, as well as diverseness in finding nodules using this approach.

In [11] the authors provide a unique multitask convolutional neural network (MT-CNN) architecture for distinguishing and finding the cancerous and non-cancerous nodules on Lung CT scans. To increase lung nodule classification performance, an image regeneration methodology is applied as an supplementary work from nine two-dimensional (2-D) images deconstructed from various angles of each nodule, this model learns the properties of 3-D tumour segmentation. characteristics. Each 2-DMT-CNNmodel has two tasks: one for nodule categorization and the other for picture reconstruction (auxiliary task).

The HSN model in [12] is a neural network that combines a lighter 3-dimensional CNN for learning deep 3D structural and dimensional information with a 2-dimensional CNN for collecting detailed semantic features of multiple slices of CT scans in a single network. Spatiotemporal-separable 3D (S3D) convolutions are employed to deal with the complex dimensional features of CT scans and reduces the cost that is required for working with 3DCNN) → To cope with the complex

dimensional characteristics of CT scans, Also, dilated convolutions in 2D CNN so as to memorize a plethora of semantic information about minor things. Moreover, to combine both 2D and 3D features effectively, a hybrid features fusion module is designed in this HSN network.

In [13], to address the tough problem of correctly categorizing nodules in this research, the authors proposed a Multi-Branch Ensemble Training architecture based upon 3D convolutional neural networks (MBEL-3D-CNN). Three fundamental concepts are combined in this method: The first step is to create a 3D-CNN that maximizes the use of structural and dimensional features of lung lesions in 3D space; the second stage is to incorporate an MBEL-3D-CNN which is well suited to lesion diversity; and the third stage is to use ensemble learning to improve the 3D- CNN model's generalization performance. In addition, the authors used offline heavy mining techniques to allow the model to handle indistinguishable positive and negative data.

3. DATASET DESCRIPTION

In this project We employ thoracic computed tomography (CT) images for lung cancer screening from the Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD) lung cancer dataset. CT scans were mostly stored as DICOM files in this collection. This dataset was compiled over a three-month period in fall 2019 from the aforementioned locations. In these two centres, oncologists and radiologists labelled the IQ-OTH/NCCD CT scan slides of patients diagnosed with various stages of lung cancer.The source of the dataset is Kaggle which is a data science and artificial intelligence platform. The dataset contains 1190 images totally. These 1190 images are the CT scan slices of 110 cases, each case approximately having 10 slices. These images are categorized into 3 classes: normal, benign, and malignant. Out of these, 40 cases are malignant cases; 15 cases are benign cases; and 55 cases are normal cases[14].

In this project we try to build the Convolutional Neural Network to classify the images into one of the three classes. Primarily the nodules with a diameter less than 3mm are considered as non-nodule and tiny nodule and are not taken into consideration since they have no clinical significance. and nodules with a diameter greater than or equal to 3mm were taken into consideration.

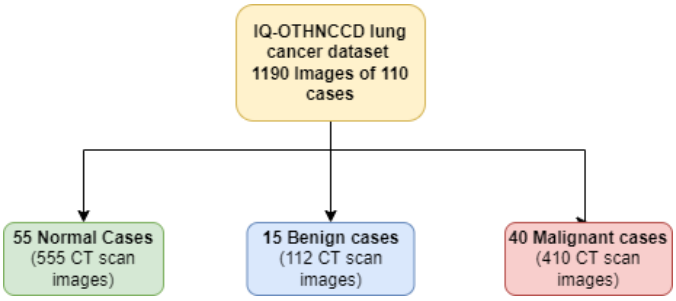

Fig -6: Lung nodule categories in our dataset.

4. PROBLEM STATEMENT

The goal of this project is to create a computer-aided design system that takes lung CT scans as input to help in the early detection of lung cancer by classification of lung cancer by accurately classifying lung nodules as normal, benign, or malignant using a variety of deep learning approaches. In this project, we will create an application, a lung cancer detection system, to assist clinicians in making better and more informed judgments when diagnosing lung cancer. This will aid in early identification of lung cancer, which will reduce the number of fatalities caused by tumour severity.

5. PROJECT OBJECTIVES

1) The primary goal of this project is to develop a reliable lung nodule detection system that will aid in the faster and timely detection and diagnosis of lung cancer.

2) To create a system that reliably predicts cancer by accurately diagnosing tumors of all sizes, especially the small lesions that go undiagnosed most of the time, as well as spotting the tumour site in the lungs.

3) To enhance the accuracy of identifying the cancerous nodules in the lungs in clinical assessment with CT scans by effectively pre-processing the raw input images and applying right classification algorithms.

4) To implement efficient and accurately predicting deep neural network models using Convolutional Neural Networks.

6. PROPOSED SYSTEM METHODOLOGY

Our lung cancer detection system is firstly fed with the IQ-OTH/NCCD lung cancer dataset that contains CT scans to enable computer-assisted systems in the identification, categorization, and quantification of lung nodules. We take the radiography images provided in the dataset and perform image processing on it to obtain the sliced image of the pulmonary lung nodule. Here the essential crucial information is the image of the tumour in the lungs which is given as input to our convolutional neural network

model. Alongside performing image processing we also perform image enhancement where we play suitable filters to the input image to remove unnecessary noises so as to prevent misleading results that may occur in subsequent processes. Then we apply OpenCV and Numpy functions on the input and separate the .png image files and image data arrays. The image data arrays obtained are then pre-processed and are made suitable for the use of classification. The pre-processed data is then fed to the Deep Learning Model. The deep learning model in our project contains a number of convolutional, RELU, pooling, dense and dropout layers[15]. The sliced image of the pulmonary lung nodule which is fed to our deep learning model goes through many layers of the neural network where convolution takes place. A convolution is the preliminary process where we apply a suitable filter to an input to produce an activation[16]. When the same filter is repetatively applied to an input, we obtain a feature map, which displays the various supporting and contrasting feature in an input, in our case the feature is the nodule which is either malignant benign. The output from the convolutional layers reflects high-level characteristics in the input after going through a sequence of recurrent convolution and pooling layers. This output is then flattened and a vector matrix is obtained and connected to the output layer by adding a FC layer. The network's ultimate levels are known as fully - connected layer. The result from the end Pooling or Convolutional Layer, which is flattened and then put into the fully - connected layers, is the input to the fully connected layer. Flattening is the process of unrolling all of the values in an N-dimensional matrix into a vector. Following the FC levels, the last layer employs the soft-max activation function to decide whether the input data is more likely to belong to the malignant or benign classes (classification)[17]. These findings, as well as the cancerous lesion areas, are shown to physicians in order to detect malignant cells, which assists in therapy.

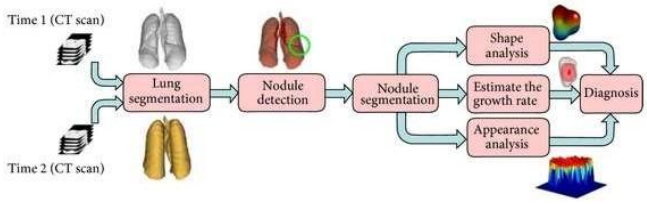

Fig -7: Processes that are involved in building a CAD system for lung nodule detection

7. IMPLEMENTATION

Lung cancer classification based on chest CT images using advanced CNN. In this project we have built our own advanced CNN models which constitutes of a number of convolutional, ReLu, pooling, dense and droupout layers. Additionally, we have worked on employing various pre-trained models such as Resnet50 and Xception to extract more critical features from chest CT scans, which will aid in the rapid detection of whether a tumour is present in the patient's lungs or not, and if it is, whether it is a normal nodule, benign nodule, or malignant nodule. This work is carried out in stages. To begin, we acquired the IQ-OTH/NCCD lung cancer dataset from Kaggle, a data science and artificial intelligence platform. Following that, we examined our dataset. During our study, we discovered that our dataset had three sub-directories, one for each class. Our dataset includes three distinct classes: normal, benign, and malignant. Normal class accumulated the most CT scan images. It contained about 555 of the dataset's 1190 images. Then there's the malignant class, which includes around 410 images of all CT scans with evidence of malignancy. The benign class has the fewest CT scan image samples. In the benign class, there were around 112 CT scan images. As a result of dataset exploration, we deduced that our dataset was skewed. To account for the dataset's imbalance, we employ a few techniques while developing our CNN models that would certainly balance the dataset and bring in some consistency. We use the Kaggle notebook where we trained the various CNN models. After choosing out dataset, we installed and imported all the necessary libraries that are required to build our models. We then loaded the dataset from dist onto kaggle. We then concentrated on pre-processing the pictures in the dataset, as raw images would produce less accurate outputs when given to the CNN models. We pre-processed and visualized the images in the dataset before proceeding with lung area extraction from CT scans. Following the extraction of the lung region, each slice in that area is segmented to determine the location of the tumours. These tumours might be benign, malignant, or normal in nature. Then the various CNN models are trained on the segmented lesion regions. In this project we develop 5 models among which 3 models are advanced CNN models built from scratch incorporating various strategies to tackle the imbalance in our dataset. The two main techniques used to balance the dataset are SMOTE-Synthetic Minority Oversampling Technique and Class weighted approach. SMOTE is a strategy that balances the class distribution by randomly adding minority class samples through replication, ensuring that the model is trained evenly on all classes. We begin by selecting a minority class example x and then traversing the feature space to find its k-nearest minority class neighbours. We then choose an example y from the k-nearest neighbours and draw a line connecting them. Then we generate a synthetic instance of minority class along that line. In this manner, we replicate minority samples to ensure that the dataset is balanced. These new examples, which are generated from existing samples, contribute no extra information to the model but aid in training it evenly on all classes[18]. On the other hand, when we balance the samples in a dataset using a class weighted technique, we do not duplicate the instances but instead apply weights to

each class. We set class weight in such a way that it is automatically balanced in order to modify weights in inverse proportion to the class frequencies in the input data[19]. The other 2 CNN models utilise pre-built architectures RESNET50 and Xception. We created the training and testing splits on the dataset. About 75% of our data is used for training and 25% of the data is used for validation. The CNN models are trained on the training data and the performance of the models are evaluated on test data and predictions are made on new data samples.

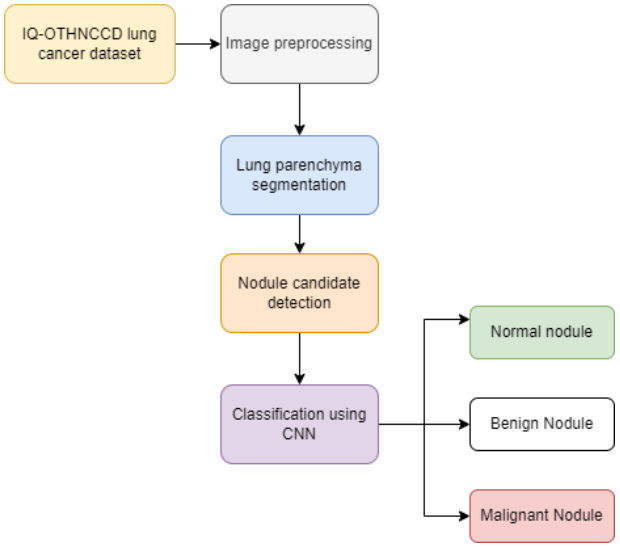

Fig -8: The pipeline of working methodology.

8. RESULTS AND DISCUSSIONS

8.1. Discussions

In this project we have demonstrated 3 scenarios where we build the Convolutional Neural Network to classify the images into one of the three classes- Benign, Malignant or Normal. The CNN models were trained using Kaggle Notebook and the outcomes of each model's training and testing phases are summarized in this section. We constructed the models, compiled it, and trained it on 50 epochs. After training the model, we visualized the accuracy and loss graphs, as well as the classification report and confusion matrix of various CNN models. To begin, Model 1 is the advanced CNN model that is fed directly with the unbalanced pre-processed data without the use of any approach to mitigate the data imbalance. The test accuracy for Model 1 comes out to be 77.77% and test loss comes out to be 0.4553.

The test accuracy and test loss plots for model 1 is visualized below.

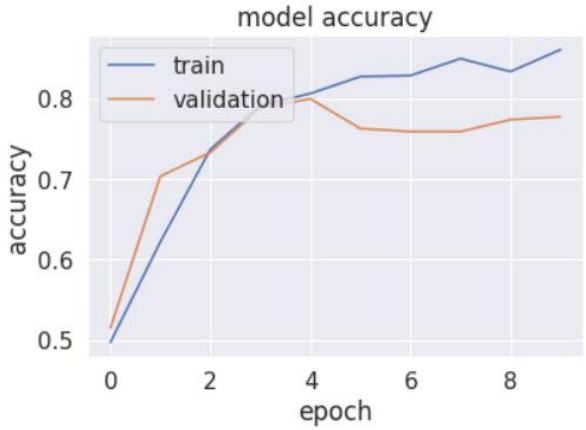

Fig -9: Training and validation accuracy plot for Advanced CNN model.

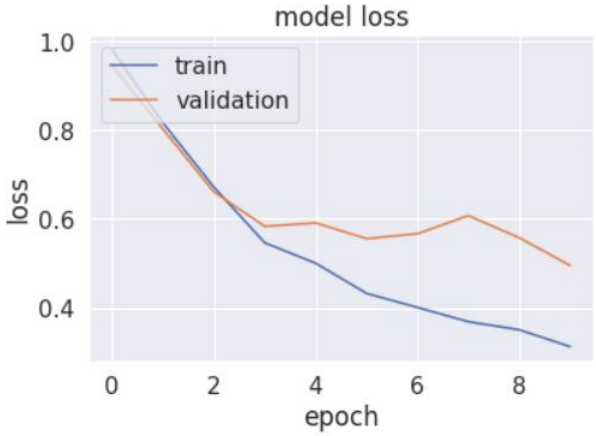

Fig -10: Training and validation loss plot for Advanced CNN model.

The model has poor accuracy and higher loss percentages because of the imbalance present in the dataset. The CNN model appears to be well trained on normal cases and is biased more towards them as the dataset contains more normal cases when compared to malignant and benign. To substantiate this, as shown in Figure -11 it is observed that when the model is fed fresh input CT scans for cross-validation, it is biased toward predicting normal instances and also falsely forecasts malignant and benign cases as normal.

```
Predictions for model1
ACTUAL: {0: 'Malignant', 1: 'Normal', 2: 'Malignant', 3: 'Benign', 4: 'Normal'}
PREDICTIONS: {0: 'Normal', 1: 'Benign', 2: 'Normal', 3: 'Malignant', 4: 'Normal'}
```

Fig -11: Displaying actual and predicted outputs of Advanced CNN model.

To assess our model's performance, we employed a variety of performance metrics, including recall, F1-score, precision, and support[20][21]. Additionally, we tabulated accuracy, micro average, and weighted average values. The classification report is used to summarise all of these values.

	precision	recall	f1-score	support
0	0.54	0.50	0.52	28
1	0.79	0.97	0.87	139
2	0.84	0.59	0.69	103
accuracy			0.78	270
macro avg	0.72	0.69	0.69	270
weighted avg	0.78	0.78	0.77	270

Fig -12: Classification report for Advanced CNN model.

Along with the classification report, we also plotted confusion matrix for every classification model to gain insight into the predictions by visualizing the accurate and erroneous(true and false) predictions made on each class[22].

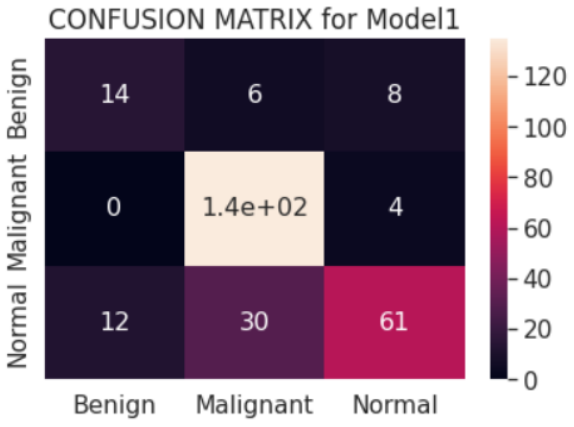

Fig -13: Confusion matrix for Advanced CNN model.

In Model 2, we employed SMOTE(Synthetic Minority Over-sampling Technique) to overcome the bias that we had in model 1. We developed a sophisticated CNN+SMOTE model and fed it a balanced dataset. Using this technique the accuracy of the model has significantly grown to 97.40% (increased by 19.63%) and loss was about 0.766.

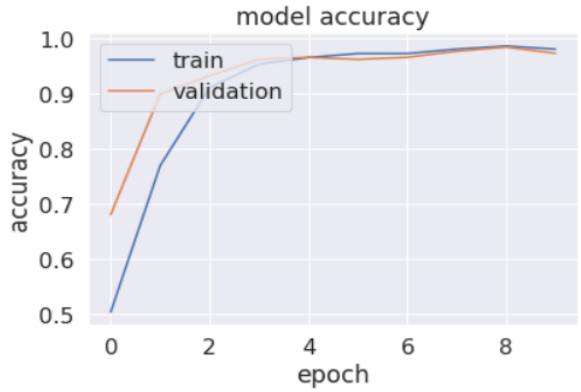

Fig -14: Training and validation accuracy plot for Advanced CNN+SMOTE model.

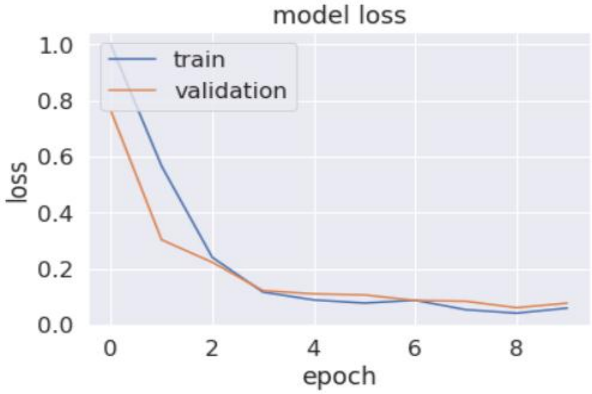

Fig -15: Training and validation loss plot for Advanced CNN+SMOTE model.

When cross-checking the performance of the model on new input test images, the predictions made are very accurate when compared with model 1.

```
Predictions for model2
ACTUAL: {0: 'Malignant', 1: 'Normal', 2: 'Malignant', 3: 'Benign', 4: 'Normal'}
PREDICTIONS: {0: 'Malignant', 1: 'Normal', 2: 'Malignant', 3: 'Benign', 4: 'Normal'}
```

Fig -16: Displaying actual and predicted outputs of Advanced CNN+SMOTE model.

	precision	recall	f1-score	support
0	0.93	0.96	0.95	28
1	0.97	1.00	0.99	139
2	0.99	0.94	0.97	103
accuracy			0.97	270
macro avg	0.96	0.97	0.97	270
weighted avg	0.97	0.97	0.97	270

Fig -17: Classification report for Advanced CNN+SMOTE model.

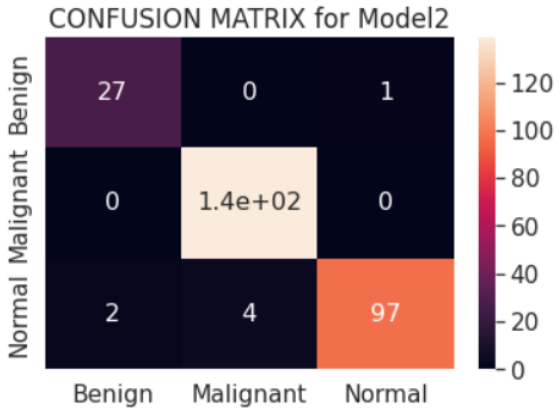

Fig -18: Confusion matrix for Advanced CNN+SMOTE model.

Another technique is being employed in model 3 which is the Class weighted approach. In this technique we do not synthesize new minority instances between existing minority instances and increases the number of instances like it was done in previous technique, instead we assign weights to every class. The model's test accuracy goes up to 95.18% and test loss is 0.1464 after applying this technique.

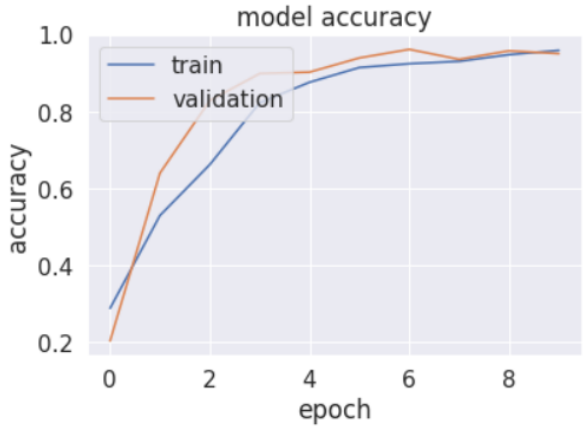

Fig -19: Training and validation accuracy plot for Advanced CNN+Class weighted approach model.

When cross-checking the performance of the model on new input test images, the predictions made are very accurate when compared with model1 and is similar to that of model2.

```
Predictions for model3
ACTUAL: {0: 'Malignant', 1: 'Normal', 2: 'Malignant', 3: 'Benign', 4: 'Normal'}
PREDICTIONS: {0: 'Malignant', 1: 'Normal', 2: 'Malignant', 3: 'Benign', 4: 'Normal'}
```

Fig -20: Displaying actual and predicted outputs of Advanced CNN+Class weighted approach model.

Fig -21: Training and validation loss plot for Advanced CNN+Class weighted approach model.

	precision	recall	f1-score	support
0	0.80	1.00	0.89	28
1	0.96	0.99	0.98	139
2	1.00	0.88	0.94	103
accuracy			0.95	270
macro avg	0.92	0.96	0.93	270
weighted avg	0.96	0.95	0.95	270

Fig -22: Classification report for Advanced CNN+Class weighted approach model.

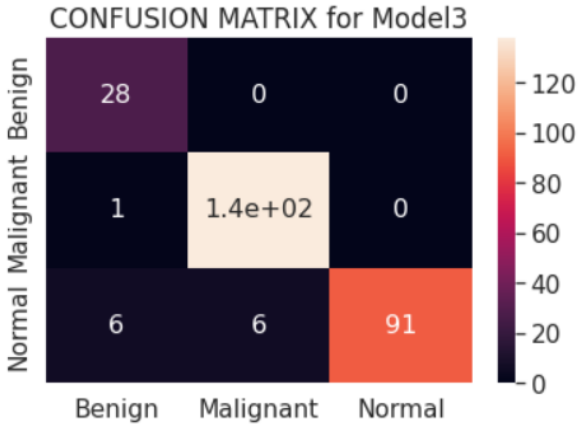

Fig -23: Confusion matrix for Advanced CNN+Class weighted approach model.

Along with these sophisticated CNN models, we created models using two pre-trained models, RESNET50 and Xception. The Xception model achieved an accuracy of 81.48% and a loss of 0.5134 during testing.

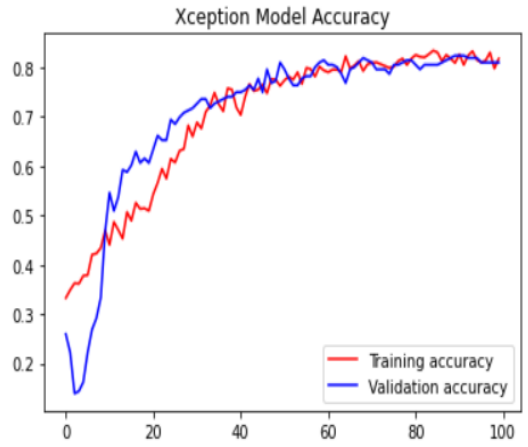

Fig -24: Training and validation accuracy plot for Xception model.

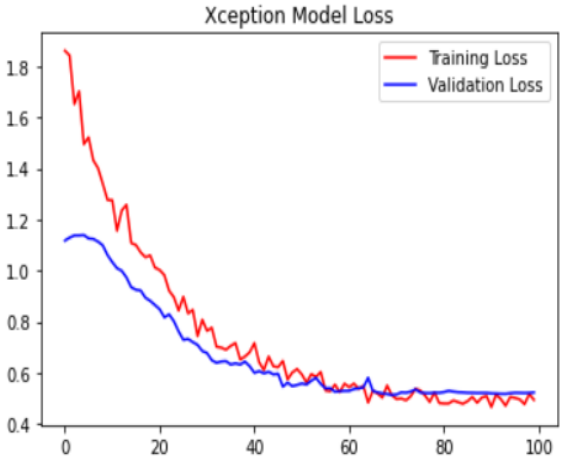

Fig -25: Training and validation loss plot for Xception model.

	precision	recall	f1-score	support
0	0.00	0.00	0.00	28
1	0.92	0.92	0.92	106
2	0.71	0.96	0.82	82
accuracy			0.81	216
macro avg	0.55	0.63	0.58	216
weighted avg	0.72	0.81	0.76	216

Fig -26: Classification report for Xception model.

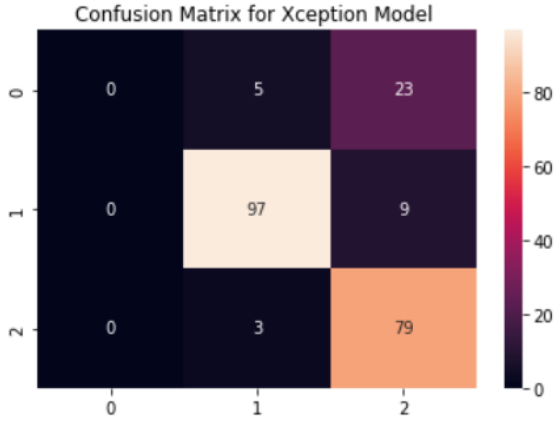

Fig -27: Confusion matrix for Xception model.

Whereas the Resnet50 model achieved an accuracy of 76.85% and a loss of 0.6357 during testing.

Fig -28: Training and validation accuracy plot for Resnet50 model.

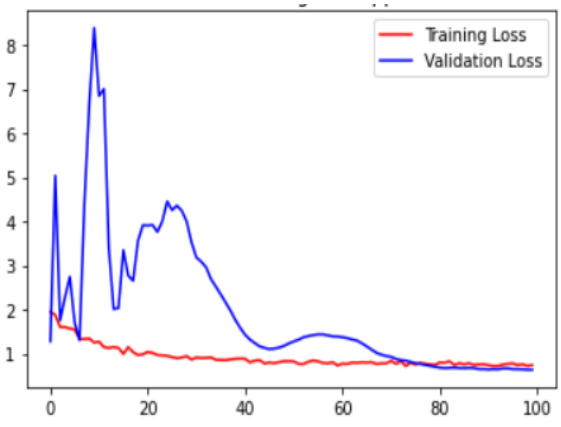

Fig -29: Training and validation loss plot for Resnet50 model.

	precision	recall	f1-score	support
0	0.00	0.00	0.00	28
1	0.96	0.77	0.86	106
2	0.62	0.99	0.76	82
accuracy			0.75	216
macro avg	0.53	0.59	0.54	216
weighted avg	0.71	0.75	0.71	216

Fig -30: Classification report for Renet50 model.

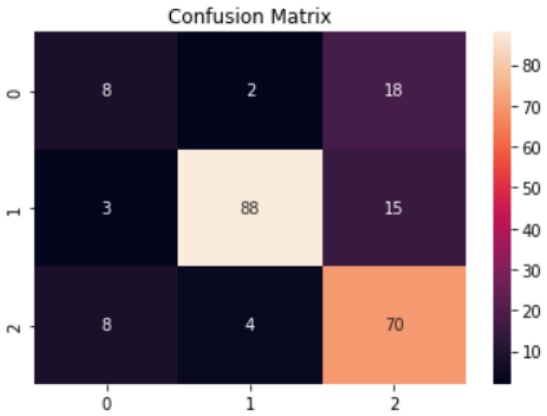

Fig -31: Confusion matrix for Resnet50 model.

8.2 Results

Table -1: Comparing accuracies and losses of various models.

COMPARING METRICES OF VARIOUS MODELS USED		
Model || Metrics	Accuracy	Loss
Advanced CNN	77.77	0.4553
Advanced CNN+SMOTE	97.40	0.0766
Advanced CNN+Class weighted approach	95.18	0.1464
Xception	81.48	0.5134
Resnet50	76.85	0.6357

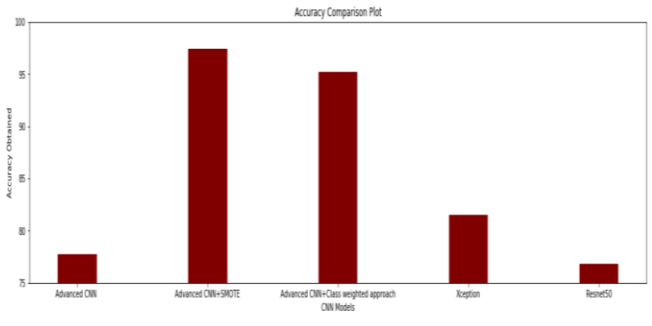

Fig -32: Bar plot showing accuracies of 5 models.

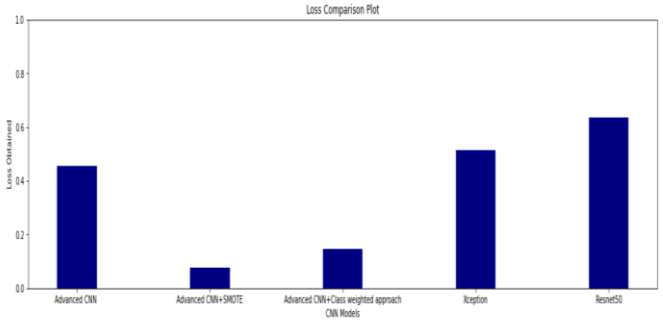

Fig -33: Bar plot showing losses of 5 models.

9. CONCLUSION AND FUTURE SCOPE

9.1. Conclusion

In this project, we study the use of image processing and deep learning techniques to predict lung cancer nodules in vulnerable patients. With the help of the research that we did, deep analysis and by gaining in depth knowledge of the scenario and its significance in the real world, we were able to develop a CNN model for lung cancer detection[23]. As a part of building our model we used various approaches and carried out image processing and classification, which leaded us to coming up with a novel system that detects lung cancer nodules with high accuracy. We were able to develop a full model that runs with more than 95% accuracy on test data. Given the difficult nature of the problem, diverseness of the data and computing difficulties we faced various challenges throughout the process of pre-processing the data so that the features in the images can be detected well and also working with the imbalanced data was a crucial step[24]. The CT scans being in hundreds of images had a memory constraint while processing and also was a time consuming process.

9.2. Future Scope

In this project we have built and trained CNN models employing various techniques like SMOTE and Class Weighted Approach to detect lung cancer nodules and were able to achieve good accuracies. Additionally, we developed CNN models using pre-trained architectures such as Resnet50, AlexNet and Xception. However, the accuracy of pre-trained models is significantly lower than that of the first three models. Thus, there is an opportunity to work on the models and determine which pre-trained model is the best match for this scenario and can be utilised for feature extraction and prediction of malignant cells with high accuracy[25]. Image classification being one of the most complicated and crucial stage of our project, to process the images rightly we can use wide range of deep learning technologies and figure out which one obtains more accuracy[26]. At each stage of the project, there a

possibility that we can use different techniques like making use of pre-trained models like Xception, Googlenet, etc, instead of using traditional CNNs for feature extraction and other modules for deep learning could be combined with deferent loss function, layers and optimization technique which would overall lead to a better model. In this project we have used the publicly available IQ-OTHNCCD lung cancer dataset which has 1190 CT scan images. There is also a scope of using different dataset which contains wide varieties of CT scans for the models to be trained on to accurately understand the nature of CT scans that are cancerous. Furthermore, we can work on the LIDC-IDRI cancer dataset[27], which is genuinely available from the cancer imaging archive and contains DICOM files containing collective information about the patients as well as multiple CT scan slices for a single patient, which will be extremely useful in detecting cancer much more precisely.

10. REFERENCES

[1] Minna, J.D., Roth, J.A. and Gazdar, A.F., 2002. Focus on lung cancer. *Cancer cell*, *1*(1), pp.49-52.

[2] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A. and Bray, F., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *71*(3), pp.209-249.

[3] Ettinger, D.S., Akerley, W., Bepler, G., Blum, M.G., Chang, A., Cheney, R.T., Chirieac, L.R., D'Amico, T.A., Demmy, T.L., Ganti, A.K.P. and Govindan, R., 2010. Non–small cell lung cancer. *Journal of the national comprehensive cancer network*, *8*(7), pp.740-801.

[4] Zerhouni, E.A., Stitik, F.P., Siegelman, S.S., Naidich, D.P., Sagel, S.S., Proto, A.V., Muhm, J.R., Walsh, J.W., Martinez, C.R. and Heelan, R.T., 1986. CT of the pulmonary nodule: a cooperative study. *Radiology*, *160*(2), pp.319-327.

[5] Javaid, M., Javid, M., Rehman, M.Z.U. and Shah, S.I.A., 2016. A novel approach to CAD system for the detection of lung nodules in CT images. *Computer methods and programs in biomedicine*, *135*, pp.125-139.

[6] Torre, L.A., Siegel, R.L. and Jemal, A., 2016. Lung cancer statistics. *Lung cancer and personalized medicine*, pp.1-19.

[7] Jemal, A., Chu, K.C. and Tarone, R.E., 2001. Recent trends in lung cancer mortality in the United States. *Journal of the National Cancer Institute*, *93*(4), pp.277-283.

[8] Albawi, S., Mohammed, T.A. and Al-Zawi, S., 2017, August. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)* (pp. 1-6). Ieee.

[9] Al-Yasriy, H.F., AL-Husieny, M.S., Mohsen, F.Y., Khalil, E.A. and Hassan, Z.S., 2020, November. Diagnosis of Lung Cancer Based on CT Scans Using CNN. In *IOP Conference Series: Materials Science and Engineering* (Vol. 928, No. 2, p. 022035). IOP Publishing.

[10] Zuo, W., Zhou, F., Li, Z. and Wang, L., 2019. Multi-resolution CNN and knowledge transfer for candidate classification in lung nodule detection. *Ieee Access*, *7*, pp.32510-32521.

[11] Zhai, P., Tao, Y., Chen, H., Cai, T. and Li, J., 2020. Multi-task learning for lung nodule classification on chest CT. *IEEE access*, *8*, pp.180317-180327.

[12] Chen, W., Wei, H., Peng, S., Sun, J., Qiao, X. and Liu, B., 2019. HSN: hybrid segmentation network for small cell lung cancer segmentation. *IEEE Access*, *7*, pp.75591-75603.

[13] Cao, H., Liu, H., Song, E., Ma, G., Xu, X., Jin, R., Liu, T. and Hung, C.C., 2019. Multi-branch ensemble learning architecture based on 3D CNN for false positive reduction in lung nodule detection. *IEEE access*, *7*, pp.67380-67391.

[14] Al-Yasriy, H.F., AL-Husieny, M.S., Mohsen, F.Y., Khalil, E.A. and Hassan, Z.S., 2020, November. Diagnosis of Lung Cancer Based on CT Scans Using CNN. In *IOP Conference Series: Materials Science and Engineering* (Vol. 928, No. 2, p. 022035). IOP Publishing.

[15] Albawi, S., Mohammed, T.A. and Al-Zawi, S., 2017, August. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)* (pp. 1-6). Ieee.

[16] Reeves, A.P. and Kostis, W.J., 2000, April. Computer-aided diagnosis of small pulmonary nodules. In *Seminars in Ultrasound, CT and MRI* (Vol. 21, No. 2, pp. 116-128). WB Saunders.

[17] Kumar, D., Wong, A. and Clausi, D.A., 2015, June. Lung nodule classification using deep features in CT images. In *2015 12th conference on computer and robot vision* (pp. 133-138). IEEE.

[18] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, pp.321-357.

[19] Patel, H. and Thakur, G.S., 2017. Classification of imbalanced data using a modified fuzzy-neighbor weighted approach. *International Journal of Intelligent Engineering and Systems*, *10*(1), pp.56-64.

[20] Erickson, B.J. and Kitamura, F., 2021. Magician's Corner: 9. Performance Metrics for Machine Learning Models. *Radiology: Artificial Intelligence*, *3*(3), p.e200126.

[21] Tatbul, N., Lee, T.J., Zdonik, S., Alam, M. and Gottschlich, J., 2018. Precision and recall for time series. *Advances in neural information processing systems*, *31*.

[22] Beauxis-Aussalet, E. and Hardman, L., 2014, November. Simplifying the visualization of confusion matrix. In *26th Benelux Conference on Artificial Intelligence (BNAIC)*.

[23] Zheng, S., Guo, J., Cui, X., Veldhuis, R.N., Oudkerk, M. and Van Ooijen, P.M., 2019. Automatic pulmonary nodule detection in CT scans using convolutional neural networks based on maximum intensity projection. *IEEE transactions on medical imaging*, *39*(3), pp.797-805.

[24] El-Baz, A., Beache, G.M., Gimel'farb, G., Suzuki, K., Okada, K., Elnakib, A., Soliman, A. and Abdollahi, B., 2013. Computer-aided diagnosis systems for lung cancer: challenges and methodologies. *International journal of biomedical imaging*, *2013*.

[25] Sathyan, H. and Panicker, J.V., 2018, July. Lung nodule classification using deep ConvNets on CT images. In *2018 9th International conference on computing, communication and networking technologies (ICCCNT)* (pp. 1-5). IEEE.

[26] Ding, J., Li, A., Hu, Z. and Wang, L., 2017, September. Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 559-567). Springer, Cham.

[27] Jacobs, C., van Rikxoort, E.M., Murphy, K., Prokop, M., Schaefer-Prokop, C.M. and van Ginneken, B., 2016. Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database. *European radiology*, *26*(7), pp.2139-2147.

[28] S.V.G.Reddy, K.Thammi Reddy, V. ValliKumari "Optimization of Deep Learning using Various Optimizers, Loss Functions, and Drop out", International journal of Recent technology and Engineering (IJRTE) Scopus, Volume 7, issue 4S2, pages 448-455, 2018.

11. AUTHOR'S PROFILES

Dr. S.V.G.REDDY completed M-Tech(CST) from Andhra University and has obtained a PhD in Computer Science and Engineering from JNTU Kakinada. He is working as Associate Professor, Department of CSE, GIT, GITAM University. His area of research work is data mining, machine learning and deep neural networks. He has guided various B.Tech, M-Tech projects and has publications in several journals. His areas of interest are drug discovery, computer vision, brain computer interface, climate change and waste management.

V. Bhuvaneshwari is currently pursuing B.Tech(CSE) from GITAM Institute of Technology, GITAM(Deemed to be University), Visakhapatnam. Her areas of research work are machine learning and deep learning. Her areas of interest are computer vision and data science.

Aniket Kumar Tikariha is currently pursuing B.Tech(CSE) from GITAM Institute of Technology, GITAM(Deemed to be University), Visakhapatnam. His areas of research work are machine learning and deep learning. His areas of interest are data mining and data privacy.

Yagna Sriram Amballa is currently pursuing B.Tech(CSE) from GITAM Institute of Technology, GITAM(Deemed to be University), Visakhapatnam. His area of research work is machine learning. His areas of interest are deep learning and data privacy.

Balumuri Sesha Sahith Raj is currently pursuing B.Tech(CSE) from GITAM Institute of Technology, GITAM(Deemed to be University), Visakhapatnam. His area of research work is machine learning. His areas of interest are data analysis and data mining.