Lung Cancer Classification Techniques
Prarthana. K. R¹, Bhavani. K²
PG Scholar¹, Assistant Professor²
Department of Information Science
Dayananda Sagar College of Engineering, Bangalore, India

Abstract:
Diagnosing lung cancer with high accuracy is most critical to make a significant change in survival rate. For diagnosing lung cancer different imaging techniques are used by radiologists such as Magnetic Resonance Imaging (MRI), Computer tomography (CT) and X-ray. These techniques help to detect the benign or malignant or normal nodules present in the lungs but with certain limitations to the human eye. This study proposes to build a classification system that can identify the benign and malignant nodules and provide better accuracy for lung cancer detection. An “IQ-OTHNCCD” lung cancer dataset of 1190 images representing CT scan slices of 110 cases is used in this research. In this study by applying classifications methods such as Convolution Neural Network (CNN) and Artificial Neural Network (ANN) to classify the malignant and benign lung nodules and finally compare all the obtained results. This study finds the accuracy of all the applied classifier models. From all these algorithms CNN outperforms with an accuracy of 98%. ANN accuracy is observed to be 71%. This approach is successful for identifying the malignant, normal and benign lung nodules with a greater number of images than previous studies.

Keywords: Lung cancer, nodule, CT scan, CAD, Deep Learning, CNN and ANN.

I. INTRODUCTION

Background
Cancer is the most treacherous disease for human beings. Lung cancer is one of the most lethal cancer types and it is responsible for more deaths than combined death count of colon, prostate, ovarian and breast cancer. The main factor causing lung cancer is smoking and the duration of smoking is directly proportional to the person getting affected with cancer. Thousands of people are infected with this type of cancer, and if they do not discover it in the early stages of the disease, then the chance of surviving of the patient will be very poor. To detect lung cancer manually is a very tedious and risky job even for specialists. For the suggested reasons above and to help in overcoming this terrible, early diagnosis with the assistance of artificial intelligence procedures most needed. For diagnosing lung cancer different imaging techniques are used by radiologists such as Magnetic Resonance Imaging (MRI), Computer tomography (CT) and X-ray. These techniques help to detect the benign or malignant nodules present in the lungs but with certain limitations to the human eye. To gain deeper insights and identification of lung cancer in early stages, different machine leaning methods are used in image classification. By applying deep learning techniques such as CNN and other classification method (ANN), an automated system can be built which can perform with higher accuracy rate and helps in accurate classification.

Motivation
Figure 2 represents the deaths that occurred in 2019 with the types of cancer in the worldwide population. This death count is increasing even if the patient is treated with proper healthcare. One of the research states that, if the lung nodules malignancy to be detected in earlier stage then the survival rate of the cancer patient could be increased drastically. Cancer can be properly handled only in primary stages and the diagnosis of cancer in primary stage is difficult. In this case the machine learning techniques such as convolution neural network and classification algorithms can be used to classify the benign, malignant and normal lung nodules. Deaths occurred by type of cancer is explained in the figure 2.

Research Objective
The focus of this research is to build an efficient classification model which gives a higher accuracy output. Thus, the deep learning model such as CCN is considered prime in this research and the output of other classification model ANN is compared...
with the same. The aim of the research is as follows:
1. Preprocessing the image dataset:
2. Applying classification algorithms: Convolution Neural Network (CNN) and Artificial Neural Network (ANN).
3. Comparison and evaluation of the applied algorithm with the best performing model.

**Research Question**
This research addresses the problem faced in achieving higher accuracy for classifying the lung nodules from malignant to benign and which will help professionals for better decision making.

**Research Question:** How well the lung nodules can be classified from malignant to benign using deep learning techniques and classification algorithms with limited computation power for achieving higher accuracy?

**Problem Statement:** To implement a CAD system that can classify lung images as malignant or benign or normal with high accuracy by applying deep learning techniques.

**II. RELATED WORK**
Identification of lung nodules and classifying them from benign to malignant is a very crucial job and time-consuming even for skilled professionals. As a product of this, much research is done in this field for developing an automated system that will take image as an input and classify the lung nodules which would be highly reliable and time-efficient which can achieve higher accuracy. Due to the recent applications and advancement in this area, multiple approaches are considered in this study to tackle this problem providing some quality results.

In this paper [2] author S. Saskala et al. presents an approach which utilizes a Convolutional Neural Network (CNN) to classify the tumors found in lung as malignant or benign. A CNN based system was implemented to detect the malignancy tissues present in the input lung CT image. Lung image with different shape, size of the cancerous tissues has been fed at the input for training the system. The proposed system is able to detect the presence and absence of cancerous cells with accuracy of about 96%, which is more efficient when compared to accuracy obtained by the traditional neural network systems. In this proposed work, the specificity obtained is 100% which shows that there is no false positive detection. Also, the accuracy, sensitivity and specificity of the proposed system are high when compared to previously available conventional neural network based systems. The author [12] used 2 medical image dataset LIDC-IDRI and LUNGx Challenge. The classification accuracy of LIDC-IDRI -96.69%. The classification accuracy for LUNGx Challenge dataset without TL base model - 86.14% then improved to 90.91% using TL base model. The classification accuracy of this proposed texture CNN is 96.69% on LIDC-IDRI dataset. In this paper [1], author Adarsh Pradhan et al., attempted to detect lung cancer using 3D CNN on CT scan images. Here, the SPIE-AAPM Lung CT Challenge dataset is used and employed different morphological preprocessing techniques like conversion to Hounsfield Unit, removing the air region and filling the lung area to obtain the lung nodule mask. By utilizing the various metadata of the DICOM files such as slice thickness, pixel spacing, rescale slope and rescale intercept, preprocessed the data and segmented the lung nodules from the CT scans. After training the 3D CNN model utilized for lung cancer detection and obtained a very good evaluation of the model. Preprocessed dataset is divided into 60%, 20% and 20% for training, validation and testing respectively, and obtained training accuracy of 83.33%, testing accuracy of 100% and precision, recall, kappa-Score, and F-score of 1. The SPIE-AAPM Lung CT Challenge dataset is a very small dataset of about 12 GB.

In paper [4], an artificial Neural Network (ANN) for diagnose the absence or presence of lung cancer in patients was developed by author. The model validated with an accuracy of 99.01%. This study showed that neural network is able to diagnose lung cancer, so it can be used as a diagnose tool by physician. In this paper [6], author developed an Artificial Neural Network (ANN) model to detect the presence or absence of lung cancer in human body. Symptoms were used to diagnose the lung cancer such as Chest pain, Coughing, Yellow fingers, Chronic Disease, Anxiety, Wheezing, Fatigue, Shortness of Breath, Allergy and Swallowing Difficulty. This ANN model created, trained, and validated and model evaluation showed that the neural network is able to diagnose lung cancer with 96.67 % accuracy, so it can used as a diagnose tool by doctors. In this paper [7], author attempt to address several limitations of neural network models in the predication of lung cancer by proposing a new neural network model. In this, the proposed model is designed without any hidden layers to avoid data over-fitting problem and then the data over-fit problem was solved while the high prediction performance remained unaffected and thereby the entire system reliability was improved.

**III. TECHNOLOGY DECISIONS**
In this section, we can see the details of the technologies that are used for this project.

Python: Python is a high level interpreted language used for general purpose programming. It is widely used for scientific computing and can be used for a wide variety of general tasks from data mining to software development. Python is the main language used for this project.

Numpy: Numpy is a library in Python that allows for efficient numerical computing in Python. This library is highly optimized to do mathematical tasks. In the project workflow Numpy is heavily used in data pre-processing and preparation. One of the main features about Numpy is it’s highly efficient n-dimensional array (ndarray). Compared to a list in Python a Numpy array can be n-dimensions and has more features associated with the ndarray. Numpy can also perform more efficient mathematical operations compared to the math library in Python.

Pandas: Pandas is also a library in Python, like numpy is also used for data pre-processing and preparation. One of the main features about pandas is the DataFrame and Series data structure. These data structures are optimized and contain fancy indexing that allow a variety of features such as reshaping, slicing, merging, joining and etc to be available. Pandas and Numpy are extremely powerful when used together for manipulating data.

Matplotlib: Matplotlib is a Python plotting library that allows programmers to create a wide variety of graphs and visualizations with ease of use. The great feature about Matplotlib is that it integrates very well with Jupyter Notebook.
and creating visualizations is simplified. Matplotlib also works very well with pandas and numpy.

OpenCV: OpenCV (Open Source Computer Vision) is a well established computer vision library which is written in C/C++ and has been abstracted to interface with C++, Python and Java. This is a powerful tool when working with images and has a myriad of tools regarding image data manipulation, feature extraction and etc.

Tensorflow: Tensorflow is an open source deep learning library by Google. It was originally developed by Google’s engineers who were working on Google Brain and has been used for research on machine learning and deep learning. Tensorflow at its core is about computations of multidimensional arrays called tensors but what makes Tensorflow great is its ability to be flexible to deploy computations on different devices such as CPU’s and GPU’s.

Keras: Keras is also a Deep Learning Framework that abstracts much of the code in the other Frameworks like Tensorflow and Theano. Compared to the other frameworks Keras is more minimalist.

Jupyter Notebook IDE: Jupyter Notebook is an open source software IDE that allows developers to create and share documents that contain live code and more.

IV. DATASET DESCRIPTION

The Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD) lung cancer dataset was collected in the above-mentioned specialist hospitals over a period of three months in fall 2019. It includes CT scans of patients diagnosed with lung cancer in different stages, as well as healthy subjects. IQ-OTH/NCCD slides were marked by oncologists and radiologists in these two centers. The dataset contains a total of 1190 images representing CT scan slices of 110 cases. These cases are grouped into three classes: normal, benign, and malignant. Of these, 40 cases are diagnosed as malignant; 15 cases diagnosed with benign; and 55 cases classified as normal cases. Here database is introduced as a classification task instead of detection task. So its main focus is to automatic classification of lung nodules as malignant or benign in CT images. The CT scans were originally collected in DICOM formats.

V. PROPOSED WORK

Lung cancer classification based on chest CT images using CNN and ANN. In the first stage, lung regions are extracted from CT image and in that region each slices are segmented to get tumors. The segmented tumor regions are used to train CNN and ANN architecture. Then, CNN and ANN are used to test the patient images. The main objective of this study is to detect whether the tumor present in a patient’s lung is normal or malignant or benign. The trained system will able to detect the cancer in lung CT image. Steps of this work are as follows:

1. Download and explore the dataset: This model uses a dataset of about 1,097 DICOM files of lung CT scans. The dataset contains 3 sub-directories, one per class: Benign cases/ Malignant cases/ Normal cases/. Import (or install) Keras, Tensorflow and other packages on system. Load dataset from disk. Create training and testing splits. Create model. Compile and save the model. Train the model on training data (split). Evaluate the model on test data (split). Prediction on new data using trained model.

In the proposed procedure, CNNs are applied to detect and classify lung cancer CT scans of the patients collected from hospitals. Convolution Neural Networks is a sort of deep learning paradigm applied for processing data which has a grid pattern like images; it is all about using Deep Learning with Computer Vision. A good way to gain foreknowledge about this technique is to imagine a Neural Network Architecture also how it is practiced to visual tasks i.e. Video and Images. Furthermore, the Convolution Neural Networks is an important technique used for Object Recognition; create Facial Recognition, Self-Driving Cars. A Convolution Neural Network is a Deep Learning algorithm that can take in an image as input, with assigning important learnable weights and biases to various objects inside this image and be capable of differentiating one from the other. In addition, the preprocessing required for this technique is much lower if compared with other classification algorithms. The role of the CNN is for reducing the images without losing features that are important for getting a good prediction. A typical CNN consists of three types of operation layers: the convolution layer (CONV), the pooling layer (POOL), and finally the classifier layer (FC), as exemplified in the figure 5.
VI. RESULTS

The CNN and ANN model was trained using Jupyter Notebook and training Results of each model is as given in this section. These models created plots of loss and accuracy on the training and validation sets. As you can see from the plots of CNN model in figure 8 and 9, training accuracy and validation accuracy are off by small margin and the model has achieved around 98% accuracy on the validation set whereas, in figure 6 and 7, the plots of ANN model gives 71% accuracy on the validation set. Finally, as in Table 1 our model classifies an image that wasn't included in the training or validation sets. CNN model predicts each of the lung cases with high confidence whereas ANN model fails to predict one of the lung case that is Benign.

Table 1. Shows Training and Validation results of ANN and CNN model.

VII. CONCLUSION

Considering all the related work above mentioned it can be observed that building an efficient and effective classification system two different machine learning algorithm have been studied to achieve optimum results and high accuracy rate. Classification models have different approach based on the relevant and the size of the data. It can be seen that deep learning model such as CNN needs to have data in large volume to perform well whereas models like ANN. As per the
classification approach, this research aims to compare the performance in sense of high accuracy. Early prediction of lung cancer helps to treat the patients more effectively. For this purpose, this study develops an effective computer-aided diagnosis scheme for early detection of this lethal cancer and CT scans have been employed here as data input. The Deep learning methods used for classification where the system is trained with large datasets to diagnose the type of cancer with its size and shape. This Project goal is to improve a CNN deep learning model to classify lung cancer nodules successfully with high accuracy. By applying deep learning techniques such as Convolutional Neural Network (CNN), Artificial Neural networks (ANN) classify the malignant, normal and benign lung nodules and finally observed that the CNN model outperformed ANN by achieving 98% accuracy on validation set. In the future, develop innovative tools that can detect Lung cancer with high risk of sensitivity and specificity. To enhance the performance of the proposed model by increasing the number of images in the used datasets, increasing the training epochs and using other deep learning techniques such as GAN, Capsule-Network in classification.

VIII. REFERENCES


[8] Sanjukta Rani Jena, Dr. Thomas George, Dr. Narain Ponraj “Texture Analysis Based Feature Extraction and Classification of Lung Cancer” 978-1-5386-8158-9/19/$31.00 ©2019 IEEE


