LUNG CANCER PREDICTION AND DETECTION USING ML

RASHMI T N, ARCHANA B S, SUJATHA A

Lecturer, Department of Computer Science & Engineering, Government Polytechnic, Channapatna, Karnataka, India
Lecturer, Department of Computer Science & Engineering, Government Polytechnic, Ramanagara, Karnataka, India
Lecturer, Department of Electronics & Communication Engineering, Government Polytechnic, Tumkur, Karnataka, India

ABSTRACT

Lung cancer is the major cause of cancer-related death in this generation, and it is expected to remain so for the foreseeable future. It is feasible to treat lung cancer if the symptoms of the disease are detected early. It is possible to construct a sustainable prototype model for the treatment of lung cancer using the current developments in computational intelligence without negatively impacting the environment. Because it will reduce the number of resources squandered as well as the amount of work necessary to complete manual tasks, it will save both time and money. To optimize the process of detection from the lung cancer dataset, a machine learning model based on CNN is used. Using an CNN classifier, lung cancer patients are classified based on their symptoms at the same time as the Python programming language is utilized to further the model implementation. The effectiveness of our CNN model was evaluated in terms of several different criteria. Several cancer datasets from the University of California, Irvine, library was utilized to evaluate the evaluated model. As a result of the favorable findings of this research, smart cities will be able to deliver better healthcare to their citizens. Patients with lung cancer can obtain real-time treatment in a cost-effective manner with the least amount of effort and latency from any location and at any time. The proposed method gets a 97.8% of accuracy rate when comparing the existing methods.

CNN-Convolutional Neural networks.

1. INTRODUCTION

Cancer is a lethal disease that is frequently brought on by the accumulation of hereditary disorders and many pathological alterations. Cancerous cells are abnormal growths that can appear anywhere on the body and pose a threat to life. To determine what might be helpful for its treatment, cancer, also known as a tumor, must be promptly and accurately discovered in the early stages. Even while each modality has its own unique considerations, some of the major causes of mortality include difficult histories, inadequate diagnoses, and inadequate treatment. When you inhale, air enters through your mouth or nose and travels through your trachea to your lungs. The trachea splits into bronchi, which pass into the lungs and further split into smaller bronchi. These divide into bronchioles, which are smaller branches. The tiny air sacs known as alveoli are located at the end of the bronchioles. When you inhale air, the alveoli take oxygen into your blood and eliminate carbon dioxide when you exhale. The major jobs of your lungs are to take in oxygen and expel carbon dioxide.

The cells lining the bronchi and other areas of the lung, such as the bronchioles or alveoli, are where lung malignancies generally begin. The pleura, a slender layer of lining, encircles the lungs. Your pleura shields your lungs and aids in their movement back and forth against the chest wall as you breathe. The diaphragm, a slender, dome-shaped muscle, divides the chest from the belly under the lungs. The diaphragm contracts and expands while you breathe, propelling air into and out of the lungs. The purpose of is to study and analyze methods for lung cancer diagnosis using machine learning. The study demonstrates how machine learning utilizing supervised, unsupervised, and deep learning approaches is used to aid in the detection and treatment of cancer. The performance of several

state-of-the-art approaches is compared using benchmark datasets for accuracy, sensitivity, specificity, and false-positive rates. Artificial neural networks do incredibly well in machine learning. Artificial neural networks are used to classify words, audio, and images among other things. Different forms of neural networks are employed for various tasks. For example, to predict the order of words, recurrent neural networks— more specifically, an LSTM—are used. Similarly, to classify images, convolution neural networks are employed.
We're going to create the fundamental building element for CNN in the inputs are images are known as convolutional neural networks (CNN). They are used to conduct object detection within a frame, evaluate and categorise photos, and group images based on similarities. Convolutional neural networks (CNNs), for instance, are used to recognise faces, people, street signs, cancers, platypuses, and many other visual data elements Project focuses on system gaining knowledge of set of rules to stumble on most cancers the usage of CT test photo. Detection of most cancers the usage of system gaining knowledge of strategies in CT test. Convolutional Neural Network (CNN) is a category of deep neural network, maximum generally implemented to investigate visible imagery. It's determined in literature survey that CNN, KNN and SVM offers 92%,84% and 88% of performance respectively in photo processing. CNN technique may be used to stumble on the lung most cancers.

In recent years, in the field of image recognition and deep learning, especially convolutional neural networks have proved highly successful. It defeated other traditional machine learning methods when it first appeared and won the championship of the ImageNet large-scale image recognition challenge in one fell swoop, significantly reducing image recognition Error rate. However, most of the deep learning models are only applied in the field of natural image recognition, and there are few applications in the areas of medical image diagnosis. The use of deep learning technology for lung cancer CT image diagnosis can significantly reduce the diagnosis time of doctors, improve hospitals' efficiency, effectively alleviate the shortage of medical resources and other problems, and early diagnosis, early treatment, and even save lives.

There are two strategies for applying deep learning to medical image diagnosis: first, Using medical images, training convolutional neural network models from scratch. Second, transfer learning, using a pre-trained convolutional neural network model and weight parameters to extract features. However, in the diagnosis of lung cancer CT images, there is currently only a way to train from scratch, and no one has ever used transfer learning. So, this study proposes a strategy for transfer learning in lung cancer CT image diagnosis and conducts experiments.

The experimental results verify the transfer learning method has delighted results. Through learning, the model selects the accurate features from the training data so that when testing new data, it can make correct decisions. Therefore, deep learning plays a crucial role in medical image processing. In recent years, deep learning has continued to make significant progress, mainly due to the continuous improvement of computing power and the continuous increase in the amount of available data, as well as the continuous improvement of deep learning models and algorithms. Since 2006, many convolutional Neural Network architectures have developed to overcome the problems encountered earlier.

2. LITERATURE SURVEY

To create a reliable technique for detecting cancer in its earliest stages. It was meant to research the various traits of information mining techniques as well as the advantages of data processing theories in the categorization of cancer. It researches various data processing and hymenopteran colony improvement approaches for the generation of useful rules and classifications of diseases. It also offers the fundamental underpinning for improving cancer diagnosis.

Bar chart equalisation is used in this work to pre-process the photos. A dataset is used in an experimental analysis to determine the performance of various classifiers, with performance being dependent on the classifier's accurate and incorrect classifications. A performance comparison of the Support Vector Machine rule with various classification methods is conducted. It clearly demonstrates its effectiveness in comparison to other classifiers by producing the greatest TP Rate and lowest FP Rate. The same author has since used Feature Extraction and Principal Element Analysis to identify cancer in CT scan images. All researchers have sought to increase the Early Prediction and Detection system's accuracy by pre-processing, segmenting features for extraction, and classifying the resulting database. Below is a summary of the research's main contributions.

2.1 Author Name: Amira Bido Sallow, Adnan Mohisin Abdulazeez
Published year:2021
Publication: IEEE International Conference on System, Vol. 2, No. 1

Introduction: Multiple organs are simultaneously affected by cancer, and distinct forms of cancer can develop in different body organs. It's possible that the ailment goes unnoticed for a very long time. According to WHO reports, if cancer is caught early enough, it may be prevented. Whether or whether the patient is given an early prognosis, their life expectancy will be increased. The prognosis for lung cancer is poor and substantially varies according to the tumour stage at the time of diagnosis. Non-small cell lung cancer (NSCLC) and small cell lung cancer are the two subtypes of lung cancer that are clinically treated (SCLC). In reality, it is a malignant tumour with uncontrolled cell tissue growth. Long-term cigarette usage was the primary cause of lung cancer development. the results of research. Nineteen different types of cancer can attack a healthy person. The majority of these malignancies, including lung cancer, cause mortality. Over 1.7 million people every year are anticipated to pass away from this illness. Machine learning (ML) research has already advanced significantly, which helps to lessen the need for human labourers. To design algorithms that perform better when exposed to relevant data, machine learning (ML) integrates statistics and computers in the field of artificial intelligence.

Result and Discussion: Each classifier's accuracy was evaluated using the confusion matrix. According to the experimental findings, a five-attribute CNN classifier produces the best prediction ratio of 95.56 percent, and a CNN classifier produces an accuracy ratio of 92.11 percent. KNN, on the other hand, has the lowest estimation accuracy, at 88.40 percent.

Advantages: It takes relatively little time to evaluate the data and is the most preferred way for detecting lung cancer in its early stages. and it has a very good accuracy level.

Disadvantages: The dataset attribute values are high and the design is quite complex. Additionally, the input field has a sizable amount of missing data.

Future Scope: The most frequent cause of mortality and one of the most severe diseases, lung cancer is made more dangerous by the challenge of making an early diagnosis. This study aims to evaluate three classifiers to determine which one is most effective at identifying lung cancer in its early stages. The informative indicators used in this investigation were obtained from lung cancer patient databases at UCI. This research focuses on using WEKA Tool to look at the precision of classification methods. The findings demonstrate that K-Nearest Neighbour KNN and Support Vector Machine (SVM) have the best accuracy (95.56%) in detecting lung cancer in its early stages. Less accuracy (88.40%) was reported

2.2 Author Name: C. Anil Kumar, S. Harish, Prabha Ravi
Published year :2022
Publication: IEEE Sensor Journal, Vol 21 No. 18

Introduction: Lung cancer, in other words, is the main global cause of death for both men and women. Other research indicates that in 2020, the number of cancer diagnoses in the United States related to lung cancer was approximately 13%. The American Cancer Society estimates that 27% of all cancer-related fatalities are caused by lung cancer. known as local binary patterns). These labels are utilised in additional picture processing, which is often displayed as a histogram. The LBP texture operator has been employed in a variety of applications due to its ability to be precise and ease of usage

These markers are then utilised by the histogram to carry out a more detailed examination of the image. In both men and women over the past three years, cancer mortality from lung illness has remained higher than cancer mortality from prostate or breast cancer This is due in great part to the complex and systemic nature of prognosis.

Result and Discussion: The information was located in the repository for machine learning at UCI, and the dataset has 32 samples, each with 57 characteristics and an overall notional range of 0–3 attributes. By converting nominal attribute and class label data into binary form, this is made possible analysis is simpler to carry out. Data transformation from the most extensively used and standardised method for data analysis is nominal to binary form. Some items are missing. When analysing the data, care should be taken to avoid making assumptions about values

in the dataset that could affect how well the algorithm performs. The label features three distinct High, medium, and low severity levels are available. The input data contains a sizable amount of missing information.

Advantages: By using kernel properties, this approach will reduce complexity and is relatively simple to implement.

Disadvantages: This method becomes slower. And also, the number of variable increases.

Future Scope: Because cancer cells are so complicated, predicting lung cancer is one of the most difficult medical problems to solve. There are approximately 100 different cancers to be concerned about in addition to lung cancer. Delaying lung cancer therapy significantly increases the probability of dying from the disease. Cancer can be cured if it is found and treated early enough. SVM is employed in this study to forecast the onset of lung cancer. The main goal of this system is to give customers an early warning so they can save time and money. Positive results from the performance evaluation of the suggested approach show that oncologists can use SVM to help in identification of lung cancer.

2.3 Author Name: Sushmita S. Patil, Aishwarya V Budhe
Published year :2021
Publication: Research Paper, IEEE Transactions on Image Processing,16 Vol 8

Introduction: The condition known as lung cancer causes abnormal cells to grow and form tumours in the lungs. The lymph fluid that surrounds lung tissue and the bloodstream both have the potential to carry cancer cells away from the lungs. When a cancer cell leaves the site of its origin and spreads through the circulation to a lymph node or another area of the body, this is known as metastasis. Lymph is delivered by lymphatic veins to lymph nodes in the lungs and chest centre. Lung cancer that starts there is referred to as primary lung cancer. There are numerous types of lung cancer, including carcinoma, adenocarcinoma, and squamous. Lung cancer accounted for 365 cases of cancer diagnoses among Jordanians in 2008, ranking first overall**.**

Result and Discussion: Image segmentation is the process of dividing a digital image into several segments, such as groups of pixels, also referred to as super-pixels. The basic objective of segmentation is to change an image's representation into a more understandable form. Image segmentation is the process of locating objects, borders, and other features in images. Image segmentation is the process of giving each pixel in an image a label such that pixels with the same label have particular features. The outcome of image segmentation is a group of segments that together cover the full image, or a set of edges and boundaries that are recovered from the image. A certain region may contain pixels that are similar to one another in terms of colour, intensity, or texture.

Advantages: By using kernel properties, this approach will reduce complexity and is relatively simple to implement. Also, it helps to predict the lung cancer in early stages.

Disadvantages: This method becomes slower. And also, number of variable increases.

Future Scope: An automated lung cancer detection technique from a lung CT image was introduced in this paper. This developed algorithm successfully detects lung cancer in a CT scan of the lungs. This algorithm was tested on a large number of images and found to be effective in detecting lung cancer with affected percentage.

2.4 Author Name: Aashka Mohite
Published year :2021
Publication: Research Paper, IEEE 31st International Symposium on Computer Based Image Processing Vol 9

Introduction: Early identification increases the likelihood of survival and can help lower the cancer's aggressiveness. The likelihood of survival in an advanced stage is lower when compared to treatment to survive cancer therapy when diagnosed at an early stage. In order to distinguish between benign tumours, which develop gradually and do not affect other body parts, malignant tumours, which develop rapidly and readily infiltrate other body parts, and cancer-free patients, the method was created in this manner. To identify lung tumours early, a variety of imaging methods are utilised, such as computed tomography (CT), sputum cytology, chest X-rays, and magnetic resonance imaging (MRI). In order to detect tumours, they must first be divided into two categories: I non-

cancerous (benign) tumours, and (ii) cancerous tumours (malignant). Frequently, the prognosis of lung cancer is based on the knowledge of doctors, who may ignore some patients and cause problems.

Result and Discussion: Results and Discussions Confusion matrix-based performance metrics are used to evaluate the performance of all these models. These numbers correspond to the trained model's results on the testing set. Accuracy, precision, recall, and F1-score are the metrics evaluated here. True Positives (TP) indicating that the actual and predicted values are the same, and True Negatives (TP) indicating that the sum of values of all columns and rows except the values of the class for which we are calculating the values is the same, False-positives (FP) are the sum of the values in the corresponding column that are not True positive, and False-negatives (FN) are the sum of the values in the corresponding rows that are not True positive.

Advantages: This approach will reduce complexity and is relatively simple to implement. Also, it helps to predict the lung cancer in early stages.

Disadvantages: This method becomes slower. And also, number of variable increases. And this is very difficult to implement.

Future Scope: A comparison of several transfer learning architectures for detecting the type of lung cancer existing in a person was conducted in this paper. DenseNet-201 was chosen as the architecture for the system's implementation with the accuracy of 53%, recall of 43%, precision of 43% and f1-score of 43%. The created model only processes a single CT scan slide. The lungs, on the other hand, should be thoroughly checked in order to detect lung cancer more effectively. Future research on this subject should entail examining the lungs from all angles.

2.5 Author Name: Mustafa Mohammed Jassim, Mustafa Musa Jaber
Published year :2022
Publication: Research Paper, IEEE Trans med Imagining Vol 5

Introduction: Five years after diagnosis, individuals with lung cancer have a 10-to-20% chance of surviving. Early lung cancer detection by low-dose computed tomography (CT) screening can help make the disease more amenable to therapy. The likelihood that a patient will live a long life is generally stated to rise if the cancer case is discovered early, diagnosed, and treated well. Medical specialists are necessary for the analysis of medical data and the diagnosis of diseases, and the intricacy of medical images makes it common for expert opinions to diverge. In the field of medicine, artificial intelligence is crucial.

Result and Discussion: The results of the query search were 591 articles from the four databases. After that, the duplicative articles (n = 49) were removed, which means the number of articles were n = 542. The second stage of the examination was done by scanning the title and abstract, where the number of articles became n = 132; in the final filter, we read the articles' full text, but we were unable to download 10 papers due to access issues, that resulted in 122 articles and finally based on inclusion criteria, we have selected 55 related articles. After that, a complete reading of all the selected articles was done, and a suggested classification was made.

Advantages: It takes relatively little time to evaluate the data and is the most preferred way for detecting lung cancer in its early stages.

Disadvantages: This method becomes slower. And also, some time produce wrong prediction.

Future Scope: This research aimed to analyse the literature related to the field of lung cancer diagnosis. The authors' suggestions in the field of lung cancer diagnosis were either to suggest new methods of solution or to change the methods of pre-treatment and segmentation for improving the accuracy of the diagnosis. Some of them have proved their ability for lung cancer diagnosis based on new types of datasets and it was reviewed based on the authors' suggestions (dataset used, pre-processing approaches, and method used) with the results (area under the ROC curve, accuracy, precision, specificity, and sensitivity) obtained for diagnosis. This study reviewed the published datasets that are available to researchers, their sources, and size.

3. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

The current lung cancer detection and prediction system employs K-Nearest Neighbour Classifier and Support Vector Machine (SVM). The data processing times for SVM and K-NNC algorithms are both longer. It is a significant issue with this approach, and neither algorithm can generate accurate output. Both the SVM and k-NNC algorithms can occasionally produce incorrect results, which can lead to patient death.

3.2 PROBLEM IDENTIFICATION

Utilizing machine learning algorithms, this project's goal is to find lung cancer. The literature on lung cancer detection and prediction has shown numerous problems. Most scientists have employed the SVM and K-NNC algorithms. Given that the SVM and K-NNC algorithms need time to run. as it gradually requires more time to deliver the result.

Sometimes the results of both algorithms are incorrect. Additionally, processing a huge data collection would take longer due to the enormous number of values involved. In order to solve these issues, we are employing the CNN algorithm for lung cancer diagnosis and prediction.

An CNN classifier is used in this part to categorise lung cancer patients according to their symptoms. In the methodology that has been proposed, pre-processing happens before data collection. The chosen classifiers are subsequently trained and assessed on the benchmark dataset a second time using a conventional 10-fold cross validation approach. To find the most reliable way to detect lung cancer, the data are calculated and assessed.

3.3 OBJECTIVES

- To detect lung cancer using Convolutional Neural Network (CNN)
- Graph-based analysis of data set used.
- Testing the CNN algorithm on lung cancer dataset
- Comparison of result with other algorithm
- The creation of numerous graphs to discuss the outcome

Lung cancer prediction is very important in early stage to saves the life time. The objectives of the given research work is mentioned below:

- Our proposed work based on the support Convolutional Neural Network for lung cancer detection.
- It increases the accuracy, precision and recall rate.
- Reduce the execution time of the model.

The best method for predicting lung cancer is CNN. The CNN algorithm is a machine learning technique that analyses classification and regression data, but it is primarily used for classification. When using the smallest dataset, CNN is more suitable.

Between the two classes, CNN establishes a decision boundary. Different decision boundaries exist to divide the two classes, but we must identify the optimum decision boundary, also known as the CNN hyperplane. The primary goal is to identify the hyperplane with the greatest separation between the data points for two classes.

3.4 PROPOSED SYSTEM

Our proposed work based on the support Convolutional Neural Network for lung cancer detection. It increases the accuracy, precision and recall rate. Reduce the execution time of the model. The best method for predicting lung cancer is CNN. The CNN algorithm is a machine learning technique that analyses classification and regression data, but it is primarily used for classification. When using the smallest dataset, CNN is more suitable.

3.5 FLOW DIAGRAM

An algorithm known as CNN is a supervised machine learning technique that can be applied to classification or regression problems. It is primarily utilised, nevertheless, in classification issues. The CNN algorithm plots each piece of data as a point in n-dimensional space (where n is the number of features you have), with the value of each feature being the value of a certain coordinate. Then, classification is done by locating the hyper-plane that best distinguishes the two classes.

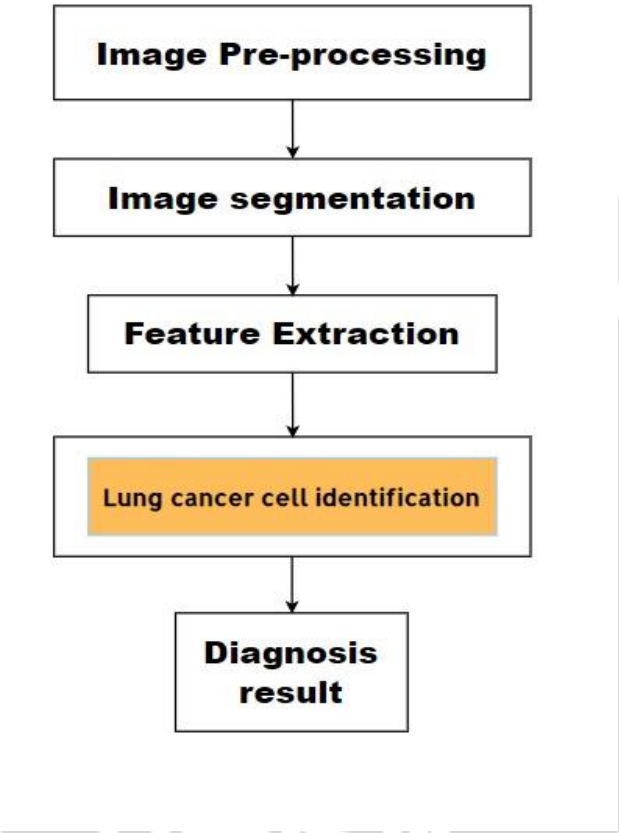

Fig 3.1: Flow of Operation

Data Acquisition: For this inquiry, we used the Lung Cancer dataset from the Kaggle web repository. The full dataset consists of 32 instances, 57 characteristics, and one class attribute. Our suggested study's main goal is to assess and contrast CNN performance against that of other technologies.

Data Pre-processing: Pre-processing data involves removing any information that is not absolutely necessary and filling in any gaps in the dataset, which is the first step in the diagnosis of lung cancer. Because of this, missing values are imputed using the K nearest neighbour strategy with three neighbours in order to boost the overall reliability of the entire dataset. Samples for both testing and training are needed.

Training and Testing Samples: A neural network, a form of artificial intelligence, is used to train the input data samples and then test them. At the start of the procedure, the weights of the neural network are generated randomly from the input data. The same dataset that was utilised for training the neural networks serves as the basis for their evaluation. To determine the frequency of errors or error rates that occur during the classification process, data is weighted. Errors are then corrected by reweighting the dataset.

Features Extraction: The dataset must be split in order to extract a significant number of features that can reduce the difficulty of identifying lung cancer. The lung tumour that has developed as a result of the proliferation of cancer

cells will be removed using this detection method (features). Utilizing the PSO programming language is the feature extraction method. Pattern recognition algorithms employ feature extraction from input data to extract the key attributes that are more valuable and nonredundant, as well as to gather information about cancer to forecast patient scenarios for interpretation and classify data using SVMs.

Classification Using CNN: The process of classifying involves grouping information logically. Both organised and unstructured data, which the system is capable of analysing, can be used as the basis for decisions. Using data mining techniques, malware and denial of service (DoS) assaults can both be avoided and mitigated.

3.6 METHODLOGY

A. Dataset

Diagnosis of lung cancer has based on clinical databases. The LIDC-IDRI dataset is the most commonly used data from this kind of datasets [1]. Dataset consists of a screening of lung cancer and CT scans. Seven research centres and eight medical imaging companies cooperated to build a dataset of 1018 cases. Furthermore, it also organizes competitions to improve the accuracy of classification. The SPIE-AAPM Lung Challenge dataset is a subset of the medical imaging conference at SPIE in 2020 with the support of American Association of Physicists in Medicine (AAPM) and the National Cancer Institute (NCI). To identify the pulmonary nodules as benign or malignant, the use of a standard dataset has proposed to test the competitors more precisely. SPIE-AAPM dataset [2] is utilized in our research. The dataset includes 70 patient CT images. Ten of these cases have been used for training and the remaining 70 for testing. In this implementation, we used the data augmentation technique to increase the number of CT image artificially from a small dataset to thousands of images.

B. Transfer Learning and CNN architectures

Transfer learning [2] is a prevalent approach in the area of computer vision because it can construct accurate models, and less time is needed. Using transfer learning is not to start learning from scratch, but to start from the model learned when solving various problems. In this way, we can use previous learning results and avoid starting from scratch. Deep CNNs are still commonly used in present-day research. They provide creative assistance to overcome many challenges relating to classification. Lack of data from training is a common issue when using deep CNN models that need a significant number of data to perform well. Furthermore, it is tedious to collect a vast dataset, and it continues even now. Thus, the transfer learning approach has generally used to solve the limited data collection issue.

Transfer learning is a method where CNN models have trained on datasets with a massive number of data, and afterward, the models have fine-tuned to train on a small required dataset. Transfer learning is a useful technique where we might use a pre-trained model like (Alex Net, ResNet18, Google net, and ResNet50) and adjust the network for the next application by making specific improvements in the architecture of the network [3]. Alex Net and the other pre-trained networks have trained for 1000 classes of real-world images. We can use this network to identify some other classes with some adjustment to the network.

The mechanism of using any pre-trained models will be described in figure 3.2. The common pre-trained models like (Alex Net, ResNet18, Google net, and ResNet50) have proposed for well-organized classification. In this study, all models have altered the last three layers to adjust the new image classification. We modified each model as in the following:

- Alex Net: the last three layers of the constructed network with a group of layers are changed fully connected layer (FC), SoftMax layer, and output layer (classification) to classify images into relevant classes.

- ResNet18: The network layers (fc1000, prob, and Classification Layer predictions) are replaced with fully connected layer, SoftMax layer, and classification output layer. Afterward, the last remaining transferred layer on the network (pool5) is linked to the novel layers.

• Google Net: also, the last three layers of the network are modified. The layers loss3-classifier, prob, and classification output layer are adjusted with a fully connected layer, SoftMax layer, and an output layer. Later, the last transferred layer still existing on the network (pool5_drop7x7_s1) is connected to the new layers.

• ResNet50: The network's (fc1000, fc1000_softmax, and ClassificationLayer_fc1000) layers are replaced with fully connected layer, SoftMax layer, and classification output layer. Afterward, the last remaining transferred layer on the network is linked to the novel layers.

• Google Net: Also, the last three layers of the network are modified. The layers loss3-classifier, prob, and classification output layer are adjusted with a fully connected layer, SoftMax layer, and an output layer. Later, the last transferred layer still existing on the network is connected to the new layers.

• ResNet50: The network's (fc1000, fc1000_softmax, and ClassificationLayer_fc1000) layers are replaced with fully connected layer, SoftMax layer, and classification output layer. Afterward, the last remaining transferred layer on the network is linked to the novel layers.

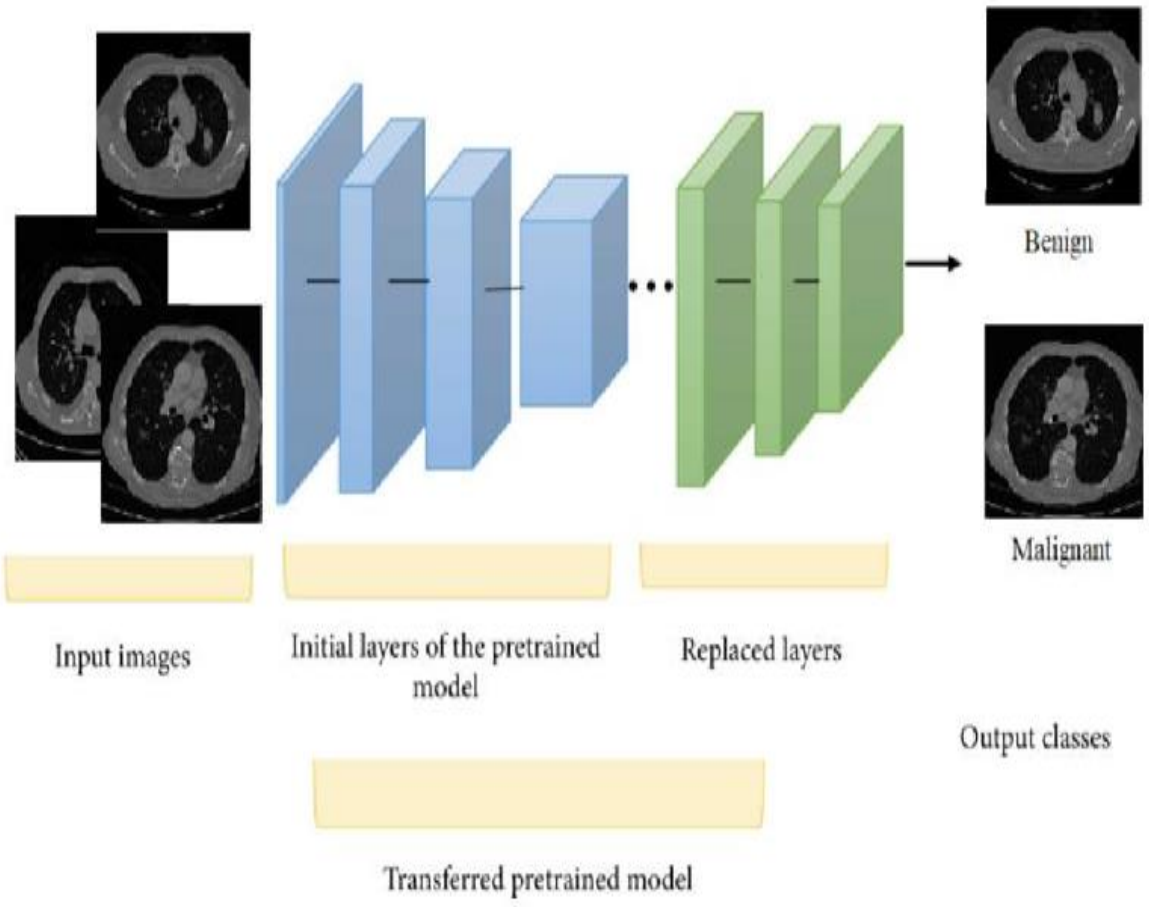

Fig 3.2: Transfer Learning Process using Pre-Trained Model.

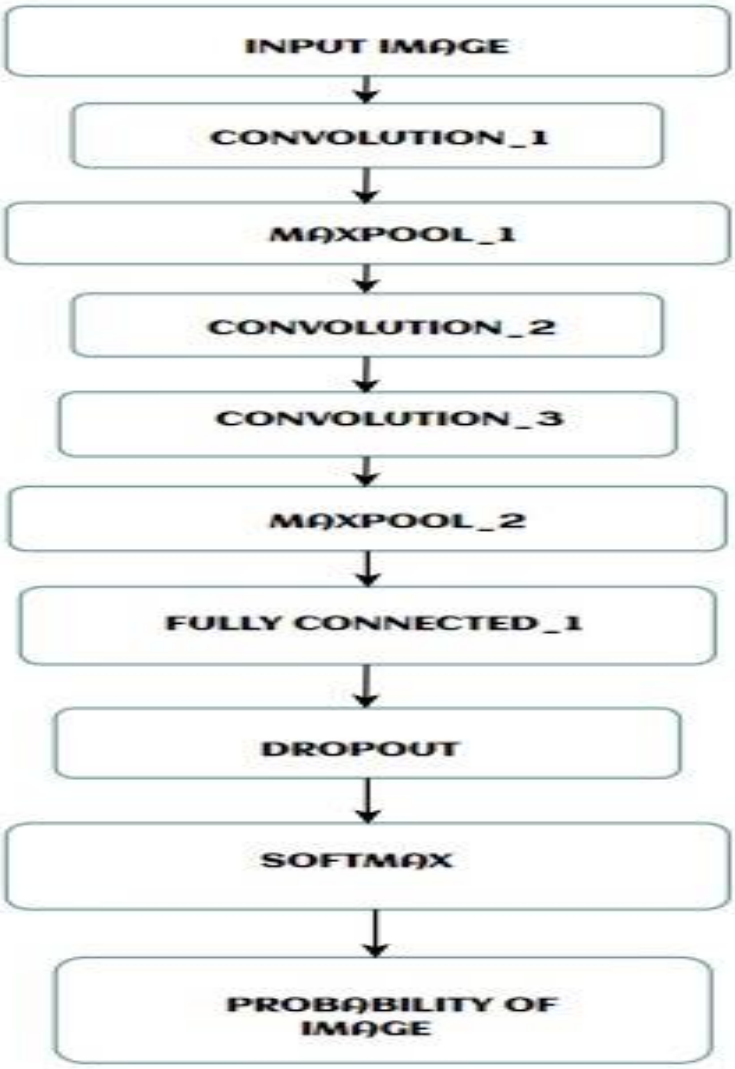

Fig 3.3: CNN Architecture

RESULT OF CNN ARCHITECTURE

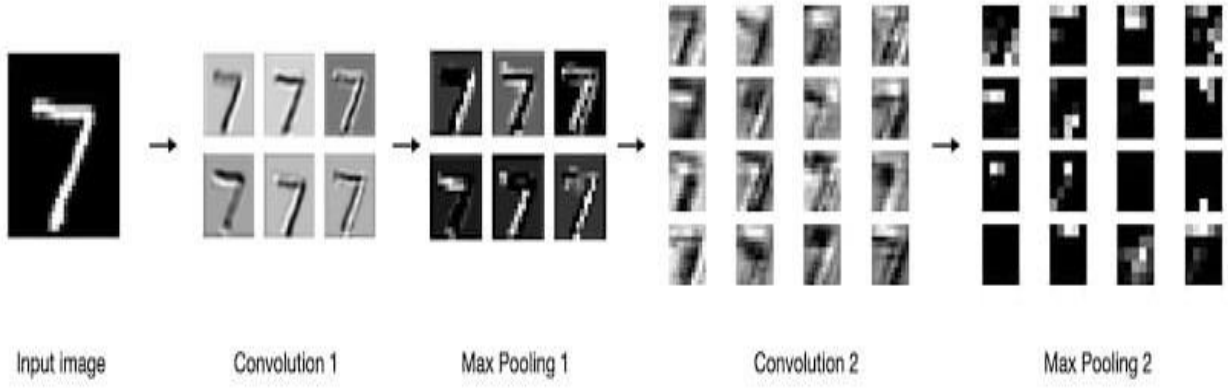

Fig 3.4: CNN Architecture Result

C. Data augmentation

Deep learning is one of the best choices in the field of image processing. However, its medical imaging application has limited the massive need for high-quality labelled images as training samples. It is costly to collect medical images, which needs a specialist to label the images. To address this problem, we took advantage of Transfer Learning and data augmentation. Data Augmentation is a technology that artificially expands the training data set by allowing limited data to generate more comparable data. It is an effective means to overcome the lack of training data and avoid overfitting issues [5]. Firstly, doctors can examine histological images of Lung cancer from different angles without impacting the process of diagnosis. Therefore, utilizing data augmentation using rotation moreover improves the dataset. Secondly, the rotation technique has enlarged the dataset's size without impacting the quality of the input images [7].

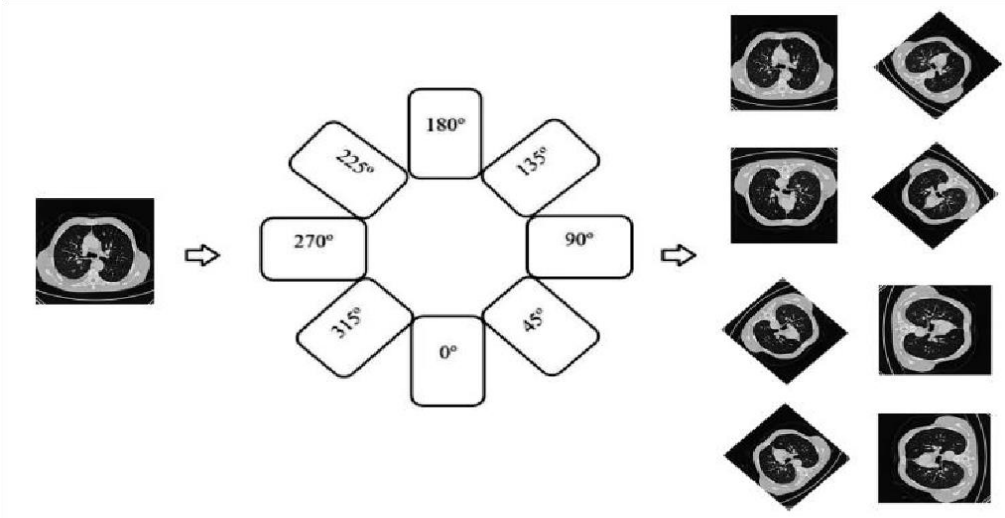

Fig 3.5: Rotation Process in Different Angle

D. Evaluation matrix

The classification process's performance was evaluated by several matrices such as confusion matrix (accuracy) equ1, recall (sensitivity) equ2, precision equ3, specificity equ4, and F1-score equ5. The confusion matrix (accuracy of the model) is the most basic, intuitive, and easiest way to measure models' accuracy [8]. The parameters and equations have defined as:
True Positive (TP): predict a complimentary class as a positive class number.
True Negative (TN): predict a negative class as a hostile class number.
False Positive (FP): predict a negative class as a positive class number.
False Negative (FN): Predict the complimentary class as a hostile class number.

$$\text{accuracy} = (\text{TP} + \text{TN})/\ (\text{TN} + \text{TP} + \text{FP} + \text{FN}) \quad (1)$$

$$Recall\ (sensitivity) = TP/\ (TP + FN) \quad (2)$$

$$Precision\ = TP/\ (TP + FP) \quad (3)$$

$$Specificity\ = TN/(TN + FP) \quad (4)$$

$$F1 - score = 2 * (Precision * Recall)\ /\ (Precision + Recall) \quad (5)$$

4 SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE

Detailed design starts after the system design phase is completed and the system design has been certified through the review. The goal of this phase is to develop the internal logic of each of the modules identified during system design. In the system design, the focus is on identifying the modules, whereas during detailed design the focus is on designing the logic for the modules. In other words, in system design attention is on what components are needed, while in detailed design how the components can be implemented in the software is the issue.

The design activity is often divided into two separate phase system design and detailed design. System design is also called top-level design [3]. At the first level focus is on deciding which modules are needed for the system, the specifications of these modules and how the modules should be interconnected. This is called system design or top-level design. In the second level the internal design of the modules or how the specifications of the module can be satisfied is decided. This design level is often called detailed design or logic design.

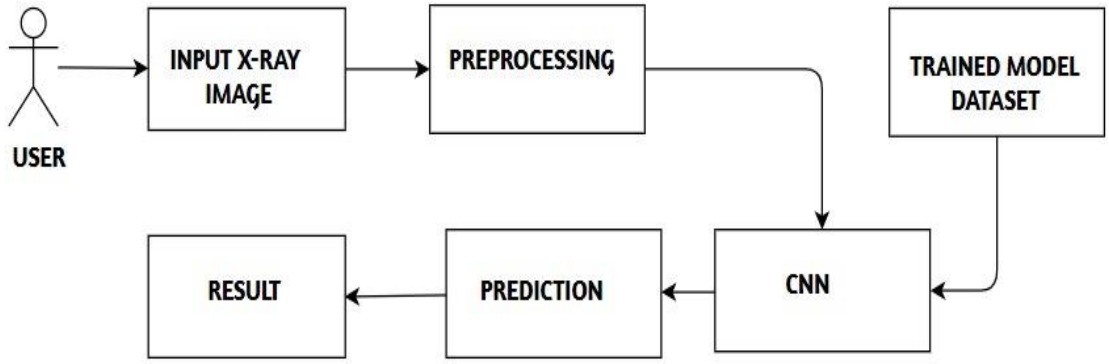

Fig 4.1: System Architecture

4.2 LOW-LEVEL DESIGN

PRE-PROCESSING

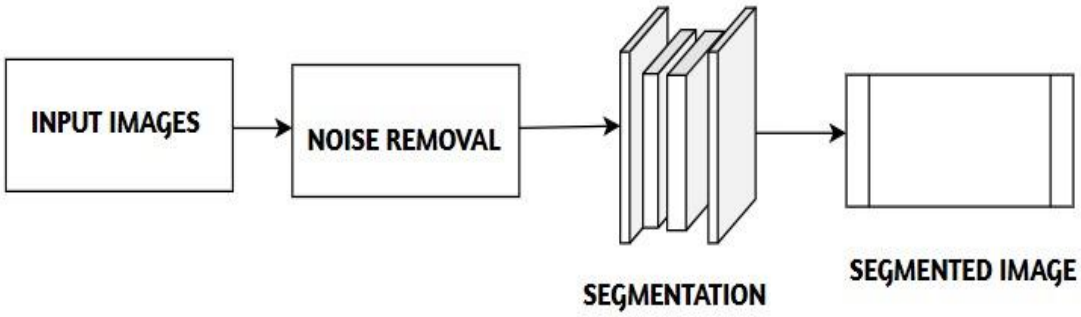

Fig 4.2: Pre-Processing

RESULT OF PRE-PROCESSING

Fig: 4.3 Result of Pre-Processing

MODEL-TRAINING

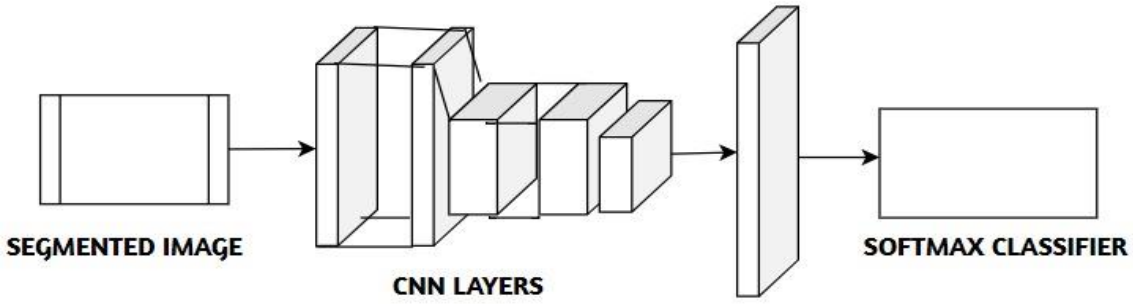

Fig 4.4: Model Training

4.3 HIGH-LEVEL DESIGN

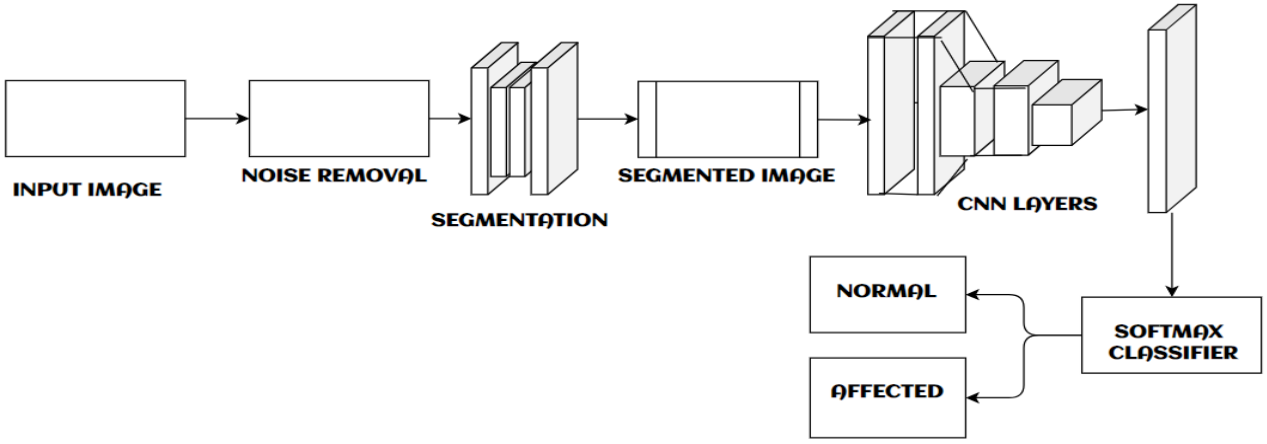

Fig 4.5: High level design

4.4 DATA FLOW DIAGRAM

DFD graphically representing the functions, or processes, which capture, manipulate, store, and distribute data between a system and its environment and between components of a system.

The visual representation makes it a good communication tool between User and System designer. Structure of DFD allows starting from a broad overview and expand it to a hierarchy of detailed diagrams. DFD has often been used due to the following reasons:

- Logical information flow of the system
- Determination of physical system construction requirements
- Establishment of manual and automated systems requirements

Basic Notation

Process: any process that changes the data, producing an output. It might perform computations sort data based on logic, or direct the data flow based on business rules. A short label is used to describe process, such as the "Submit payment."

External entity: an outside system that sends or receives data, communicating with the system being diagrammed. They are the sources and destinations of information entering or leaving the system. They might be an outside organization or person, a computer system or a business system. They are also known as terminators, sources and sinks or actors. They are typically drawn on the edges of the diagram.

Data flow: the route that data takes between the external entities, processes and data stores. It portrays the interface between the other components and is shown with arrows, typically labelled with a short data name, like "Billing details.

Data Flow Diagram Level-0

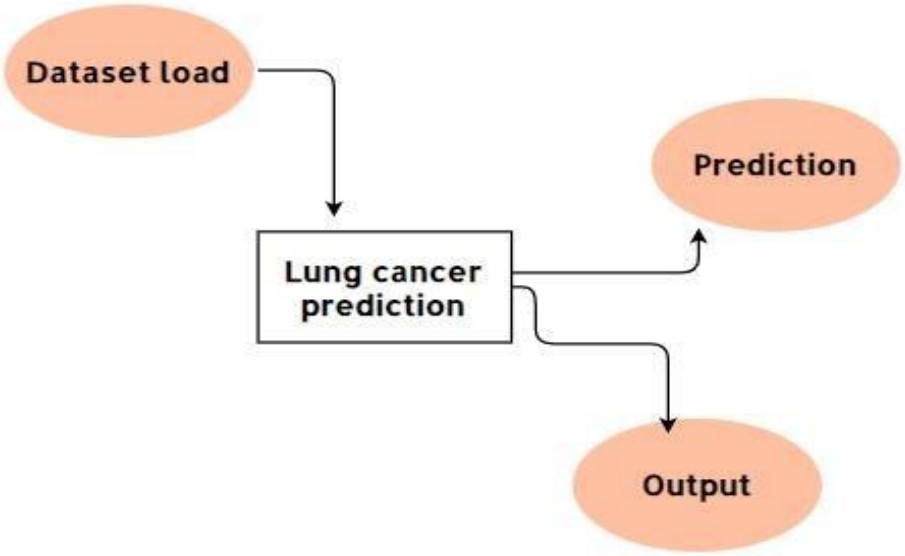

Fig 4.6 : Data Flow Diagram Level-0

Data Flow Diagram Level-1

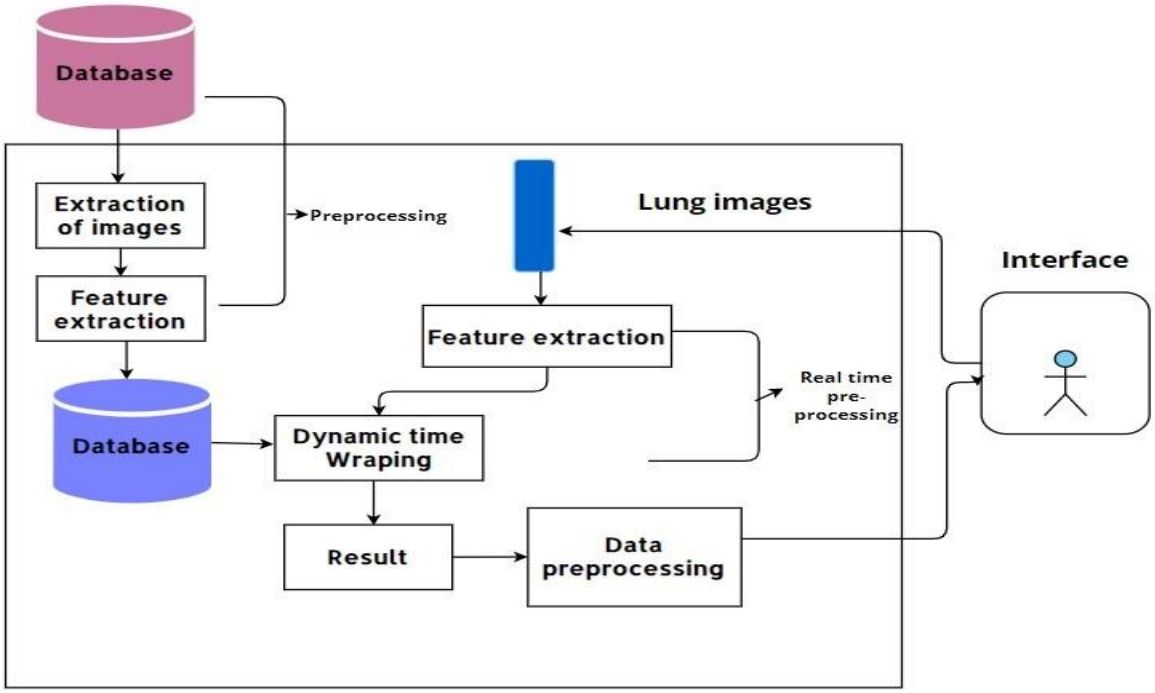

Fig 4.7: Data Flow Diagram Level-

5. CONCLUSIONS

As we all know, deep learning the training of the model relies on a large amount of data. Therefore, the two techniques used in our study are Transfer Learning and data augmentation technique. Through using dataset with fine-tuning models (Alex Net, Resnet18, Google net, and Resnet50), it could classify lung cancer data, i.e., Classification of Benign and Malignant. Our model's learning in 20 Epoch has obtained a high accuracy and very fit learning curve with suitable training time. By applying these techniques and obtaining a high performance of the model evaluation, we early confirmed lung cancer detection to protect patients from death. In the future, we would compare our model to some other models and improved the data with the proposed model.

6. REFERENCES

[1] O. Russakovsky et al., “ImageNet Large Scale Visual Recognition Challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.

[2] S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, 2010, doi: 10.1109/TKDE.2009.191.

[3] G. E. H. Alex Krizhevsky, Ilya Sutskever, “Handbook of approximation algorithms and metaheuristics,” Handb. Approx. Algorithms Metaheuristics, pp. 1–1432, 2007, doi: 10.1201/9781420010749.

[4] V. Sangeetha and K. J. R. Prasad, “Syntheses of novel derivatives of 2-acetylfuro[2,3-a]carbazoles, benzo[1,2-b]-1,4-thiazepino[2,3- a]carbazoles and 1-acetyloxycarbazole-2- carbaldehydes,” Indian J. Chem. - Sect. B Org. Med. Chem., vol. 45, no. 8, pp. 1951–1954, 2006, doi: 10.1002/chin.200650130.

[5] C. Szegedy et al., "Going deeper with convolutions," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07-12-June, pp. 1–9, 2015, doi: 10.1109/CVPR.2015.7298594.

[6] E. Cengil and A. Çinar, "A Deep Learning Based Approach to Lung Cancer Identification," 2018 Int. Conf. Artif. Intell. Data Process. IDAP 2018, 2019, doi: 10.1109/IDAP.2018.8620723.

[7] Z. Shi et al., "A deep CNN based transfer learning method for false positive reduction," Multimed. Tools Appl., vol. 78, no. 1, pp. 1017– 1033, 2019, doi: 10.1007/s11042-018-6082-6.

[8] R. Anirudh, J. J. Thiagarajan, T. Bremer, and H. Kim, "Lung nodule detection using 3D convolutional neural networks trained on weakly labeled data," Med. Imaging 2016 Comput. Diagnosis, vol. 9785, no. November 2017, p. 978532, 2016, doi: 10.1117/12.2214876.

[9] Q. Z. Song, L. Zhao, X. K. Luo, and X. C. Dou, "Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images," J. Healthc. Eng., vol. 2017, 2017, doi: 10.1155/2017/8314740.

[10] R. M. Devarapalli, H. K. Kalluri, and V. Dondeti, "Lung cancer detection of ct lung images," Int. J. Recent Technol. Eng., vol. 7, no. 5, pp. 413–416, 2019. [11] V. Makde, J. Bhavsar, S. J.