Early Detection of Lung Cancer using Deep Learning

Manoj M[1] Reeja R Rajan[2] Ramya T V[3]
[1,2,3]Assistant Professor
[1,2,3]Jawaharlal College of Engineering & Technology, India

Abstract— Lung Cancer detection at an earlier stage has become a very important and needy one for a human being. Early detection helps several patients with the best chance of recovery. The features which are used for the detection of Lung Cancer are collected from the Computed Tomography (CT scan) images. Deep Learning is an emerging technique that allows us to increase the accuracy of the result. In this paper, we have implemented cancer detection using Convolutional Neural Network (CNN). It shows promising results in terms of accuracy (94%) while comparing with other techniques.

Keywords: Lung cancer, Deep learning, Activation function, RELU, Sigmoid, CNN

I. INTRODUCTION

Lung cancer is a condition that causes cells to divide in the lungs uncontrollably [3]. This causes the growth of tumors that reduce a person's ability to breathe. Lung cancer begins in the lungs and may spread to lymph nodes or other organs in the body, such as the brain [3].

Some of the symptoms of lung cancer include a new cough that doesn't go away, coughing up blood, even a small amount, shortness of breath, chest pain, Hoarseness, losing weight without trying, Bone pain [6]. There are mainly two types of lung cancer. They are (i) Non-Small Lung Cancer, (ii) Small Cell Lung Cancer.

A. Non-Small Cell Lung Cancer (NSCLC):

About 85% to 90% of lung cancers are non-small cell. They are grouped by the kind of lung cell cancer started in and by how the cells look under a microscope [6].

B. Small Cell Lung Cancer (SCLC):

Only about 10% to 15% of people with lung cancer have small cell lung cancer. It is also called oat cell cancer. It grows and spreads more quickly than non-small cell lung cancer [6]. It often spreads to other parts of the body at an early stage. Lung cancer, like all cancers, can act differently in each person, depending on the kind of lung cancer it is and the stage it is in. But when lung cancer spreads outside the lungs, it often goes to the same places [6]. The first-place lung cancer usually spreads to is the lymph nodes in the center of the chest [14]. These lymph nodes are called mediastinal lymph nodes. Lung cancer may also spread to the lymph nodes in the lower neck. In its later stages, lung cancer may spread (metastasize) to distant parts of the body, like the liver, brain, or bones [14]. To predict these deadly diseases, deep learning technique is very useful. In this paper, we predict the cancer cells by implementing the machine learning algorithms. Section 2 depicts the literature survey related to our work and the Proposed methodology is derived in section 3. The Experiment results and discussion will be discussed in section 4 and section 5 concludes our work.

II. LITERATURE SURVEY

Earlier several methods have been proposed to detect and classify lung cancer in CT images using different algorithms.

For example, Camarlinghi et al [12], have used three different computer-aided detection techniques for identifying pulmonary nodules in CT scans. Abdulla and Shaharum

[12] used feed-forward neural networks to classify lung nodules in X-Ray images albeit with only a small number of features such as area, perimeter, and shape.

Kuruvilla et al. [3] have used six distinct parameters including skewness and fifth & sixth central moments extracted from segmented single slices containing 2 lungs along with the features mentioned in and have trained a feed-forward backpropagation.

Gunavathi proposed a methodology based on texture features using the artificial neural network (ANN), with an accuracy rate of 93.30%. Using the combination of texture and shape features for detection and classification may result in improved classification accuracy.

Hongyang Jiang et al [10]. 2016 gives the different approaches of preprocessing the lung CT scan images before providing them to CNN architecture. This resulted in better results as there are so many non-imaging regions that can reduce the accuracy of feature extraction. In 2D images, objects may overlap on each other, so that lung nodule detection may have a high false rate.

Shen et al. [6] diagnosed lung cancer on the LIDC database using a multiscale two-layer CNN and the reported accuracy was 86.84%

Based on the above literature, we need an improvement in lung cancer detection. The next section proposes our methodology to detect cancer.

III. METHODOLOGY

Our proposed work is divided into (i) Preprocessing(ii) Activation Function Selection(iii) Model Creation (iv) Prediction(v) Evaluation. Figure 1 shows the above different steps of our proposed work.

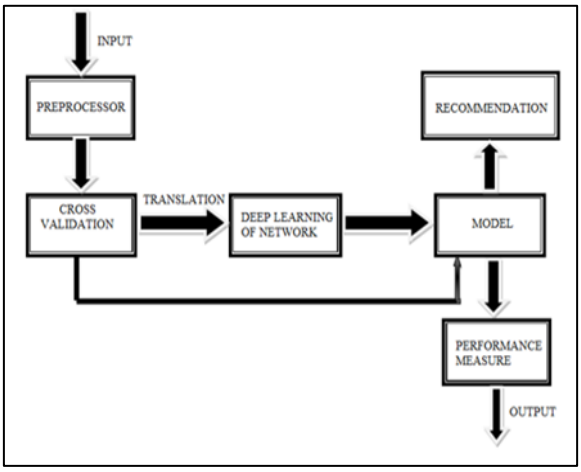


Fig. 1: Proposed System

A. *Preprocessing*

The preprocessing is a process which converts raw dataset into processed data. It includes three activities viz., data cleaning, integration and outlier analysis.

B. *Activation Function*

An activation function is a nonlinear transformation process that will do in our input data before sending it to the next layer of neurons. Figure 2 shows the familiar activation function with their equation.

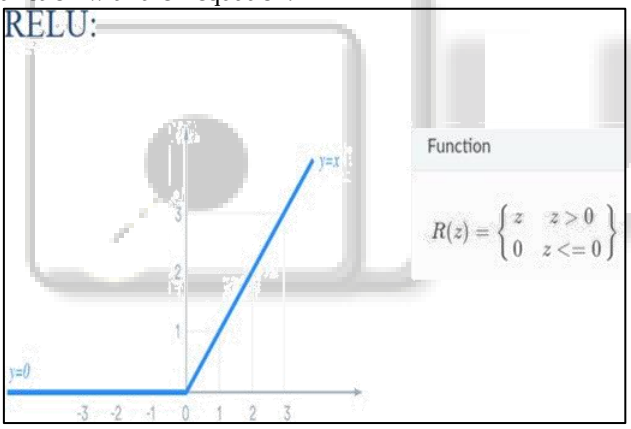


Fig. 2.1: Relu Activation

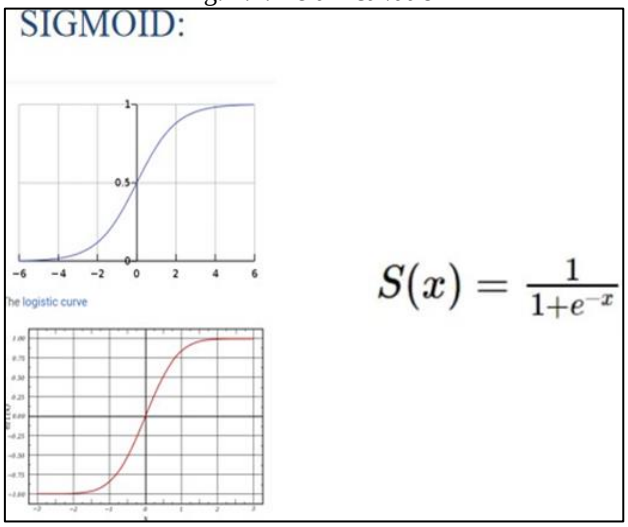


Fig. 2.2: Sigmoid Activation Function

C. *Model Creation*

The architecture diagram shows in Fig. 3.1 and Fig. 3.2 gives the exact workflow. Initially, the available dataset is feed as the input, that is done at the preprocessing stage. The given input is analyzed and cross- validated and forwarded to the deep learning network. Then the model is created and the performance measure is calculated and the output is generated.

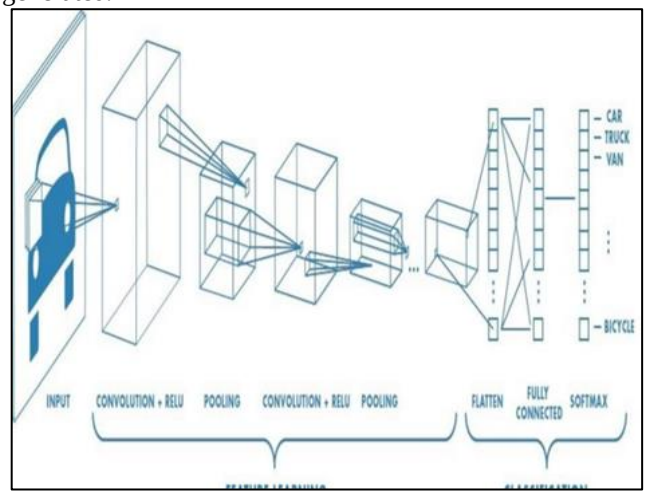


Fig. 3.1: Deep Learning Network Overview

AI is the simulation of human intelligence process by machine [12]. It is an approach to make a computer (or) robot to think like a human. It is a study of how the human brain thinks, learns, decide and works to solve problems [13]. The application of AI includes Expert system, Speech recognition, Machine vision [12]. Machine learning is an application of AI that provides system ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn themselves [13]. Artificial Neural Network (ANN) is a computational model based on the structure and functions of biological neural networks. It is considered nonlinear.

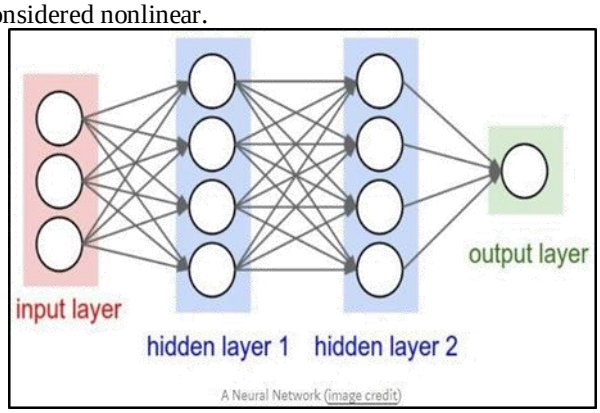


Fig. 3.2 Connectional View of ANN

ANN [12] is an interconnected group of nodes inspired by a simplification of neurons in a brain. Here the node represents an artificial neuron and an arrow represents a connection from the output of one artificial neuron to the input of another [7]. The large size of data or Big data Deep learning is an AI function that imitates the workings of the human brain in processing data and creating patterns for use in decision making [9]. Deep learning is a subset of machine

learning in AI. That has networks capable of learning unsupervised from data that is unstructured or unlabeled [8]. Deep learning learns from a vast amount of unstructured data that could normally take human decades to understand and process [7].

A Convolutional Neural Network (CNN) is a Deep Learning algorithm that can take in an input image, assign importance (weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. It mainly increases the number of hidden layers in the network which helps in producing the output.

The processing required on CNN is much lower as compared to other classification algorithms [6]. While in primitive methods filters are hand- engineered, with enough training, CNN can learn these filters/characteristics [6]. some of the tests used for Diagnosis of Lung Cancer are [6]

1) Imaging tests: An X-ray image of your lungs may reveal an abnormal mass or nodule. A CT scan can reveal small lesions in your lungs that might not be detected on an X-ray.
2) Sputum cytology: If you have a cough and are producing sputum, looking at the sputum under the microscope can sometimes reveal the presence of lung cancer cells.
3) A tissue sample (biopsy): A sample of abnormal cells may be removed in a procedure called a biopsy [6].

Convolutional Neural Network (CNN) indicates that the network employs a mathematical operation called convolution.

Convolution is a specialized kind of linear operation [5]. Convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers [2]. The architecture of a ConvNet is similar to that of the connectivity pattern of Neurons in the Human Brain and was inspired by the organization of the Visual Cortex [3]. Individual neurons respond to stimuli only in a restricted region of the visual field known as the Receptive Field [11]. A collection of such fields overlaps to cover the entire visual area [12].

Convolutional Neural Networks take advantage of the fact that the input consists of images and they constrain the architecture more sensibly [5]. In particular, unlike a regular Neural Network, the layers of a ConvNet have neurons arranged in 3 dimensions that are width, height, depth. The word depth here refers to the third dimension of an activation volume, not to the depth of a full Neural Network, which can refer to the total number of layers in a network [3]. For example, the input images in CIFAR-10 are an input volume of activations, and the volume has dimensions 32x32x3 (width, height, depth respectively) [2]. The neurons in a layer will only be connected to a small region of the layer before it, instead of all of the neurons in a fully connected manner [11].

A ConvNet can successfully capture the Spatial and Temporal dependencies in an image through the application of relevant filters [12]. The architecture performs a better fitting to the features of the image dataset due to the reduction in the number of parameters involved and the reusability of weights [5]. In other words, the network can be trained to understand the sophistication of the image better [3]. Label Encoder has been used here to convert the data available in the string to integer [2]. Activation functions used here are Rectified Linear Unit (RELU)which is used for feeding the input since it considers the negative input value as zero and if positive input is given it takes the same value which was provided.

Then the sigmoid activation function is used here for the output whatever the input is given maybe it takes the input in the range 0 and 1and thus provides the output either 0 or 1[11]. Thus, it is very helpful to detect whether the patient is affected by lung cancer or not [12].

D. Prediction

The prediction is the main module in our proposed algorithm. The dependent variable is calculated by using the trained model.

E. Evaluation

There are different measures to evaluate our proposed model. They are (i) Mean Absolute Error (MAE), (ii). Root Mean Squared Error, (iii) Mean Squared error. Among them, in our work, we have used Mean Absolute Error (MAE). It is the difference between the actual value and the predicted value. It is shown in equation 1.

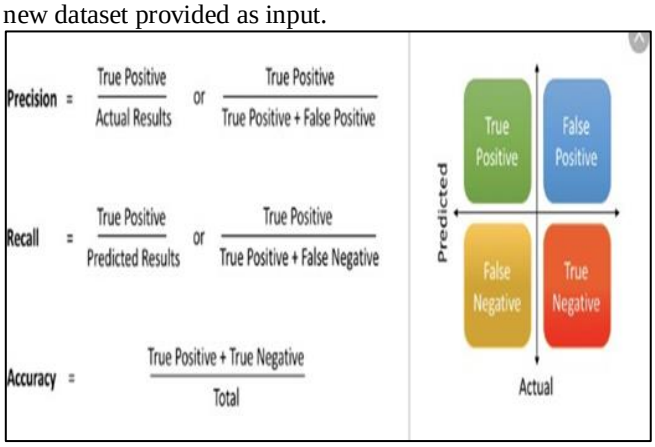
$$MAE = \frac{1}{N}\sum_{u,i}\left|p_{u,i} - r_{u,i}\right| \qquad (1)$$

IV. IMPLEMENTATION

The lung cancer dataset consists of 16 columns. The first 15 columns are taken as the input and the value in the 16th column is obtained by analyzing the values in the 15 input features.

In the Preprocessing step input dataset consisting of several features are analyzed. For processing the input dataset RELU (Rectified Linear Unit) activation function has been used since it ranges the input value in between 0 and 1 for whatever input value provided as the input. Sigmoid activation has been chosen for displaying the output. Then the model is created using Keras. Fit and transform functions have been used for training the input dataset.

Now the model can predict the output based on the input features provided. Thus, the evaluation is made for the new dataset provided as input.



The Precision, Recall and F1score values are calculated for the given dataset and the results are obtained with 90% accuracy.

Precision means the percentage of your results which are relevant. On the other hand, recall refers to the percentage of total relevant results correctly classified by the algorithm. Precision and recall both indicate accuracy of the model.

The below histogram shows the analysis of each and every feature for the dataset provided. The convolution neural network obtains the good result mainly because the convolution layer operation may obtain the characteristic from the shape and the texture of two different dimensions existing. The Histogram is displaying the analysis of the input features that are shown in the graphical representation. This helps in easy prediction of results. The input features include smoking, alcohol consumption, coughing, yellow fingers, etc. Two graphs are generated, one for comparing the accuracy value with the epoch and the other for comparing loss value with the epoch.

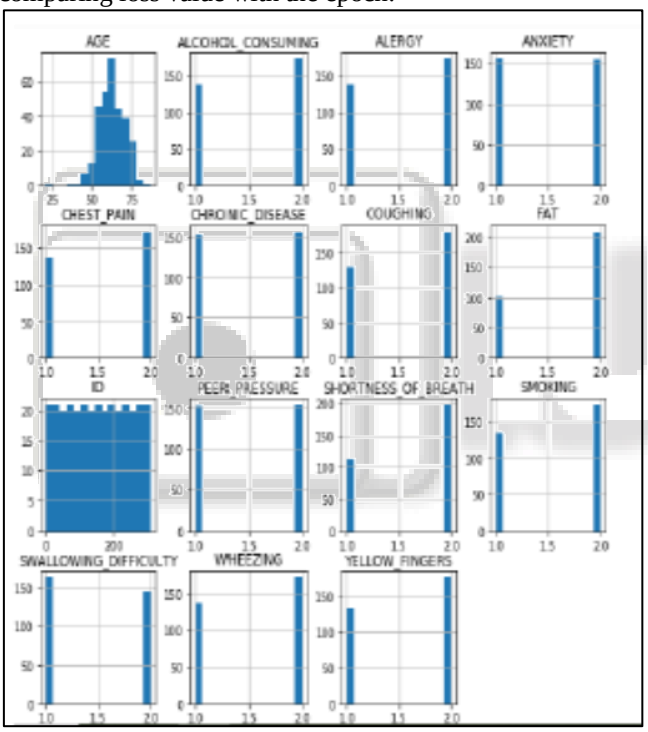


Fig. 4: Histogram of input features

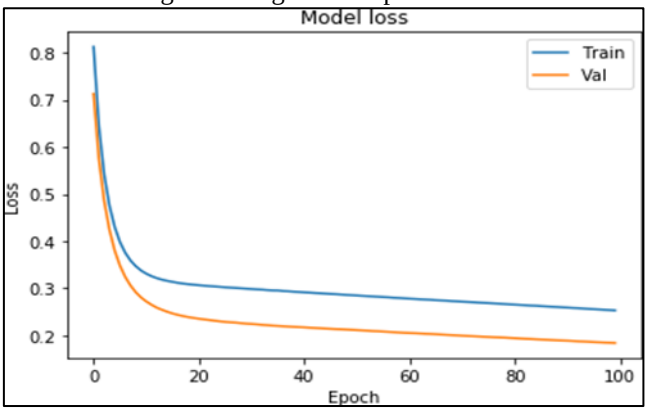


Fig. 5.1: Loss Vs Epoch

The above figure displaying the loss value with respect to the epoch. The blue line in the figure indicates the trained data. The orange line in the figure displaying the loss value obtained.

The below figure displaying the accuracy value with respect to the epoch. The blue line in the figure indicates the trained data. The orange line in the figure displaying the accuracy value obtained.

In different convolution kernels according to different weights for different image characteristics, a convolution kernel shared parameter in the whole process of convolution. Due to the introduction of more hidden layers in the convolutional neural network, the accuracy of the result is increased above 90%. Thus, it helps in predicting Lung cancer at an earlier stage and so doctors can give treatments at an earlier stage to make the patient recover.

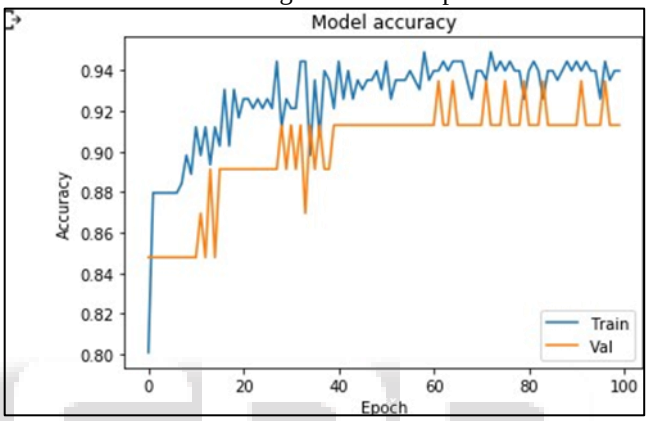


Fig. 5.2: Accuracy Vs Epoch

The above graphs are displaying model loss and model accuracy. Model loss and accuracy have been calculated by analyzing the available data values. Some steps need to be followed to prevent lung cancer are to avoid smoking, to avoid secondhand smoke, test your home for radon, avoid carcinogens at work, eat a diet full of fruits and vegetables, exercise most days of the week [3]. Thus, by analyzing the features the detection is done.

V. Conclusion

The deep learning technique is a familiar artificial intelligence technique used in many domains. We have implemented the deep learning technique for our lung cancer dataset. The proposed model is trained using the trained dataset and the unknown values are predicted. The model performance is measured and shows better results (94%) when comparing with the existing results.

References

[1] E. Cengil, A. Çinar, and Z. Güler. "A GPU-based convolutional neural network approach for image classification." Artificial Intelligence and Data Processing Symposium (IDAP), 2017 International. IEEE, 2017.

[2] P. Ai, D. Wang, G. Huang, and X. Sun, "Three-dimensional convolutional neural networks for neutrinoless double-beta decay signal/background discrimination in high-pressure.

[3] J. Kuruvilla and K. Gunavathi, "Lung cancer classification using neural networks for CT images," Computer Methods and Programs in Biomedicine, vol. 113, no. 1, pp. 202–209,2014.
[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "I magnet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, pp. 1097–1105, 2012.
[5] Convolutional Neural Network http://cs231n.github.io/convolutionalnetworks/
[6] Key Statistics for Lung Cancer, https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/key-statistics.html.
[7] https://medium.com/analyticsvidhya/understanding-basics-of-deep-learning-by- solving-xor-problem-cb3ff6a18a06
[8] https://colab.research.google.com/drive/1QIivAJbrrcV7ZqVpf7KsDzVhLvkpwr_F
[9] https://www.google.com/search?q=kaggle+datasetoq=kagg&aqs=chrome.1.0l2j69i57j0l5.2754j0j7&sourceid=chrome&ie=UTF-8
[10] Hongyang Jiang, He Ma, Wei Qian, Mengdi Gao, and Yan Li, "An Automatic Detection System of Lung Nodule Based on Multi-Group Patch-based Deep Learning Network", IEEE Journal of Biomedical
[11] Roth, H.R., Lu, L., Seffff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., E., Summers, R.M.: A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part I. LNCS, vol. 8673, pp. 520–527. Springer, Heidelberg (2014)
[12] Suzuki, K., Li, F., Sone, S., Doi, K.: Computer- aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose ct by use of massive training artificial neural network. IEEE Trans. Med. Imaging 24(9),1138–1150 (2005)
[13] van der Walt, S., Sch¨onberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T.: Scikit-image: Image processing in python. Technical report, PeerJ PrePrints (2014)
[14] Way, T.W., Hadjiiski, L.M., Sahiner, B., Chan, H.P., Cascade, P.N., Kazerooni, E.A., Bogot, N., Zhou, C.: Computer-aided Diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours. Med. Phys. 33(7), 2323–2337 (2006).

