Lung Cancer Classification and Detection Using Deep Learning Technique

Dr.K.Sudha Rani [1†], Dr.A.Suma Latha [2††], Dr.S.Sunitha Ratnam [3†††], Dr.J.Bhavani [4††††], Dr.J.Srinivasa Rao[5†††††] N.Kavitha Rao[6†††††]

Sudharani_k@vnrvjiet.in, sumaakunuri@vrsiddhartha.ac.in, sunita.ratnam.s@gmail.com, bhavani_j@vnrvjiet.in, srinivasarao_j@vnrvjiet.in, nkavitha_ece@mvsrec.edu.in

[1]Assoc. Prof. EIE Dept. VNRVJIET Hyderabad Telangana, India
[2]Assit. Prof. EIE Dept. VRSEC Vijayawada Andhra Pradesh, India
[3]Assoc. Prof. Dept.of Physics MRECW Hyderabad Telangana, India
[4]Assoc. Prof. EEE Dept. VNRVJIET Hyderabad Telangana, India
[5]Assoc. Prof. EEE Dept. VNRVJIET Hyderabad Telangana, India
[6]Assist.Prof. ECE Dept. MVSREC Hyderabad Telangana India

Abstract
Lung cancer is a complex and frightening disease that typically results in death in both men and women. Therefore, it is more crucial to thoroughly and swiftly evaluate the malignant nodules. Recent years have seen the development of numerous strategies for diagnosing lung cancer, most of which use CT imaging. These techniques include supervisory and non-supervisory procedures. This study revealed that computed tomography scans are more suitable for obtaining reliable results. Lung cancer cannot be accurately predicted using unsupervised approaches. As a result, supervisory techniques are crucial in lung cancer prediction. Convolutional neural networks (CNNs) based on deep learning techniques has been used in this paper. Convolutional neural networks (CNN)-based deep learning procedures have produced results that are more precise than those produced by traditional machine learning procedures. A number of statistical measures, including accuracy, precision, and f1, have been computed.

Keywords:
convolutional neural networks (CNN), CT Scans, Deep learning, Lung Cancer, Prediction.

1. Introduction

Since the early 1800s, lung cancer has been known to frequently affect people who have had prolonged exposure to tobacco smoke and other toxic chemicals. Today, anyone with access to anything from the most basic medical diagnostic tools to the most sophisticated systems may identify cancerous cells and lung tumors in a variety of ways. Several medical imaging procedures, including chest X-rays, CT (Computed tomography), biopsies, and magnetic resonance imaging, can detect lung tumors (MRI). However, the patient survival rate has been declining daily.

In the present times, the evolution of technology in the domain of medical diagnosis has madeit so easy to detect various malignant diseases and at the same time many advanced treatments improving day by day are benefiting people enormously.

Cancer is a potentially fatal illness that involves abnormal cell proliferation and has the potential to invade or spread to many bodily parts. When the body's natural regulating mechanisms fail, aged cells continue to expand uncontrollably and give rise to aberrant new cells instead of dying. A tumor is a mass of tissue made up of these extra aberrant cells. According to Guobin Zhang et al. [3], P. Mohamed Shakeel et al. [4], lung cancer develops when cells in the body's respiratory organs divide uncontrolled. This enables tumor growth. Eventually lowering a person's capacity for breathing and spreading to other body components.

. Lung cancer, as described by Suren Makajua et al. [1], is discovered using a variety of imaging methods, such as Computer Tomography (CT), chest X-ray, and resonance imaging (MRI). Patients with malignant neoplastic disease have a lower survival rate on a daily basis. According to the ICMR Report, a recent study with a survival rate of 19% people diagnosed with cancer survive for five years or more [10].

Major issue behind this downfall of survival rate is late identification of malignant tumors among the internal organ that grow whereas not any management damaging the lungs very badly. Considering this situation, there' a pressing necessity to develop a durable system which can diagnose malignant neoplastic disease accurately. In

Manuscript received October 5, 2024
Manuscript revised October 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.10.9

recent decades, Deep Learninghas been tested to be a really powerful domain for its ability to handleimmense amounts of data Kazuhiro Suzuki et al. [2]. The usage of hidden layers has surpassed several ancient techniques, considerably in pattern recognition.

R. Linder et al. [5] and S. Kalaivani Pramit et al. [18]. Ying-HweyNai et al. [6] have been discussed about CNN technique, which is very potent technology in recent years to tests on how well it handles massive volumes of data. Many outdated techniques have been overtaken by the use of hidden layers, particularly in pattern recognition.

The identification of malignant elements among the lungs has only been up to par, despite the existence of a variety of techniques, such as those used by Murat Karahatak et al.[9] for the detection of carcinoma through CT scans as shown in Fig. 1. Carcinoma accounts for 25% of all cancer fatalities and is the second most frequent disease in people. The causes are restricted access and wide value ranges. Therefore, we created the proposed system while considering basic accessibility and affordability difficulties as well as the advancement of technology. Convolutional neural networks have claimed a significant position in pharmacological science like tumor segmentation due to the rise in medical image analysis.

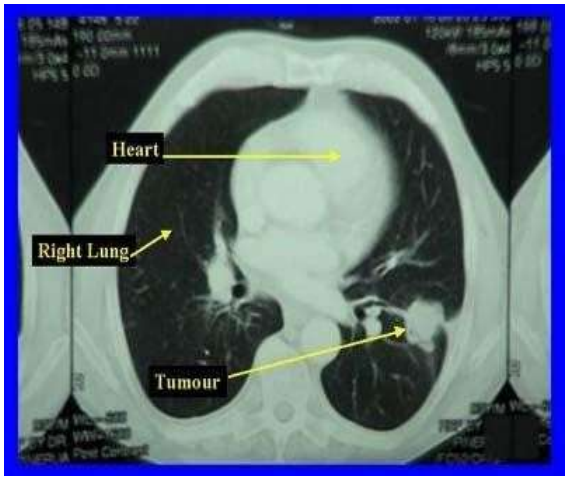


Fig.1 CT scan image depicting the view of tumor inside the left lung of a patient

There are 2 major varieties of lung cancer and that classified as

2.1.1 Non-small cell lung cancer (NSCLC)

NSCLC accounts for 80% to 85% of cases of respiratory organ cancer. Adenocarcinoma, big cell cancer, and epithelial cell carcinoma are the three primary subgroups of NSCLC. In terms of viewpoint and therapy, these subtypes are comparable. So, non-small cell lung cancer and other diseases were then resorted, P.Shakeel Mohamed et al.[4].

Adenocarcinoma: The cells that typically release materials that resemble mucus are where adenocarcinomas start. This type of lung cancer is most common among current or former smokers, but it is also the most prevalent type of lung cancer in non-smokers. It is more common in women than in males, and it is likely to target younger demographics. Adenocarcinoma is discovered in the respiratory organ's outer components and may be identified before it spreads.

(i) Squamous cell carcinoma: Epithelial cell carcinomas begin in flat cells referred to as squamous cells, these are primarily present in the internal lining of lung airways. They're originate in the central a part of the lungs, close to a main airway called bronchus. This disorder is especially according in maturity smokers.

(ii) Large cell (undifferentiated) carcinoma: It may appear anywhere in the lung. It can grow and develop very quickly, making treatment difficult. Large cell system carcinoma is a subtype of gigantic cell cancer that has a potentially lethal growth rate.

(iii) Different subcategories: Aden squamous carcinoma and sarcomatous carcinoma are other less ordinarily occurring subtypes.

2.1.2 Small cell lung cancer (SCLC)

The SCLC, also known as oat cell carcinoma, accounts for 10% to 15% of all respiratory organ malignancies. Unlike NSCLC, this type of lung cancer develops and spreads rapidly. According to calculations, about 70% of people with SCLC as depicted in Fig. 2 are estimated to have cancer that has spread at the time of diagnosis.

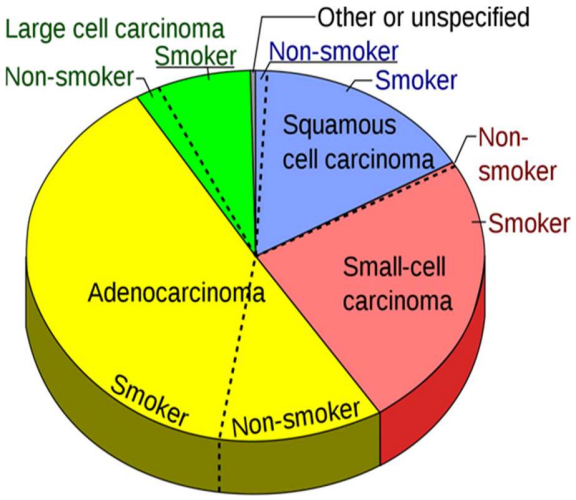


Fig.2 Small cell lung cancer

This cancer responds effectively to treatment and radiation therapy because it spreads quickly Sang Min Park et al. [7]. Along with the above-mentioned carcinoma types, a few completely different kinds of tumors also arise, including:

(i) Lung tumor: Tumors grow slowly Yongqian Qiang et al. [8], and occur fewer than five percent.

(ii) Other respiratory organ tumors: Different varieties of lung cancer resembling sarcomas, adenoid, cancer and cuckoo cystic carcinomas occur terribly rarely.
Sometimes, cancers beginning in other organs may also unfold to lungs. Cancerous parts at skin, kidneys can spread to lungs.

According to a recent study, the survival rate is only 19%, meaning that only nineteen out of every 100 people who are diagnosed with cancer will survive for five years or more (American cancer study et al. Harleen Kaur et al.[13] and Vijaya et al. [12].

The main causes of this failure are the delayed diagnosis and the failure of the medical staff to recognize the existence of tumors in the reports. Given this circumstance, it is urgently necessary to create a long-lasting system that can accurately diagnose malignant carcinoma disease. Deep Learning has been demonstrated in recent decades to be a particularly potent technique due to its capacity to manage enormous volumes of data. The use of hidden layers has significantly outperformed a number of outdated pattern recognition approaches.

An approach to diagnose lung cancer using machine learning classification and other techniques has been examined in a comparative study on the subject. According to J.R. Quinlan et al.and L. Breiman et al., Osman et al. [14, 15,17], SVM, also known as support vector machine, K-Nearest Neighbor, Nearest Neighbors(NN), and logistic regression, may categorize medical data sets with a higher degree of accuracy with the drawback of requiring more preprocessing for quicker results. However, Convolutional neural network is the most popular deep learning network that has an improbable ability of rough imitation of human vision. So exploiting CNNand developing a system which could detect carcinoma disease is going to be advantageous inovercoming various disadvantages currently.

Another method developed by S. Kalaivani Pramit et al.[18] employed image processing and a backpropagation artificial neural network (ANN) training technique to determine whether an irregularity was malignant. SVM is used to classify X-ray images, after which the M3 filter is suggested for preprocessing and connected component labelling (CCL) is used to assess the region of interest. Statistical moments are then used to extract the features, which are then used to classify the images.

This paper serves various objectives including Enablingimmediate identification of lung cancer and then benefiting the patient involved treatment saving his/her life, low-cost affordability and larger accessibility is another added advantage for the planned method and robust and properly trained model implemented on python software makes the Radiologist's job easier to identify the cancerous nodule once invisible or unidentified because of its size and place present.

As explained in another method titled "Lung cancer detection by employing artificial neural networks and fuzzy clustering approaches," the early identification of lung cancer may require a complex strategy due to the morphology of the cancer cells, the majority of which are stacked on top of one another. There is a method presented for segmenting images colored by phlegm to detect lung cancer in its early stages using segmentation techniques like Hopfield Neural Network (HNN) and a Fuzzy C-Mean

(FCM) bunch algorithm. The manual analysis of the phlegm samples typically requires more time than other methods and highly skilled personnel for results that are error-free.

2. Methodology:

The identification process of lung cancer disease is depicted in the flow diagram in Fig. 3. The cancer region from the CT scan images is identified and provided to the pre-processing procedure.When the data set of CT scan images is used as an input and applied to the segmentation process. This method involves resizing all of the images, array conversion, and feature selection processing. The resulting dataset will be divided into a training dataset and a testing dataset for use in the selection process. Additionally, the classification algorithm is used to forecast lung cancer as described by Osmar et al. [11], R. D'az-Uriarte et al. [16],and provides results based on accuracy, classification report, and confusion matrix at the conclusion .

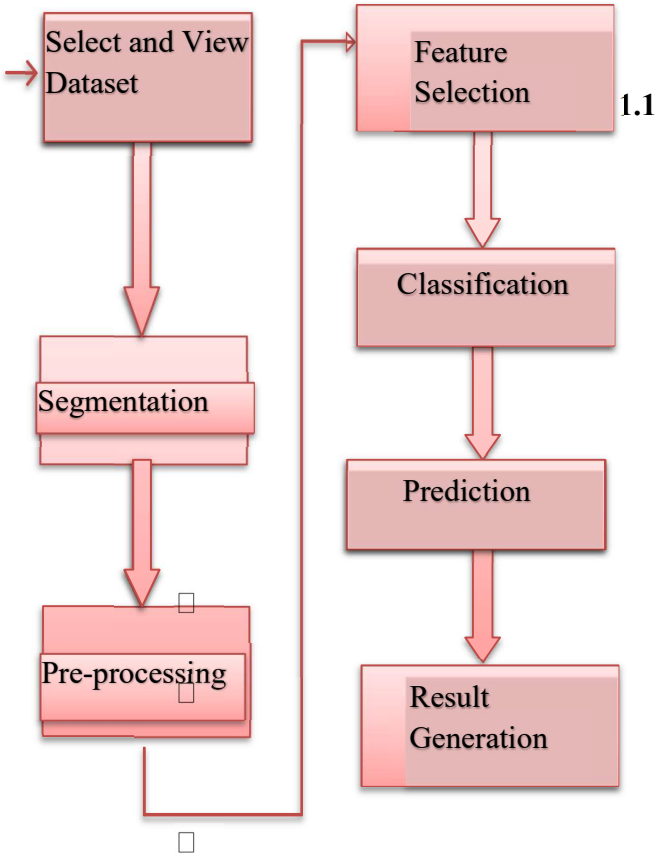


Fig.3. Flow Chart

Despite the presence of assorted techniques for detection of carcinoma through CT scans and completely different advanced equipment, the identification of cancerous components amongthe lungs has been merely up to mark. Lesser access and higher value ranges are the reasons behind it. So, keeping the evolved technology and thought of basic accessibility, affordabilityissues in hand, we have developed the projected system. Owing to the increase in medical image analysis, the convolutional neural networks has contended a heavy role in the medical science resembling tumor segmentation, Histopathological cancer classification, Cerebral micro bleeds, Anatomy-specific classification of medical images and plenty of more. Similarly, CNN goes to play a heavy role in the projected research by serving a wonderful purpose of detection of respiratory organ cancer following the below steps.

2.1 Data Selection and Loading

The data selection can be defined as the process of selecting the data for cancer prediction. 524 chest CT scan images from LUNA16, a large-scale study of automatic nodule detection methods for chest CT, have been gathered as a data set for this project. The images in this dataset are both without cancer and cancer.

2.2 Image Segmentation

Segmentation is the process of differentiating and segmenting data into different groups. Thisis the method of separating a digital image into various segments (pixels). The aim of segmentation is properly modifying the appearance of CT scan image into a significant and simpler of something to analyze. Image segmentation is actually considered to detect borders and objects in images with the same label share certain features. Under various segmentation techniques available, we have considered binary Threshold segmentation as shown in fig.4.

Threshold binary: If intensity of pixel is higher than the set threshold, the value is set to 255, else set to 0 (which is black).

Threshold Binary, Inverted: It is quite opposite implementation of Threshold binary

Truncate: If the intensity of available pixel value is larger than threshold, it is truncated down to the value of threshold. The pixel values are set to be the same as the threshold. And all other values remain the same.

Threshold to Zero: Pixel intensity is made 0, for all those pixel intensities which are smaller than the threshold value.

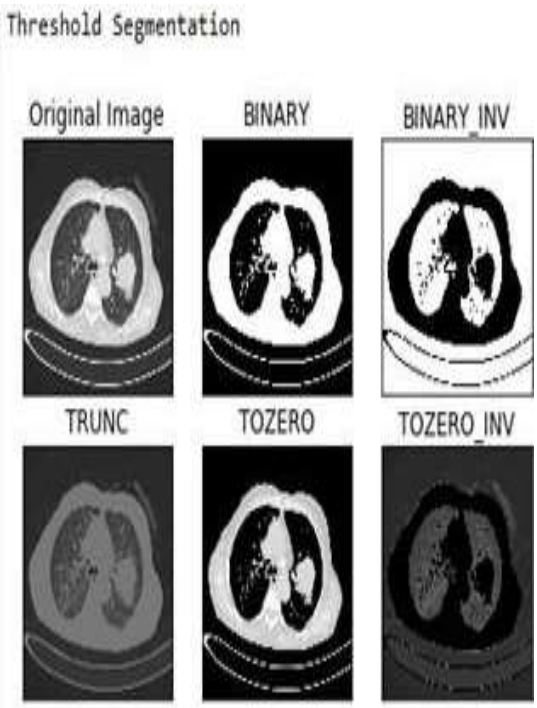


Fig.4 Threshold segmentation on the original image of lung CT scan image

2.2 Data Preprocessing

Image data pre-processing is resizing of the data. Resize image dataset: Rescaling the grey scale chest CT scan image size into 196 is done and then categorical data is defined as variableswith a finite set of rescaled values. Most deep learning algorithms require array input and output variables. In data preprocessing, image enhancement like brightness adjustment, contract adjustment haven been applied, as shown in Fig.3 (a) and (b).

Fig3: (a) Original image (b) enhanced image

Splitting data set

Data splitting is a simple act of separating current image dataset into two parts, usually for cross-validation purposes. A predictive model is developed using a component of the picture data, and the performance of the model is assessed using a second portion. A greater portion of the data is used for training and a smaller portion is used for testing when we divide an image data set into training sets and testing sets. We used 30% of the data as a testing set and 70% of the data as a training set. From a mathematical perspective, we used 197 photos for testing and 3 2 7 images for training.

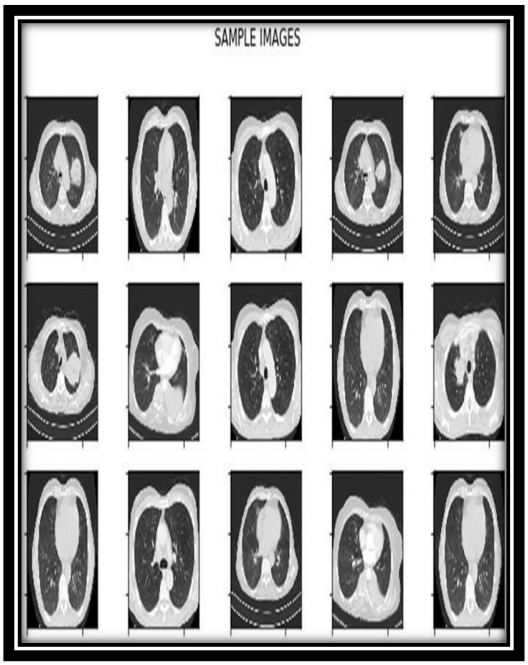


Fig4: Sample images from Data set

From the data set of CT scan images as shown in **Fig.4** seventy percent of images are used into training set and thirty percent of images are utilized as testing set.

2.4 Classification

This is the most important phase of the whole process, convolutional neural networks have been used in this research. An image can be given into a convolutional neural network, which is a deep

learning technique, and different parts of the image can be given weights and biases that can be learned to help discriminate between them. Additionally, this technique needs far less pre-processing than earlier categorization algorithms. CNN's goal is to streamline the visuals without losing important details that are essential for making precise predictions (32). The three distinct working layers that comprise a typical CNN include the convolutional layer (CONV), the pooling layer (POOL), and finally the classifier layer (FC).

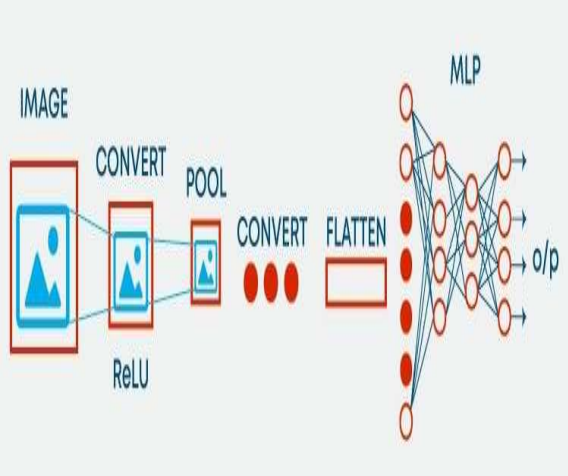


Fig. 5 CNN Architecture

(i) Input flow will be broken into various layers
(ii) The features inside these layers are used to construct gradients and edges.
(iii) Textures and patterns are created from the above step.
(iv) Objects are built from textures and patterns. These objects are utilized to build other objects.

The below **Fig.6** shows a pictorial representation of classification undergoing by CTscan images.

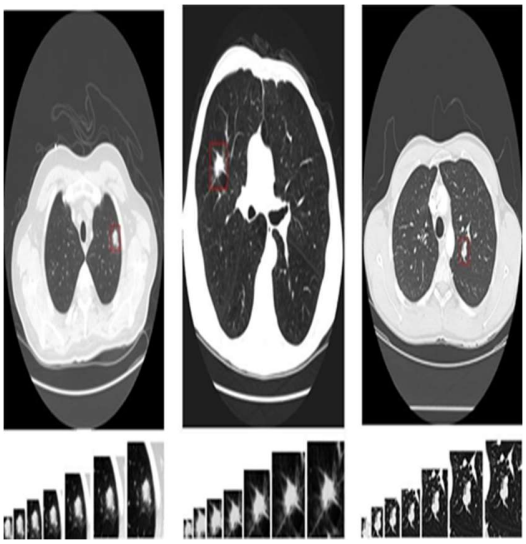


Fig.6: Picture depicting classification of CT scan images

Accuracy: It is the most inherent performance detection and it's simply a magnitude relation of properly expected observation to the overall observations. Accuracy is ratio between sum of TP and TN to all the remaining cases.

True Positive (TP) measures whether the model precisely forecasts the positive class

True Negative (TN) measures how well the model precisely forecasts the negative class.

False Positive (FP) measures how much the model gives the improper forecast of the negative class.

False Negative (FN) measures how the model mistakenly predicts the positive class**Precision:** Among the positive predictions, what percentage is really positive.

$$Precision = \frac{TP}{TP + FP}$$

Recall: It is the measure to predict the percentage of total positive samples

$$Recall = \frac{TP}{TP + FN}$$

F measure: It is the precision and recall harmonic mean.

$$F1\ score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

Confusion Matrix: It is a matrix of size 2×2 for binary classification with actual values on one axis and predicted on another. It is a matrix of size 2×2 for binary classification with actual values on one axis and predicted on another. The results of confusion matrix shows the true samples and false samples are correctly identified by the model and very few samples are predicted wrong. Fig.7 and 8 depict the graphs measuring accuracy and losses while training and testing.

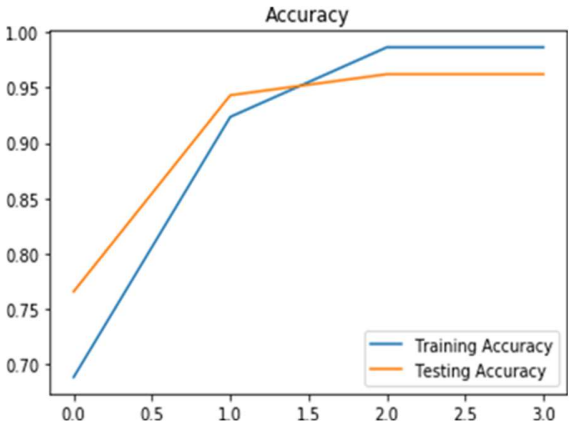


Fig .7: Accuracy graph for model's ability in training and testing

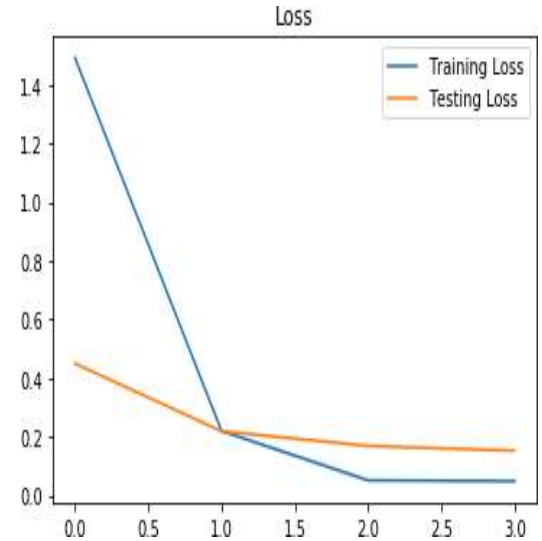
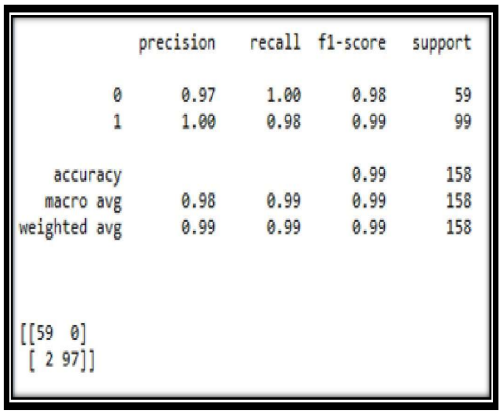


Fig.8: loss graph for model's ability in training and testing

Fig.9 (a) & (b) shows the performance metrics obtained. All the metrics have achievedconsiderably good score for the model implemented.



	precision	recall	f1-score	support
0	0.97	1.00	0.98	59
1	1.00	0.98	0.99	99
accuracy			0.99	158
macro avg	0.98	0.99	0.99	158
weighted avg	0.99	0.99	0.99	158

```
[[59  0]
 [ 2 97]]
```

Fig.9 (a): Performance metrics obtained while experimenting results.

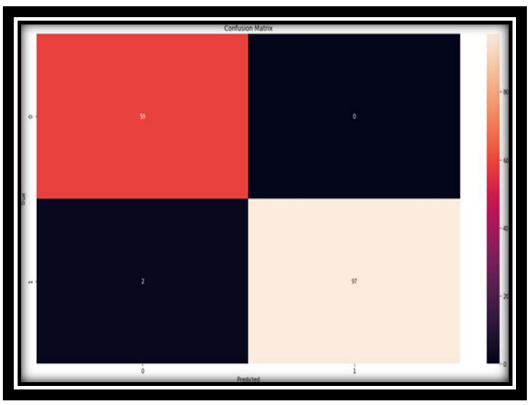


Fig.9 (b): Confusion matrix

In the final prediction as shown in **Fig10,** images are selected to find whether the selected CTscans normal or has cancer. From the **Fig.11**, the model has predicted that the image has no presence of cancer.**Fig.12** shows another output image which has been predicted as cancerousone.

Fig10: Final prediction

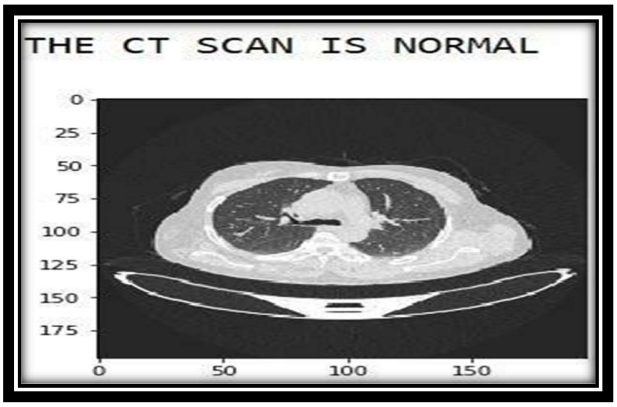


Fig.11: output image without cancer

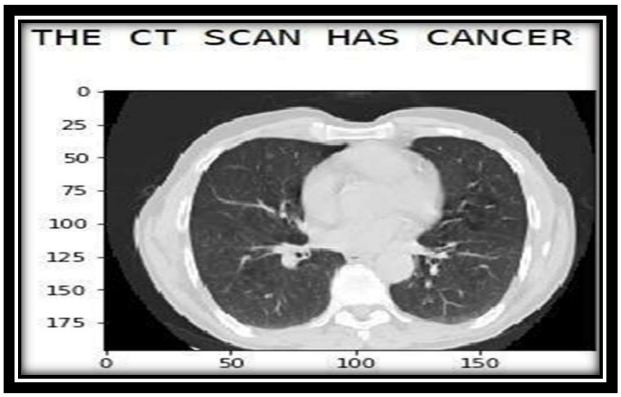


Fig12: Output image with cancer

References

[1] Suren Makajua , P.W.C. Prasad, AbeerAlsadoona , A. K. Singhb , A. Elchouemic, "LungCancer Detection using CT Scan Images", 6th International Conference on Smart Computing and Communications, ICSCC 2017, Kurukshetra, India, 7-8 December 2017.

[2] Kazuhiro Suzuki, MD, PhD, Yujiro Otsuka, BSc, Yukihiro Nomura, RT, PhD, Kanako K. Kumamaru, MD, PhD, Ryohei Kuwatsuru, MD, PhD, Shigeki Aoki, MD, PhD, "Development and Validation of a Modified Three-Dimensional U-Net Deep-Learning Model for Automated Detection of Lung Nodules on Chest CT Images From the Lung Image Database Consortium and Japanese Datasets" Volume 29, pp.S11-S17, February 2022.

[3] Guobin Zhang, Shan Jiang , Zhiyong Yang, Li Gong, Xiaodong Ma, Zeyang Zhou, Chao Bao, Qi Liu "Automatic nodule detection for lung cancer in CT images: A review" Computers in Biology and Medicine , volume 103 , pp.287–300, 2018.

[4] P. Mohamed Shakeel , M.A. Burhan Uddin , Mohamad Ishak Desa "Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks" Volume 145, pp. 702-712, October 2019.

[5] R.Linder, T. Richards, and M. Wagner, "Microarray data classified by artificial neural Networks", Methods in Molecular Biology, Volume, pp.382, 345-72, 2007.

[6] Ying-Hwey Nai , Josh Schaefferkoetter , Daniel Fakhry-Darian a , Sophie O'Doherty , John J. Totman , Maurizio Conti , David W. Townsend , Arvind K. Sinha , Teng-Hwee Tan, Ivan Tham , Daniel C. Alexander ,Anthonin Reilha "Validation of low-dose lung cancer PET-CT protocol and PET image improvement using machine learning" Physica Medica, Volume 81, pp. 285–294, 2021.

[7] Sang Min Park, Min Kyung Lim, Soon Ae Shin,Young Ho Yun, "Impact of pre diagnosis smoking, Alcohol, Obesity and Insulin resistance on survival in Male cancer Patients: National Health Insurance corporation study" Journal of clinical Oncology, Volume 24, Issue 31,pp. 5017-24, November 2006.

[8] Yongqian Qiang, YouminGuo, Xue Li, Qiuping Wang, Hao Chen, Duwu Cuic, "The Diagnostic Rules of Peripheral Lung cancer Preliminary study based on Data Mining Technique", Journal of Nanjing Medical University, volume 21,issue 3,pp.190-195, 2007.

[9] Murat Karabhatak , M.CevdetInce, "Expert system for detection of breast cancer based on association rules and neural network" Expert systems with Applications, Volume 36, Issue 2, pp. 3465-3469,March 2009.

[10] ICMR Report 2006. Cancer Research in ICMR Achievements in Nineties, 2006.

[11] Osmar R.Zaïane, " Principles of Knowledge Discovery in Databases", Available:webdocs.cs.ualberta.ca/~zaiane/courses/cmput690//notes/Chapter1/ch1.pdf, 1999.

[12] Vijaya. Gajdhane Prof. Deshpande, "Detection of Lung Cancer Stages on CT scanImages by Using Various Image Processing Techniques ",IOSR Journal of Computer Engineering, Volume 16, Issue 5, pp. 28-35, September 2014.

[13] Harleen Kaur and Siri Krishan Wasan, "Empirical Study on Applications of Data MiningTechniques in Healthcare", Journal of Computer Science, volume 2, Issue 2, pp. 194-200, 2006.

[14] J.R. Quinlan, "Induction of decision trees. Machine learning", Volume 1, issue1, pp.81–106, 1986.

[15] L. Breiman, "Random forests", Machine learning, Volume 45, Issue 1,pp.5–32, 2001.

[16] R. D´ıaz-Uriarte, A. de André's, "Gene selection and classification of microarray datausing random forest" BMC bioinformatics, volume 7, issue 1, 2006.

[17] R.S. Michal ski and K. Kaufman, “Learning patterns in noisy data: The AQ approach.Low dose Learning and its Applications”, Springer Verilog, pages 22–38, 2001.
[18] S.Kalaivani, Pramit Chatterjee, Shikhar Juyal, Rishi Gupta , “Detection Using Digital Image Processing and Artificial Neural Networks”, International Conference on Electronics, Communication and Aerospace Technology ICECA, Coimbatore, India, April 2017.

Dr.K.Sudha Rani, received B.Tech from Bhoj Reddy Engineering Colleg for Womene, JNTUH in 2001, and M.Tech from Andhra University in 2006 . Awarded Ph.D. from JNTU Kakinada, in 2018.

Now working as Asoociate Professor in VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad. She has 20 years of teaching experince. She has published around 28 publications in reputed International Journals/book chapters. She published one patent and one book. She is a member of ISOI,IAENG.

Dr. Sumalatha Akunuri**,** received B.Tech from Andhra University, in 2003. M.tech from Andhra University, Visakhapatnam in 2006.

Awarded Ph.D. from Andhra University, Visakhapatnam in 2020.Now working as Assistant Professor in Velagapudi Ramakrishna Siddhartha Engineering College,Vijayawada. She has 16 years of teaching experince. She has published around 15 publications in reputed International Journals/confeences. She published one patent. She is a member of International Society of Automation (ISA), International Associati of (IAENG)

Dr.S.Sunitha Ratnam, received MSc from sri krishnadevaraya University in 1993.Awarded Ph.D. from Osmania University IN 2016.

Now working as Professor in Malla Reddy Engineering College For Women, Hyderabad. She has 10 years of teaching experince. She has published around 10 publications in reputed International Journals/book chapters.

Dr.J.Bhavani, Associate professor at EEE department, VNRVJIET ,Hyderabad has got 22 years of teaching and Research experience at various reputed institutions

She was graduated from SV University affiliated college in 2000, post graduation in power Electronics from JNTUH in 2005 and she has received doctorate in 2015 from JNTUH, hyderabad. She has published 18 international journals and 14 international conferences .Her area of specializations are PWM schemes, DC-DC Converters and Renewable energy

systems.she has received best paper award and best young women scientist award .

Dr.Jalluri.Srinivasa Rao, received B.Tech from JNTUK in 2005, and M.Tech from JNTUH in 2007 . Awarded Ph.D. from JNTU H, in 2015.

Now working as Asoociate Professor in VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad. He has 16 years of teaching experince. He has published around 29 publications in reputed International Journals/Conferences/book chapters. He published two patents. He is a member of ISTE.

N.Kavitha, received her B.Tech and M.Tech Degree from JNTUH. She is working as an Assistant Professor in ECE Department of Maturi Venkata Subba Rao (Autonomous) Engineering College, Nadergul, Hyderabad.

since 2008.She is a life member of IETE, ISTE, and CSI. Her area of interest are image & signal processing, communications.