A Project Review-1 Report

on

Lung Cancer Detection Using Machine Learning

Submitted in partial fulfillment of the
requirement for the award of the degree of

B. Tech in Computer Science

Under The Supervision of
Name of Supervisor : Ms. Sonia Kukreja

Submitted By

Anish Kumar (19SCSE1010275)
Ghufran Ahmad Khan (19SCSE1010238)

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
March, 2023

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **"Lung Cancer Detection Using Machine Learning** in partial fulfillment of the requirements for the award of the **B.TECH** submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the January 2023 – May 2023, under the supervision of **Ms. Sonia Kukreja, Assistant professor**, Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the project has not been submitted by us for the award of any other degree of this or any other places.

Anish Kumar19SCSE1010275
Ghufran Ahmad Khan (19SCSE1010238)

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Supervisor Name : **Ms. Sonia Kukreja**
Designation: **Assistant professor**

CERTIFICATE

The Project Viva-Voce examination of **Anish Kumar(19SCSE101275)** and **Ghufran Ahmad Khan (19SCSE1010238)** has been held on 28th March 2023 andhis/her work is recommended for the award of B.TECH.

Signature of Examiner(s) **Signature of Supervisor(s)**

Abstract

Uncontrolled cell growth in lung tissues is a possible cause of lung cancer. One of the primary causes of cancer-related deaths worldwide is lung cancer. Recovery from lung cancer requires an early diagnosis. The major cause of cancer-related deaths this century has been lung cancer, and this trend is expected to continue in the decades to come. Lung cancer is treatable if the disease's symptoms are found at an early stage. Many reliable systems for the treatment of lung cancer that are simple to use and lower in cost have been developed using the most popular data science technology. This research presents a comparative study of several machine learning-based methods from the last three years for the identification of lung cancer. There are an excessive number of methods now available to detect lung cancer, most of which depend on CT scans and others on x-ray images. To detect lung cancer, almost all of them are using image classification methods to find lung cancer nodules. This is combined with several segmentation techniques and a wide range of classifier algorithms to get a more accurate result.

Machine learning-based lung cancer prediction models have beenproposed to help clinicians manage incidental or screen-detected indeterminate pulmonary nodules. These systems may reduce variability in nodule classification, enhance decision-making, andreduce the number of benign nodules that are needlessly followedup or worked on. We also discuss some ofthe challenges in the development and validation of such techniques and outline the route to clinical adoption.

We discuss challenges in the development and validation of suchtechniques and outline the path to clinical adoption. ML-based lung cancer detection models have the potential to improve the accuracy of diagnosis, reduce costs, and reduce the number of false-positive results. To achieve this, researchers must continue to refine and validate these techniques, considering the complexities of the data and the wide range of clinical scenarios in which they may be applied. With further research, ML-based lung cancer detection models may become a valuable tool to improve lung cancer detection and management.

List of Table

List of Figures

Acronyms

B.Tech.	Bachelor of Technology
M.Tech.	Master of Technology
BCA	Bachelor of Computer Applications
MCA	Master of Computer Applications
B.Sc. (CS)	Bachelor of Science in Computer Science
M.Sc. (CS)	Master of Science in Computer Science
SCSE	School of Computing Science and Engineering

1. Introduction

When a few of the body's cells grow out of control and spread to other internal organs, it becomes cancer. Cancer may appear almost anywhere in the trillions of cells that make up the human body. If not treated properly, tumors may cause several illnesses. Tumor cells multiply uncontrolled, unlike normal cells that replace old or damaged ones[7]. They sometimes suffer modifications (mutations). Damaged cells that grow uncontrolled and expand into enormous amounts of tissue become tumors. Tumors may be either benign (noncancerous) or malignant (cancerous). The growth and spread of benign tumors are often sluggish. Malignant tumors have the ability to spread throughout the body, develop quickly, infect neighboring normal tissues, and do great damage as show in the fig: 1. Modern technologies like low-dose computed tomography and other methods for early lung cancer identification make it feasible to treat lung cancer sooner [12].

Lung cancer is caused by an abnormal tissue development termed a nodule, which arises from cells in the airways of the respiratory system. These cells look spherical on chest X-rays and are always in contrast. If lung nodules are detected early, patients may have a better chance of survival. Evaluation of raw chest X-ray pictures is laborious and difficult because lung nodules are hard to see [10]. Computer-aided diagnostics must find a small nodule in a big 3D lung CT image [5]. The CT image is overloaded with background noise from air and bone and other nearby muscles, fat, organs, and blood vessels; therefore, this noise must first be reduced in order for the CAD systems to search efficiently. Our classification pipelines include nodule candidate, image pre-processing, and malignancy classification detection. Machine learning may help identify and treat lung nodules in AI-generated CT images. There are multiple lung cancer detection methods available today. A general block diagram of the machine learning approach for lung cancer detection is shown in Fig. 2. We evaluated new cancer detection methods to choose the best[10].

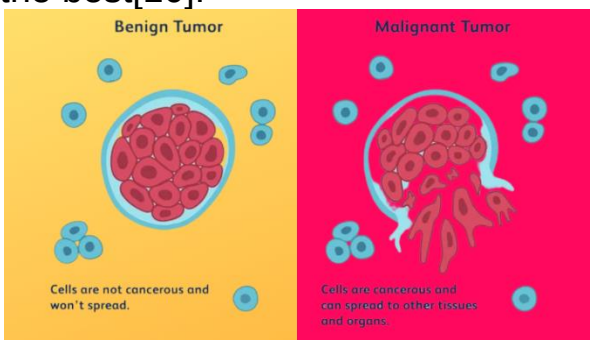

Figure 1: Benign Tumor and malignant tumor

2. Literature Survey

In order to enhance work, business, etc., researchers have successfully used statistical and machine learning approaches. Therefore, machine learning methods are also being applied in the area of medicine to identify diseases and treat them appropriately. Additionally, prediction models for lung cancer are being created using machine learning approaches. Image processing methods have previously been used to detect lung cancer. Researchers have attempted to apply deep learning and neural network methods in order to forecast lung cancer more precisely. Using methods from machine learning and conventional neural networks, some researchers have tried to categories and diagnose lung cancer. Deep learning methods for lung cancer detection have recently been tested by certain researchers.

Several researchers published a study in October 2019 based on a supervised machine technique called as "random forest" that may aid in the early identification and diagnosis of lung cancer. The major goal of this research is to use the Random Forest Algorithm to categorize data related to lung cancer. The most accurate learning algorithm is Random Forest, which is used by many researchers in the healthcare industry. They used two datasets to test the model, and based on the findings, they observed that the recommended method, which was based on the random forest algorithm, obtained an accuracy of 100% for dataset 1, and 96.31% for dataset 2.

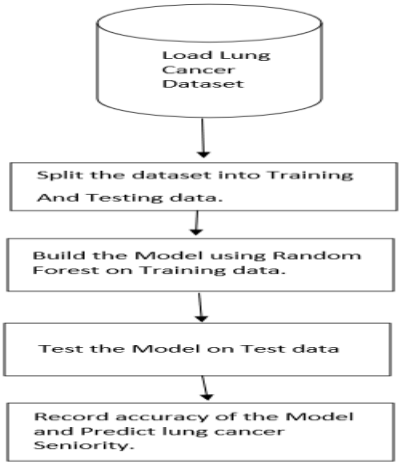

Figure 2: Lung cancer detection using random forest technique

Several researchers published a study in 2020 that was based on some

statistical methods for early identification and diagnosis of lung cancer. The paper states that the capacity of a computer-aided diagnostic (CAD) system to recognize lung nodule cancer is highly beneficial to medical practitioners and may lower the rate of death. For the purpose of early lung nodule cancer diagnosis, a CAD system is created in this work. For the purpose of enhancing the image's contrast, a method known as Contrast Limited Adaptive Histogram Equalization (CLAHE) is used. Otsu thresholding is used to segment lung tumors, and morphological filters are then used to remove the background and other geometrical objects. In order to de-noise the generated image, the discrete wavelet transform (DWT) is applied. The Gray Level Co-occurrence Matrix (GLCM) is used to extract features such as correlation, energy, and other properties. Principal component analysis (PCA) is used to choose features. In order to evaluate whether an image is benign (non-cancerous) or malignant (cancerous), Support Vector Machine (SVM) is utilized. By using both the normal and malignant pictures at the same time, this study presents a comprehensive and totally automated lung cancer categorization method. These approaches enable accurate and competent lung nodule identification from CT scan pictures. The automated nodule identification accuracy of this model was 97.34% for the LIDC dataset and 96.55% for the ELCAP dataset[6].

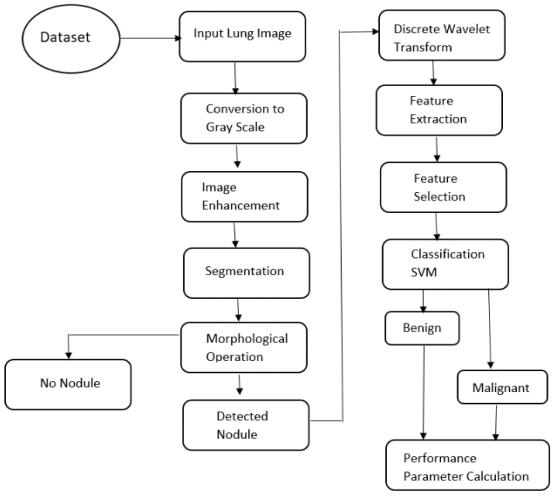

Figure 3:Lung cancer detection using SVM

Several researchers presented a study based on a deep-learning machine learning approach that aids in the identification of lung abnormalities utilizing chest X-ray and lung CT scan pictures for an analysis of lung abnormalities. This method has the potential to assist in the detection and diagnosis of lung cancer at an earlier stage. In order to detect lung cancer and pneumonia in their earliest stages, this research relies wholly and exclusively on the framework of deep learning (DL). This work proposes two separate DL methods, one for doing an internal analysis of the abnormalities and they are as follow: (i) it is proposed that a modified Alex Net (MAN)-based initial deep learning strategy be used to classify chest X-ray photos into normal and pneumonia categories. In order to increase accuracy rate while assessing lung cancer, the second Deep Learning effort integrates a combination of handmade and learnt characteristics in the MAN. This DL framework has an accuracy rate of 97.27% when it is validated using typical lung cancer CT images from the LIDC-IDRI[7].

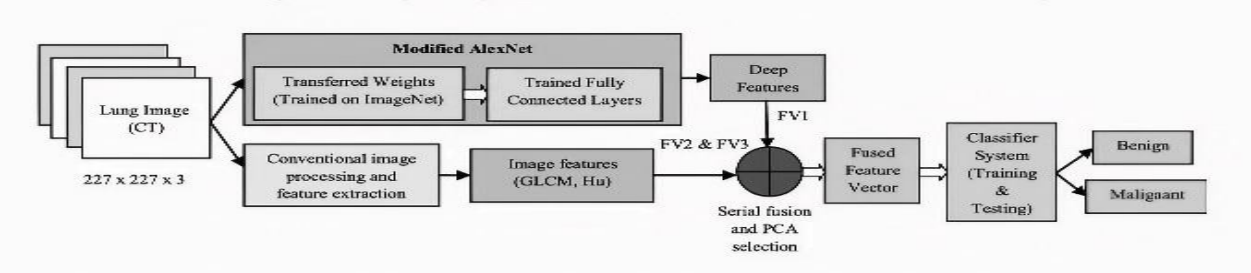

Figure 4: Lung cancer detection using Deep learning

In 2021, Gopichand Shelke, Shraddha Patil, and Smita Raut presented "Lung Cancer Detection Using Machine Learning Approach." They combine machine learning methods and digital image processing to find the tumour in the photos in their suggested system. Digital image processing and machine learning rules are the two components that make up the overall model. Image capture, grayscale conversion, noise reduction, picture binarization, segmentation, characteristic extraction, machine learning, and the last phase, most cancers mobility recognition, are the eight processes in digital image processing. A machine learning algorithmic decision tree is used in the second stage to forecast the results.

50 photos were used to train the model. The result specifies the tumor's malignancy or benignity. And 78% accuracy of the design was discovered[8].

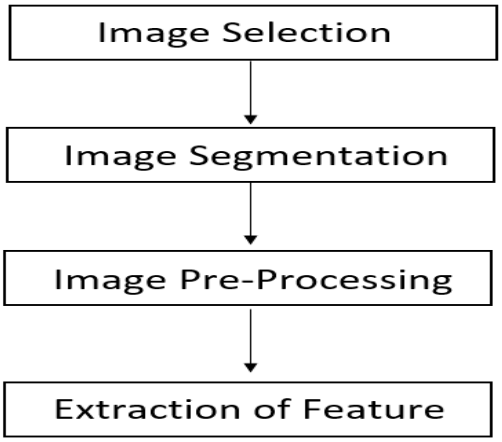

Figure 5: lung cancer detection using Decision tree machine learning algorithm

A group of researchers collaborated in 2022 to submit a study that was based on a method known as support vector machine (SVM) analysis, which helps in the detection of lung abnormalities by making use of a text dataset for an evaluation of lung abnormalities. In this particular research project, an SVM-based machine learning model was used in order to enhance the process of lung cancer detection. Patients suffering from lung cancer are sorted into different categories according to the symptoms they exhibit with the use of a support vector machine (SVM) classifier, and the model is refined with the help of the computer language Python. Different tumors may be diagnosed using a variety of techniques. However, there are only a few distinct methods for figuring out what populations are present in them. This essay will provide a technique for not only finding malignant tumors but also doing the necessary calculations to determine their size, shape, and location. Thus, in addition to counting and winning, the kind of tumor may also be determined. The development of lung cancer is predicted in this study using SVM. If the prognosis turns out to be true, the physician may be able to devise a more efficient treatment plan for the patient and arrive at a conclusion about their condition sooner.

In this study, the recommended model was contrasted using the SVM and SMOTE methodologies. When compared to the current approaches, the suggested method has a 98.8% accuracy rate[10].

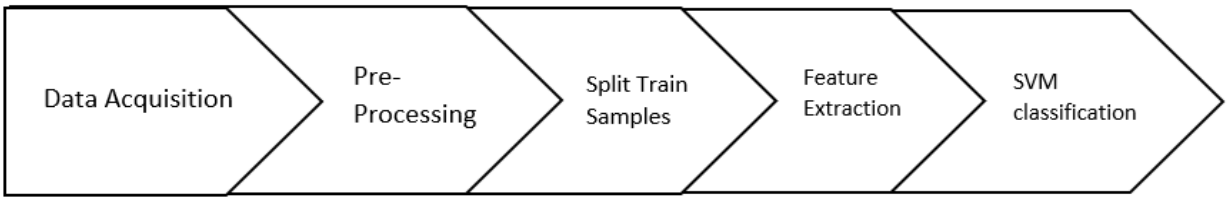

Figure 6:lung cancer detection using SVM and SMOTE techniques

The most recent literature review makes it clear that lung cancer has received significant recognition from the scientific community. The majority of solutions have relied on traditional machine learning and neural network techniques. Others used deep learning techniques to both types of photos (CT and X-ray). Therefore, the goal of this research is to examine different techniques and identify the most effective way for identifying lung cancer.

Table 1: A table providing a comparison of the five separate review articles published between 2019 and 2022.

Ref.	Year	Methods	Results
1	2019	Random Forest Algorithm to categorize data	The recommended approach achieved 100% accuracy for dataset 1 and 96.31% for dataset 2.
6	2020	Support Vector Machine (SVM)	This model's accuracy for LIDC and ELCAP was 97.34% and 96.55%, respectively.
7	2020	Chest X-ray pictures are classified using MAN, while lung CT images are classified using EFT.	DL accuracy for X-rays is 96% and with CT, 97.27%.
8	2021	Decision tree machine learning algorithm	The accuracy of the model was 78%.
10	2022	SVM and SMOTE techniques	The accuracy of this model was 98.8%

3. Implementation

SYSTEM ARCHITECTURE

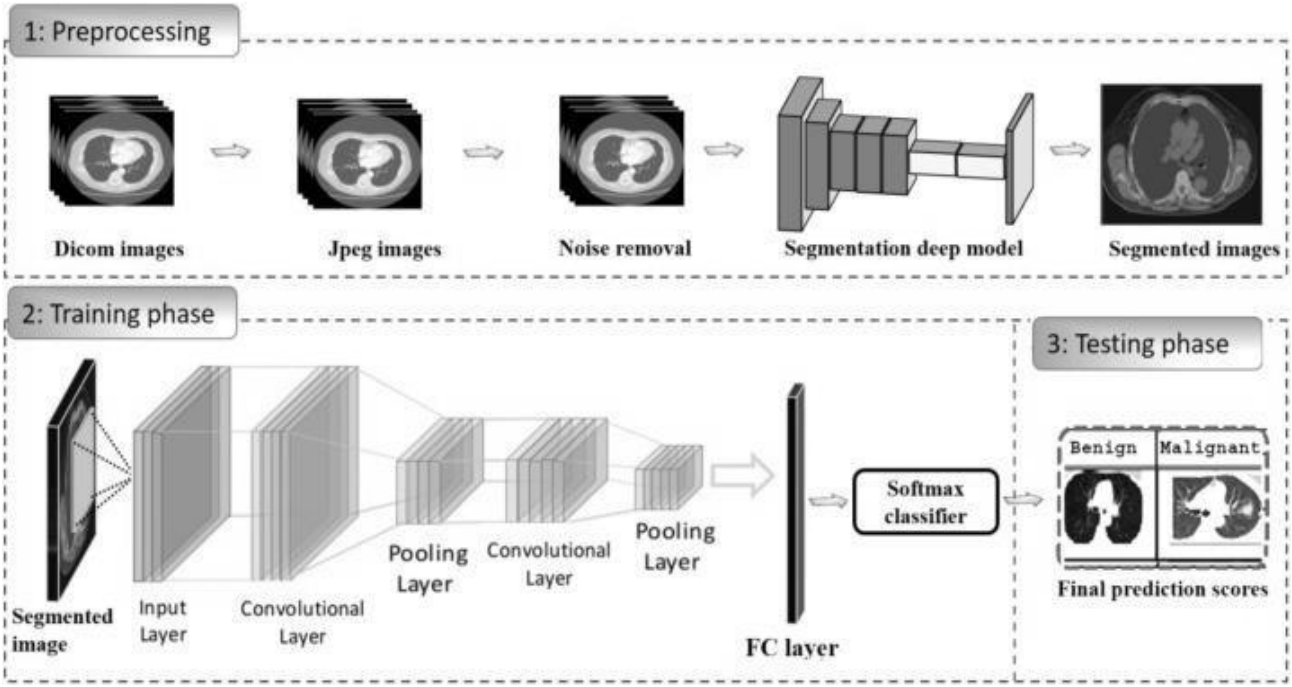

Figure 7:SYSTEM ARCHITECTURE

Steps

- Download CT Scan Images
- Load Dataset into python
- Create a pandas data frame
- Map annotations to image filenames
- Extract images
    - Understand the 3D image data
    - Visualize images
    - Segment lungs & cancer
    - Save Segmented image into a file.npy
- Upload the segmented images into a file.npy
- Train the U-Net Model
- Evaluate model
- Save model as .hdfs
- Download model from floyhub

- Create flsak backend
- Create front end
- Load model into web application
- Deploy

4. Model Descriptions

4.1. Data Set

The database used is obtained from Lung Image Database LUNA16,Data Science Bowl 2017 This is a lung nodule classification database containing the scans of a total of 1018 patients. Each patients' CT scan in turn is comprising of around 150 to 550 dicom format images. The database provides four classifications namely-(i)Unknown, (ii)Benign, (iii)Malignant, and (iv)Metastatic

4.2. Convolutional Neural Network

The convolution layer of a produces a feature map by convolving different sub regions of the image with a learned kernel. Further, non-linear activation functions such as a sigmoid, tanh or rectified linear (ReLu) can also be applied. Another method for reducing computations is the pooling layer, where a region of the image/feature map is chosen and the maximum among them is chosen as the representative pixel. Hence, a 2x2 or 3x3 grid can be reduced to a single scalar value. traditional fully connected layer can also be used in conjunction with the convolutional layers and are usually used towards the output stage.

4.3. Convolution Layer 1

The data in 3-D hdf5 format forms the input to the first convolution layer. This layer of kernel size 50x50 with a stride of 6. The output of this layer produces 78 features. The weight filler is set to a 0.01 Gaussian distribution change and the bias is set at constant zero. This output is then fed to the Rectified Linear (ReLu) layer to bring all the negative activations to zero. The primary application of this layer is to detect the lowest level features, e.g., whether there is classification in some area of the image.

4.4.Convolution Layer 2:

The first Convolution layer output is fed into the second having a kernel size of 3x3 and a stride of 1. This layer pads the data with one enclosure of zeros. The weight filler is the same as convolution layer 1 and the bias is set to a constant value of 1. Also, this layer is followed by a ReLu layer. This layer is intended to make use of the information predicted from the previous layer and detect the pattern of classification - e.g., popcorn, diffuse etc. From the training phase, it will hence learn as to which among the patterns are benign, and which are malignant. In this way the CNN achieves two objectives - it learns features hierarchically, and it eliminates the need for specific feature engineering.

4.5.Max-pooling Layer:

After the convolution layer 2 comes the max-pooling layer where the most responsive node of the given kernel is extracted. The kernel size used in the proposed network is 13x13 with stride shift of 13. This is primarily intended to reduce the computational effort. Since each CT scan composes of 500 images, if we have a batch size of 50, the number of required computations can be significantly large, leading to frequent memory overload. The max-pooling layer is used particularly to ease memory and data bottlenecks by reducing the image dimensions.

4.6. Dropout layer:

The dropout layer is used in the network to prevent over- fitting. This is done by switching off random neurons in the network. Our proposed network uses a dropout layer with a drop ratio of 0.5. The intent of this layer is to improve the classification quality on test data that has not been seen by the network earlier.

4.7. Fully connected layer:

A fully connected layer which provides two outputs is used. It uses Gaussian weight filler of 0.5 and a constant bias filler of 0. The two output neurons from this layer gives the classification of benign or malignancy. This layer is mainly intended to combine all the features into one top level image and will ultimately form the basis for the classification step.

6. REFERENCES

[1] A. Rajini and M.A. Jabbar “Lung Cancer Prediction Using Random Forest” In oct 2019.

[2] Haron, H., Zeebaree, D. Q., Zebari, Abdulazeez, A. M., and D. A. Trainable Model Based on New Uniform LBP Feature to Identify the Risk of the Breast Cancer. In 2019 International Conference on Advanced Science and Engineering (ICOASE).

[3] Yu, K. H., Lee, T. L. M., Yen, M. H., Kou, S. C., Rosen, B., Chiang, J. H., and Kohane, I. S. “Reproducible machine learning methods for lung cancer detection using computed tomography images: Algorithm development and validation” Journal of medical Internet research, In 2020.

[4] Jothilakshmi, R., and SV, R. G. “Early Lung Cancer Detection Using Machine 6 Learning And Image Processing” In Journal of Engineering Sciences, in 2020.

[5] Tanzila Saba “Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges” In Journal of Infection and Public Health, in 2020.

[6] Waseem Abbas, Khan Bahadar Khan, Muhammad Aqeel, Muhammad Adeel Azam Muhammad Hamza Ghouri, Fawwad Hassan Jaskani,“Lungs Nodule Cancer Detection Using Statistical Techniques” IEEE 23rd International Multitopic Conference (INMIC) 2020.

[7] Abhir Bhandary, G. Ananth Prabhu, V. Rajinikanth, K. Palani Thanaraj, Suresh Chandra Satapathy, David E. Robbins Charles Shasky Yu-Dong Zhang, João Manuel R.S. Tavares, N. Sri Madhava Raja, “Deeplearning framework to detect lung abnormality – A study with chest X-Ray and lung CT scan images” In 2020.

[8] Smita Raut, Shraddha Patil, Gopichand Shelke, “Lung Cancer Detection using Machine Learning Approach”, International Journal of Advance Scientific Research and Engineering Trends(IJASRET) in 2021.

[9] Dakhaz Mustafa Abdullah, Adnan Mohsin Abdulazeez, & Amira Bibo

Sallow “Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques” In May, 2021 in Qubahan Academic Journal.

[10] Anil Kumar, S. Harish, Prabha Ravi, Murthy SVN, B. P. Pradeep Kumar, V. Mohanavel, Nouf M. Alyami, S. Shanmuga Priya, and Amare Kebede Asfaw “Lung Cancer Prediction from Text Datasets Using Machine Learning” In 2022.

[11] Srinivas Arukonda, *S.Sountharrajan Investigation of Lung Cancer detection Using 3D Convolutional Deep Neural Network(2020)

[12] Rohit Y. Bhalerao A novel approach for detection of Lung Cancer using Digital Image Processing and Convolution Neural Networks.(2019)

[13] Lilik Anifah Cancer Lungs Detection on CT Scan Image Using Artificial Neural Network Backpropagation Based Gray Level Coocurrence Matrices Feature(2017).

[14] Prajwal Rao∗, Nisha Ancelotti Pereira†, and Raghuram Srinivasan Convolutional Neural Networks for Lung Cancer Screening in Computed Tomography (CT) Scans(2016).