PROJECT REPORT (14_11_2022).docx

WORD COUNT	5705	TIME SUBMITTED	15-NOV-2022 02:03PM
		PAPER ID	92665033

A PROJECT REPORT ON

DIAGNOSIS AND STAGE DETERMINATION OF CT SCANNED IMAGES OF LUNG CANCER USING ML

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE

IN THE PARTIAL FULFILLMENT FOR THE AWARD OF THE DEGREE

OF

BACHELOR OF ENGINEERING
IN
INFORMATION TECHNOLOGY

BY

ATHARVA MISAL (BEITB201)

RAHUL BADGUJAR (BEITB226)

OMKAR RASKAR (BEITB232)

AKSHATA SAPTASAGAR (BEITB236)

UNDER THE GUIDANCE OF

MRS. SANDHYA WAGHERE

DEPARTMENT OF INFORMATION TECHNOLOGY

PIMPRI CHINCHWAD COLLEGE OF ENGINEERING

Sector no.26, Pradhikaran, Nigdi, Pune - 411 044

2022-23

CERTIFICATE

This is to certify that the project report entitled

Diagnosis and Stage Determination of CT Scanned Images of Lung Cancer using Hybrid Model

Submitted by

ATHARVA MISAL (BEITB201)

RAHUL BADGUJAR (BEITB226)

OMKAR RASKAR (BEITB232)

AKSHATA SAPTASAGAR (BEITB236)

is a bonafide work carried out by them under the supervision of Prof. Mrs. Sandhya Waghere and it is approved for the partial fulfillment of the requirement of Savitribai Phule Pune University for the award of the Degree of Bachelor of Engineering (Information Technology)

This project report has not been earlier submitted to any other Institute or University for the award of any degree or diploma.

MRS. SANDHYA WAGHERE	**DR. SONALI PATIL**
Internal Guide	Head of Department of Information Technology

Place:

Date:

CONTENTS

LIST OF ABBREVIATIONS

1. SVM - Support Vector Machine
2. CT Scan - Computed Tomography Scan
3. UML - Unified Modeling Language

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The major cause of cancer-related mortality globally is lung cancer. Even though anybody can get lung cancer, several risk factors, like smoking and being around smoke inhalation, raise the likelihood. The type of tumor and the cancer's stage at the time of discovery will determine the course of treatment. Radiation, surgery, immunotherapy, and chemotherapy are all possible forms of treatment. When the illness is discovered early on, it is typically curable. So, in order to enable patients to receive fast treatment, specialists are constantly developing novel methods for early detection of lung cancer. The use of technology here is beneficial. There are various existing models which focus on the diagnosis and determination of stages of various cancer and other related diseases. This paper focuses on the diagnosis and stage determination of Lung cancer by compiling various Ml models. This paper proposes a model that will not only diagnose the presence of disease but will also help the medical faculty in knowing the particular stage of the disease. Also, advanced analysis is provided by the models which give a brief overview of the disease and highlights the stakeholders about the curability of this disease. The model uses various algorithms and compiles some of the best-suited algorithms which will provide results with higher accuracy and precision.

INTRODUCTION

Continuous use of alcohol, tobacco, and other hazardous substances has been linked to a higher risk of developing cancer as well as various disorders. Additionally, bad eating habits, little to no exercise, and air pollution have emerged as concerning contributors to a number of ailments. Cancer is a condition in which cells continue to multiply unchecked. Lung cancer is the term used to refer to this cellular proliferation when it occurs in the lungs. Additionally, lung cancer can develop when cells from other tissues migrate to the lungs. Lung cancer is the second most common cancer globally, according to recent studies. Around the world, there will be 2.2 million new cases of lung cancer until 2020. Lung cancer can also be brought on by a history of lung disease, chemical exposure, or household air pollution. Since it might be mistaken for the influenza in its initial stages, lung cancer, also known as lung malignancy, is typically difficult to detect. However, this condition can be rectified by using multiple ML models to identify lung cancer in its earliest stages. In accordance with the American Cancer Society, there were about 236,740 lung cancer cases identified in 2022, resulting in close to 130K fatalities.

Lung Cancer is mostly divided into two phases namely SCLC and NSCLC. SCLC is Small Cell Lung Cancer which is deadly serious and is found in the inner layers of the walls of bronchitis. NSCLC is Non-Small Cell Lung Cancer which is less serious as compared to SCLC and also is most commonly found across the world. In 2020, 2.2 million people were diagnosed with Lung Malignancy out of which 1.7 million people lost their lives to cancer. This rate can be reduced by the detection of lung Malignancy at its early stages. Various Machine Learning models can be used to diagnose the disease.

One such is the Support Vector Machine technique which involves classifying the dataset into several classes. The support vector classifiers help the model to divide the datasets into multiple classes providing high accuracies for small datasets. Decision trees are one of the models which help to segregate the data into various subsets and the number of decision trees can be coupled into the Random Forest model generating high accuracies for complex datasets. Enhanced

Classifiers integrate the analysis of images of various models and based on voting classifiers the results can be outsourced which can then be further used to treat the disease.

1.1 MOTIVATION

- Trust in the referring clinician.
- Benefits of early-detection of lung cancer.
- Low or limited harm from LDCT scan perception.
- Experiences of friends or family with advanced cancer.

1.2 OBJECTIVES

- To enhance, filter and segment dataset.
- To extract the necessary features from the processed dataset and classify the extracted features into classes.
- To further classify the abnormal set of dataset into stages.
- To provide the curability status of the acquired stages.

1.2 PRINCIPLE:

Diagnosis involves specialists making decisions about a person or group of individuals. It considers the whole lifetime of an individual including the individual's past life, experiences, style of living, attitude, and interests. Diagnosis requires knowledge, skill and the ability to synthesize

and evaluate large portions of data. Diagnosis is a team effort which requires effective communication. Determination involves detecting the anatomic extent of a particular disease. It helps in measuring to which extent cancer has spread across the human body. This can be done with the help of going through a process of scanning as well as biopsies and other related tests. The determination can be done in the form of numbered stagings in the TNM format. Stages involve the numbering severity of cancer cells on the count of 0 to 4. TNM includes a detailed analysis of tumours, nodes and metastasis.

LITERATURE SURVEY

A. Analysis of CT Scan Images to Predict the Stages of Lung Cancer Using Image Processing Techniques

[1]Lung cancer, also known as lung carcinoma, is a malignant lung tumour that is characterised by unchecked cell proliferation. Due to many reasons, detection of lung cancer earlier is important. CT images are used widely as they have been proven more effective than X-rays for diagnosing lung cancer at an early stage. The main advantage of selecting digital image processing is the superiority of image data and optical ability over other techniques for lung cancer staging. Evaluating image cells and obtaining information is easier due to picture processing. The statistical parametric approach and image analysis based on the gray-level co-occurrence matrix (GLCM) are used to eliminate other features. GLCM gives a pairing of types that are related by second order. The GLCM has more characteristics than the statistical approach. Whereas the statistical approach has less complexity than GLCM.

Firstly we obtain the images for preprocessing. It includes various techniques as mentioned below. The features are then extracted with the help of the GLCM technique and statistical approach for determining the values of the features. At last, classifiers are used to partition cancer into two stages, one is limited and the other is extensive. Then the classifier performance is calculated

1) Image acquisition: CT Scanned images of the patients are obtained. These images are to be preprocessed for the reduction of noise.

2) Pre-processing:

a) Smoothing: To eliminate the noise from the pictures smoothing technique is used,

b) Enhancement: To make the picture quality better, enhancement is used. For getting desired results through enhancement, the Gabor filter is used. It is proven better than auto enhancement.

c) Segmentation: To partition the image into several segments, Segmentation is used. A global threshold and Otsu's technique are used.

d) Morphological Opening: In this technique, erosion and dilation are used.

3) Feature extraction: Using the grey level for analysis of the texture of the matrix co-occurrence is done. The texture of the picture is quantified with the help of metrics known as image texture. This helps to know that the colour intensities are spread spatially in a picture. Then the GLCM is used to obtain the values of the grey level pixel.

4) Classifiers: Support vector machine (SVM): SVM is used for classification as well as regression, but the most widely used is the classification problem.

Two methods were used to obtain the features which were the statistical parametric approach and the other was GLCM. The second approach has fewer features but is better than other methods as the dimensions in it exceed the tenth power equivalent to total features and increase with it. This study helped to preprocess the image and improve the precision of detecting the stage of lung cancer. By using SVM the statistical approach yields an accuracy of 78.95 percent.

B. Automatic Detection of Lung Cancer using CT Scan images

In this research study, Hoque et al[1] proposed an automated approach where CT Scan gray-scale images were incorporated for cancer detection. Lung cancer images were taken as input and after being processed by medical image processing methods such as pre-processing and post-processing output images were generated containing the region only. The preprocessing consists of enhancement, filter operation, and segmentation. The post-processing consists of feature extraction and identification.

The flow of the proposed approach is represented in Fig. 1. The first step in this method is image acquisition. This consists of two processes: preprocessing and post processing. During the preprocessing process of image enhancement, filtration and segmentation is done. During the post-processing feature extraction and finally identification is done.

A. Image Acquisition : For the implementation of the proposed method, a Lung cancer dataset [1] has been used. A MATLAB of size 450× 450 is used to store the images in image format and are shown as RGB gray scale images with the entries ranging from 0 to 1.These are changed into HSV (hue saturation value) format for processing them.

B. Pre-processing :

- First stage is image enhancement. It is done using contrast stretching which is proven to work better on gray-scale images[2]. Fig 2 shows the difference between original and enhanced image.
- Second stage is Filter Operation which is done to improve sharpness and edge enhancement. Here the Median filter is used which showed better performance than the mean filter or the Gaussian filter[3]. Fig 3 shows the before and after of the Filter Operation.
- The third and final step in preprocessing is Segmentation. It basically splits the image into parts based on their similar characteristics. Thresholding based Otsu's technique is used based on its wide uses to segment an image for further processing like feature analysis as well as to do the binary transformation of an image.

Post-Processing :

- Feature Extraction: For each shape in the image, the MATLAB operation region-propos returns a number of characteristics. Out of these, area, circularity (roundness and diameter), and solidity are used in the proposed model.
- Identification : To identify the ROI, is the purpose of the identification process.. The ROI is detected by using the features extracted in the feature extraction section.

The ROI is identified successfully when a region is contained by each feature's value length

Result Analysis of the proposed method : The following table 1 shows the result analysis of the proposed method. The number of images in the dataset, the number of true positive and true negative images, the number of false positive and false negative images, and the accuracy of those datasets are all included. It demonstrates that for batches 1, 2, and 3, there are, correspondingly, 24, 23, and 24 true positive values and also 1, 1 and 1 true negative values for the same respectively. It demonstrates that the suggested approach correctly detects 26, 26, and 26 pictures from batches 1, 2, and 3, which contains a total of 78 images. It gives an accuracy of 96.15%, 92.30% and 96.15% for all 3 batches respectively. This study, therefore, achieved pathology approved accuracy of 95%.

The possible experimental outcomes of this study proposed method are three images in axial, coronal and edged orientation from three different datasets.Images inputted are shown in the first column. The second column shows the implementation of image enhancement. The third column shows the filtered image obtained with the use of the median filter. In the fourth column, white spots that are presented in the image are the result of the threshold segmentation which are basically areas where intensity values are higher than a specified threshold. Finally, only the cell is found in the final column.

In this paper, they have proposed the use of Support Vector Machine (SVM) classifier and feature selection in detection of CT scan images of lung cancer. This study's main goal is to identify lung cancer precisely and with the greatest degree of accuracy possible. Compared to prior studies, this research not only streamlines methods that save time but also increases accuracy. However, the accuracy rate can still be improved, which means that there is still a chance to work with grayscale photos.

C. Early Stage Prediction of Lung Cancer Using Various Techniques of Machine Learning

[3]Lung cancer is chiefly activated due to cigarette smoking. Lung cancer is one of the most dangerous forms of cancer and is currently diagnosed in abundance. 2.09 million cases have been diagnosed to date. This diagnosis of lung cancer mostly includes a variety of complex processes. The World Health Organization also stated that by 2018, most deaths occurred regarding Lung Cancer. There are two forms of Lung Malignancy and they are Small-Cell Lung Cancer(SCLC) and Non-Small-Cell Lung Cancer (NSCLC). SCLC is deadly serious and is mostly found in the bronchial wall's inner layers. SCLC spreads rapidly as compared to NSCLC. NSCLC is less dangerous as compared to SCLC in the early stages. Catalog Classification is used to classify whether a person possesses lung cancer or not. Several Machine learning algorithms like Support Vector Machines, K-Nearest Neighbour, Clustering can be used to solve such problems. Machine learning mostly deals with statistical models eliminating the need for instructions which depend on patterns and inferences

ICD9 demonstrative codes can help to differentiate a particular chronic disease from the rest of the chronic diseases. But here several visits are needed by the patient to make the algorithm familiar. Fuzzy-C-Means or Fuzzy-Possibilistic-C-Means can be used but the accuracy only notes at 80.36 per cent. Also, it was discovered that to achieve better accuracy, one has to make use of hybrid models. Neural Networks like ANN and CNN were used to diagnose cancer but the output acquired from these models provided output with less accuracy compared to that of ML models[3]. And the accuracy of these network models can be increased by coupling them with advanced neural technology but implementing those needed with high knowledge and eventually, the concluded model resulted in greater complexities.

The proposed work in the paper consists of the model that estimated the performance of various other models.

The first step in building the model comprises Data pre-processing which includes Importing the required libraries and importing the necessary datasets.

Successive steps to importing datasets include Redressing Missing data followed by Converting data into Categorical data where all the independent categorical features are converted into numerical features.

Feature selection was done to extract the necessary features eliminating the rest which might have led to future computational complexities. further data gets split up into train, test and validation sets.

The algorithms used for comparison include the following:

- Support Vector Machine: SVM is a supervised machine learning model which uses a classifier named Support Vector Classifier. This classifier assists in model training using training dataset and predicts the output into several classes.
- Random Forest: It involves creating multiple decision trees and the output is predicted by combining the votes of n-decision trees.
- K-Nearest Neighbor: KNN uses similarity measures to calculate the difference between actual and observed data points. Critical estimations involve k vectors to be odd but this model uses 2 classes issues to wear off the tie between the classes.
- Neural Networks: Using an ANN network, various forms of data are fed to the input layer which then processes the data and transfers the processed output to the hidden layer. Further, activation functions are applied in the hidden layer which helps in re-tuning the weights within the input layer and hidden layer and transferring the result to the output layer.
- Voting Classifier: Hard voting is implemented which generally takes majority votes by the predicted class labels. This model is generally a hybrid model as it uses SVM, Random Forest and KNN to train the classifier.

The results involve parametric accuracies of all 5 models with 0.8 and 0.2 as training and testing datasets respectively. Accuracy of SVM rates for 95 per cent. Random Forest gives 97.5 percent accuracy whereas KNN gives 97 per cent accuracy. Neural networks account for 95.99 per cent accuracy. The voting classifier gives the highest accuracy of all which is 99.5 percent.

The model helps in predicting early-stage cancer in people. The model gave a brief description of famous ML classification models and also compared the accuracies of SVM, KNN, Random Forest, Neural Networks and Voting Classifiers. Thus, the voting classifier is considered to be best suited for predicting cancer at its early stages. Also, the scope can be extended by using Logistic regression models or Extra trees classifiers or by using Boosting methods.

D. Application of Machine Learning Methods for Determining the Stage of Cancer

Cancer is one of the most prevalent causes of death in the globe. Cancer is among the main causes of mortality worldwide, contributing for 20% of fatalities in the European region, as reported by the WHO (World Health Organization). The models created using machine learning techniques can aid in the process of identifying a patient based on their physical symptoms. Predicting the stage of a cancer's particular symptoms is the aim of study.

The term "staging" is used to define the tumor's location, size, kind, lymphadenopathy distribution, and existence of metastases. The TNM method is being used to better the information flow between doctors and to aid patients in understanding their illness.

The TNM system uses the letter T together with another letter or number to designate the tumor's size. When it is impossible to analyze the adjacent lymph nodes, a value of X is applied. When there is no malignancy in the adjacent lymphoid tissue, the value is 0. When describing the size, position, and numerous neighboring lymph nodes in which cancer has spread, N is combined with one, two, or three. When malignancy has spread towards other parts of the body, the letter M is used to denote this.

When the sickness has not spread to other parts of the body, the zero value is employed. Forecasting a patient's oncological disease stage using the TNM system is the first research problem.

Techniques: A lot of data is processed and analyzed using machine learning techniques. Three techniques were used to generate regression models: Decision trees, Support Vector Machines, Ensemble Algorithms, and 3) Measures To analyze the models, four measures are used.

A) Mean Squared Error

B) Mean Absolute Error

C) R-Squared

D) Root Mean Squared Error

Experimental procedure: The initial training sample contains information on 18329 patients with breast cancer, whereas the testing set contains 6112 entries.

The second sample's training data set contains information on 22732 skin patients with cancer and 5150 records from the test set. Ten-fold cross-validation is performed on a set of training data to improve accuracy of the model and avoid retraining. Nine Fine Tree, Medium Tree, and Coarse Tree models are produced using the Decision Tree approach. Six models are built using SVM and six using Ensemble Algorithms.

The minimal leaf size is the fundamental variable that is utilized to build various tree models. For the Fine Tree Model, Medium Tree Model, and Coarse Tree Model, respectively, the model parameters are 4, 12, and 36. The metrics used to evaluate the models have very slight differences in their values. The simulation results indicated that the models built using the Decision Tree have the best value of R-squared 0.99. The Decision Tree approach was shown to require the least training time, under 2 seconds, but the Support Vector Machine method requires training times ranging from 14. 691 to 177. 3 sec. Models Fine Tree, Medium Tree, and Coarse Tree are the quickest.

PROBLEM STATEMENT

Diagnosis and Stage Determination of CT Scanned Images of Lung Cancer using Hybrid model

- Lung Cancer is one the leading death causes globally.
- Thus there needs to be a model which detects the disease as well as determines the various stages of lung cancer.
- Earlier there were separate models each used for a particular purpose like one for diagnosing and the other one for determining stages.
- Our aim is to build a composed model which will serve both the purposes of detection of disease presence as well as determination of stages.

PROJECT REQUIREMENT SPECIFICATION

4.1 SOFTWARE AND HARDWARE REQUIREMENTS :

- **HARDWARE REQUIREMENTS:**

1. Processor: Intel Core 5
2. Minimum RAM: 8GB
3. Hard Disk: 500GB

- **SOFTWARE REQUIREMENTS:**

1. Anaconda Navigator
2. Google Collab
3. Kaggle dataset

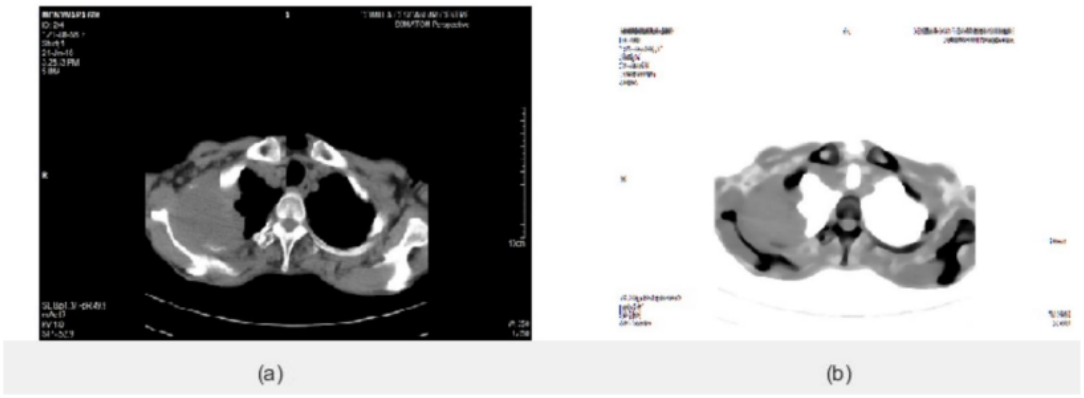
SYSTEM PROPOSED ARCHITECTURE

5. 1 IMAGE ENHANCEMENT:

Image enhancement is the process of adjusting digital images so that the results are more suitable for display or further image analysis. For example, to remove noise, sharpen, or brighten an image, making it easier to identify key features.

For this process Contrast Stretching is used, as it performs better on gray scale images. Contrast stretching (often called normalization) is a simple image enhancement technique that attempts to improve the contrast in an image by 'stretching' the range of intensity values it contains to span a desired range of values, the full range of pixel values that the image type concerned allows.

Figure 1: Contrast stretching image enhancement technique



a) Before Enhancement ***b) After Enhancement***

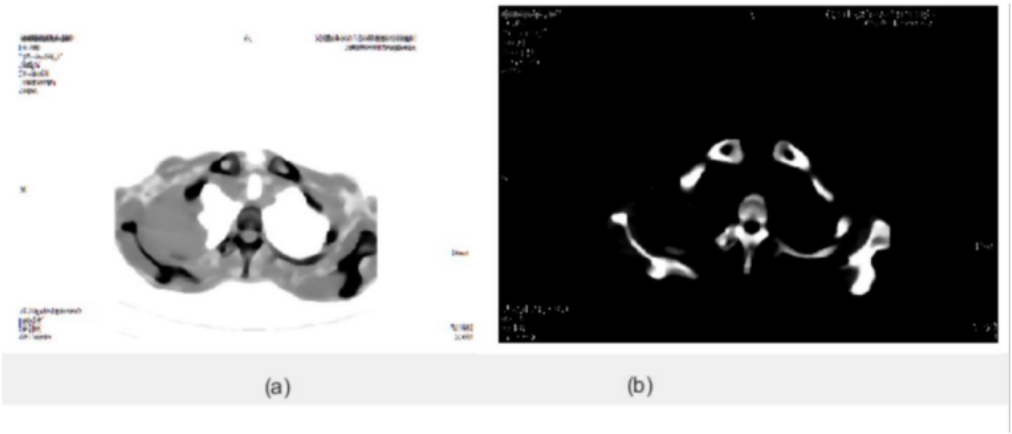
5.2 IMAGE FILTRATION:

Filter operation is performed on the image to increase the smoothness, sharpness as well as edge enhancement. It is a neighborhood operation, in which the value of any given pixel in

the output image is determined by applying some algorithm to the values of the pixels in the neighborhood of the corresponding input pixel. A pixel's neighborhood is some set of pixels, defined by their locations relative to that pixel.

Median filter applied in our planned method instead of mean filter or Gaussian filter. It is very widely used in digital image processing because, under certain conditions, it preserves edges while removing noise.

Figure 2: Median Filtration technique for image filtration



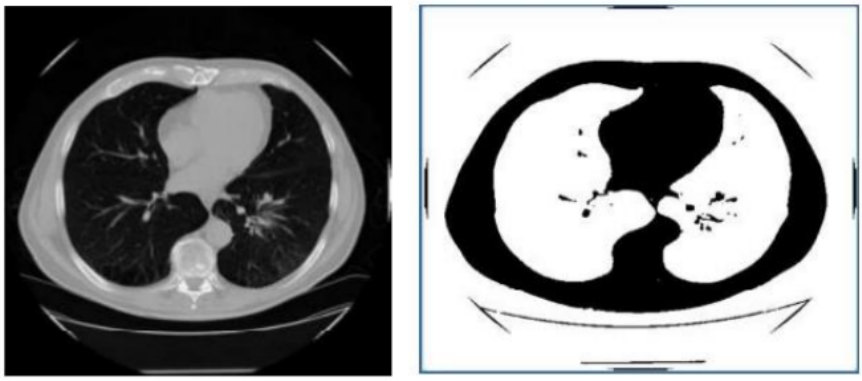
a) Before Filtration ***b) After Filtration***

5.3 IMAGE SEGMENTATION:

Segmentation is a technique to partition an image into multiple segments, used to partition an image into multiple parts or regions, often based on the characteristics of the pixels in the image.

For segmentation purposes, Otsu's method (global thresholding method) is used. It takes a smaller amount of time to compute the threshold value than other techniques. Converts images into binary images.

Figure 3: Otsu's method for image segmentation



a) Before Segmentation *b) After Segmentation*

5.4 FEATURE EXTRACTION AND CLASSIFICATION:

SVM will be used for both extraction and classification. SVM can be used for both feature extraction as well as classification basically by specifying a hyperplane. Mostly PCA and KDA as well as autoencoders, neural networks are used for feature extraction but using them separately and then combining them with SVM model leads to complexity in model.

Thus, we specified SVM as a single model which can be used for both extracting the features as well as classification. As this will directly import the extracting outputs to the classification step. This will help in reducing the complexity and overall weight of the model.

Figure 4: Extraction and Classification of image using SVM

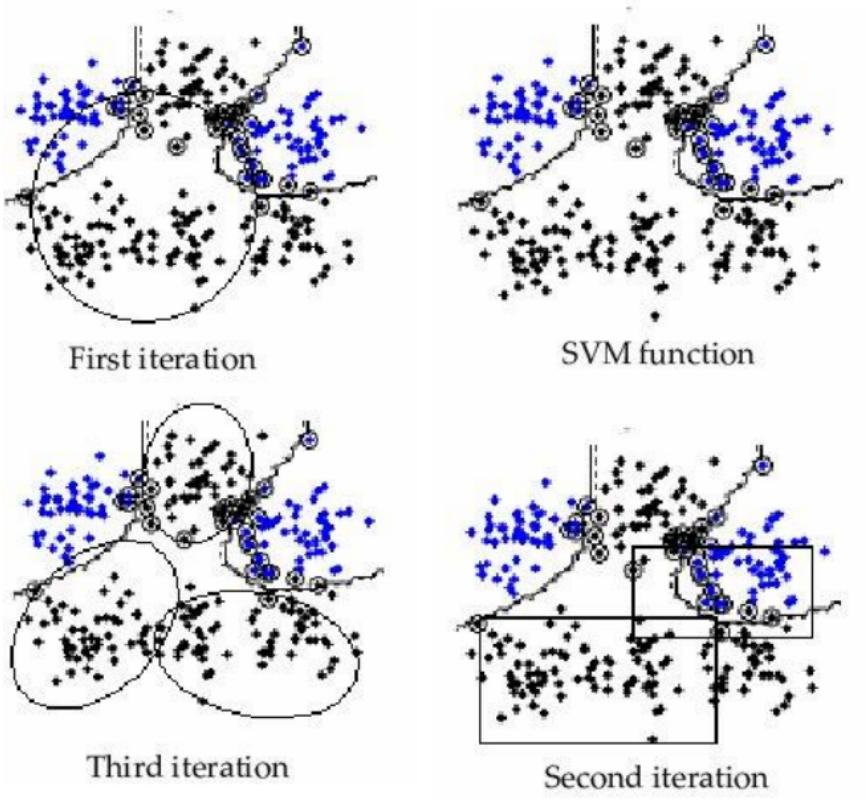


Figure 4.1: Extraction technique using SVM

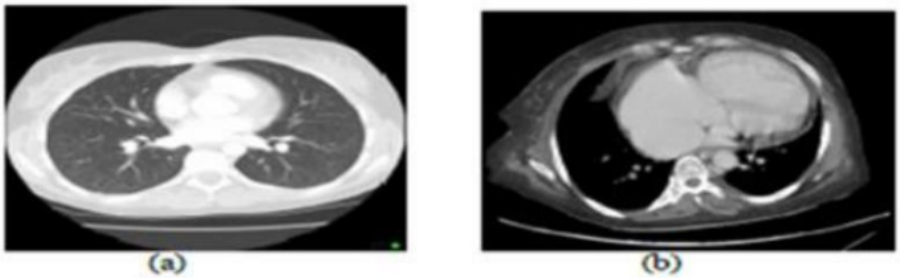


Figure 4.2: Classification using SVM technique

5.5 SUPPORT VECTOR MACHINE :

Support Vector Machines are a subpart of classification in the supervised learning set of Machine Learning Algorithms. Regression problems can also be solved using this method. The main aim of this algorithm is to identify a hyperplane in a particular space where it classifies the data points into classes.

SVM is also used for feature extraction where defining classes and hyperplanes one can extract the necessary data points as well as get the output in a segregated class format.

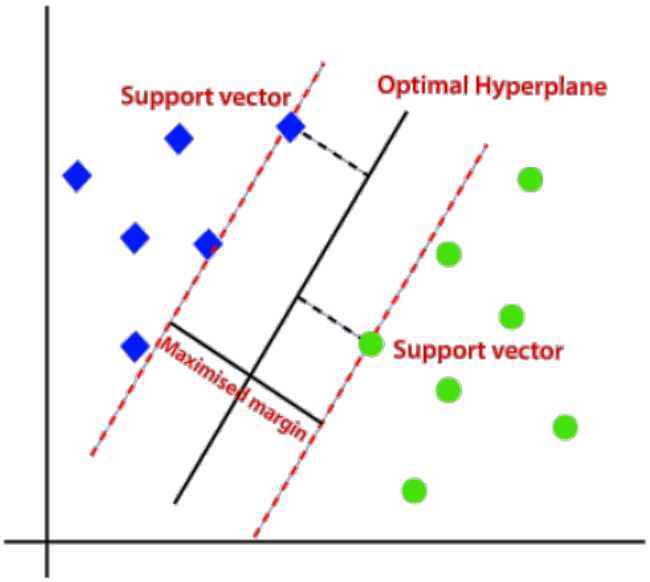


Figure 5: Support vector model specified an optimal hyperplane which divided data sets into 2 classes

5.6 DECISION TREES :

Decision Trees are mainly used for classification wherein the model classifies a complex instance into detailed and atomic instances. The root node is the main instance which further gets classified into branches basically called child nodes.

The Fine Tree Model is a subset of decision trees which uses threshold values for further detailed classification. It can range upto 100 splits in a particular go.

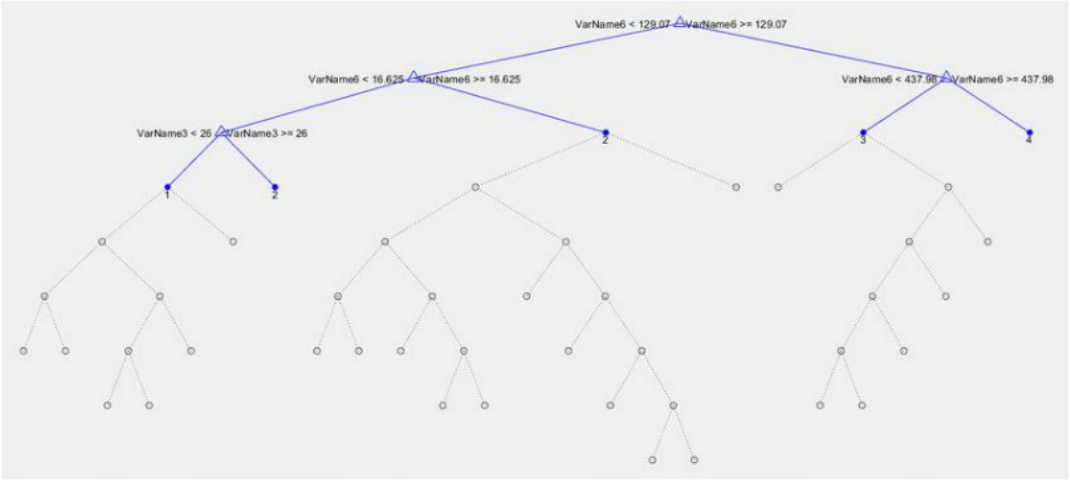


Figure 6: Fine Tree Decision Tree model segregating an instance with a defined threshold value

5.7 WEIGHTED AVERAGE :

Weighted average is a technique of calculation which estimates the varied important degrees of numbers or variables in a particular dataset. In this technique, each instance is multiplied by a weight which is predetermined and then final calculations are made and output is predicted.

This method accounts for greater accuracy as compared to simple average technique where all the instances are multiplied with a single identical weight.

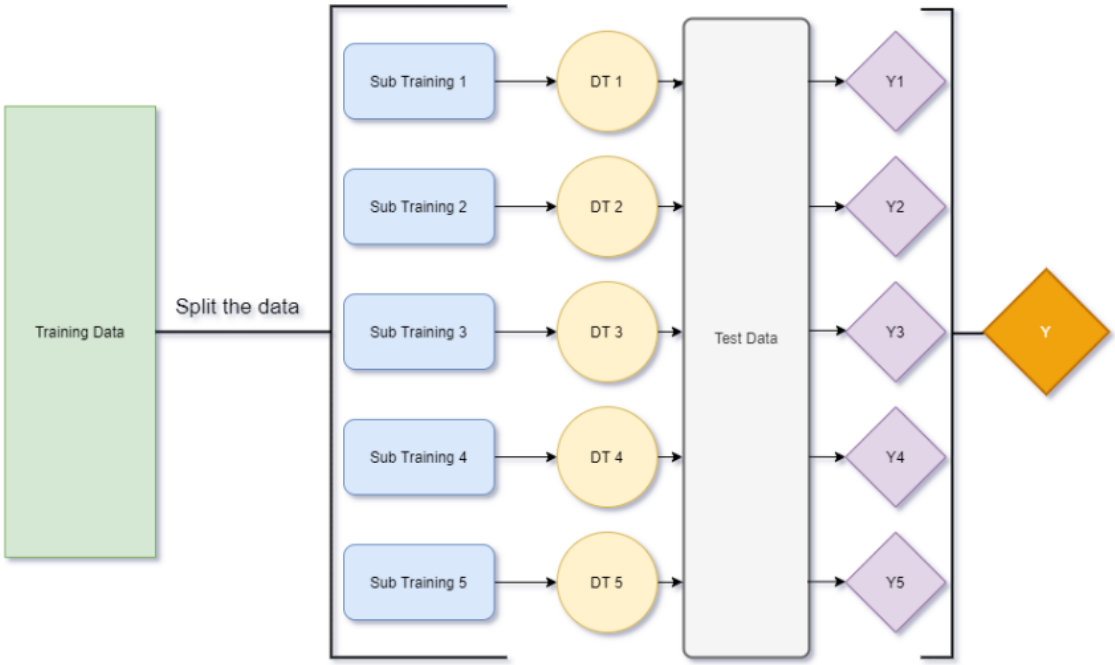


Figure 7: Weighted Average technique - Ensemble Model

HIGH LEVEL DESIGN OF THE PROJECT

6.1 SYSTEM ARCHITECTURE:

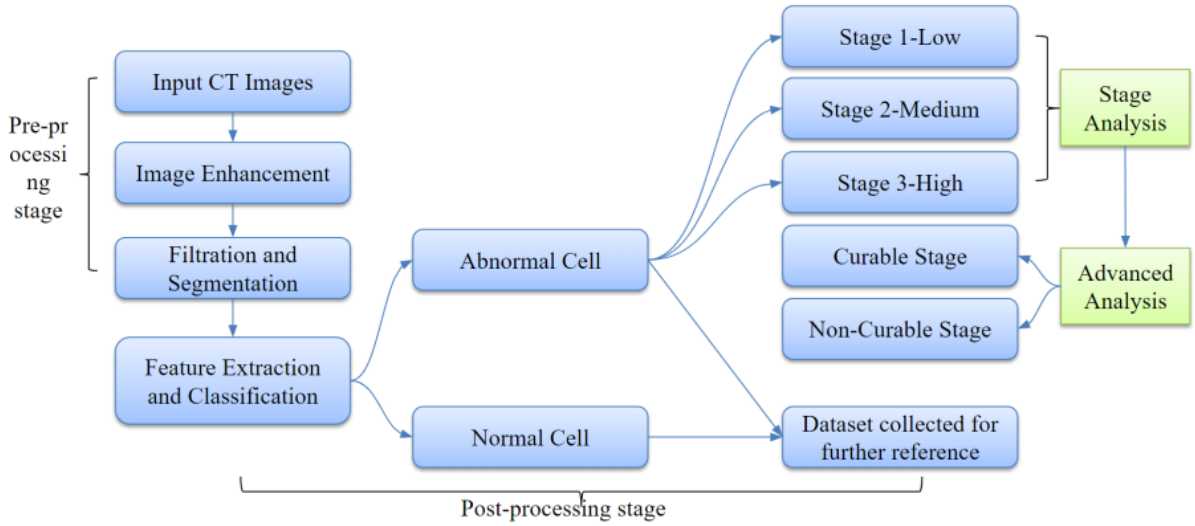


Figure 8. System Architecture

SYSTEM IMPLEMENTATION

7.1 FLOW OF SYSTEM:

1. Input data of CT images will be fed to the model.
2. During pre-processing, the model will enhance the images using Contrast Stretching. The enhanced images will be filtered as well as segmented using Median filter and Otsu's method respectively.
3. Feature extraction will be applied to extract certain images from the dataset as well as classify the interested regions on the basis of priority. Then classification of images as Normal and Abnormal Cells will be done. For both the Feature Extraction and Classification SVM will be used.
4. For the abnormal cells classified images, the image will be segregated on the basis of stages using decision classifier technique
5. Advanced segregation and mapping will also provide the curability extent of the disease.

7.2 UML DIAGRAMS

Unified Modeling Language is a standard language for writing software blueprints.The UML may be used to visualize,specify,construct and document the artifacts of a soft intensive system. UML is process independent,although optimally it should be used in process that is use case driven,architecture-centric,iterative,and incremental.

A Number of UML Diagrams are available.

- Use case Diagram.
- Activity Diagram.
- Sequence Diagram.

- Class Diagram.

7.2.1 USE CASE DIAGRAM :

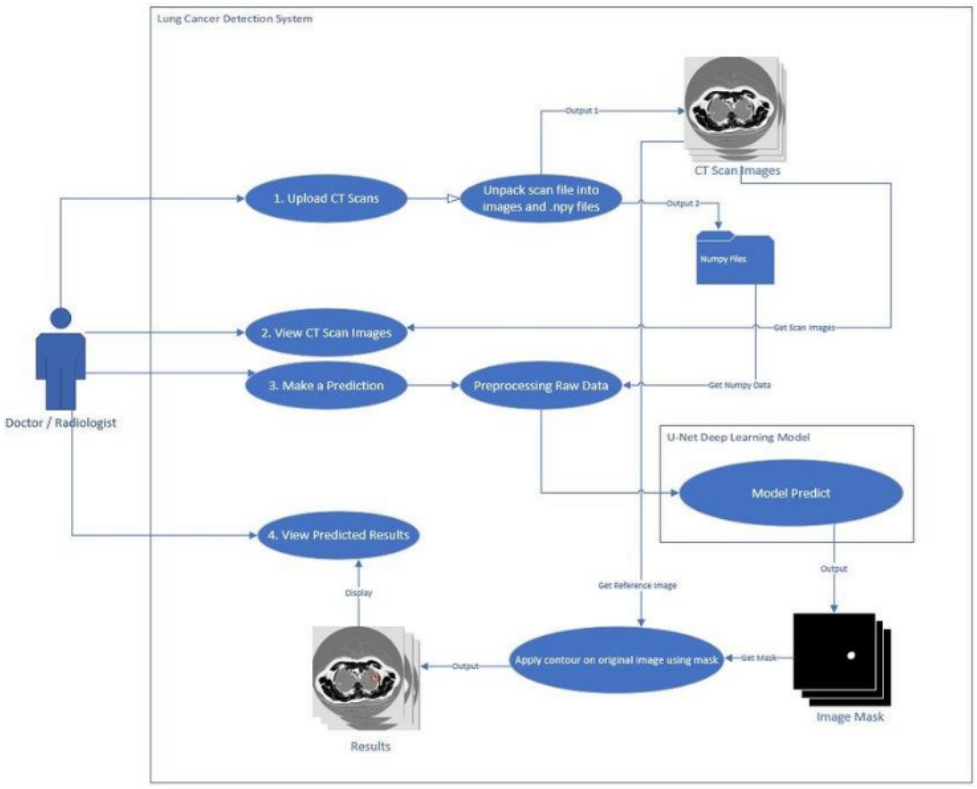


Figure 9: Use Case Diagram

A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved. A use case diagram can identify the different types of users of a system and the different use cases and will often be accompanied by other types of diagrams as well. The use cases are represented by either circles or ellipses.

7.2.2 ACTIVITY DIAGRAM:

Figure 10: Activity Diagram

Activity diagrams are graphical representations of workflows of step wise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams are intended to model both computational and organizational processes (i.e workflows), as well as the data flows intersecting with the related activities. Although activity diagrams primarily show the overall flow of control they can also include elements showing the flow of data between activities through one or more data stores.

7.5 SEQUENCE DIAGRAM:

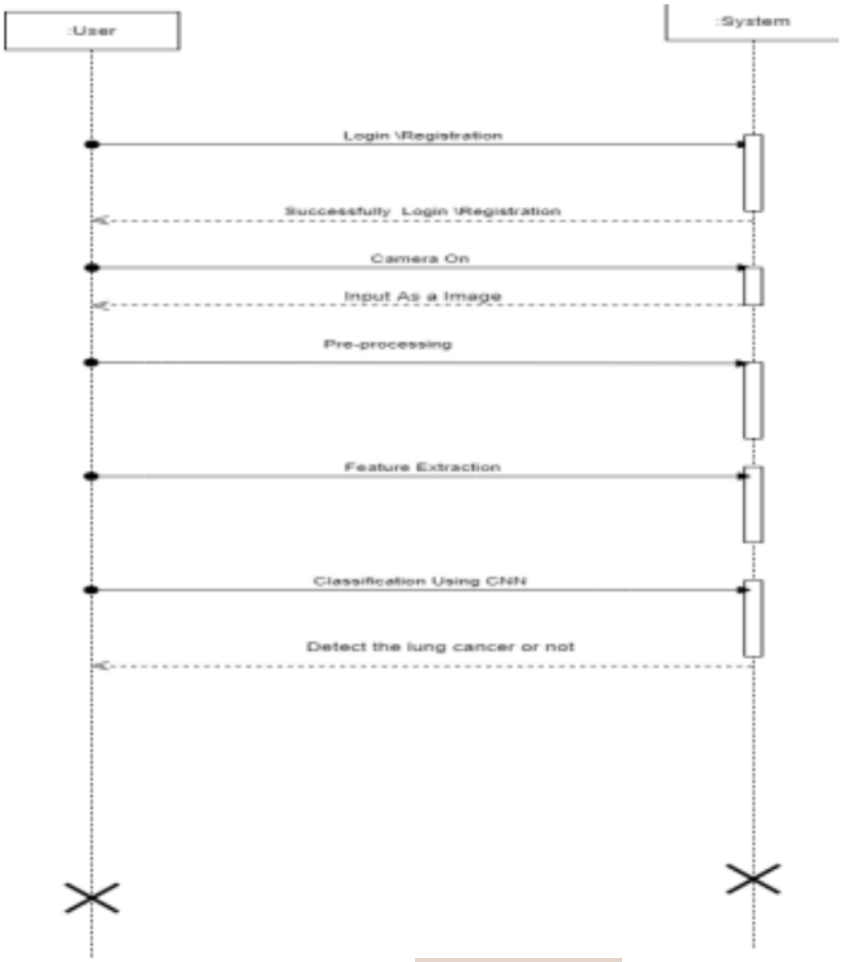


Figure 11: Sequence Diagram

A sequence diagram shows object interactions arranged in time sequence. It depicts the objects involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case realizations in the Logical View of the system under development. Sequence diagrams are sometimes called event diagrams or event scenarios.

7.2.4 CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects.

The class diagram is the main building block of object-oriented modeling. It is used for general conceptual modeling of the structure of the application, and for detailed modeling translating the models into programming code. Class diagrams can also be used for data modeling.[1] The classes in a class diagram represent both the main elements, interactions in the application, and the classes to be programmed.

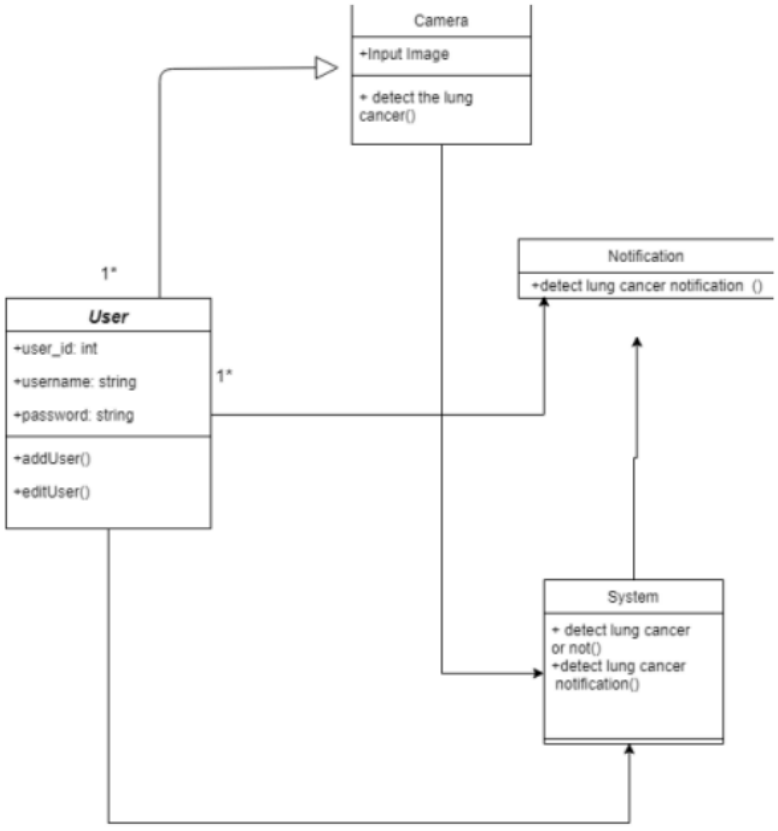


Figure 12: Class Diagram

7.3 DATASET DESCRIPTION :

Chest CT-Scan images Dataset [12]

No. of images: 1000

No. of classes: 4

Table 1: Dataset Description

Specification	Training	Testing	Validation
Adeno Carcinoma	195	120	23
Large Cell Carcinoma	115	51	21
Squamous Cell Carcinoma	155	90	13
Normal	148	54	15
Total	613	315	72

CONCLUSION AND FUTURE SCOPE

Support Vector Machines, Decision Trees and Enhanced Classifiers will be used in combination to generate precise results. Here, Image Enhancement is done using Contrast Stretching and Filtration is done using Median Filter. For segmentation, Otsu's method(Global Thresholding method) is used. SVM is used for Feature Extraction and Classification. Further Fine-Tree model which is a subset of the decision tree model is used for stage classification eventually helping in stage determination. Additionally enhanced classifiers like the weighted average technique will be applied to the output of decision trees thus classifying the curability of the disease. Also, the complexities of models are taken into consideration. The produced results will be detailed as well as highly accurate. The model building can be used for various other disease detection mechanisms by only changing the feature sets. This will help in providing a complete overview of the disease for a particular human being along with treatments and medicinal aids attached to certain stages.

REFERENCES

[1] M. Islam, A. H. Mahamud and R. Rab, "Analysis of CT Scan Images to Predict Lung Cancer Stages Using Image Processing Techniques," 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2019, pp. 0961-0967, doi: 10.1109/IEMCON.2019.8936175.

[2] A. Hoque, A. K. M. A. Farabi, F. Ahmed and M. Z. Islam, "Automated Detection of Lung Cancer Using CT Scan Images," 2020 IEEE Region 10 Symposium (TENSYMP), 2020, pp. 1030-1033, doi: 10.1109/TENSYMP50017.2020.9230861.

[3] C. Thallam, A. Peruboyina, S. S. T. Raju and N. Sampath, "Early Stage Lung Cancer Prediction Using Various Machine Learning Techniques," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1285-1292, doi: 10.1109/ICECA49313.2020.9297576.

[4] M. Todorova, "Application of Machine Learning Methods for Determining the Stage of Cancer," 2020 International Conference Automatics and Informatics (ICAI), 2020, pp. 1-4, doi: 10.1109/ICAI50593.2020.9311355.

[5] Juan Cui, Fan Li, Guoqing Wang, Xuedong Fang, J David Puett, and Ying Xu. Gene-expression signatures can distinguish gastric cancer grades and stages. PloS one, 6(3):e17819, 2011.

[6] P. Basak and A. Nath, "Detection of different stages of lungs cancer in ct-scan image using image processing techniques", International Journal of Innovative Research in Computer and Communication Engineering, vol. 5, pp. 9708-9719, 2017.

[7] Va Dominic, Dr. Deepa Gupta, Sangita Khare, and Aggarwal, Ab, "Investigation of chronic disease correlation using data mining techniques", in 2015 2nd International Conference on Recent Advances in Engineering and Computational Sciences, RAECS 2015, 2015.

[8] James D. Brierley BSc, MB, FRCP, FRCR, FRCPC, Mary K. Gospodarowicz MD, FRCPC, FRCR (Hon) Christian Wittekind MD, "TNM Classification of Malignant Tumours Eighth Edition", Union for International Cancer Control (UICC), Pages:1-241, 2017.

[9] A. D. Gunasinghe, A. C. Aponso and H. Thirimanna, "Early Prediction of Lung Diseases," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019, pp. 1-4, doi: 10.1109/I2CT45611.2019.9033668.

[10] D. Jayaraj and S. Sathiamoorthy, "Random Forest based Classification Model for Lung Cancer Prediction on Computer Tomography Images," 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), 2019, pp. 100-104, doi: 10.1109/ICSSIT46314.2019.8987772.

[11] Kaggle Dataset (Chest CT-Scan images Dataset) https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images

PROJECT REPORT (14_11_2022).docx

ORIGINALITY REPORT

32%
SIMILARITY INDEX

PRIMARY SOURCES

1	Ariful Hoque, A.K.M. Ashek Farabi, Fahad Ahmed, Md. Zahidul Islam. "Automated Detection of Lung Cancer Using CT Scan Images", 2020 IEEE Region 10 Symposium (TENSYMP), 2020 Crossref	285 words — 4%
2	"Computational Science – ICCS 2021", Springer Science and Business Media LLC, 2021 Crossref	252 words — 4%
3	www.academicscience.co.in Internet	170 words — 3%
4	Maya Todorova. "Application of Machine Learning Methods for Determining the Stage of Cancer", 2020 International Conference Automatics and Informatics (ICAI), 2020 Crossref	132 words — 2%
5	www.coursehero.com Internet	100 words — 2%
6	sjcit.ac.in Internet	80 words — 1%
7	repository.smuc.edu.et Internet	79 words — 1%
	hitechsolutions.info	

8	Internet	78 words — 1%
9	iosrjen.org Internet	61 words — 1%
10	www.ijartet.com Internet	56 words — 1%
11	link.springer.com Internet	50 words — 1%
12	anapub.co.ke Internet	49 words — 1%
13	Jayakumar K, Namdev Parth Deendayal, Gurnehmat Kaur Dhindsa, Agrim Nagrani, Vinay Bali. "CT Intensity Segmentation of Lungs", 2022 2nd International Conference on Intelligent Technologies (CONIT), 2022 Crossref	44 words — 1%
14	ijircce.com Internet	42 words — 1%
15	"Second International Conference on Image Processing and Capsule Networks", Springer Science and Business Media LLC, 2022 Crossref	40 words — 1%
16	ebin.pub Internet	40 words — 1%
17	www.sciencegate.app Internet	40 words — 1%
18	sciencescholar.us Internet	38 words — 1%

19	www.amrita.edu Internet	35 words — 1%
20	www.ijcaonline.org Internet	27 words — < 1%
21	www.jetir.org Internet	27 words — < 1%
22	www.ncbi.nlm.nih.gov Internet	27 words — < 1%
23	primalgrowpro.shop Internet	21 words — < 1%
24	"Research on Different Classifiers for Early Detection of Lung Nodules", International Journal of Recent Technology and Engineering, 2019 Crossref	19 words — < 1%
25	iats17.firat.edu.tr Internet	19 words — < 1%
26	Mahmudul Islam, Al Hasib Mahamud, Raqeebir Rab. "Analysis of CT Scan Images to Predict Lung Cancer Stages Using Image Processing Techniques", 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2019 Crossref	18 words — < 1%
27	"Data Analytics and Learning", Springer Science and Business Media LLC, 2019 Crossref	17 words — < 1%
28	"Computational Intelligence in Pattern Recognition", Springer Science and Business Media LLC, 2022 Crossref	15 words — < 1%

29	issuu.com Internet	15 words — < 1%
30	Sima Sarv Ahrabi, Alireza Momenzadeh, Enzo Baccarelli, Michele Scarpiniti, Lorenzo Piazzo. "How much BiGAN and CycleGAN-learned hidden features are effective for COVID-19 detection from CT images? A comparative study", The Journal of Supercomputing, 2022 Crossref	14 words — < 1%
31	easychair.org Internet	13 words — < 1%
32	repository.maranatha.edu Internet	13 words — < 1%
33	smart-jobz.blogspot.com Internet	13 words — < 1%
34	Chinmayi Thallam, Aarsha Peruboyina, Sagi Sai Tejasvi Raju, Nalini Sampath. "Early Stage Lung Cancer Prediction Using Various Machine Learning Techniques", 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020 Crossref	12 words — < 1%
35	gdeepak.com Internet	12 words — < 1%
36	Gur Amrit Pal Singh, P. K. Gupta. "Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans", Neural Computing and Applications, 2018 Crossref	10 words — < 1%
37	eprints.uthm.edu.my Internet	

10 words — < 1%

38 iict.bas.bg
Internet
10 words — < 1%

39 www.ec.tuwien.ac.at
Internet
10 words — < 1%

40 "Advances in Intelligent Computing and Communication", Springer Science and Business Media LLC, 2021
Crossref
9 words — < 1%

41 "International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications", Springer Science and Business Media LLC, 2018
Crossref
9 words — < 1%

42 Chunxi Zhang, Weijin Wu, Jia Yang, Jiayuan Sun. "Application of artificial intelligence in respiratory medicine", Journal of Digital Health, 2022
Crossref
9 words — < 1%

43 O. Jiménez, M. A. García, M. L. Marina. "Neural Network Capability for Retention Modeling in Micellar Liquid Chromatography with Hybrid Eluents", Journal of Liquid Chromatography & Related Technologies, 2006
Crossref
9 words — < 1%

44 P. S. S. Madhulika, Nalini Sampath. "Chapter 42 An Elaborative Approach for the Histopathological Classification of the Breast Cancer using Residual Neural Networks", Springer Science and Business Media LLC, 2022
Crossref
9 words — < 1%

45 T. M. Shahriar Sazzad, K. M. Tanzibul Ahmmed, Misbah Ul Hoque, Mahmuda Rahman.
9 words — < 1%

"Development of Automated Brain Tumor Identification Using MRI Images", 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019
Crossref

46	dokumen.pub Internet	9 words — < 1%
47	repository.ntu.edu.sg Internet	9 words — < 1%
48	thescipub.com Internet	9 words — < 1%
49	www.cnn.com Internet	9 words — < 1%
50	www.degruyter.com Internet	9 words — < 1%
51	123dok.com Internet	8 words — < 1%
52	Lecture Notes in Electrical Engineering, 2016. Crossref	8 words — < 1%
53	lib.buet.ac.bd:8080 Internet	8 words — < 1%
54	yarriambiack.vic.gov.au Internet	8 words — < 1%
55	Imen Hentati, Mohamed Ali Fourati, Nadia Trabelsi, Ibtissem Triki, Saïd Sassi, Moncef Zairi. "Decision Support System Development to Groundwater Management and Aquifer Vulnerability Assessment: Hydrogeological	6 words — < 1%

Information System of Monastir (HISM)", Journal of Geographic Information System, 2019

Crossref

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	OFF	EXCLUDE MATCHES	OFF