International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT22812

Stage Classification of Lung Cancer using the Comparative Analysis of the Machine Learning Techniques

[1]V. Deepa, [2]P. Mohamecl Fathimal

[1]Research Scholar, SRM Institute of Science and Technology, Vadapalani, Chennai, Tamil Nadu, India

[2] Assistant Professor, SRM Institute of Science and Technology, Vadapalani, Chennai, Tamil Nadu, India

Article Info
Volume 8, Issue 1
Page Number : 32-41

Publication Issue :
January-February-2022

Article History
Accepted : 02 Jan 2022
Published : 12 Jan 2022

ABSTRACT

Stage classification today is widely used in the fields of biological sciences and engineering, The idea of the stage classification is to perform a clinical analysis of the spread of the disease . Lung cancer which is termed as Lung Carcinoma is a highly dangerous lung tumour which is defined by the uncontrolled growth of cells in the tissues of the lung. This growth of cells leading to the tumour identifies the different stages of cancer. The tumours are identified based on the probability density function. The goal is to design models for the stage classification of the of cancer patients . The description of the extent of a tumour consists of three components: T for extent of the primary tumour, N for involvement of lymph nodes, and M for distant metastases. Each T, N, and M component is divided into several categories (eg, T1, T2). This study proposes to build a classification system that can identify the stage classification using the lung cancer dataset for better accuracy. An "IQ-OTHNCCD" lung cancer dataset of 1190 images representing CT scan slices of 110 cases is used in this research.

Keywords: Lung Cancer, Stage Classification, Machine Learning Comparative Analysis

I. INTRODUCTION

After a patient is diagnosed with lung cancer the doctors will try to find out the extent of the spread is called **as staging.** The stages of a cancer helps in identifying the level of spread in the body. Doctors make use of cancer's stage in generating the survival data of the cancer patients. The important key parameters in cancer disease is the extent of the size of the tumour, the spread of the lymph node, the spread of the cancer called as metastasis, estrogen status and the progesterone status. We define the survival time as a period between the diagnosis of disease and death of the patient. The survival analysis is predicted at the last stage of the cancer in which the disease is aggressive.

1. 1 Category of patients with stage I Cancer:

In stage I cancer it is associated with decreased survival rate of the patients. The chemotherapy was done at the less survival level of the patients. For patients over age 50 years old with stage I cancer the treatment is mandatory to reduce the hazard.

1. 2 Category of patients with stage II Cancer:

For younger patients with stage II cancer, surgery and radiation therapy are significant. The greatest risk is related to the radiotherapy method. For patients older than 50 years with stage II cancer, surgery and radiation therapy are important when the risks associated with the radiation therapy method used are highest. In this category patients are associated with a decrease in patient survival.

1. 3 Category of patients with stage III Cancer:

For stage III patients younger than 50 years, surgery and cancer treatment are important. The best risks associated with surgery helps the patient reduce the risk. For patients older than 50 years with stage IIII cancer, variable surgery is of great importance. The optimal risk associated with surgery is percentage of people undergoing surgery is lower . Each treatment and surgery is linked to improving the survival rate of patients with stage III cancer. But the risk rate of person undergoing surgery at the United Nations agency is reduced.

1. 4 Category of patients with stage IV Cancer:

For the patients above the fifty with stage IV cancer the mortality rate is more. The Hazard ratio is more for the cancer patients in the last stage. Cancer therapy and surgery are used to improve the survival rate of the patients. The stages of cancer can be classified as the limited stage and the extensive stage. The limited stage includes the cancer found in the one side of the chest and the lymph nodes. The extensive stage includes the cancer spread along the lung and lymph nodes. Lungs are sponge like structure available in the chest area. when we breathe inside the air enter through the mouth and goes into the lungs through windpipe(trachea). The alveoli absorbs the oxygen from the inhaled air. Lung cancers starts at the cell lining called as bronchi . The lung cancer can be classified in three ways localized namely Regional and distant. Based on the incidence and the aggressiveness the lung cancer is classified as Lung adenocarcinoma, Squamous cell lung carcinoma, Large cell lung carcinoma. The stages of cancer can be divided as the TNM stages.

1) T (tumour),
2) N (node)
3) M (metastasis):

The different stages of cancer have been represented in tabular format. The x-ray provides the better analysis of the tumour. The Hazard ratio is more for the aged people. The mortality is more at the last stage of the cancer patient. the noise observed in the CT-Images can be handled using the various techniques like smoothing. The classification algorithms are used to improve the clarity of the cancer images.

Table 1. Stages of cancer and spread level

S. no	Stages	Spread level
01	Stage I	cancer located at lungs not spread in the lymph nodes
02	Stage II	cancer located in the lungs and nearby lymph nodes
03	Stage III	cancer located in the lungs and nearby lymph nodes and in the middle of the chest denoted as IIIA, IIIB
04	Stage IV	Advanced stage of lung cancer where the fluid spreads around the all parts of the organs.

II. Dataset Description

The Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD) lung cancer dataset includes CT scans of patients diagnosed with lung cancer . The dataset contains of a total of 1190 images representing CT scan slices of 110 cases.

These cases are grouped into three classes: normal, benign, and malignant. The CT scans were originally collected in DICOM formats.

III. Proposed work

The aim of the research is to preprocess the image , applying classification algorithms such as Convolution Neural Network (CNN) and Artificial Neural Network (ANN). , usage of the Comparison and evaluation techniques for the applied algorithm with the best performing model. The first step is to acquire a ct scan image of the lung cancer in which a lot of noise are observed in the CT -scan Image. In order to improve the clarity of the picture preprocessing techniques are used . Hence various techniques like smoothing and enhancement is done.

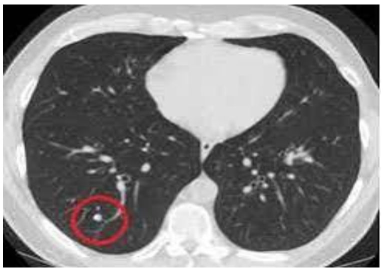


Fig 1. CT-scan of the Lung cancer Image

The purpose of the pre processing is to improve the image data and unwanted distortion. Filtering techniques are used for removing the salt and pepper noise. Image enhancement is done in order to get a better quality of the picture . Image enhancement is classified on to spatial and frequency domain.

Feature Extraction

Feature extraction is an important part to decide the stage of the cancer . The features extracted from the lung nodules are area, perimeter and eccentricity

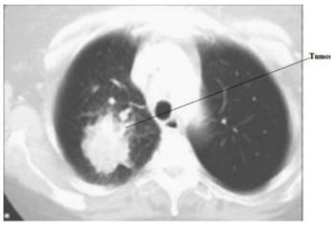


Fig 2 . Tumour in Lung CT Image

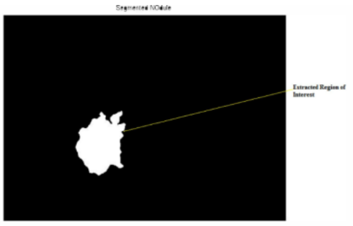


Fig 3 Segmented Nodule for feature extraction.

The summation of the white pixels represents the extracted region of interest.

Stage Identification :

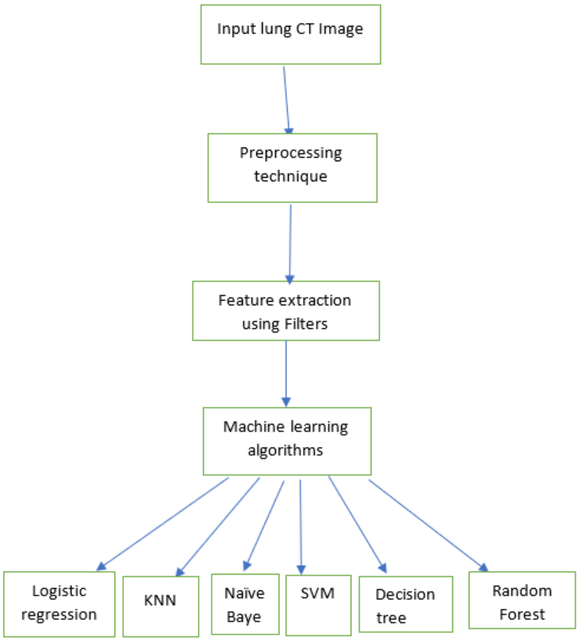
Depending on the areas of the malignant nodule the stage classification of the cancer can be identified. Stage classification is a very important since it helps the doctor to predict the extent of the disease. In TNM staging T denotes the size of the tumour;N denotes the spread of the lymph nodes and M denotes metastasis. Based on the values obtained from the feature extraction the stage classification of the cancer can be done. The stage of the cancer increases if the values of the feature metrics are more.

IV. Methodology

Initially the Lung CT Images are preprocessed and Segmented. The next stage is feature extraction using the filters. The third stage is classification using the optimisation algorithm.

The objectives of the TNM classification are

- An id treatment planning,
  Provide an indication of prognosis,
- Assist in the evaluation of treatment results,
- Facilitate the exchange of information between treatment centres,
- Contribute to continuing investigations of human malignancies,
- Support cancer control activities, including through cancer registries.



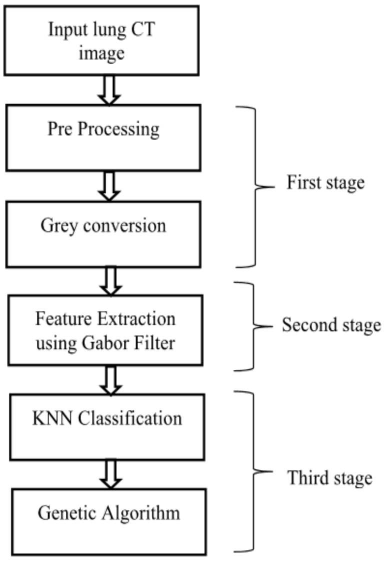


Fig 3 Stages of proposed Algorithm

V. Discussion

In the first stage the lung regions are extracted from CT image and in that region each slices are segmented to get tumours. The segmented tumour regions are used to train the CNN models. The main objective of this study is to detect whether the tumour present in a patient's lung is normal or malignant or benign . The image segmentation partitions the image in to different segments. It helps in identifying foreground and background marking of the data. The different features extracted are the area, perimeter and eccentricity based on the complexity of the tumour. The Edge based segmentation is the most commonly used in medical Analysis. The Stage classification on lung cancer can be done based on the size of the tumour. Tumour is classified into different sub groups. T1(<=3cm) is again divided into T1a and T1b, T2 (>3 cm) is again divided into T2a and T2b which is again divided in to T3. The subdivision of the T1 tumour size into T1a and T1b does not affect the stages of the lung cancer. cases with tumour size greater than 3 cm and no lymph node involvement or distant metastases that were classified as stage IB. When tumour size is greater than 3 cm with lymph node involvement but no distant metastases they are classified as T2a, T2b, or T3. Initial stages are identified as Stage I. In the case of the multiple tumour nodules were subdivided T4 to T3. Multiple stages are classified as stage II, III. Distant metastasis(M) is defined as M1a and M1b. Both M1a and M1b are considered stage IV.

VI. Experimental Setup

The Image segmentation appropriate for medical Images is the canny Edge detection algorithm. The purpose of the edge detection is to reduce the amount of data in the image. It helps in identifying the image brightness from the typically organised curved line segments . The following fig shows the original Lung cancer ct image and the Edge detected ct Image .

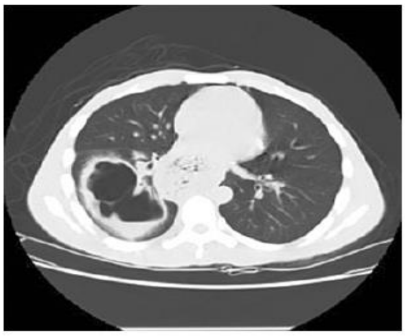


Fig . Original CT -Image

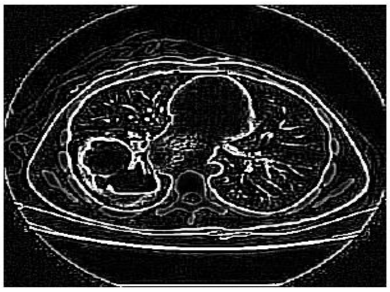


Fig . Canny Edge detected Image

In this paper canny Edge detection for the medical Images is proposed . Canny edge detection algorithm is applied to produce the final Image . It is the optimal and widely used as edge detection technique in research.

Results

The TNM stages of the Lung cancer represents the Lymph node spread of the cancer cells. The lymph node denotes the control activity of the cancer registry. When the tumour size is greater it is classified as T1a, T2b. It represents the cancer cell description in the neighbouring cell damage in the inner lining of the lungs. It facilitates the evaluation of the treatment plans . It is indication of the treatment centres . The Lymph node of the evaluation of the Treatment of the control activity of the Logistic regression of the registry. The lymph node of the control activity of the Naive Bayes . Indication of the Prognosis of the Treatment data.

Description of the TNM Descriptors

Stage	Stage Grouping	Stage Description
Stage 0	Tis N0 M0	Cancer cells are only in the inner lining of your lungs.
Stage 1	T1 mi N0 M0	The cancer is not spread to the lymph nodes or to other parts of your body.
Stage 2	T2b N0 M0	The tumour is between 4 and 5 cm across the membrane of the lungs
Stage 3	T2a/T2b N2 M0	The tumour may be between 3 and 5 cm across. lungs.

Comparative analysis of the Classifier accuracy obtained for the different algorithms

Sno	Algorithm	Accuracy
01	**Logistic regression**	**97%**
02	**K-nearest neighbour**	**75%**
03	**Naïve Bayes method**	**88%**
04	**Support vector machine**	**96%**
05	**Decision tree**	**93%**
06	**Random forest**	**82%**

Techniques used for survival analysis

Non-Parametric, Parametric and Semi-parametric Analyses:

Non-parametric survival analysis is used to analyze the data avoiding assumptions for the different distributions. This kind of analysis restricts the data from occurrence of errors. One of the commonly used non-parametric estimator is Kaplan-Meier estimator which is also called as product limit estimator. The disadvantage of the non-parametric analysis is for comparing the survival functions for limited number of groups.

Non-Parametric, Parametric and Semi-parametric Analyses:

Non-parametric survival analysis is used to analyze the data avoiding assumptions for the different distributions. This kind of analysis restricts the data from occurrence of errors. One of the commonly used non-parametric estimator is Kaplan-Meier estimator which is also called as product limit estimator. The parametric analysis can be carried out with regression parametric model and Proportional hazard model. Cox regression models or PH models are used for the estimation of survival time which helps in making assumptions to hazard function in the given formula. The disadvantage of the non-parametric analysis is for comparing the survival functions for limited number of groups. The survival analysis is done using the SEER Data. The population set is computed using the statistical analysis method.

Experimental Setup

ON-LINE LUNG CANCER OUTCOME CALCULATOR

We use a online tool to predict the lung cancer outcome calculator . The goal of the tool is to make use of the 5 outcome variables and to remove the redundant attributes. The calculator uses the 13 Input variables to estimate the mortality of lung cancer. The original attributes used in the SEER database is represented as follows:

1. **Age at diagnosis:** The age of the patient at the time of diagnosis for lung cancer should be given as numeric value.
2. **Birth place:** The birth town of the patient. There are 198 options available to select for the attribute from the SEER database
3. **Cancer grade:** A description of how the cancer cells may grow or spread. the different options available are well differentiated, moderately undetermined.
4. **Diagnostic confirmation:** It denotes the confirmation of the lung cancer by different methods like positive histology, positive cytology, positive microscopic confirmation , positive laboratory test and other clinical diagnosis.
5. **Farthest extension of tumour**: The spread of the tumour from the local region to the metastasis region . There are 20 options available. it can be represented as the localised or the lymphatic region. The name of the attribute is represented as 'EOD extension'.
6. **Lymph node involvement:** The highest specific lymph node chain that is involved by the tumour. There are 8 options available for this attribute. The name of the attribute is represented as 'EOD Lymph Node Involv'.
7. **Type of surgery performed**: The first step of the therapy was to removes or destroys cancerous tissue of the lung. There are 25 options available for this attribute.
8. **Reason for no surgery**: The different reasons why surgery was not performed . The Available options are - surgery performed, surgery not recommended . This helps in reducing the surgery operation.
9. **Order of surgery and radiation therapy:** It denotes the the order in which surgery and radiation therapies were available. the different options are available.
10. **Scope of regional lymph node surgery**: It describes the surgical procedures like removal of lymph nodes at the time of surgery, biopsy. There are 8 options available for this attribute.
11. **Cancer stage:** This attribute describes the spread of the cancer such as size of the tumour, Region spread like localised or distant . The cancer stage helps in detecting the spread of the disease at different places
12. **Number of malignant tumors in the past:** An integer denoting the total number of malignant tumors in the patient's lifetime so far. It denotes the both numeric and categorical values for both malignant and benign tumours within a single attribute. This attribute is mainly used for detecting the complexity of the tumour.

13. **Total regional lymph nodes examined**: An integer denoting the total number of regional lymph nodes

that were computed in order to identify the complexity of the tumour. The data is represented in the data from one node to another this attribute is used for the extension of the tumour region.

S. No	Model Name	Data Set Source	Data used for Analysis	Approach used	Data Preprocessing Technique	Analysis of What	Parameters used	Outcome
1	Pyspark Dataset	Lung cancer dataset	Image	AI approach	Clinical Data	Analysing the Lung cancer CT Images using the SVM method	Parameters used are Curvature,perimeter and NC ratio	Achieved 86% accuracy using the SVM method
2	COSMIC and TCGA Dataset	Lung cancer dataset	Text	AI approach	Clinical data	Analysis of the mutated gene using the classification model	The different parameters are accuracy, sensitivity, specificity, G-mean, F-score value	The comparision analysis of the different ml algorithms were done .The proposed method increased the survivability of the lung cancer .The best results were obtained using the SVM,RBF Models
3	Boston Lung cancer cancer Dataset	Lung cancer Patient data collected from the t Massachusetts General Hospita	Image	AI approach	Clinical data	Analysis of the Tumour history using the Lung cancer	The different parameters used AUC , Accuracy, Specfcity Sensitivity p	The usage of CNN in predicting histology in early-stage NSCLC patients. Lesion analysis using the CNN.
4	IASLC Lung Cancer Staging	National cancer database from the United states	Image	AI approach	Clinical data	Analysing the physical examination, laboratory tests, and imaging to determine the stage and classify it to TNM classification	Tumour size,Lymph node,Metastasis and Age groups	The different modification were done in the TNM classification of the Lung cancer
5	LIDC-IDRI (888 samples)	Three NSCLC datasets for the clinical stages of the Lung tumour	Image	AI approach	Clinical data	Multi stage classification of the Lung cancer	The different parameters used are the AUC , Accuracy, Specificity Sensitivity p	It focusses on the T-Staging of the Lung cancer
6.	Luna -16 Dataset	Lung Nodule analysis of the Lung cancer	Image	AI approach	Clinical data	Analysis of the lung nodules are classified and malignancy level is detected with the accuracy of 95% and log loss of 0.38	The different parameter accuracy,Log loss and Dic coefficient	Improving the efficiency of the Lung nodule Detection and malignancy level using the CT -Images
7	LCDS System	A proposed model	Image	AI approach	Clinical data	Analysis of the TNM stage classification	Image proprocessing, Image segmentation and Feature extraction	TNM stage classification from the CT SCAN Images using the DWT algorithm using the 98% accuracy
8.	CT scan Images	Image processing	Image	AI approach	Clinical data	Analysis of the nodule classification from the proposed system such as Malignant ,Benign,Normal	Accuracy,sensitivity, specification,F-Score	Acuracy of 96% was achieved using the SVM method.The proposed method perfomes the stage classification of the Lung cancer.
9.	Lung cancer Survival using the neural network classifier	Proposed system of the Lung cancer	Image	AI approach	Clinical data	Analysing the Lung cancer Images using the Binary Thresholding technique	Gray scale conversion, Normalization,Noise reduction,Binary Image , removal of the unwanted portion of the Image .	The proposed system introduce a binary thresholding technique, strong feature extraction method and compare to other existing system in order to improve the better performance of the Lung Cancer system.

11	LIDC Dataset	Collection of the Lung cancer CT Images in DICOM Image format	Image	AI approach	Clinical data	Analyse the early detection of the Lung cancer using the CT Scan Images	Accuracy ,sensitivity and specificity	CT scan image is used as input image, is processed and early stage lung cancer is detected using an SVM (support vector machine) algorithm as a classifier in the classification stage for the purpose of improving accuracy, sensitivity and specificity.
12	A Novel hybrid model is proposed	Swarm Optimisation techniques	Image	AI approach	Clinical data	To analyse the lung cancer mortality. CT scan images are being used they are analyzed by radiologists to recognize and identify nodules into malignant and benign nodules	Sensitivity,specificity, accuracy, Entropy and corelation	A new strategy to early detection, prediction and diagnosis has been introduced in order to improve patient safety and mitigate using the SVM Method
13	Bayes Theorem for the Lung cancer analysis	Collection of the Lung CT Images	Image	AI Approach	Clinical Data	Texture analysis Lung cancer extraction and classification based on feature extraction	Accuracy,sensitivity and Specificity	The performance is evaluated by CT lung images. Then the maximum likelihood classifier is used for prediction. The classification accuracy is 95%.
14	LIDC Dataset	Collection of the Lung CT Images	Image	AI Approach	Clinical Data	An image processing techniques has been used to detect early stage lung cancer in CT scan images. The CT scan image is pre-processed followed by segmentation of the ROI of the lung.	Accuracy,sensitivity and Specificity	The accuracy is detected using the SVM Method. A Receiver Operating Characteristics (ROC) curve is used to analyze the performance of the system. Overall the system has accuracy of 95.16%, sensitivity of 98.21% and specificity of 78.69%.
15	Statistical analysis of the Histopathological study	Study population of the Lung cancer patients	Text	AI approach	Clinical Data	Analysis of the patient characteristic	The size and histological type of the tumour, its l extent of the performed surg number of metastatic lymp and presence of skip metastas	5 year survival analysis was done using the N2 features using the Histopathological study.
16	LUNA -16 Dataset	Lung cancer Dataset	Text	AI approach	Clinical Data	Analysis of the patient-related information such as scans like CT-Scan, X-Ray, MRI Scan, unusual symptoms in patients or biomarkers, etc.	The different parameter area,perimeter and eccentricity was analysed	The proposed method helps in detecting cancer using the SVM method and accuracy was obtained .
17	TCGA Dataset	Information collected from Cnacer Genome atlas information	Image	AI approach	Clinical Data	Our analysis demonstrates the potential to provide significant prognostic information in multiple cancer types, and even within specific pathologic stages	To compute the survival loss ,function cox propotional hazard model was used .	Comparing survival rates in low and high risk groups was done for lung cancer .
18	LIDC Dataset	Dataset on the lung nodules	Image	AI approach	Clinical Data	To analyse the Randomly selected samples using the bagging technique	Pixel Intensity,mean value, wavelet feature,region property	Detecting the nodules to be cancerous or non-cancerous. Best results achieved were 90.73% for accuracy, 90.8% for specificity and 90.67% for sensitivity. the ML algorithm used is the Random forest which is the best Decision tree algorithm

19	Dicom Image Dataset	The CT-Scan images of lung cancer patients	Image	AI approach	Clinical Data	To analyse the location of the tumour using the lung cancer dataset.	Sphericity,Roundness ,Indenations.	It is to classify the types of lung tumours for extracted and selected features using learning algorithms. classify the lung tumours as benign, malignant for extraction and selected features using Learning algorithm.
20	LIDC Dataset	Lung nodule detection dataset	Image	AI approach	Clinical Data	To analyse the lung nodules using the size of the tumour	Size ,complexity of the cancer cells	Different classification algorithms was used to check the accuracy such as the K-nearest neighbour,Naïve Bayes and Support vector machine .

VII. CONCLUSION AND FUTURE WORK

As the detection of the Lung cancer increased the CT scans tend to increase across the patients.The performance of the medical image can be analysed based on different parameters like accuracy,loss and computation time. Accuracy:It is a parameter used to measure the accuracy of the model. loss Function : The error in the neural network is called as the loss function. computation time : The time required to complete the process is called as the computation time. Machine learning has made an enormous impact in the health industry with the usage of the various algorithms. In this paper, we conclude that better accuracy can be achieved with the less time.

VIII. REFERENCES

[1]. Adarsh Pradhan, Bhaskarjyoti Sarma, Bhiman Kr Dey "Lung Cancer Detection using 3D Convolutional Neural Networks" 2020 International Conference on Computational Performance Evaluation (ComPE) North-Eastern Hill University, Shillong, Meghalaya, India. Jul 2-4, 2020.

[2]. S. Sasikala, M. Bharathi, B. R. Sowmiya "Lung Cancer Detection and Classification Using Deep CNN" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-2S December, 2018

[3]. ParnianAfshar, AnastasiaOikonomou, Farnoosh Naderkhani, Pascal N.Tyrrell, Konstantinos N. Plataniotis, Keyvan Farahani & Arash Mohammadi "3D-MCN: A 3D Multi-scale Capsule Network for Lung Nodule Malignancy Prediction" Scientific Reports | (2020) 10:7948

[4]. Ola Mohammed Abu Kweik, Mohammed Atta Abu Hamid, Samer Osama Sheqlieh, Bassem S. Abu-Nasser, Samy S. AbuNaser "Artificial Neural Network for Lung Cancer Detection". International Journal of Academic Engineering Research (IJAER) ISSN: 2643-9085 Vol. 4 Issue 11, November – 2020

[5]. Chun-Hui Lin, Cheng-Jian Lin, Yu-Chi Li, Shyh-Hau Wang "Using Generative Adversarial Networks and Parameter Optimization of Convolutional Neural Networks for Lung Tumor Classification" Appl. Sci. 2021, 11, 480.

[6]. R. Sujitha and V. Seenivasagam "Classifcation of lung cancer stages with machine learning over big data healthcare framework" 30 April 2020 © Springer-Verlag GmbH Germany, part of Springer Nature 2020

[7]. Varsha Prakash and SmithaVas.P "Survey on Lung Cancer Detection Techniques" 2020 International Conference on Computational Performance Evaluation (ComPE) North-Eastern Hill University, Shillong, Meghalaya, India. Jul 2-4, 2020, ©2020 IEEE

[8]. Mehdi Hassan Jony, Fatema Tuj Johora, Parvin Khatun and Humayan Kabir Rana “Detection of Lung Cancer from CT Scan Images using GLCM and SVM” (ICASERT 2019), ©2019 IEEE.

[9]. Ahmet Kadir Arslan, Şeyma Yaşar and Cemil Çolak “An Intelligent System for the Classification of Lung Cancer Based on Deep Learning Strategy”, 2019.

[10]. Ibrahim M. Nasser, Samy S. Abu-Naser “Lung Cancer Detection Using Artificial Neural Network”. International Journal of Engineering and Information Systems (IJEAIS) ISSN: 2000-000X Vol. 3 Issue 3, March – 2019

[11]. YAN KUANG Unsupervised Multi-Discriminator Generati ve Adversarial Network for Lung Nodule Malignancy Classification”-2020.

[12]. Preeti Katiyar, Krishna Singh “A Comparative study of Lung Cancer Detection and Classification approaches in CT images” 2020, 7th International Conference on Signal Processing and Integrated Networks (SPIN)

Cite this article as :

V. Deepa, P. Mohamecl Fathimal , "Stage Classification of Lung Cancer using the Comparative Analysis of the Machine Learning Techniques", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 1, pp. 32-41, January-February 2022. Available at doi : https://doi.org/10.32628/CSEIT22812
Journal URL : https://ijsrcseit.com/CSEIT22812