An Effective Early Stage Detection of Lung Cancer Using Fuzzy Local Information cMean and GoogLeNet

Sobia Shafiq
Department of Computer Science
National University of Modern Languages
(NUML), Rawalpindi 46000, Pakistan.
sobia.shafique@numl.edu.pk

Muhammad Adeel Asghar
Department of Computer Science
National University of Modern Languages
(NUML), Rawalpindi 46000, Pakistan.
adeel.asghar@numl.edu.pk

Muhammad Emad Amjad
Department of Computer Science
National University of Modern Languages
(NUML), Rawalpindi 46000, Pakistan.
emad.amjad@numl.edu.pk

Jawwad Ibrahim
Department of Computer Science
National University of Modern Languages
(NUML), Rawalpindi 46000, Pakistan.
jawwad.ibrahim@numl.edu.pk

***Abstract*—Cancer is one of the main causes of death worldwide, accounting for an incredible 5 million fatalities per year. In this article, innovative machine learning algorithms are used to detect lung cancer at an early stage. To extract features, computed tomographic scan images were used. In the initial stage of lung nodule, preprocessing is accomplished for data cleaning and resizing of dataset. In the second stage, a set of features was recovered from the preprocessed images using Fuzzy Local Information cMean (FLIcM). Aside from this, deep features were retrieved and merged together for improved performance using GoogLeNet. To detect small cell lung cancer (SCLC), scans with no tumours after categorization using Support Vector Machine (SVM) were enhanced using Contrasted Limited Adaptive Histogram Equalization (CLAHE) to recognise small cell lung cancers. Other than simple nodules, which are noncancerous cells, the suggested model has shown to be the most effective at detecting SCLC; as a result, we were able to reach a classification performance of 91.5%. The suggested model improves classification performance by 3% when employing a diffused feature set for early stage detection of SCLC, compared without using CLAHE.**

***Index Terms*—Lung Cancer Detection, Support Vector Machine, Fuzzy Local Information cMean, GoogLeNet, Small cell Lung Cancer (SCLC).**

978-1-6654-6477-2/22/$31.00 ©2022 IEEE

I. Introduction

A staggering 5 million people die from cancer worldwide each year, making it one of the major causes of mortality [1]. Doctors find it difficult to identify cancer cells in the early stages using CT scans since they develop and show symptoms in the later stages, and the majority of lung cancer deaths are due to late discovery.

The following cancers have been identified thus far:

(A) Small cell lung cancer (SCLC): AKA: Oat-cell cancer because, as the name suggests, cancer cells resemble oats under a microscope. It starts to grow inside the bronchi and typically spreads to other body regions. This kind of lung cancer affects less than 20% of people and is primarily brought on by smoking. Because it is so aggressive, this type of cancer requires rapid attention [2].

(B) Non-Small Cell Lung Cancer (NSCLC): Approximately 85% of patients with lung malignancy have this form of cancer. NSCLC grows more slowly than SCLC. In this kind of cancer, lung tumours are often smaller than a golf ball. These can be categorised into three categories [3].
Some of the previously employed procedures have been proven to have low accuracy, while others have slightly

2022 16th International Conference on Open Source Systems and Technologies (ICOSST) | 978-1-6654-6477-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICOSST57195.2022.10016866

higher accuracy—but still fall short of 100%. Therefore, in comparison to other programmes out there, our study of cancer aims to raise this degree of accuracy as much as feasible.

A. Lung Nodule vs Small Cell Lung cancer (SCLC)

A lung nodule is a tiny, developing cell that may or may not be a cancerous growth. Rarely, a little nodule is cancer, which makes it very challenging for the radiologist to diagnose. A spherical area of the lung that is denser than typical lung tissue is known as a lung nodule. X-rays and computed tomography (CT) scans frequently reveal it. Lung nodules are rather typical. In reality, up to 50% of individuals with chest X-rays had lung nodules. The majority of nodules are not malignant, although they could represent an early stage cancer cell. Figure 1 illustrates the difference. The solid and partially solid nodules are more likely to be cancerous cell. From figure a) part shows the solid nodule, b) partially solid and c) is non solid region.

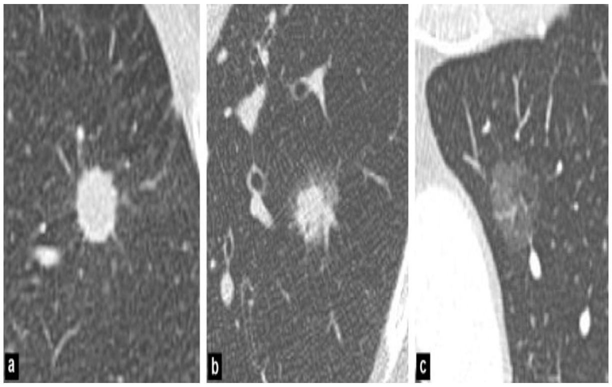

Fig. 1. Lung Nodule and SCLC [4]

The objective for designing this technology is to detect this specific type of lung cancer, which affects the majority of the world's population. Furthermore, current hospital systems do not give such efficient and promising systems for detecting lung growth or nodules with the accuracy and efficiency as this system does. The major goal of designing this technology is to aid in the diagnosis of malignant cells in lung cancer before the situation becomes catastrophic and the patient becomes critically ill. The following are the primary goals of this research:

- Use GoogLeNet and FLIcM to improve imaging clarity in CT scans for cancer detection and differentiation between lung nodules.
- The proposed method incorporates basic lung cancer processing, as well as the use of a green spectrum image using the CLAHE methodology for sharp differentiation. Extra tumult is introduced, which converges with the picture's ordinary clamour, and this combined turbulence is sorted using two medians. The obtained result is then converted to grayscale, resulting in a more refined image.
- The ideal higher-dimensional picture details are recovered using GoogLeNet's efficient model and transported to SVM and KNN, where the intracranial cancer is recognised. This method yields an accurate position of the malignant cell, making cell removal easier.
- Using this method, radiologists and clinicians can now clearly distinguish between an early stage malignant cell and a lung nodule (noncancerous spot on lung).

To determine if the patient has infected lung nodules or not, the system will examine the provided Computer Tomography scans of the lungs. GoogLeNet and FLIcM deep learning techniques will be used to make it operational. Features from both methodologies are combined to choose the best features, and then redundant features are removed. As a result, it will facilitate the easy detection of lung nodules and their removal using SVM classifier. It will be able to save the lives of the millions of people who contract the disease each year.

The entirety of this research paper's specifics are as follows: The modern methods used in lung cancer detection are covered in Section II. Section III offers a thorough examination of the datasets. The methodology section is covered in Section IV, which contains the feature extraction using GoogLeNet and FLIcM. The classification using KNN and SVM as well as image enhancement using CLAHE are both explained in Section V. The outcome of the proposed technique is described in Section VI, and the conclusion is given in Section VII.

II. Literature Review

To identify lung cancer from CT scan images, numerous AI and machine learning techniques have been developed. However, a precise technique that can distinguish between a lung nodule and a malignant cell has yet to be developed. Additionally, tiny cancer cells at an early stage of the disease are exceedingly challenging to evaluate. Few studies have explained the key difference between a small lung cancer and a nodule, despite the fact that many researchers use deep and machine learning algorithms to automatically detect lung tumours. Since the difference is so slight, it is exceedingly difficult for a machine to identify malignant cells just on the spot-on lung that was discovered. In this work, a few of the most recent methods were discussed.

Artificial Neural Networks (ANN) is used to identify lung cancer. The method for detecting lung cancer uses the person's information and symptoms as input variables for the system (ANN) [5]. The proposed method has a 96.67% accuracy rate for cancer detection. ANN employs a mathematical model that combines the structural and operational characteristics of neural networks. An ANN network is trained on a dataset in the first phase, and the weights of connections between neurons are fixed in the second phase. Corresponding with reference [6], the authors' primary attention is on the image enhancement and low processing technique based on Gabor Filer within Gaussian rules; as a result, a more effective feature extraction is produced. In comparison to other procedures, the suggested technique produces effective results. Pixel percentage mask-labeling with high accuracy and reliable operation are the major aspects that provide accurate picture comparison.

Authors proposed a method to diagnose lung cancer by lowering misclassification and detecting lung cancer by enhancing the quality of the lung picture [7]. Mean histogram equalisation is utilised in the first phase to remove noise from the images, and the Improved Profuse Clustering Technique (IPCT) is employed in the second phase to improve the image quality. In the final stage, features are investigated using a deep learning instantly trained neural network for lung cancer prediction. The accuracy of the suggested system is 98.42%. In [8], researchers employ deep residual networks to extract information from CT scan pictures in order to predict lung cancer. Lung anatomy and nodules are identified for the classifier during feature extraction. For predicting lung cancer, the classifiers XGBoost and Random Forest are used. The proposed method achieves 76% accuracy with ResNet+XGBoost and 84% accuracy with UNet+RandomForest. Russian software company Botkin AI [9]. The three main parts of this technology are ”automatic learning,” which uses massive quantities of photos as inputs to assess the risk of getting lung cancer, ”hybrid intelligence,” which uses ”hybrid intelligence,” which determines specific treatment scenarios. Thousands of photos provided by hospitals and clinics testing Botkin's algorithm are ”fed” into it. AI.

III. Materials

A. Dataset Description

The LUNA16 [10] data set was used in this study (Lung Nodule Analysis 2016). The collection consists of 1186 photos that have 888 CT scans labelled on them. It includes annotations that were gathered over the course of a two-phase annotation process by four qualified radiologists. Each radiologist labelled the lesions they determined to be non-nodules, nodules between 3 mm and greater than 3 mm, and nodules.

B. Preprocessing

In this phase, CT scan results from the relevant material are transformed into 227×227 dimensions, and the configuration is also altered to make it work with MATLAB. In order to maintain a consistent format, all of the photographs have also been converted to grayscale versions. The new data was recorded and kept because it will be useful for a number of MATLAB activities. Additionally, the dimensions were modified to 227×227 to provide a standard size for GoogLeNet feature extraction.

IV. Methodology

The suggested method consists of just two easy steps. The first stage is to extract features using FLIcM and GoogLeNet, which are then combined to create a high-quality feature set. The Euclidean Distance technique is then used to eliminate redundant characteristics. Images classified under No category were improved using the CLAHE approach to improve performance in the diagnosis of early stage lung cancer. Figure 2 presents the model's overall structure.

A. Feature Extraction

1) GoogLeNet: GoogLeNet is a well-known and well-described deep neural network for organising, classifying, and identifying digitised images. The two techniques of one-on-one convolution and average pooling, which make GoogLeNet intense and divergent, are what make it diversified and sustainable. With the aid of 1x1 convolutions, the overall framework's variable aggregate is diminished. The reduction of pre-trained variables aids average pooling in increasing accuracy rate by 0.6 percent. The sent images with 224×224 dimensions are received by GoogLeNet, a 22-layered neural network. The use of different sized convolutional spiral filters in the core component of GoogLeNet improves the numerous scale operation. These fundamental elements are piled on top of one another to interact with the input images, and this collateral functioning aids in improving the productivity of the numbers. The final features from GoogLeNet have size of 1186×1000.

2) Fuzzy Local Information cMean: This study uses a fuzzy Local Information cMeans (FLIcM) [11] algorithm to extract hand-crafted features from raw scan pictures. The method uses a novel fuzzy approach to combine both the local spatial information and the grey level information. The method involves grouping all of the photos together to extract related objects from the dataset. It would be simple to distinguish between a SCLC and a non-cancerous lung nodule using these features.

Dunn [12] first developed the FcM clustering technique, which was later expanded by Bezdek [13]. The procedure, an iterative clustering technique, minimises the weighted within group sum of squared error objective function to obtain an ideal division $Fmean_w$

$$Fmean_w = \sum_{i=1}^{N} \sum_{j=1}^{c} D_{ij} d^2(x_i, p_j) \tag{1}$$

Where, w in $Fmean$ is the weight of the fuzzy component, N is the notal number of images in the dataset, c is the number of clusters, D is the degree of cluster of x_i and p_j component. x_i is the valus of x for ith cluster and p_j is the prototype of cluster center for jth cluster.

The degree for each cluster is calculated as ;

$$D_{ij} = \frac{\sum_{i=1}^{N} (D_{ij})^w . x_i}{\sum_{j=1}^{c} (D_{ij})^w} \tag{2}$$

This will give us the set of attributes for all number of images.

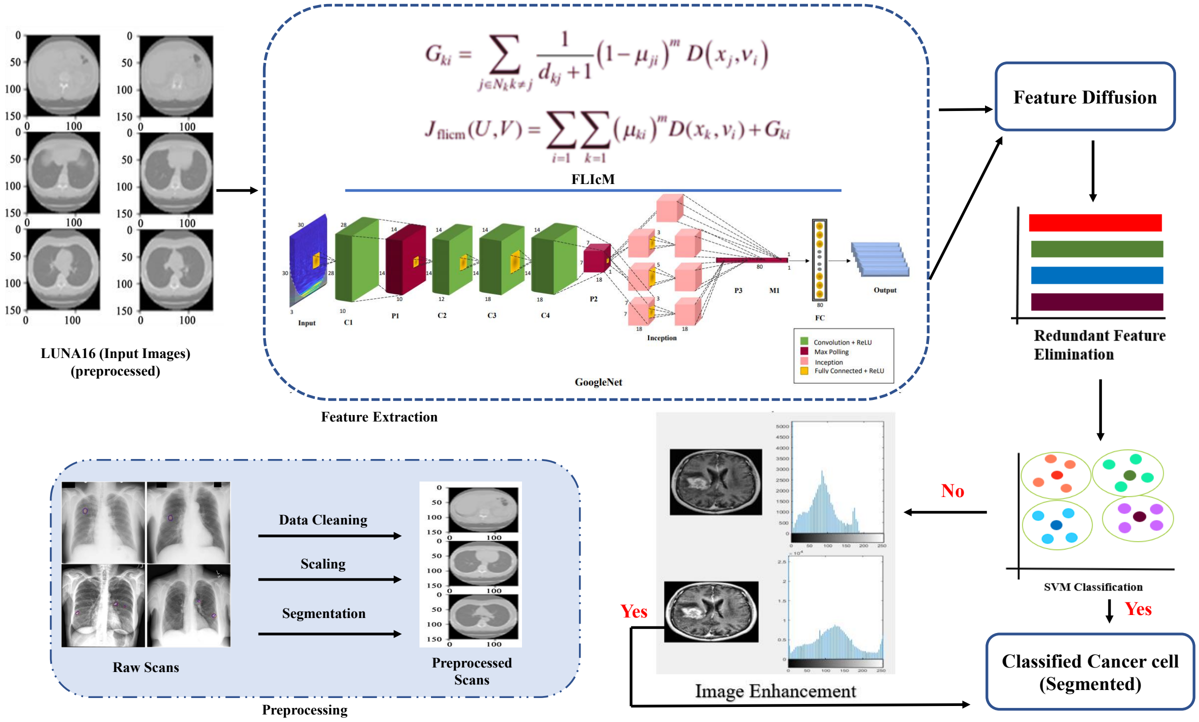

Fig. 2. Small Cell Lung Cancer Detection Methodology using FLIcM and GoogLeNet

3) Feature Diffusion: A crucial component of this work that combines the collection of related features obtained from FLIcM and GoogLeNet is feature diffusion. In order to produce a set of features with excellent quality, both feature sets are diffused together. In this approach, a handcrafted feature and a rich feature set collaborated to offer a large number of high-quality features. Unrelated sets of features are eliminated from the scans since they are deemed to be of little use. To carry out this activity, a straightforward method is constructed in MATLAB. Equation 3 illustrates how we used the Euclidean distance formula to obtain the high-quality features.

$$d(x,y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2} \tag{3}$$

The quality of the selected features is then further evaluated using a probablity function, which formulates their mutual information using equation 4. Mutual information is measured by calculating the combined probability density function (PDF).

$p(X)$ and $p(Y)$ is the probability of two variables then its PDF will be

$$MI(X;Y) = \sum \sum p(x,y) log(\frac{p(x,y)}{p(x)p(y)})) \tag{4}$$

Two vectors from different classes with the same feature are X and Y. The significance of the discrepancy between the two attribute values increases with decreasing MI value. As a result, the final attribute vector is created without attribute values. Selected attribute values only have high MI values. This technique evaluates all emotion data while testing high-quality features. The computational expenses likewise reduce as the attribute vector does.

4) Redundant Feature Elimination: Each model's characteristics are extracted to create a feature set. Since each model contains 1000 attributes, we have linked features with the formula $CFV = F \times class$. By removing extraneous components, redundant feature elimination (RFE) decreases the size of features.

B. Image Enhancement

For further refinement in contrast to no cancer cell images, technique CLAHE [15] was first applied to the Green Channel Images in this section of our proposal. Then, two subsequent filters called the Median and Morphology filters are used to remove the integrated noise. Images from the No category were improved using the CLAHE technique [16].

V. Classification

Images of lung cancer are classified using a support vector machine (SVM), which demonstrates the precision of the training model. The chosen time period was separated into several folders using a sliding control. This technique selects N distinct sets, one of which is utilised for training and the other for validation. Sets with N-1 members are used. An endless number of repetitions of the process results in the creation of Confusion Matrix.

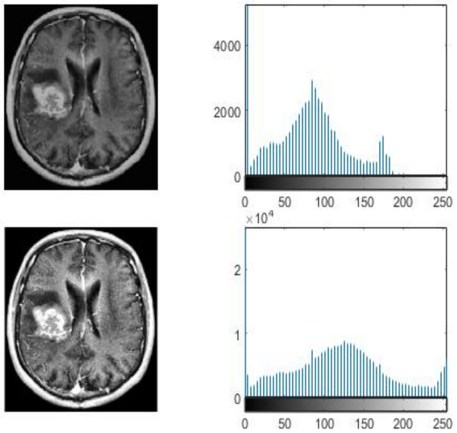

Fig. 3. Image Enhancement using CLAHE

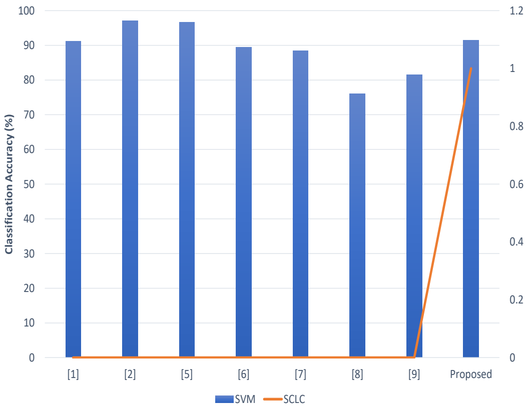

Fig. 4. Comparison with previous methodologies using SVM

TABLE I
RESULTS OF CLAHE ON IMAGES WITH AND WITHOUT NODULES

	Unprocessed Images		**Processed Images**	
	With Cancer	Without Cancer	With Cancer	Without Cancer
HIGHEST MSE	3.316	4.1	1.739	3.394
LOWEST MSE	0.11	0.2	0.054	0.183
AVERAGE MSE	0.6	0.5	0.681	0.83
HIGHEST PSNR	41.2	77.486	57.762	85.708
LOWEST PSNR	35.329	35.3	59.2	21.6
AVERAGE PSNR	34.3	56.2	44.2	63.6

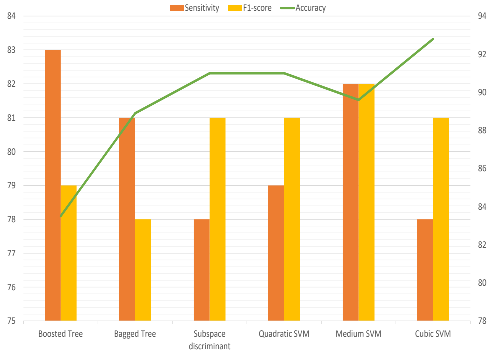

Fig. 5. Classificatiopn Performance with SVM classifier

As it is depicted in figure 4 it is clearly shown that the proposed model performs well and only this model is validated for detecting SCLC (orange line in the figure).

A. Support Vector Machine

The classification method [14] divides several classes into two or more groups. In the preprocessing stage, SVM builds an initial prototype and divides the segments using a technique called hyperplane separation. To improve image clarity, the hyperspace margin is raised. In order to distinguish between healthy and afflicted individuals, the GoogLeNet model is fitted into the SVM divisor. As a kernel-based model, SVM offers non-linear observation-based information. It also plays a crucial role in bringing linearity into separation when the dimensionality of the dataset is tilted.

VI. RESULTS AND DISCUSSION

CLAHE approach produces effective outcomes by the addition of noises, commotion, and filtration of the noise to create the final image. In fig 4, the outcomes of the proposed procedures are contrasted with state-of-the-art techniques found in the literature review.

With the help of this technique, we were able to determine that the Mean Square Error (MSE) values for the lowest and highest samples were 0.0512 and 1.121, respectively, and that the high and low Signal to Noise ratio values for the reference tab III table are respectively 39.732 and 51.211. Peak Signal To Noise (PSNR) and MSE computation median averages are reported as 0.566 and 59.51, respectively. Figure compares the original and enhanced grayscale versions side by side. 3.

There are many different methods for identifying and diagnosing lung cancer, and in this section, the conclusion and justification for the proposed proposal's present methodology are explained along with a comparison of the two methods.

$$P = \frac{T_c}{T_c + F_c} \tag{5}$$

where T_c the tumor is expected to be there and it is present there indeed
F_c is the portion where it is estimated that the tumor will be existing but it is not present

$$R = \frac{T_c}{T_c + F_{NC}} \tag{6}$$

TABLE II
RESULTS OF SVM AND K-NN CLASSIFICATIONS

Classification Type	**Kernal Type**	**Accuracy (%)**	**Precision**	**Recall**	**F-score**
SVM	*Linear*	91.45	0.9974	0.9878	0.9926
	cubic	92.12	0.9993	0.9698	0.9843
	Quadratic	89.77	0.9987	0.9704	0.9843
KNN	*Weightage*	92.41	0.9556	0.9826	0.9689
	Cubic	91.34	0.9331	0.9685	0.9505
	Medium	92.45	0.9554	0.9351	0.9451

TABLE III
CONFUSION MATRIX

Images	**With Cancer**	**No Tumor**
Total Images	1186	1521
Correct forecast of Images	1600	1231
Incorrect forecast of Images	21	81

Recall or sensitivity is the fragmentation of the total aggregate of the significant incidents that were actually recovered. Also, F_{NC} is the portion where the tumor isn't expected to be present and it is not really present there.

$$fscr = 2 \times \frac{P - R}{P + R} \tag{7}$$

where P is the precision, R is the recall, F_c is the false positive means the false detection of cancer cell similarly T_c is the true detection of cancer cell. After this step the true with no cancer cell (T_{NC})were sent for image enhancement using CLAHE. f-score f_{scr} is the average of the accuracy and recall.
spatial features are weighed up with other features using unique methodologies of classification as shown in table II. SVM is utilized as a classifier in many anterior works for their corresponding techniques. Comparison of accuracy of our technique utilizing the same classifier can be optically discerned in fig. 4.

VII. CONCLUSION

Authors have created a method because it may be difficult to distinguish between a small malignant cell and a lung nodule. A technique makes use of a distributed feature set comprising fuzzy local data and GoogLeNet and filters out undesirable features. By excluding the features of nodules, this diffusion aids in the selection of solely malignant cell features. When trained on the LUNA16 dataset after classification using SVM, we achieve a classification accuracy of 89.6%. The images that were not yet classified for early stage SCLC detection were improved utilising CLAHE to determine whether or not there are cells available. After improvement, our accuracy was 91.5%.The system performs well for detecting SCLC, and the proposed methodology makes it simple to distinguish between a SCLC and a lung nodule.

REFERENCES

[1] Makaju S, Prasad PW, Alsadoon A, Singh AK, Elchouemi A. Lung cancer detection using CT scan images. Procedia Computer Science. 2018 Jan 1;125:107-14.

[2] Wu Q, Zhao W. Small-cell lung cancer detection using a supervised machine learning algorithm. In2017 international symposium on computer science and intelligent controls (ISCSIC) 2017 Oct 20 (pp. 88-91). IEEE.

[3] Wahbah M, Boroumand N, Castro C, El-Zeky F, Eltorky M. Changing trends in the distribution of the histologic types of lung cancer: a review of 4,439 cases. Annals of diagnostic pathology. 2007 Apr 1;11(2):89-96.

[4] Lederlin M, Revel MP, Khalil A, Ferretti G, Milleron B, Laurent F. Management strategy of pulmonary nodule in 2013. Diagnostic and interventional imaging. 2013 Nov 1;94(11):1081-94.

[5] Nasser IM, Abu-Naser SS. Lung cancer detection using artificial neural network. International Journal of Engineering and Information Systems (IJEAIS). 2019 Mar;3(3):17-23.

[6] Al-Tarawneh MS. Lung cancer detection using image processing techniques. Leonardo Electronic Journal of Practices and Technologies. 2012 Jan;11(21):147-58.

[7] Shakeel PM, Burhanuddin MA, Desa MI. Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks. Measurement. 2019 Oct 1;145:702-12.

[8] Bhatia S, Sinha Y, Goel L. Lung cancer detection: a deep learning approach. InSoft Computing for Problem Solving 2019 (pp. 699-705). Springer, Singapore.

[9] Bulten W, Kartasalo K, Chen PH, Ström P, Pinckaers H, Nagpal K, Cai Y, Steiner DF, van Boven H, Vink R, Hulsbergen-van de Kaa C. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nature medicine. 2022 Jan;28(1):154-63.

[10] Armato III SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical physics. 2011 Feb;38(2):915-31.

[11] Krinidis S, Chatzis V. A robust fuzzy local information C-means clustering algorithm. IEEE transactions on image processing. 2010 Jan 19;19(5):1328-37.

[12] Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.

[13] Bezdek JC. Pattern recognition with fuzzy objective function algorithms. Springer Science & Business Media; 2013 Mar 13.

[14] Asghar MA, Khan MJ, Rizwan M, Shorfuzzaman M, Mehmood RM. AI inspired EEG-based spatial feature selection method using multivariate empirical mode decomposition for emotion classification. Multimedia Systems. 2022 Aug;28(4):1275-88.

[15] Majeed AR, Awan WA, ul Hassan N, Asghar MA, Khan MJ. Retinal Fundus Image Refinement with Contrast Limited Adaptive Histogram Equalization, Noise Filtration and Intensity Adjustment. In2020 IEEE 23rd International Multitopic Conference (INMIC) 2020 Nov 5 (pp. 1-6). IEEE.

[16] Oyelade ON, Ezugwu AE. A deep learning model using data augmentation for detection of architectural distortion in whole and patches of images. Biomedical Signal Processing and Control. 2021 Mar 1;65:102366.