Available online at: https://proceeding.researchsynergypress.com/index.php/ihsatec
Proceeding on The International Halal Science and Technology Conference (IHSATEC)
ISSN 2828-1500 (Online) | 2828-1985 (Print)
Volume 1 Number 1 (2021): 213-222

The Model Development for Early Lung Cancer Analysis by Using Image Processing and Neural Network

Sirirat Promduang[1], Pongpisit Wuttidittachotti[2]
[1] Information Technology Management; [2] Data Communication and Networking
Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

Abstract

Lung cancer is the leading mortality disease for men and women compared to other organ cancers. The leading cause is smoking until the patient begins to show symptoms. Therefore come to see a doctor determine that the disease has spread in the last stage, causing cumbersome and complex treatment. Therefore, early screening of patients is very important to allow patients to enter receive treatment in a timely manner and have a chance to recover from the disease. This research has developed a model for early lung cancer analysis by using a CXR image that can screen a large number of patients when they are asymptomatic. Let's improve the image enhancement to reduce the noise with median filter and then go into image processing by image segmentation with Active contour algorithm, image edge detection with Laplacian of Gaussian (LoG) algorithm, and image extraction with Shape and GLCM in combine with data classification with a neural network using MLP compared against SVM classifiers. Training and testing the performance of the model by the result of MLP provides a better time and up to 99% accuracy.

Keywords: *Image Processing, Lung Cancer, Neural Network, Support Vector Machine*

INTRODUCTION

Lung cancer has the highest mortality rate in both men and women compared to other organ cancers. The main cause is smoking, together with other environmental factors, including the inevitable increase in air pollution. Although evolution continues to aid the diagnosis of early and metastatic lung cancer, the mortality rate did not decrease because patients' access to primary lung cancer screening is limited, and the symptoms of early-stage lung cancer are unclear. Patients visit their doctor when they have symptoms. The disease has spread so much that it is difficult to treat. There are cases where doctors accidentally detect nodules in the lungs when the patient goes for a check-up. Therefore, proactive screening is important to allow asymptomatic patients to be screened early if abnormalities are found in the early stages. Doctors will be able to cure the disease in time to prevent spread to other organs, and the patient has a chance to recover.

From studying and exploring the problems that arise, Innovative information technology has been used to develop a model for the analysis of lung X-ray images of lung cancer patients and lung cancer patients with image processing and neural networks. Let's start by using CXR images, which are the

Corresponding author
Sirirat Promduang
siriratpr.it@gmail.com
DOI: https://doi.org/10.31098/ihsatec.v14i1.504

primary source of important information. Many diseases can be screened to improve image quality, known as image preprocessing to reduce noise. The resulting image is processed using image segmentation using Active Contour algorithm, image edge detection using Laplacian of Gaussian (LoG) algorithm, and image feature extraction using Shape and GLCM combined with neural network classification using MLP classifier and the most popular is Support Vector Machine (SVM) classifier to compare performance evaluation results. The objective is to develop a chest X-ray image analysis model using image processing techniques and neural networks for the early screening of lung cancer abnormalities. It will be an important guideline for physicians to plan treatment and follow-up in the next steps.

LITERATURE REVIEW

Nadkarni et al. (2019) detect lung cancer by image feature extraction of shape only and classification with SVM. No other classification data is compared. Jena et al. (2019) detect lung cancer by image extraction of shapes and textures combined with an SVM classifier. The acquisition of primary image processing data by image segmentation does not exist, and no other classification data is compared. M. Tech et al. (2018) detect lung nodules by image feature extracting with GLCM for texture only classified by MLP vs. KNN. Chellan et al. (2018) Detect lung cancer with Region Base Active Contour image segmentation, GLCM feature extraction for texture data only, no comparison of other classification data. Kasinnathan et al. (2017) Detect lung tumor with watershed segmentation and GLCM feature extraction for texture only combined with classification SVM, and no other classification data is compared. All of the above research uses data from CT scans that have limited access to patient screening.

In this study, an easily accessible lung X-ray image source was used for the initial screening not only to monitor treatment but also to cover the entire lung segmentation process as well as the extraction of image characteristics with both spatial and surface characteristics and distinguishing them with popular classifiers, both MLP and SVM classifiers, in order to obtain the optimal classifier with the best accuracy and time.

RESEARCH METHODOLOGY

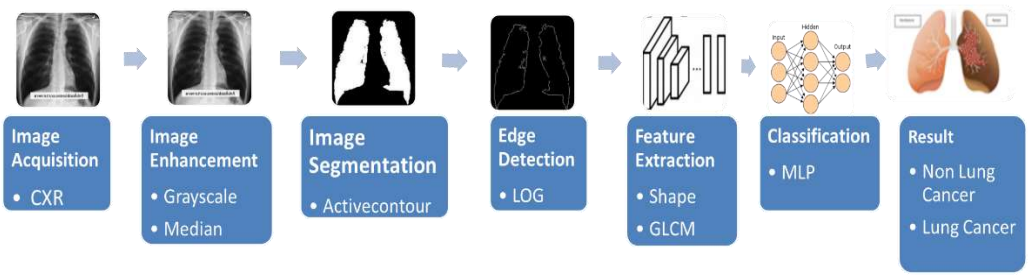

Figure 1. Early Lung Cancer Analysis Model

ISSN 2828-1500 (Online) | 2828-1985(Print)

This research is a model designed to develop the analysis of Chest x-ray images. To be screened for early lung cancer using image processing and neural network as shown in Figure 1. Each step is detailed as follows.

A. Preparation Data

1*) Data Collection* start collecting CXR images from JSRT Digital Image Database by downloading Nodule 154 images file and NonNodule 93 images file. The original image file is in the format of.IMG with image size 2084*2084 pixels, save all data to a computer for further review.

B. Image Pre-Processing

1) Image Acquisition is the process of importing data both nodule and nonnodule CXR images, image size 2048*2048, grayscale, selecting 100 samples into the computer database as JSRT Dataset file name, enter to MATLAB R2021a program when import data convert the original.IMG file to .jpg file for use in the program's processing with the open file, read the file, write file, close file, and then the image is saved in the workspace for the next process.

2) Image Enhancement after going through image acquisition and then image enhancement with Median Filter when A is an input image, grayscale also filter the image with Median Filter to reduce noise that has the appearance of Salt and pepper well with the command B = medfilt2(A) where A is a 2D variable represents a pixel grayscale image 2048*2048*uint8 8-bit integer format. Since the received image is large, it is scaled down to speed up processing with the command B = imresize(A, scale) when the scale is the scaling ratio is 0.25. After resizing, the output image will be a pixel grayscale image 512*512*uint8 format, 8-bit integer, the resulting image after filtering will be more smooth and preserve the details without loss of data.

C. Image Processing

1) Image Segmentation with Active Contour. Once the image has been enhanced, it goes into image segmentation with Active Contour algorithm with function of bw = activecontour(A,mask,n) segments the 2-D grayscale image A into foreground is lung and background regions is boundary of non-pulmonary using active contour segmentation when the mask is a binary image same size as A image 512*512*double that specifies the initial state of the active contour represent the boundaries of lung regions (white) with function of mask = zeros(size(A)) replace value with 0 is background (black) in mask define the initial contour position used for contour evolution to segment the image as the object of interest (lung) with function m(140:155, 140:155)=1;m(375:390, 140:155)=1; replace value with 1 is foreground (white) both left and right of lung, n is maximum number of iterations = 800 and the output image bw is a binary image where the foreground(lung) is white and the background is black obtain faster and more accurate segmentation results, specify an initial contour position that is close to the desired object boundaries.

ISSN 2828-1500 (Online) | 2828-1985(Print)

2) Image Edge Detection with Laplacian of Gaussian (LoG) is a necessary process to find the edge image using the Laplacian of Gaussian algorithm of the image interested (lung). The image edge detection represents an area of the edge of an image using function bw = edge(I, method, threshold) when I am an input variable. This research is the output image from the Active Contour algorithm. The technique used is Laplacian of Gaussian ('log'), and the threshold value used is 0.001. The output image (bw) is 512*512*logical, black and white logical type shown in Figure 2.

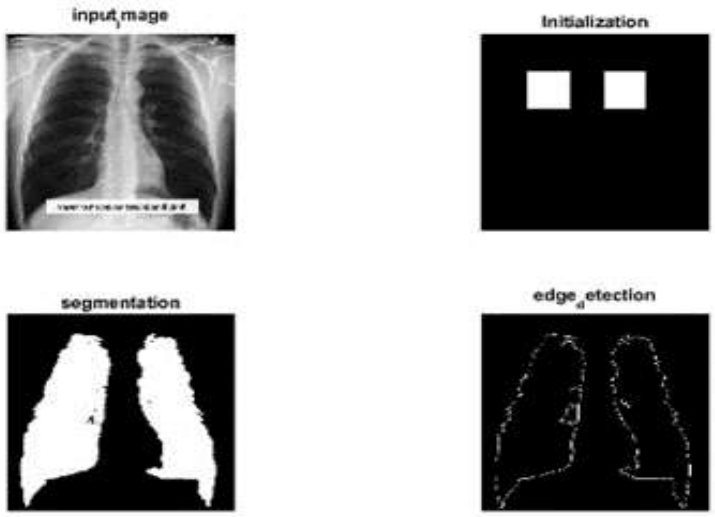

Figure 2. (Input image), mask (initialization), image segmentation, image edge detection

3) Image Feature Extraction with Shape and GLCM. After edge detection, it goes into image feature extraction for extracts important information of image in the Shape section by using with function stats = regionprops (BW,properties) when BW represents a binary input image variable and properties represents a group of connected pixels with similar is selected 4 attributes: Area, Perimeter, Eccentricity, Solidity and then texture analysis using the GLCM (gray level co-occurrence matrix) by begin generating a GLCM of the same size as the input image using the function glcm = graycomatrix(I) when I represents the input binary image, graycomatrix generates value GLCM by calculating the frequency of pixels with grayscale values (Gray level intensity) After that, use function stats = graycoprops(glcm, properties) to calculate the statistics specified in the properties from glcm to analyze the texture these statistics provide information about the texture of an image feature the selected values have 4 attributes: Contrast, Correlation, Energy and Homogeneity. When in collected total 8 attribute values in 1 image were obtained saved to workspace for data classification in next process

ISSN 2828-1500 (Online) | 2828-1985(Print)

D. Classification

After going through the process of image feature extraction then create A input variables were collected from image feature extraction is numerical type included all 8 attributes consist of Area, Perimeter, Eccentricity, Solidity, Contrast, Correlation, Energy, Homogeneity and target variables (predictor) represents of the lung cancer is 1 and non-lung cancer is 0 include total 100 samples in the file name DATASETA. The variable names are column aligned, and the observation part is in the row. When entering neural network processing, specify the name of the input variable (ip01) by converting dataset array with input variable display in a row aligned, and observation values display in a column aligned and specify the name of the target variable (tg01) by converting dataset array with target variable display in a row aligned and observation values display in a column aligned after that the input (ip01) and target (tg02) variables are stored in the workspace enter to the Apps. Neural Net Pattern Recognition for MLP classifier and classification learner for SVM classifier for next process.

1) Neural Network with the MLP (multilayer perceptron) classifier, the dataset of lung cancer and non-lung cancer patients is imported all 100 samples obtained from the image feature extraction of all 8 attributes, and the target result is lung cancer and non-lung cancer. Then the network begins to train data to create a network model to practice learning the network throughout validation of the data to find the best model until an acceptable error value is obtained; therefore, stop learning and use that best model to test the network by importing the same dataset as training after that to train data until is obtained appropriate value and can evaluate network the accuracy performance with a confusion matrix.

a) Select Data to define the input data and target data values of the neural network saved in the workspace, replace the input data with ip01 and the target data with tg01 of the network and select the dataset as matrix columns.

b) Validation and Test Data select a percentage at random validation 15%, testing 15% to be used to evaluate performance and 70% as training data, learners should have values covering the range of Validation and Testing for creating a system of model shown in Figure 3.

c) Network Architecture customize the hidden layer of the network where this panel provides the default number of hidden neurons is 10 after training the network if it still does not work well can become back to this panel to change the number of neurons until the result of neural network processing provides highest accuracy is obtained considering the acceptable error value after training data shown in Figure 4.

d) Train the Network after customizing the number of hidden neurons of the network will start to train for creating a network to practice learning the network throughout validation the data to find the best model until an acceptable error value is obtained therefore stop learning shown in Figure 5.- 6.

ISSN 2828-1500 (Online) | 2828-1985(Print)

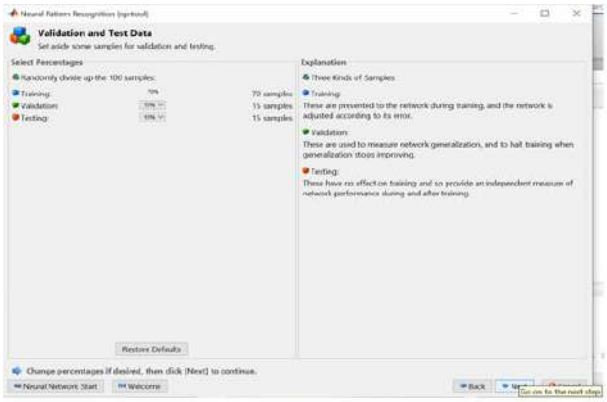

Figure 3. Validation and Test data

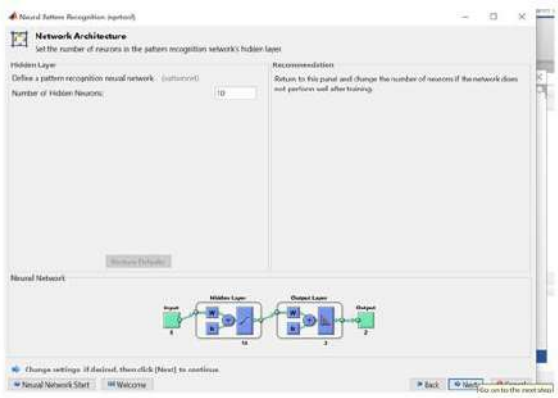

Figure 4. Network Architecture

Figure 5. Train Network

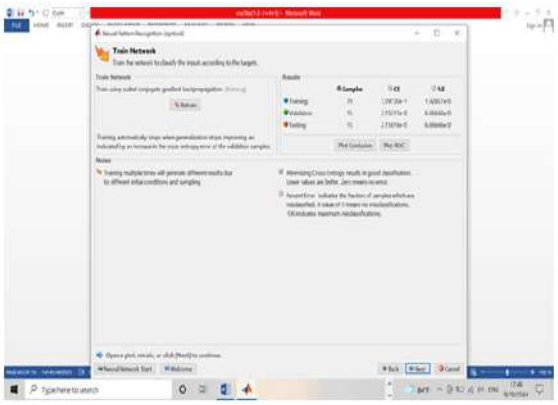

Figure 6. The error value is minimal (%E)

e) Performance Evaluate Network after training the network to the best model, the network performance is assessed for accuracy by testing the error-measurement network independent of all previously used segmented data before. When importing the same sample dataset as the training dataset, both the input data and the target sample data samples are in matrix column format for network testing. Error-values are calculated, and the test performance evaluation results are displayed with Confusion Matrix tables can be found as shown in Figure 7.

ISSN 2828-1500 (Online) | 2828-1985(Print)

Figure 7. Confusion Matrix Test Neural Network

2) Support Vector Machine to start imported datasets from the numeric matrix format of the DATASETA file. Select the cross-validation format. Then select a model type that contains all the SVM training data to find the best kernel as linear at the training accuracy values are shown in Figure 8. The model is then tested by importing the same dataset. To determine the accuracy obtained from the test, as shown in Figure 9.

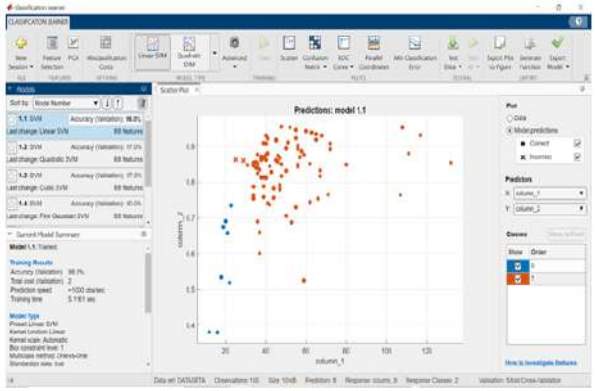

Figure 8. Training SVM

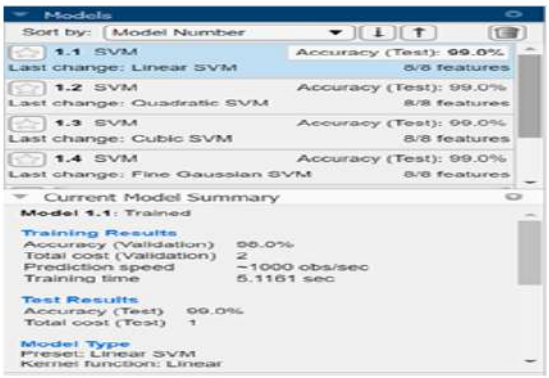

Figure 9. Testing SVM

3) Performance Evaluation CXR image classification of lung cancer and non-lung cancer patients with the best classifier of a neural network and SVM, the results of the correct probabilities are evaluated using the Confusion Matrix method to show the test validity values shown in Table 1.

ISSN 2828-1500 (Online) | 2828-1985(Print)

Table 1. Confusion Matrix

	Predict Positive	Predict Negative	
Actual	True Positive (TP)	False Negative (FN)	Positive
Actual	False Positive (FP)	True Negative (TN)	Negative

Accuracy is the amount of data that accurately predicts the data in all group classifications

$$= \frac{\text{true positive + true negatives}}{\text{true positive + false positive + false negative + true negative}} \quad (1)$$

where,
true positive (TP) is the result of the prediction that it is in persons with a disease
true negative (TN) is the result of the prediction that it is not in persons without a disease
false positive (FP) is the result of the prediction that it is in persons without a disease
false negative (FN) is the result of the prediction that it is not in persons with a disease

FINDINGS AND DISCUSSION

A study for CXR images of patients with lung cancer and non-lung cancer number of 100 samples using MATLAB R2021a program result of image processing shown in Table 2.

Table 2. Input Variable by Image Feature Extraction and Target Variable by Predictor

NO.	Area	Eccentricity	Solidity	Perimeter	Contrast	Correlation	Energy	Homogeneity	Predictor
1	39	0.812907	0.322314	46.198	0.013615	0.352011	0.96556	0.993192729	1
4	40	0.9073443	0.333333	47.422	0.01282	0.2973912	0.969099	0.993590234	1
5	40	0.9295226	0.344828	43.741	0.007	0.3687	0.9821	0.9965	1
6	36	0.8245825	0.36	40.764	0.008004	0.3879577	0.978984	0.995998196	1
7	56	0.8434234	0.264151	64.748	0.013683	0.330124	0.966077	0.993158329	1
8	62	0.7441517	0.295238	67.138	0.008478	0.3456748	0.978638	0.995761222	1
9	37	0.8348882	0.293651	41.194	0.010908	0.3449775	0.972557	0.994545774	1
10	44	0.7818214	0.321168	50.49	0.014073	0.3182831	0.965481	0.992963399	1
11	18	0.5332149	0.4	21.172	0.012567	0.316557	0.969202	0.993716365	0
12	25	0.8621657	0.431034	29.616	0.009609	0.401672	0.974424	0.995195542	0
14	20	0.6916028	0.4	22.95	0.014241	0.2647902	0.966591	0.992879311	0
15	16	0.3788861	0.421053	19.212	0.009158	0.3457139	0.976929	0.995421049	0

From Table 2. It shows the numerical statistics of the property selected from the image feature extraction of all 8 attributes is input variable consists of: area, perimeter, eccentricity, solidity, energy, contrast, correlation, homogeneity including target variable predictor value 1 represents lung cancer patients and 0 represents is non lung cancer patients total of 100 observation. All statistical data were collected in the DATASETA file name

ISSN 2828-1500 (Online) | 2828-1985(Print)

to enter the classification the neural network MLP and the SVM classifier to further assess the validity of the two classifiers, as shown in Table 3.

Table 3. The classification was tested with neural network MLP and classification learner SVM based on the accuracy method.

Classifier	MLP	SVM
True Positive	89	90
False Positive	0	1
False Negative	1	0
True Negative	10	9

From Table 3. Show that MLP neural network, together with the image feature extraction data, was able to classification CXR images correctly TP up to 89 images, misclassification FN 1 image from total 90 lung cancer images, and classifiers images correctly TN 10 images, misclassification FP 0 image from total 10 non-lung cancer images, while SVM in combine with image extraction data, was able to classification CXR images correctly TP up to 90 images, misclassification FN 0 image from total 90 lung cancer images, and classifiers images correctly TN 9 images, misclassification FP 1 image from total 10 non-lung cancer images. If considering the Confusion Matrix values in the classification of MLP and SVM neural networks, the test results are shown in Table 4.

Table 4. Confusion Matrix values from MLP and SVM classifiers.

Classifier	**Accuracy**	**Precision**	**TPR**	**TNR**	**FPR**	**FNR**	**Time/sec**
MLP	99%	100%	98.89%	100%	0%	0.01%	0
SVM	99%	98.9%	100%	90%	0.1%	0%	5.1161

Table 4. Show confusion matrix values of the neural network MLP classifier and SVM classifier, showing various efficacy values obtained from result predicting for lung cancer and non-lung cancer patients both the MLP classifier provides accuracy 99% , precision 100%, TPR recall or (sensitivity) 98.89%, TNR(specificity)100%, FPR 0%, FNR 0.01% and time 0 seconds, while the classifier SVM provides accuracy 99%, precision 98.9%, TPR 100%, TNR 90%, FPR 0.1%, FNR 0% and time 5.1161 seconds. When comparing the two classifiers for 100 observation, this study chose a classifier MPL is used in a model to optimize system operation which provides very high performance at the better time.

This research has been found that when image segmentation with Active contour, both lung positioning and appropriate mask size were used to accurately contour the lung area, and upon edge detection, this study found with (LoG), the edge locating efficiency is greatly improved in the detection of pulmonary nodules. In neural network classification of lung cancer and non-lung cancer patients, the MLP classifier compared to the SVM classifier gave similar values, where MLP had a better time.

ISSN 2828-1500 (Online) | 2828-1985(Print)

CONCLUSION

This research presents the development of a model for early lung cancer screening from CXR images by image processing and Neural Network. The input variables consisted of area perimeter eccentricity solidity energy correlation contrast homogeneity and target variables predictive lung cancer and non-lung cancer. Select an appropriate model, use Cross-Validation, start training, validation, and testing the model for evaluating efficiency from the Confusion Matrix. The results of Neural Network accuracy prediction depend on the selection of the variables in the study. Selection of good information to train the network and choosing a suitable model. Limitations of this research are the number of samples and setting the proper position of the mask for Active Contour segmentation when the data sources are different, and in the future, further disease stages can be analyzed.This model to analyze CXR images of patients with lung cancer and non-lung cancer by image processing and Neural Network MLP classifier with Confusion Matrix value TPR = 98.89% TNR = 100% FPR = 0% FNR = 0.01% Accuracy = 99%.

REFERENCES

Gonzalez, R. C. & Woods, R. E. (2002), *Digital Image Processing*, 2nd edn, Prentice Hall, New Jersey.

Gonzalez, R. C. & Woods, R. E. (2008), *Digital Image Processing*, 3rd edn, Prentice Hall, New Jersey.

Gonzalez, R. C. & Woods, R. E. (2018), *Digital Image Processing*, 4th edn, Pearson.

Kass, M., Witkin, A., & Terzopooulos, D. (1988), *Snake: Active contour models*, *International Journal of Computer Vision 1*, pp.321-331.

Nadkarni, N.S. & Borkar, S. (2019), "Detection of Lung Cancer in CT Images using Image Processing", *International Conference on Trends in Electronics and Informatics (ICOEI)*, pp.863-866, IEEE database, doi: 10.1109/ICOEI.2019.8862577

Jena, S.R., George, T., & Ponraj, N. (2019), "Texture Analysis Based Feature Extraction and Classification of Lung Cancer", *International Conference on Electrical, Computer and Communication Technologies (ICECCT),* IEEE database, doi: 10.1109/ICECCT.2019.8869369

Potghan, S., Rajamenakshi, R., & Bhise, A. (2018), "Multi-Layer Perceptron Based Lung Tumor Classification", *International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp.499-502, IEEE database.

Chellan, T.D. & Chellappan, A. (2018), "Novel computer-aided diagnosis of lung cancer using bag of visual words to achieve high accuracy rates", eISSN 2051-3305, pp.1941-1946.

Kasinnathan, G. & Selvakumar, J. (2017), "Lung tumor area recognition and classification using EK-mean clustering and SVM", *International Conference on Nextgen Electronic Technologies: Silicon to Software (ICNETS2)*, IEEE database, doi: 10.1109/ICNETS2.2017.8067906

Avinash, S., Manjunath, K., & Kumar, S.S. (2017), "An improve image processing analysis for the detection of lung cancer using Gabor filters and watershed segmentation technique", *International Conference on Inventive Computation Technologies (ICICT)*, IEEE database, doi: 10.1109/INVENTIVE.2016.7830084

MathWorks, (2021), Available: **<https://www.mathworks.com>.** Accessed 15 December 2021.

ISSN 2828-1500 (Online) | 2828-1985(Print)