Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766144

DETECTION OF LUNG CANCER USING IMAGE PROCESSING AND MACHINE LEARNING APPROACHES

K. Ramkumar[1] and M. Natarajan[2]

[1]Research Scholar, Department of Computer and Information Science, Faculty of Science, Annamalai University, Annamalainagar, Tamil Nadu, India
Email: ramau13@gmail.com

[2]Department of Computer and Information Science, Faculty of Science, Annamalai University, Annamalainagar, Tamil Nadu, India
Email: mind.2004@gmail.com

Abstract

Lung cancer is one of the major and most dangerous diseases in the world and early diagnosis and treatment can save lives. CT scan image is the best technique in the medical field for finding various diseases, it is difficult for finding the diseases for doctors to interpret and identify the cancer using CT scan images. In these cases, computer-assisted diagnosis can be helpful for doctors to identify the cancer cells. In recent research numerous computational models and algorithms using image processing techniques and machine learning approaches have been implemented. In this paper, to proposed different computational techniques to analyze the images using the best technique at the moment and find out its limitations and disadvantages, and finally propose a new model which is used to improve the accuracy performance. The image processing techniques were analyzed at each step and disadvantages were identified. In this research, found that some of the instances have low accuracy, and some have higher accuracy. Numerical illustrations are also provided to prove the results and discussion with research aims to increase accuracy.

Keyword: CT Scan Image, Lung Cancer Detection, Image Processing, Pre-processing, and Machine Learning.

1. Introduction

Cancer is one of the most serious and widespread diseases, responsible for a large number of deaths every year. Among all the different types of cancer, lung cancer is the most common cancer with the highest mortality rate. Computed tomography scans are used to identify lung cancer because they provide a detailed picture of the tumor in the body and track its growth. Although CT is preferred over other imaging modalities, the visual interpretation of these CT scan images can be an error-prone task and lead to delays in lung cancer detection. Therefore, image processing techniques are widely used in the medical field for detecting early-stage lung tumors [1]. Image Processing techniques help enhance images for human interpretation. Information can be processed and extracted from images for machine interpretation. The pixels in the image can be manipulated to any desired density and contrast. Images can be easily saved and accessed. Machine learning is a growing technology that enables computers to automatically learn from past data. Machine learning uses various algorithms to build mathematical models and make predictions based on historical data or information. It is used for various tasks like image recognition, voice recognition, recommender system and many more.

The Best Imaging Technique CT imaging is dependable for lung cancer diagnosis because it can uncover any suspected and unsuspected lung cancer nodule [2]. However, variations in intensity in CT images and misjudgment of the anatomical structure by physicians and radiologists can lead to issues in labeling the cancer cell [3]. To help radiologists and physicians to precisely detect cancer, computer-assisted diagnosis has lately become a complementary and promising tool [4]. Lung cancer cannot be seen with the naked eye and its symptoms are often masked by other symptoms of the disease, such as bronchitis, asthma and cough. Lung cancer is usually detected when an X-ray or CT scan of the patient's chest is performed for another valid reason [5]. Lung cancer usually spreads toward the center of the chest cavity; This is due to the natural flow of lymph flowing outwards from the lungs and inwards towards the center of the chest [6].

General methodology of the lung cancer detection system, which consists of five basic stages. The first stage is image acquisition, which represents the collection set of images related to the body part. For this work, we obtained DICOM CT scan images of the lungs from an online database [7]. In recent years, many machine learning approaches, especially neural networks, have been widely used to detect lung cancer from medical images. Many of the proposed approaches have achieved a high accuracy rate [8]. Image preprocessing technique known as Contrast Limited Adaptive Histogram Equalization (CLAHE). They used a grayscale co-occurrence matrix (GLCM) to extract the image features, which also provides information about the position of those pixels that have similar grayscale values. A GLCM can contain a variety of statistical features that can be extracted from the matrix for analysis purposes. Authors have used automatic feature selection algorithms to determine the best features. A study to predict short-term breast cancer risk using the image data set of mammograms. The proposed scheme is based on the four image processing modules such as image preprocessing, segmentation, feature extraction and classification to calculate image feature asymmetry. [11] Processed the image using (i) image preprocessing and (ii) feature extraction methods. The image pre-processing step consists of two main segments: image enhancement and image segmentation. For image enhancement, they tested three algorithms known as Gabor filter, automatic enhancement, and fast Fourier transform. Similarly, watershed techniques to perform image segmentation [12]. Feature extraction laws to extract features from the region of interest [13]. Many other image processing techniques for image enhancement, pre-processing, segmentation and feature extraction have been proposed by [8]. Simple Otsus method for image segmentation, together with morphological aperture method, with periodic line as fixed size structuring element [14].

Analyzing the literature searches, the system [15] is currently the best solution based on the accuracy and the advantages of the steps used. Proposed a system that classifies lung cancer as benign or malignant. The system uses the Priori information and HousefieldUnit (HU) to calculate the Region of Interest (ROI). Shape features such as area, eccentricity, circularity, fractal dimension, and texture features such as mean, variance, energy, entropy, skewness, contrast, and smoothness are extracted to train and classify the support vector machine to identify whether the node is benign or is malignant [16].

2.0 Proposed Model

The proposed model as shown in Figure 2 below. In this model using difference image pre-processing stages and enhancements techniques to be used like Gabor filter, median filters and Gaussian filters were implemented in this phase. The first stage called pre-processing which is used to enhance the image quality. The processed image is segmented using watershed segmentation called second stage. In this case, the image with cancer area marked. In addition to features such as area, perimeter, and eccentricity, features such as centroid, diameter, and mean pixel intensity were extracted in the feature extraction phase for the detected cancer.

The best model ends after detecting the cancerous area, extracting the features, and calculating the accuracy using different machine learning approaches. However, classification as various stages like beginning or malignant has not been implemented. Therefore, an additional phase of cancer nodule classification using the Support Vector Machine (SVM) was performed. Extracted features are used as training features and a trained model is generated.

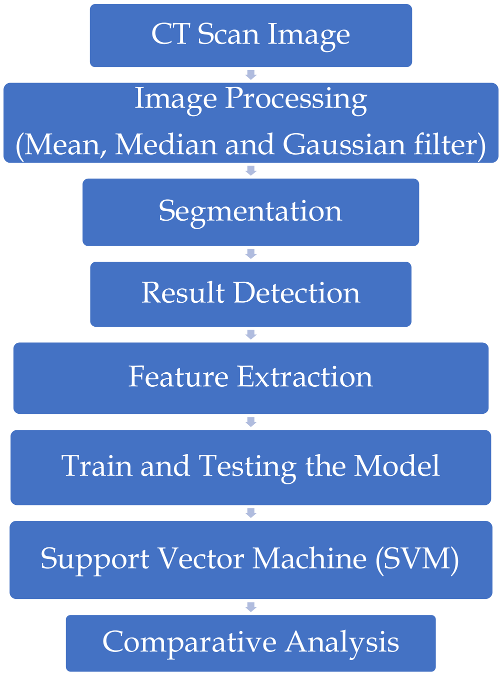

Fig. 1. Proposed System Architecture

3.0 Image Preprocessing

The proposed model considers chest cancer image [17], the data folder consists of train, test and valid folders each folder contains 3 folders of different chest cancer types of namely adenocarcinoma, large cell carcinoma, squamous cell carcinoma and another folder contain normal CT-Scan images (normal). The train folder contains the training images using train folder, the test folder contains the testing images and valid folder contain the validating images.

Image preprocessing uses a gray scale image (figure 2) converted into mean filter for grayscale image of CT scan images (figure 3). Some sounds are embedded in CT images at the time of the image acquisition process, contributing to false detection. Noise can sometimes be recognized as a cancerous nodule. Therefore, for accurate detection of cancer, these sounds must be removed. The median filter removes salt and pepper noise from CT images (figure 4).

A Gaussian filter is implemented after the median filter. It smoothes the image and removes speckle noise from the image (figure 5).

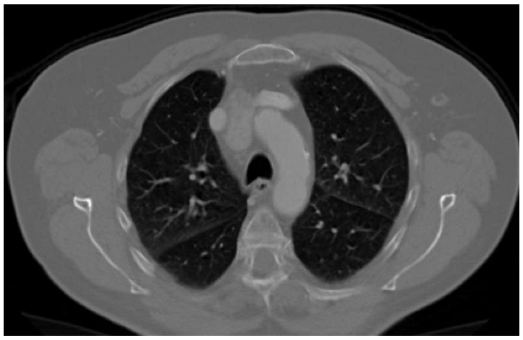

Fig. 2: Grayscale CT-Scan image

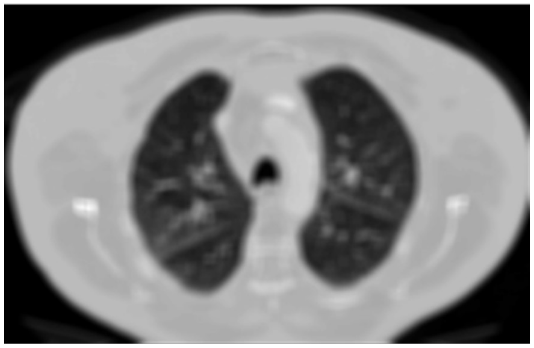

Fig. 3: Mean Filter CT-Scan image

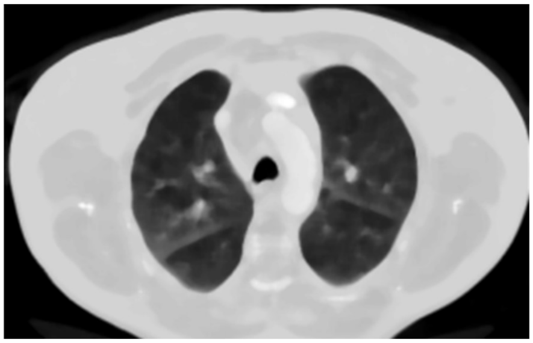

Fig. 4: Median Filter CT-Scan image

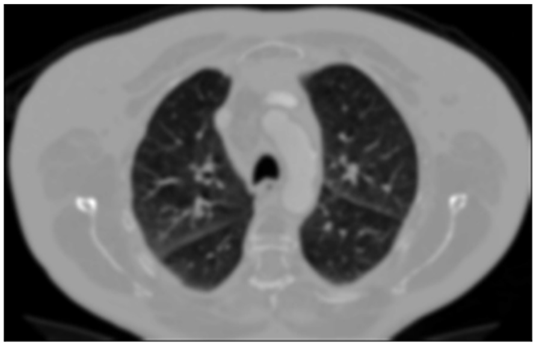

Fig. 5: Gaussian Filter CT-Scan image

3.1. Segmentation

Segmentation is the process of removing region of interest from given image. Region of interest containing each pixel similar attributes. Here we are using maximum entropy thresholding for segmentation. First of all, we have to take gray level of original image then calculate histogram of gray scale image then by using maximum entropy separate foreground from background. After maximum entropy we obtained binary image that is black and white image. In the proposed model watershed segmentation is implemented. Its main feature is that it can separate and identify the touching objects in the image. This feature helps in proper segmentation of cancer nodules if it is touching to other false nodules.

3.2. Features extraction

Feature extraction is very different from Feature selection: the former consists in transforming arbitrary data, such as text or images, into numerical features usable for machine learning. The latter is a machine learning technique applied on these features. Feature extraction plays an important role in extracting information present in the given image. In this stage, features like area, perimeter, centroid, diameter, eccentricity, and Mean intensity. These features later on are used as training features to develop classifier.

3.3. Classification

This stage classifies the detected nodule as malignant or benign. Support vector machine (SVM) is used as classifier.It is supervised machine learning method. SVM defines the function that classifies data into two classes [16]. The function is defined as $D(x)=w^{T}x_{i}+b$ where x_i are training inputs, $w^{T}$is m dimensional vector, and b is bias term. Here, i=1…. M.

$$D(x)=w^{T}x_{i}+ b \geq 1 \text{ for } y_{i}=1$$
$$D(x)=w^{T}x_{i}+ b \leq -1 \text{ for } y_{i}=-1$$

Main strengths of the proposed model are pointed as Increase in accuracy of cancer nodule detection than the best current model, Classifies the detected lung cancer as malignant or benign and Removes salt-pepper noises and speckle noise that creates false detection of cancer.

4. Result and Discussions

In this research for implementation, the real patient CT scan images are taking into consideration from Kaggle archive [17]. It is the open-source database of lung cancer screening CT images for model development with training, testing and validation and evaluation of image pre-processing and computer assisted diagnostic methods for lung cancer detection. The dataset consists of three type of lung cancer CT images namely adenocarcinoma, large cell carcinoma and squamous cell carcinoma which includes 1009 cases of dataset. Images are in PNG format with size 512*512 pixel. PNG format is difficult to process, in this case those images are converted to JPEG Gray scale image using Python packages. The proposed model is then developed in Python. Python is one of the tools for research development and analysis with open-source license. Various stages of image pre-processing and both the detection and features extraction are implemented using Python and classification is implemented using machine learning approaches. The results and discussions include various levels of pre-processing and classification based on dataset. In these cases, figure 6 indicate the raw lung cancer CT image grayscale image converted into two types of conversion namely median and gaussian filter which is used to enhance the quality of the image, the results shown in figure 7 and figure 8. A binary image consists of pixels that can be one of exactly two colors, usually black and white. Binary images (figure 9) are also called bi-level or two-level; pixel art made up of two colors is often referred to as 1-bit or 1-bit. This means that each pixel is stored as a single bit as 0 or 1. Finally the original grayscale image converted into different states and finally converted into cancer marked image, the result shown in figure 10.

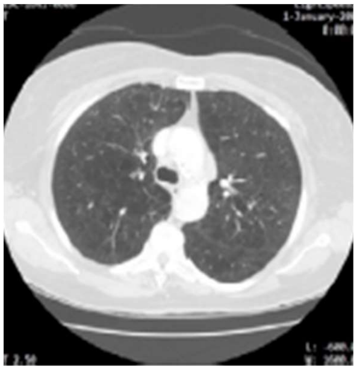

Fig. 6: Gray scale CT image

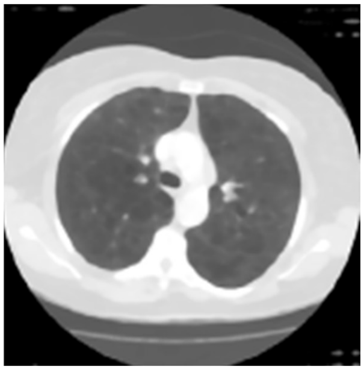

Fig. 7: Median Filtered Image

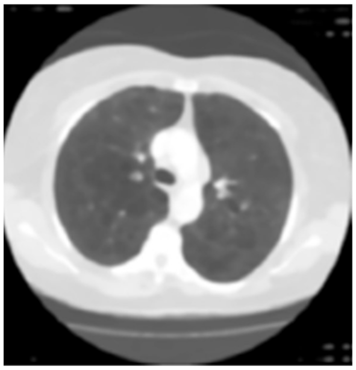

Fig. 8: Gaussian Filtered Image

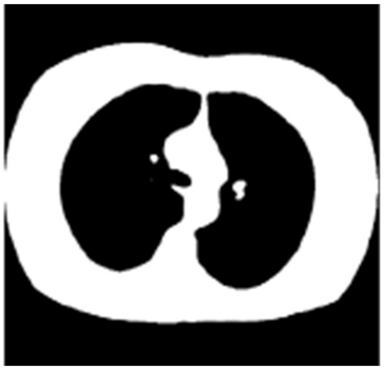

Fig. 9: Grayscale Image to Binarized Image

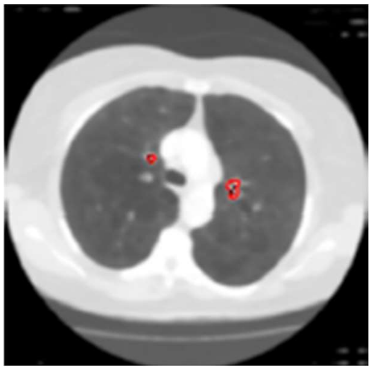

Fig. 10: Cancer Marked Image

In the above, image processing techniques are completed using various stages namely original grayscale image, median filtered image, Gaussian filtered image, binarized image, segmented image and cancer marked image respectively. Result of all CT scan images and its accuracy shown in table 1 and figure 11. In this case number of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) using the following equations (1), (2) and (3). In medical diagnostics, test sensitivity is the ability of a test to correctly identify those with the disease (TP rate), while test specificity is the test's ability to correctly identify those without the disease (TN rate). The overall dataset includes three type of cancer ST images. The dataset split namely total number of images, number of detected images (cancer) and number of normal images. The proposed work which is used to train the model as 70% of images and 30% of images for testing purpose. The detailed distribution mentioned in table 1.

Sensitivity and specificity describe mathematically the accuracy of a test that indicates the presence or absence of a disease. If people who have the disease are considered positive and those who don't are considered negative, then sensitivity is a measure of how well a test can identify true positives, and specificity is a measure of how well a test can identify true negatives. Sensitivity (true positive rate) is the probability of a positive test result given that the person is truly positive. Specificity (true-negative rate) is the probability of a negative test result, given that the person is truly negative. The terms sensitivity and specificity were introduced in 1947 by the American biostatistician Jacob Yerushalmy [18]. The performance of the model was measured by several factors including sensitivity, specificity, accuracy, and F1 score. False positive fractions and true positive fractions were used to demonstrate the ROC curve using table 1.

Table 1. Performance of the test statistics

Predicted Model	**Disease Detection**	**Normal**	**Total**
Predicted Model **Positive**	True Positive (TP)	False Positive (FP)	TP + FP
Predicted Model **Negative**	False Negative (FN)	True Negative (TN)	FN + TN
Total	TP + FN	FP + TN	TP + FP + FN + TN

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad \ldots (1)$$
$$\text{Sensitivity} = TP / (TP + FN) \quad \ldots (2)$$
$$\text{Specificity} = TN / (TN + FP) \quad \ldots (3)$$

Table 2: Training and Testing using CT Scan Images

CT Scan Images	**Total CT scan Images**	**Training (70%)**	**Testing (30%)**
Adenocarcinoma	347	243	104
Large Cell Carcinoma	187	131	56
Squamous Cell Carcinoma	260	182	78
Normal	215	151	64
Total CT Scan Images	1009	706	302

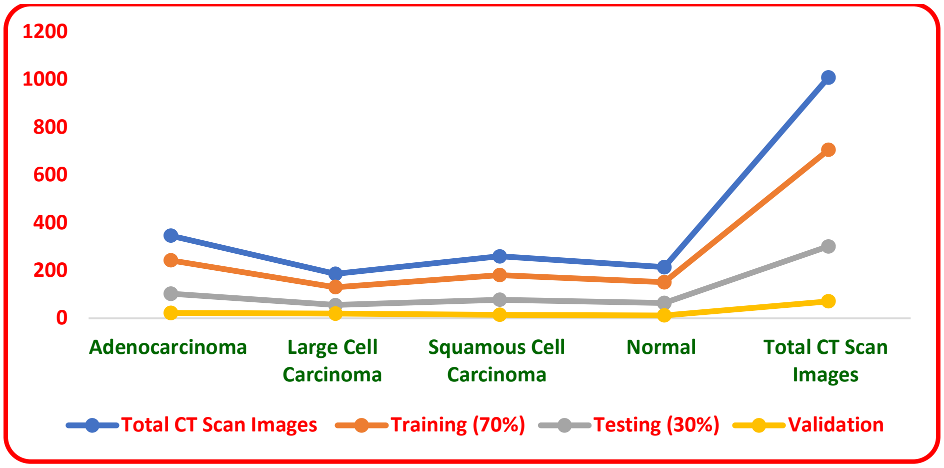

Fig. 11: Training and Testing using CT Scan Images

The result and discussion, comparative studies are one of the most important stage about the interpretation. In this case, easy to compare various literature with the proposed system using accuracy, sensitivity, and specificity equation 1, 2 and 3.

Different image processing techniques implemented using Google CoLab cloud environment with python. The configuration of the system Python 3 Google compute engine background with RAM: 1.10 GB/12 GB and Disk: 23.56 GB/107.72 GB. The following table 3 and figure 12 indicate the processing time for different image processing techniques.

Table 3: Image processing techniques and its computational time

Image Processing Techniques	Computational Time (In seconds)
Median Filtering	0.2277
Gaussian Filtering	0.3370
Binarized Image	1.2442
Cancer Marked Image	1.3605
Total time	3.1694

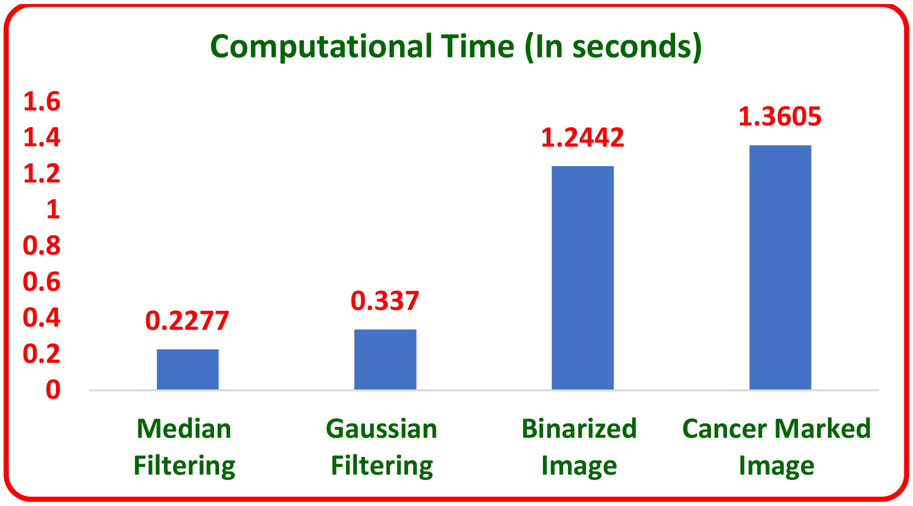

Fig. 12. Image processing techniques and its computational time

Table 4. Comparison analysis of proposed and existing model

Test Statistics Performance	Existing Model	Proposed Model
Total Number of Detected Images	58	58
Number of True Positive (TP)	50	52
Number of True Negative (TN)	3	2
Number of False Positive (FP)	2	2
Number of False Negative (FN)	3	2

Table 4. Test statistics performance of proposed and existing model

Statistic	Existing Model	Proposed Model
Sensitivity (%)	94.34	96.30
Specificity (%)	60.00	65.00
Disease Prevalence (%)	91.38	93.10
Positive Predictive Value (%)	96.15	96.30
Negative Predictive Value (%)	50.00	45.00
Accuracy (%)	91.38	93.10

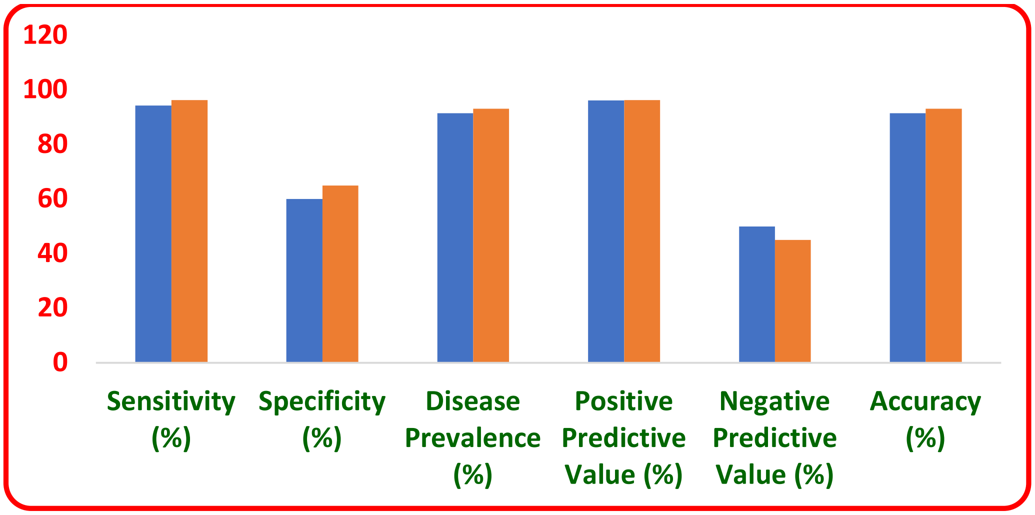

Fig. 14. Test statistics performance of proposed and existing model

The accuracy of proposed model, it can be seen that there is progressive increase in accuracy from 91.38% to 93.10%, the sensitivity increase from 94.34% to 96.30% and specificity increased from 60% to 62%. The related results shown in table 4, table 5 and figure 14. From the detected cancer, features like area, perimeter, centroid, diameter, eccentricity, and mean intensity of the Pixels were extracted. Extracted features were used to train support vector machine and trained model was developed. Training time for classification learner was 3.1694 seconds with finger 14.

6. Conclusion

The existing model has satisfactory result of accuracy and does not classify the degree of cancer of detected nodules. Therefore, a new system has been proposed. The proposed system is used to detect the cancerous nodule from the lung CT scan image using watershed segmentation for detection and SVM for classification of nodule as malignant or benign. Therefore, future scope improvement in this can be done by implementing classification in different stages. Also, further accuracy can be increased by proper pre-processing and eliminations of false objects.

References

1. Nadkarni, N.S. and Borkar, S., 2019, April. Detection of lung cancer in CT images using image processing. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 863-866). IEEE.
2. Gindi, A., Attiatalla, T.A. and Sami, M.M., 2014. A comparative study for comparing two feature extraction methods and two classifiers in classification of earlystage lung cancer diagnosis of chest x-ray images. *Journal of American Science*, *10*(6), pp.13-22.
3. Suzuki, K., Kusumoto, M., Watanabe, S.I., Tsuchiya, R. and Asamura, H., 2006. Radiologic classification of small adenocarcinoma of the lung: radiologic-pathologic correlation and its prognostic impact. *The Annals of thoracic surgery*, *81*(2), pp.413-419.
4. Xiuhua, G., Tao, S. and Zhigang, L., 2011. Prediction models for malignant pulmonary nodules based-on texture features of CT image. In *Theory and Applications of CT Imaging and Analysis*. IntechOpen.
5. Dimililer, K., Ugur, B. and Ever, Y.K., 2017. Tumor detection on CT lung images using image enhancement. *The Online Journal of Science and Technology*, *7*(1), pp.133-138.
6. Al-Tarawneh, M.S., 2012. Lung cancer detection using image processing techniques. *Leonardo Electronic Journal of Practices and Technologies*, *11*(21), pp.147-58.
7. Aerts, H.J.W.L., Velazquez, E.R., Leijenaar, R.T., Parmar, C., Grossmann, P., Cavalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D. and Hoebers, F., 2015. Data from NSCLC-radiomics. *The cancer imaging archive*.
8. Gonzalez, R.C. and Woods, R.E., 2002. Digital image processing. upper saddle River. *J.: Prentice Hall*.
9. Dwivedi, S.A., Borse, R.P. and Yametkar, A.M., 2014. Lung cancer detection and classification by using machine learning & multinomial Bayesian. *IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)*, *9*(1), pp.69-75.
10. Sun, W., Zheng, B., Lure, F., Wu, T., Zhang, J., Wang, B.Y., Saltzstein, E.C. and Qian, W., 2014. Prediction of near-term risk of developing breast cancer using computerized features from bilateral mammograms. *Computerized Medical Imaging and Graphics*, *38*(5), pp.348-357.
11. Chaudhary, A. and Singh, S.S., 2012, September. Lung cancer detection on CT images by using image processing. In *2012 International Conference on Computing Sciences* (pp. 142-146). IEEE.
12. Pratap, G.P. and Chauhan, R.P., 2016, July. Detection of Lung cancer cells using image processing techniques. In *2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)* (pp. 1-6). IEEE.
13. Bhusri, S., Jain, S. and Virmani, J., 2016, March. Classification of breast lesions based on laws' feature extraction techniques. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1700-1704). IEEE.
14. Kuruvilla, J. and Gunavathi, K., 2014. Lung cancer classification using neural networks for CT images. *Computer methods and programs in biomedicine*, *113*(1), pp.202-209.
15. Ignatious, S. and Joseph, R., 2015, April. Computer aided lung cancer detection

system. In *2015 Global Conference on Communication Technologies (GCCT)* (pp. 555-558). IEEE.

16. Rendon-Gonzalez, E. and Ponomaryov, V., 2016, June. Automatic Lung nodule segmentation and classification in CT images based on SVM. In *2016 9th International Kharkiv Symposium on Physics and Engineering of Microwaves, Millimeter and Submillimeter Waves (MSMW)* (pp. 1-4). IEEE.
17. Chest Cancer Detection - Transfer Learning. (2023, January 1). https://www.kaggle.com/code/abdelghaniaaba/chest-cancer-detection-transfer-learning/data
18. Sensitivity and specificity. (2023, January 27). In *Wikipedia*. https://en.wikipedia.org/wiki/Sensitivity_and_specificity.