PERIODICO di MINERALOGIA Volume 91, No. 3, 2022
ISSN: 0369-8963

Implementation of Machine Learning Techniques for detection of Lung Cancer

ARUNARASI JAYARAMAN[1)*], RAVULA SONALI[2)], MONISHA SANKAR[3)], DIVYA MURUGANATHAN[4)]

[1]*Electronics and Communication Engineering, Sri Sairam Engineering College, Chennai.*

[2]*Electronics and Communication Engineering, Sri Sairam Engineering College, Chennai.*

[3]*Electronics and Communication Engineering, Sri Sairam Engineering College, Chennai.*

[4]*Electronics and Communication Engineering, Sri Sairam Engineering College, Chennai.*

Abstract

An uncontrollable growth of body cells in lungs is called lung cancer. Lung cancer is ranked fourth among all the cancers over the world. Diagnosing of lung cancer is a vital step at early stages for increasing the survival rate. To detect lung cancer, CT (computed tomography) is an effective method to track lung cancer. In this proposed system, the software model is well - trained with different advanced CT images which are accurate and distinguished from the existing one. This proposed lung cancer model includes methods for segmenting the lung regions out of which GLCM parameters are collected and used for categorization of tumour stages using CNN. This proposed work has an efficiency of 98 percent.

Key words: *Computed tomography, Lung cancer, Haralick features, Haar wavelet transform, Convolutional neural network (CNN).*

Introduction

Lung cancer is a cancer that starts in the lungs and expands to the rest of the body. It is one among the world's top causes of death. Active smokers have the highest risk of developing cancer. Even non-smokers who are exposed to cigarettes for an extended amount of time are harmed. The risk of developing this condition rises with the number of cigarettes smoked or the amount of time spent in the presence of smoke. Pollutants in the air have also been linked to lung cancer. Cancer cells reproduce quickly and expand in a nonlinear way, resulting in tumour formation because these cancer cells do not have any indications or symptoms in their early stages, it is difficult to diagnose the condition. This results in development of the tumour to a severe stage where curing is difficult. Detecting lung cancer early, can raise the chances of survival by 50-70 percent.

Aim and Intention

The principle objective is to provide a system that will assist clinical experts in cross-confirming their expected lung cancer outcomes. Since the current diagnosis procedure is time-consuming, tedious, and expensive, this deep learning-based technology can, for the most part, detect cancer growth and forecast stages. As this is a computerized technique based on image processing and Artificial Intelligence, it reduces the amount of time it takes a human to forecast the existence of cancer cells from an image.

 https://doi.org/10.37896/pd91.3/91323

ISSN: 0369-8963

Literature Review

It is suggested that pre-processing of CT scans of the lungs, which includes converting RGB to gray value and gray scale to binary pictures in order to obtain an accurate and precise image by decreasing interference [1]. The data is segmented using watershed segmentation and thresholding [9][16]. Binarization and masking are used to finish feature extraction. The images are utilized as feed to the convolutional neural network classifier to determine whether the CT image is malignant or non-cancerous [10].

To diagnose lung cancer, [2][5] used ANN Back-propagation based GLCM features. The lung data comes from the Cancer Imaging Archive Database, which includes 50 computed tomography images. Image pre-processing, segmentation, feature extraction, and cancer progression detection utilizing a three-layer neural network back propagation approach are all steps in this strategy. The System can discriminate between healthy lung and lung disease with an efficiency of 80% or greater, according to the findings[4].

The common method for detecting and assessing lung cancer is depicted in [3]. Expert evaluations on various parameters characterizing a nodule's morphology and shape are frequently used in clinical practice to assess the malignancy of lung nodules, but these criteria are primarily objective and randomly defined [13][15].Various Image Processing techniques using MATLAB[14] as a platform, a number of studies on Lung Cancer Recognition and Classification have been conducted for CT images as input [7].A wide evaluation of the most essential algorithms utilized in the CAD application for lung tissue diagnostics, with emphasis on each algorithm's performance[6].

Some systems to utilize Artificial Neutral Networks to detect lung cancer (ANN) to provide insufficient precision [5]. It is relatively simple to use and excellent for complex/abstract issues such as picture identification, although increasing efficiency by a few percent can increase the scale by several orders of magnitude[11]. The implementation of the system required a larger area before using it in the clinic [8][17].Different phases of analyzing the system are involved by pre-processing, segmentation, feature extraction and classification [18]. The pre-processing includes using of median filter followed by watershed segmentation [12]. The results show improvement in the classifier. It relatively focuses on the modelling techniques that are applied to two feature sets.

Proposed System

A set of images is used to train the suggested system. The test picture is given as an input to the system. To categorise an image as malignant or non-cancerous, a series of processes are taken. In case of a cancerous image, the area affected by the cancerous cell is separated by a yellow boundary. The test images are taken from kaggle. The test images are the CT images of the human lung. In this system, CT images are used rather than x-ray images because they have better image quality. After the data collection, the image is proposed using two different filtering method. The filtering methods used are median filter and the Gaussian filter. The median filter removes salt and pepper noises from the image, whereas the Gaussian filter reduces image noise and fuzzy areas.

 https://doi.org/10.37896/pd91.3/91323

ISSN: 0369-8963

Next after pre-processing image is sent to the segmentation domain. The system is given the test image as an input. Segmentation is the process of separating an image data into several segments and determining the image's limitations. Once the image is segmented classification of the image is done. In Classification the proposed work uses CNN classifier. A class of deep neural network that recognizes and classifies particular features in an image is KNN classifier. It is used widely in analysing visual images. Once the image is processed in a CNN classifier the image is identified to be cancerous or noncancerous image. The area affected by the cancer cell are bounded by a yellow.

Methodology and Implementation

The step-by-step flow of the proposed work is explained in the flow chart. The detailed explanation of each step is followed below

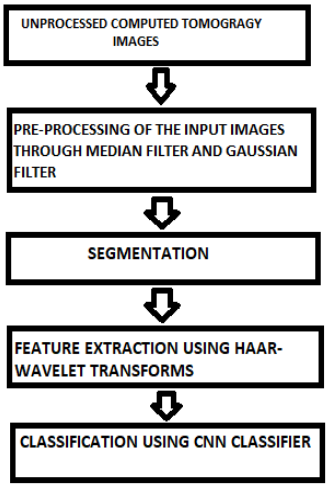

Figure1. Proposed Methodology

Data Collection

The primary work in this task is to get a CT scan image of the user. CT scans are used as input because Computed Tomography images are less noisy than X-ray and Magnetic resonance imaging. These photos are used to improve accuracy and reduce distortion. The input photos are taken from kaggle which has around 200 lung images or more from affected and non-affected patients CT scan reports. More noise can be seen in the obtained photos.

Figure 2. Input CT image

https://doi.org/10.37896/pd91.3/91323

ISSN: 0369-8963

The input given to the model is the CT scan of a patient 's lung. Figure 2 represents an example of a lung CT image of a person. The purpose of using a CT image is that it has better detailing towards the tissues, lung, bones and blood vessels. In general CT images have a short study time with high quality images.

Pre-processing

Resizing the image is done in the first phase to remove any unwanted areas of the image. After that, the images are submitted to median filters, which can be used to eliminate the salt and pepper noise. The Gaussian Filter is used to minimize image noise and fuzzy areas. Canny Edge Detector is a multi-level edge detection operator that detects many types of edges in an image.

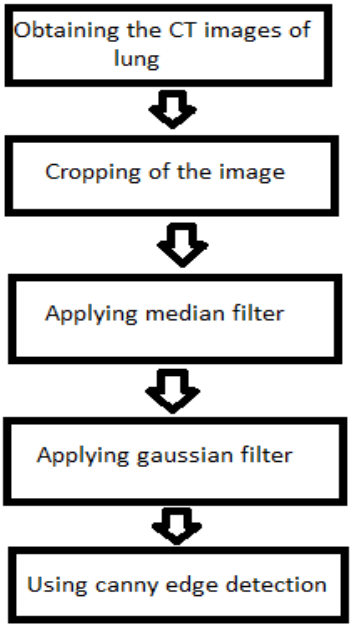

Figure3. Flow diagram of the Image prepossessing

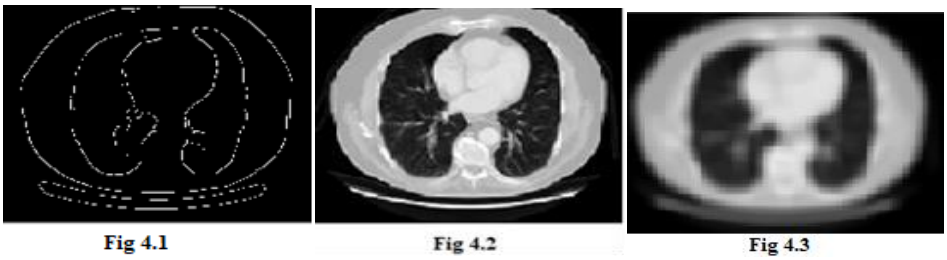

Figure 4 .1 Represents the edges in the image after Filtering **Figure 4.2** Median filtered image **Figure4.3** Gaussian Filtered image

https://doi.org/10.37896/pd91.3/91323

ISSN: 0369-8963

The above figures represent the different types of filters used in the system. Filtering of an image is essential to reduce the higher frequency numbers and to have a better image quality for the model to differentiate between a cancerous cell and a non - cancerous cell.

Segmentation

In the domain of medical imaging, image segmentation is used as a part of the screening procedure. This is a technique for segmenting a digital photo and defining the borders between objects and images. The main purpose of segmentation is to analyze the information in CTscan images and transform them into more informative and effective ones so that they can be examined in more detail.

Figure 5. The figure depicts the segmented image of the CT lung scan

The system's segmentation process consists of steps such as k means and active contour means divides the unlabelled dataset into various clusters. The active contour segmentation approach can be used to analyse both dynamic and 3D picture data.

Feature Extraction

Feature extraction aids in the extraction of significant data items that are fed into the classifier as input. Fig 6 shows the flow of the Feature extraction process step-by-step. The image is primarily resized into 3 distinct resolutions, then HWT are implied to these pictures. The Haar wavelet is an orthogonal wavelet that is the most commonly used in image processing. Haar wavelets use less memory and, unlike other wavelets, are exactly reversible. This wavelet calculates pair wise averages and differences by solely reflecting changes between nearby pixel pairs. Following that, the GLCM is constructed in four different ways, with seven features retrieved from each.

 https://doi.org/10.37896/pd91.3/91323

ISSN: 0369-8963

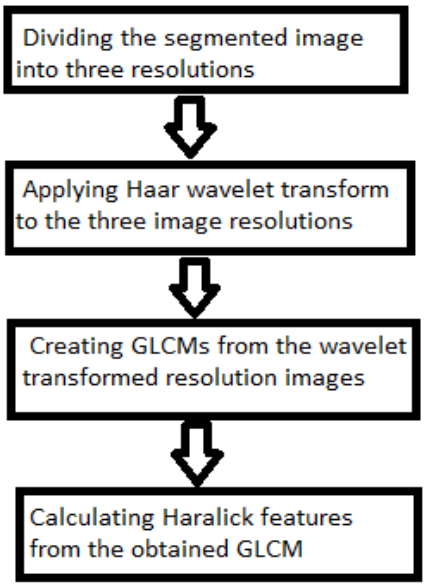

Figure 6: Flow chart for Feature Extraction

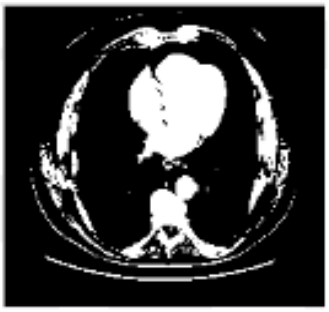

Figure 7. Gray scale image

The GLCM function determines an image's texture by counting how many pixel combinations share the same value of the pixel and spatial relationship occur. After the gray-level co-occurrence matrix has been designed, second order statistical characteristics, also known as Haralick features, must be extracted.

Classification

CNN is a powerful pattern recognition and image processing method. The main features of CNN are that it has a simple structure, less training parameters and greater adaptability. Picture identification, object detection, and segmentation are a few tasks that CNN is employed for. The primary advantage of CNN over ANN is that without any human intervention, it automatically recognises the key traits.

Results and Discussion

The result in the proposed work is the testing accuracy determined to be 98% on average. The area surrounded by yellow line is determined to be the area affected by the cancer cells.

https://doi.org/10.37896/pd91.3/91323

ISSN: 0369-8963

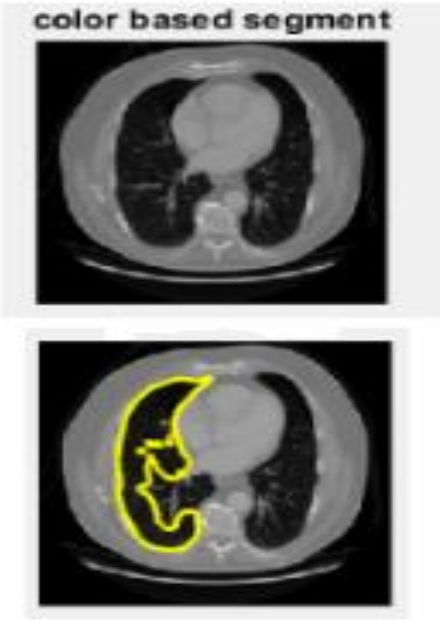

Figure 8. The image shows the area affected by the cancer cells.
The area affected is bounded by a yellow line

Through the figure 8. we infer that the person is suffering from lung cancer and the part at which the cancer cells are present are detected.

Future Scope

More images, such as X-rays, CT scans, MRIs can be used to improve accuracy, allowing healthcare professionals to provide rapid treatment at a cheap cost. Since this disease has a negative economic impact, more research should be conducted to identify information gap on disease prevention and detection techniques, which will assist in the design of vaccines and other control measures.

Conclusion

The proposed system strategies attempt to provide a computer algorithm for lung cancer detection. The use of a median Gaussian filter to remove impulsive noise from photographs was successful. The removal of impulsive noise from images using a median, Gaussian filter proved successful. The morphological techniques are also important contribution to successful outcomes in the process of segmentation. CNN has shown to be an effective tool and a classifier that is accurate enough.
The goal of this method is to increase the lung cancer detection system's accuracy, speed and also aid in the detection of lung cancer at an earlier stage.

 https://doi.org/10.37896/pd91.3/91323

ISSN: 0369-8963

References

.1. BHALERAO, R. Y., JANI, H. P., GAITONDE, R. K., & RAUT, V, *Advanced Computing & Communication Systems (ICACCS) ,* pp. 577-583. IEEE. (2019)

2. YUTONGXIE, *IEEE,* (2018).

3. LILIKANIFAH, HARYANTO, RINAHARIMURTI, *IEEE,* (2017).

4. ALAM, J., & ALAM, S, *Computer, Communication, Chemical, Material and Electronic Engineering*, (pp. 1-4). *IEEE,* (2018).

5. RAOOF, S. S., JABBAR, M. A., & FATHIMA, S. A. In *2020 on innovative mechanisms for industry applications* pp. 108-115, *IEEE,* (2020).

6. ARUNARASI J, INDUMATHY PUSHPAM, *European Journal of Scientific Research*,53(2), 2011.

7. RIQUELME, D., & AKHLOUFI, M. A. *AI*, *1*(1), 28-67, (2020).

8. WU, Q., & ZHAO, W. (2017, October). In *2017 International symposium on computer science and intelligent controls (ISCSIC)* (pp. 88-91). IEEE, (2017).

9. RADEEP, K. R., & NAVEEN, N. C. *Procedia computer science*, 132, 412-420.,(2018)

10. SIM, J. A., KIM, Y., KIM, J. H., LEE, J. M., KIM, M. S., SHIM, Y. M., ...& YUN, Y. H. *Applications of machine learning*. *Scientific reports*, 10(1), 1-12, (2020).

11. SHANTHI, S. *Big Data Innovation for Sustainable Cognitive Computing* (pp. 255-266). Springer, Cham, (2021).

12. YASSIN, N. I., OMRAN, S., EL HOUBY, E. M., & ALLAM. H, *Computer methods and programs in Biomedicine*, 156, 25-45, (2018).

13. SUNDARASEKAR, R., & APPATHURAI, A, *Fluctuation and Noise Letters*, 21(03), (2022).

14. ARUNARASI J, INDUMATHY PUSHPAM, *Journal of Theoretical and Applied Information Technology*, 52(1), (2013).

15. KALYANI, R., SATHYA, P. D., & SAKTHIVEL, V. P. *International Journal of Intelligent Engineering and Systems*, 14(2), (2021).

16. ALZUBAIDI, A. K., SIDESEQ, F. B., FAEQ, A., & BASIL, M. *Information & Communications Technology applications*, pp. 219-224, IEEE, (2017).

17. VAIYAPURI, T., DUTTA, A. K., PUNITHAVATHI, I. H., DURAIPANDY, P., ALOTAIBI, S. S., ALSOLAI, H., ... & MAHGOUB, H. *In Healthcare*, 10(4), pp- 677, (2022).

18. CHEN, X., HAO, B., LI, D., REITER, R. J., BAI, Y., ABAY, B., ... & FAN, L, *Journal of pineal research*, 71(2), (2021).

 https://doi.org/10.37896/pd91.3/91323