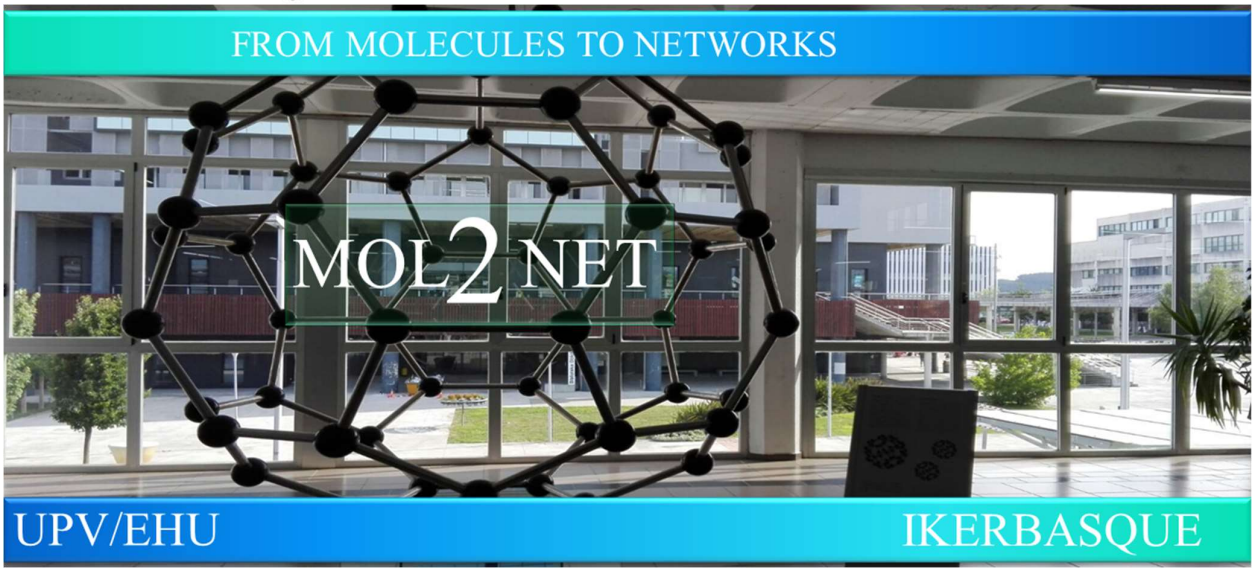
MOL2NET'22, Conference on Molecular, Biomedical & Computational Sciences and Engineering, 8th ed.



Machine Learning-Based Classification of Lung Cancer Using CT Scan Images

Aqib Ali [, a], Samreen Naeem [a], Sania Anam [b], and Muhammad Zubair [b]*

[a] *College of Automation, Southeast University, Nanjing, China.*
[b] *Govt Associate College for Women Ahmadpur East, Bahawalpur, Pakistan.*

. * *Corresponding author: aqibcsit@gmail.com*

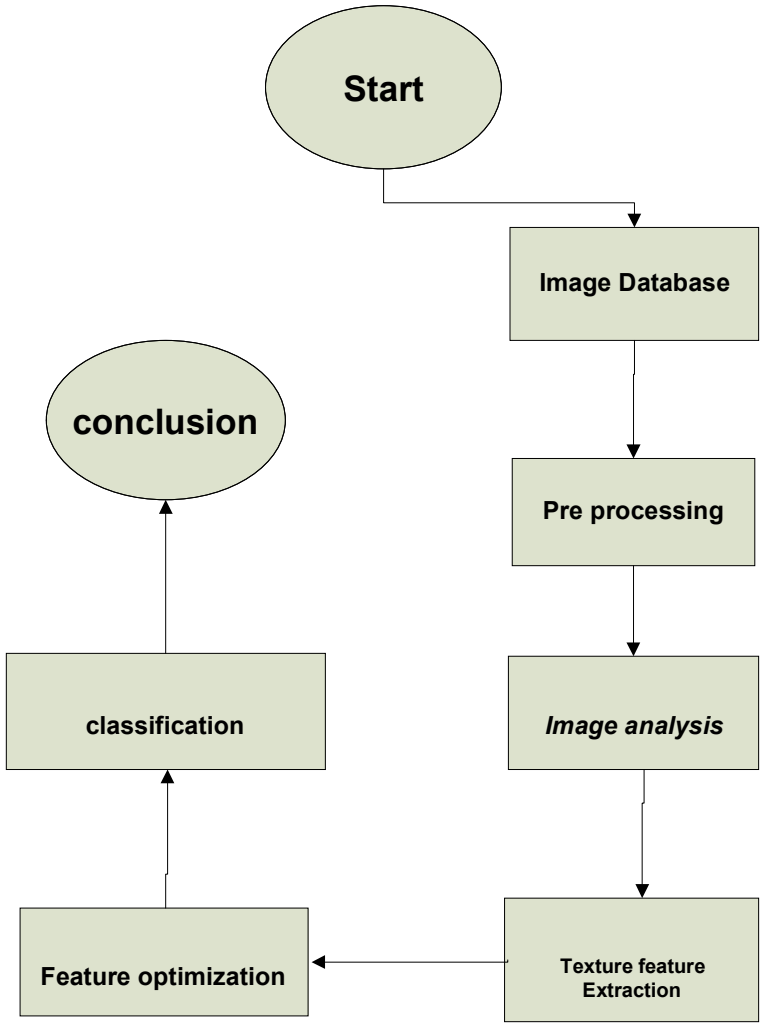
Abstract.

Lung cancer is one of the most precarious dysfunctions to humankind species and amongst the leading causes of human life expiration, especially in developing countries. Mycobacterium Tuberculosis bacterium is a causative agent of lung cancer. The highly aerobic physiology of M. tuberculosis requires suitable oxygen for survival, which is why Lung is its habitat. Lung cancer is fatal because its detection is challenging, especially in smokers. This study presents a machine vision-based approach for lung cancer detection through CT (computerized tomography) scan images. The study aims to ensure reliable, precise, and accurate detection of lung cancer through texture features extracted from CT scan images (acquired from Bahawal Victoria hospital Bahawalpur, Pakistan). Pre-processing techniques (grayscale conversion, filtration, etc.) played an influential role in removing noise, which might reduce accuracy. Mazda tool has been used for feature extraction and identification of 30 optimized features using three techniques F (Fisher) + PA (probability of error + average correlation) +MI (mutual information). The data mining tool Weka has deployed different classification

algorithms with ten cross-validation folds. Artificial Neural Network (ANN: n class) showed comparatively better and probably best accuracy of 95.66 %, respectively.

Keywords: Lung Cancer, Machine Learning, Optimized Features, Artificial Neural Network

Graphical Abstract



Introduction

Lung cancer is a fatal and life-threatening disease for men and women. The main reason is smoking. The survival rate from this disease is very low, but one thing can save the patient if cancer detection is done at early stages. Usually, the human body grows slowly, but cancer cells grow fast without any ratio. This abnormal division process of cells is called cancer. Cancer cells divide abnormally or can damage other related cells, so early detection and control are challenging for doctors. National Cancer Institute (NCI) 2014 reports that 224210 lung cancer cases were found in which 159260 people died due to lung cancer [1]. It is difficult to detect lung cancer early on with the naked eye, so a particular

technique or method must be needed to diagnose it accurately with early symptoms. So, Computer Tomography (CT) is one of the most popular ways to detect lung cancer by using CT images of the patient. Computer Tomography (CT) images are low in cost and give a precise result [2].

According to [2], images are used as input, and this input is used for enhancement. After enhancing the image, it is segmented and further used for feature extraction, so the result extracted is helpful for cancer detection. So, like other systems, it is also beneficial for the early detection of lung cancer [3]. Lung cancer is a dangerous and death-causing disease worldwide. According to the World Health Organization (WHO), 64 million people worldwide have been infected, and the death rate can lead to one-third in 2030. According to another report by WHO, 1.5 million people died in 2012 due to lung cancer, so according to that reason, it is becoming life-threat to human beings [3]. This disease can only be engulfed by detecting accurate results, or proper medical treatment can lead the patient to a healthy, safe life. So, by using CT scan images, we can detect it early and improve survival rate. We are trying to detect lung cancer by collecting new data from different laborites and hospitals via online systems next step is analyses that are necessary to organize data.

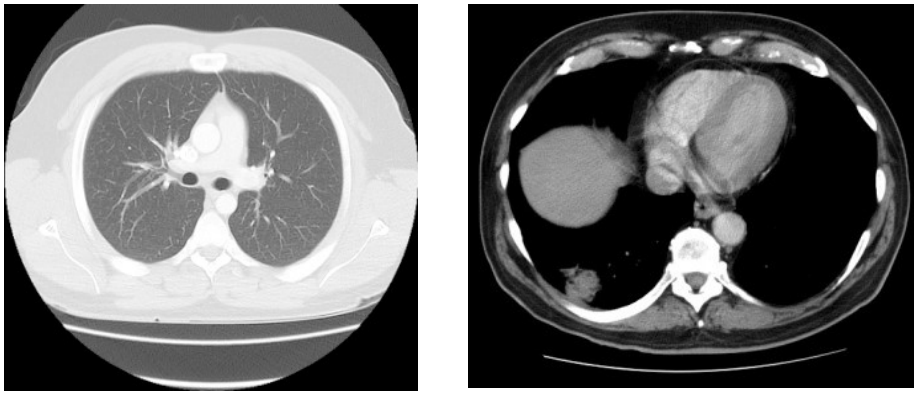
Lung Cancer is a dangerous disease. Different methods or techniques are implemented in this field to diagnose this disease early. Every field of science is doing hard work to overcome it. Computer science is also putting his part in it. Computer science is a very vast field, so we selected Artificial Intelligence to detect Lung cancer at an early stage by using a different methodology [4]. Artificial Intelligence is also a waste field. Many authors have done work on it by using different methodologies. I selected a different methodology to diagnose it by taking 100 CT scan images. In which 50 images are collected from regular patients, 50 scan images are abnormal, which may be Benign and malignant cancer. We use CVIP tools and Weka software to enhance feature extraction, feature reduction, classification, and image processing. Approximately 100 images are collected and examined in different ways to test lung cancer. The primary methodology proceeds with image cropping, resizing, median filters, and different algorithms like BF (Best first), Greedy CFS, and classifiers. Feature Optimization of CT scan images is our next step. After that, features are calculated in proper fomite, and results are obtained in different varieties. Ultimately, the classification of cells and data will differentiate between cancer and normal cells [5].

Lung cancer is a gender-independent disease and increasing day by day in the whole world. The patient has no physical symptoms or signs until several conditions are reached. Million people are dying due to it. Considerable research has been done on it, and every field of science plays an essential part in overcoming this disease to save a life. According to NCI, 159,260 people from 224210 died from lung cancer in 2016 [6]. American Cancer Society reported on its site in 2018 that 1685,210

people are affected by cancer, and approximately 595,690 are dead due to that disease per day rate of deaths is 1630 due to cancer of different types. Artificial Intelligence plays a significant role in diagnosing different types of cancers. Different methods have been introduced that are implemented using CT scan images that are mostly cheap, easy to use, and understandable. MRI images are mainly used for brain tumors. KNN, ANN, FMNN, MATLAB, CVIP tools, Sobel edge detection, Gaussian method, CAD system, and XRAY images are also crucial in cancer detection [7].

Materials and Methods

An image database is a collection of CT scan images of Lung cancer that are a collection of images collected from different sources. These Images are collected and arranged in a well-managed and disciplined way, used when necessary, and easily accessible when processing starts. We collected different CT scan image datasets of Lung cancer for processing and early detection of the disease. These data sets are collected from two leading websites (www.cancerimagingarchive.net) and second is (radiopaedia.org) [8]. These are two popular sites for collecting any cancer dataset. Data set is an Image processing able to be distinct because the act of an image appearance a quantity of basis. Image dispensation is forever the primary footstep in the exertion stream series since, with no image, no dispensation is likely. I have acquired two types of samples of lung CT scan images. CT scan images of a normal person and CT scan images of persons having cancer. Each type has acquired 50 image data sets. I have collected 100 images for data sets of two different categories.



Human Lung (Normal) **Human Lungs (Cancer)**

Figure 1: CT Images of Human Lung **[4]**

The implementation of different methods and techniques in image preprocessing to extract limited features and the best algorithm is applied to get the best results. In this chapter, results are obtained and explained using the proposed methodology with the help of CVIP tools version 5.6e and Weka [9]. These are the best result generation and feature selection software in image processing,

mainly in medical science. First, the data set consists of 6 patients ' CT scan images in two groups, standard patient CT scan images and patients with lung cancer, with 50, 50 images. A complete 100 CT scan images are collected in a dataset. In the next step, these images are used in CVIP tools after cropping and resizing the web's small size in an image editor. Cropped images are converted into grayscale levels in utilities. All 100 images are cropped and resized equally, then one by one, changed into color to gray, and created a circle to select the best location. It is an ROI method to select specific parts. Each image has created five circles with (512*512). Width and height. Colum and row are started with (128*128), and radius wand blur radius with (32*32).

Each image created 3 ROI by the same width, height, and radius, and the blur radius of the row and Colum changed in each image with the addition of 16 values. So, a total of 300 ROI is created from 100 images. This ROI is saved by selecting binary, histogram, and Texture features. These are a total of 23 features. After creating a text file, this file is converted into a CSV file, and then an arff file is created from weka software to generate other results. Weka files are opened, and two main search methods are applied for optimization to extract the best features [10].

Results and Discussion

.

- Classification model: Artificial Neural Network (ANN: n class)
- Time taken to build the model: 0.34 seconds
- Test mode: 10-fold cross-validation

Table 1: ANN Classifier Summary

Total Number of Instances	300	
Correctly Classified Instances	287	**95.6667 %**
Incorrectly Classified Instances	13	**4.3333 %**
Kappa statistic	0.9133	
Mean absolute error	0.0563	
Root mean squared error	0.1906	
Relative absolute error	11.2592 %	
Root relative squared error	38.1158 %	

Table 2: ANN Classifier Detailed Accuracy

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	Class
0.947	0.033	0.966	0.947	0.956	0.914	0.979	**Normal**
0.967	0.053	0.948	0.967	0.957	0.914	0.979	**Abnormal**
0.957	**0.043**	**0.957**	**0.957**	**0.957**	**0.914**	**0.979**	**Weighted Avg.**

Table 3: Confusion Matrix using ANN Classifier

Classified as	A	B
A = Normal	142	8
B = Abnormal	5	145

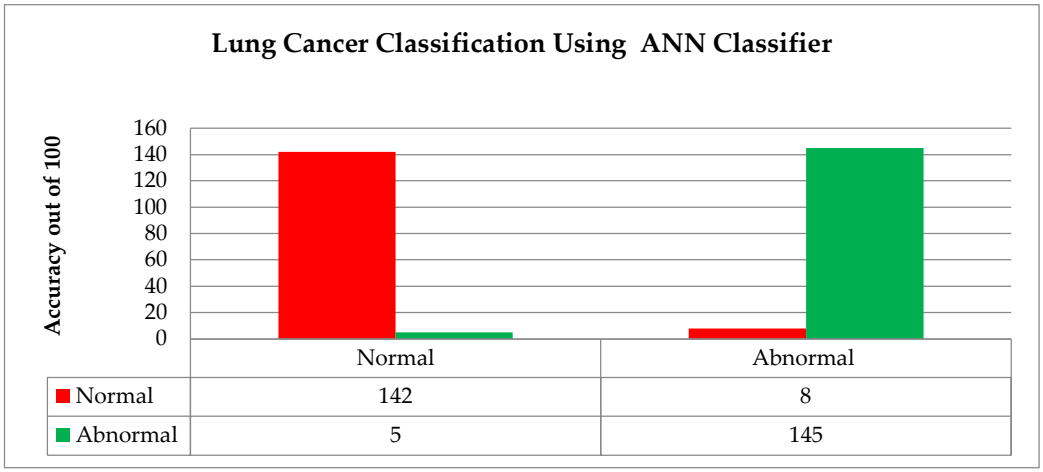


Figure 2: Accuracy of Dataset using ANN Classifier (10-Fold)

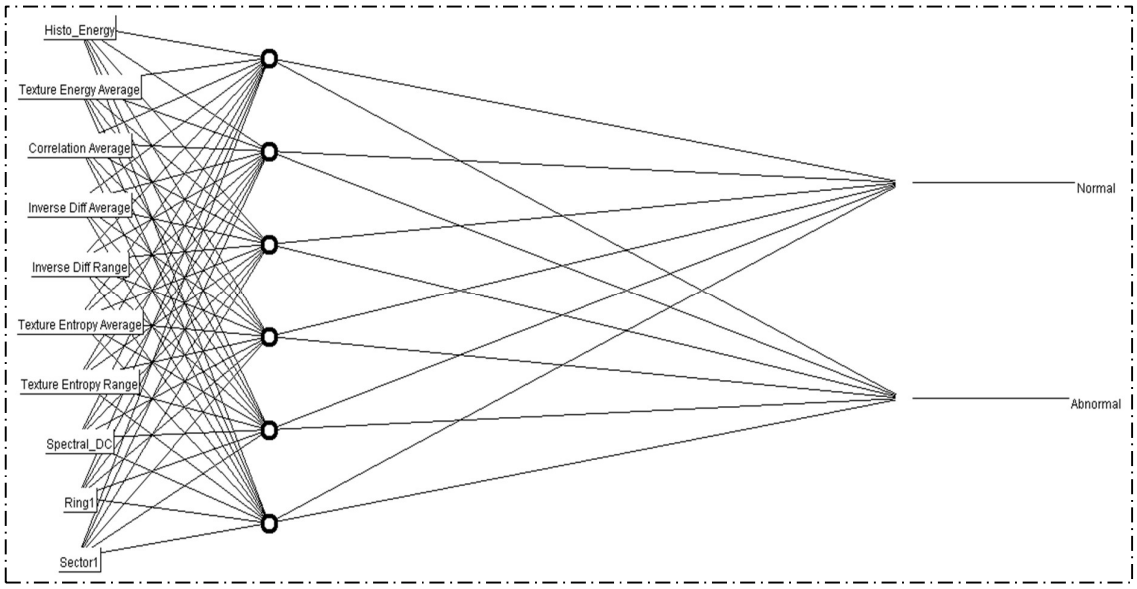


Figure 3: Graphicly User Interface of ANN Classifier Result (10-Fold)

Conclusions

In this research, our significant idea is to distinguish two main categories of lung cancer CT scan images of the lung that texture analysis and texture element are calculated beginning every ROI {512x512}, {128x128}, {32x32} using CVIP tool software, and results are generated using weka software. Different combinations of filters and features are used to generate different results. So, these methods are helpful for different radiologists to diagnose lung cancer. We achieved 95.66 % as the maximum correctness for this work by using ANN classifier.

References

[1]. Thandra, K. C., Barsouk, A., Saginala, K., Aluru, J. S., & Barsouk, A. (2021). Epidemiology of lung cancer. Contemporary Oncology/Współczesna Onkologia, 25(1), 45-52.

[2]. Woodman, C., Vundu, G., George, A., & Wilson, C. M. (2021, February). Applications and strategies in nanodiagnosis and nanotherapy in lung cancer. In Seminars in cancer biology (Vol. 69, pp. 349-364). Academic Press.

[3]. ALI, A.; NAEEM, S.; ZUBAIR, M. Machine Learning Based Classification of Chronic Kidney Disease Using CT Scan Images, in Proceedings of the MOL2NET'22, Conference on Molecular, Biomedical & Computational Sciences and Engineering, 8th ed., 26–31 December 2022, MDPI: Basel, Switzerland, doi:10.3390/mol2net-08-13903

[4]. Howlader, N., Forjaz, G., Mooradian, M. J., Meza, R., Kong, C. Y., Cronin, K. A., ... & Feuer, E. J. (2020). The effect of advances in lung-cancer treatment on population mortality. New England Journal of Medicine, 383(7), 640-649.

[5]. Salehi-Rad, R., Li, R., Paul, M. K., Dubinett, S. M., & Liu, B. (2020). The biology of lung cancer: development of more effective methods for prevention, diagnosis, and treatment. Clinics in chest medicine, 41(1), 25-38.

[6]. Xie, Y., Meng, W. Y., Li, R. Z., Wang, Y. W., Qian, X., Chan, C., ... & Leung, E. L. H. (2021). Early lung cancer diagnostic biomarker discovery by machine learning methods. Translational oncology, 14(1), 100907.

[7]. Naeem, S., Ali, A., Qadri, S., Khan Mashwani, W., Tairan, N., Shah, H., ... & Anam, S. (2020). Machine-learning based hybrid-feature analysis for liver cancer classification using fused (MR and CT) images. Applied Sciences, 10(9), 3134.

[8]. Richards, T. B., Soman, A., Thomas, C. C., VanFrank, B., Henley, S. J., Gallaway, M. S., & Richardson, L. C. (2020). Screening for lung cancer—10 states, 2017. Morbidity and Mortality Weekly Report, 69(8), 201.

[9]. Manoharan, S. (2020). Early diagnosis of lung cancer with probability of malignancy calculation and automatic segmentation of lung CT scan images. Journal of Innovative Image Processing (JIIP), 2(04), 175-186.

[10]. ALI, A.; NAEEM, S.; ANAM, S.; ZUBAIR, M. Machine Learning-Based Automated Detection of Diabetic Retinopathy Using Retinal fundus images., in Proceedings of the MOL2NET'22, Conference on Molecular, Biomedical & Computational Sciences and Engineering, 8th ed., 26–31 December 2022, MDPI: Basel, Switzerland, doi:10.3390/mol2net-08-13906