Lung Cancer Detection By Using Deep Learning

Vignesh Kumaran.S[1], Dr.K.A. Jaya Balaji[2]
[1]Dept of Computer Science
[2]Assistant Professor
[1, 2] Sri Krishna Arts And Science College, Coimbatore

Abstract- *Cancer is a grave and widespread affliction that claims numerous lives each year. Among the plethora of cancer variants, lung cancer stands as the most prevailing, boasting an alarming mortality rate. Employed for the discernment of lung cancer, computed tomography scans offer intricate depictions of tortuous within the body, enabling the monitoring of their progression. Notwithstanding its preference over alternative imaging modalities, the visual interpretation of CT scan images can be susceptible to errors, potentially leading to delays in lung cancer identification. Consequently, the realm of medicine extensively employs image processing methodologies to facilitate the early-stage detection of lung tumours. This endeavour introduces an automated framework for the identification of lung cancer within CT scan images. The algorithm for detecting lung cancer introduces techniques such as median filtering for initial image pre-processing, succeeded by the segmentation of the lung region of interest through mathematical morphological operations. Geometric attributes derived from the extracted region of interest are then harnessed to discern and categorize CT scan images as either normal or abnormal, facilitated by a Convolutional Neural Network.*

Keywords- Lung Cancer, Machine Learning, Artificial Intelligence, Deep Learning, Image Analysis, Medical Imaging, CT scan, Classification, Support Vector Machines (SVM)

I. INTRODUCTION

Image processing is one of the core areas used in various domains. It is used to identify cancer affected regions in lung images. Identification of cancer affected regions in lung is mainly initiated with image processing techniques such as noise removal, feature extraction, identification of affected regions and possible comparison with historical data of lung cancer. Usually, digital image processing follows many techniques to unite different shapes in an image into a single unit. In this work, it follows a clever technique for identifying a particular region in the lung image. The region identified through the segmentation technique can be viewed from different angles and with different lighting. The basic advantages in choosing the technique are to identify the colour difference between cancer affected regions and those parts which are not affected, with the intensity of images

In the 20th century, digital image processing is done through optical device. In 1960s and 1970s, image processing was introduced in the medical field. The information related to the image are represented in the form of 2D image with X and Y co-ordinates, to calculate the amplitude of the image. The function values are computed and the image is said to be digital in case all those values are finite [9].

Many existing image sensing tools are feasible through digital image processing. The data processed digitally retrieves pixel information and stores in a repository. The visual images are gathered in digital format and unnecessary data are removed. The purpose to do so is to compute minimum processes which are not applicable for large data. If process and resource are limited, the data makes the study of the visual easy. To remote sensing data, digital image processing does not have any limitations over these data [5]. Image processing techniques are developed with the use of X-ray and body scanning image devices in the medical field. Low level processing is used to get information for the image processed. Image process techniques demand an extremely powerful processor because the software package has to processes larger images [5].

II. ARTIFICAL INTELLIGENT MEACHINE LEARNING

AI/ML, an abbreviation for artificial intelligence (AI) and machine learning (ML), marks a significant advancement in the realm of computer science and data handling, heralding a rapid revolution across numerous sectors.

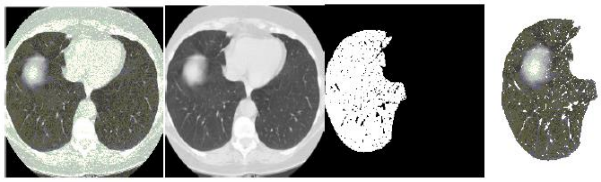
In the midst of businesses and various establishments embracing digital metamorphosis, they encounter a mounting surge of data that holds immense value yet progressively becomes more challenging to gather, handle, and interpret. There is a pressing demand for innovative techniques and tools to effectively oversee the copious amounts of accumulated data, extract meaningful perspectives from it, and take decisive actions upon unearthing those perspectives.

III. LUNG CANCER TYPES AND STAGES

Lung cancers are broadly categorized into two primary groups: Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC). NSCLC comprises 85-90% of lung cancer cases, while SCLC constitutes the remaining 10-15%. The progression and dissemination of NSCLC occur at a gradual pace, whereas SCLC is an aggressive cancer that rapidly metastasizes to other parts of the body. Smoking is the predominant factor contributing to SCLC cases, whereas NSCLC is managed through a combination of surgical intervention, chemotherapy, and radiotherapy, dependent on the stage of cancer diagnosis. In the context of SCLC, chemotherapy primarily serves as the treatment approach. A comparison between SCLC and NSCLC is detailed in Table 1. NSCLC manifests in four distinct types, each warranting specific treatment strategies:
Epidermoid / Squamous Cell Carcinoma: This type of cancer originates in the lining of the bronchial tubes and is more prevalent in males.Adenocarcinoma: Arising in the mucus glands of the lungs, this cancer is most commonly observed in females and non-smokers.

Bronchioalveolar Carcinoma: A rare variant of adenocarcinoma, this cancer develops near the air sacs of the lungs.

4Large-Cell Undifferentiated Carcinoma: Emerging close to the lung surface or its outer boundaries, this type exhibits rapid and extensive spread



(a) (b) (c) (d)

IV. DATA SET

There are different types of datasets for lung cancer detection, depending on the source, format and purpose of the data. For example, you can find:

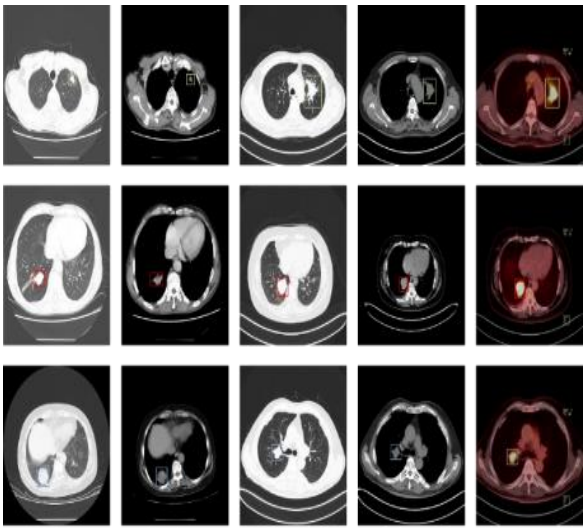
- A large-scale CT and PET/CT dataset for lung cancer diagnosis with XML annotation files that indicate tumor location with bounding boxes.
- A lung cancer prediction system dataset that contains information about symptoms, smoking history, family history and diagnosis.
- A text dataset for lung cancer prediction from medical reports using machine learning.

This dataset consists of **CT and PET-CT DICOM images** of lung cancer subjects with XML Annotation files that indicate tumour location with bounding boxes. The images were retrospectively acquired from patients with suspicion of lung cancer, and who underwent standard-of-care lung biopsy and PET/CT.

V. DATA PRE-PROCESSING

Data preprocessing is an important step in machine learning for lung cancer detection. It involves improving the quality of the input data, such as images or text, by removing noise, enhancing contrast, normalizing, segmenting, extracting features and selecting relevant variables. Data preprocessing can help reduce the dimensionality of the data, improve the accuracy of the machine learning models and speed up the training process.

The methodology section outlines a precise approach for the classification and prediction of lung cancer through the utilization of machine learning and image processing technology. Initially, the process involves the acquisition of images, followed by a preprocessing stage that incorporates the geometric mean filter. This step is instrumental in enhancing the quality of the images. Subsequently, the images undergo segmentation facilitated by the K-means algorithm.



VI. MODEL DEVELOPMENT

In model development there is a brain is known as model it will identify whether the patient is healthy or cancer

patient by using of matrix method if there is matrix table 255*255 it will have some numbers like 0 and 1, it may over write the image to the matrix table

If they have the empty space or some holes are denoted as 0 other the space are denoted as 1

- The images are resized to particular resolution, cropped to a specific size
- Then, converted to a matrix like tensors to read the features of image inside the neural network model.

VII. IMAGE SEGMENTATION

The process of separating out required region of interest from the image is known as segmentation. Mathematical morphological operations are powerful tools in acquiring lung region from binary images. In our methodology, first the pre-processed Gray scale images were converted to binary images Morphological opening operation was performed to the binary image with disk structuring element for removal of unwanted components from the image. The opened image was then complemented and clear border operation was performed to it. The lung masks were obtained by filling the holes and gaps present in the lungs. Finally exclusive OR operation was performed to lung mask output and clear border output to give us the segmented tumour region. Image segmentation is to facilitate the representation of an image into something that is more meaningful and easier to analyse. Image segmentation is used to allocate objects and boundaries. It is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. The result of image segmentation is a set of segments that collectively cover up the entire image or a set of contours extracted from the image. Each pixel of the image in a region is similar with respect to some characteristic or computed property, such as colour intensity, or texture: adjacent regions are significantly different with respect to the same characteristics. In this Project, we used the marker-controlled watershed image segmentation approach

VIII. SYSTEM ARCHITECTURE

DATA SET DESCRIPTION:

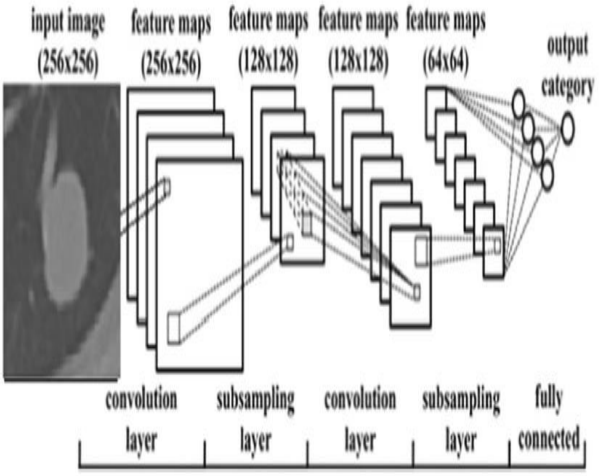
TCIA uses a standards-based approach to non-detection of images stored in the Digital Imaging and Communication Medicine (DICOM) format to release protected health information (PHI). The reproduction and repetition of the measurement of the three radiologists were high (all cc, ≥0.96) [12].

MODULE DESCRIPTION:

The computer-aided measurement was more repetitive (all ccc, 1.00). There was a 95% limit for computer-aided unidimensional, two-dimensional, and two-dimensional scanning volume measurement (-7.3%, 6.2%), (-17.6%, 19.8%), and (-12.1%, 13.4%) respectively.

IX. TRAIN THE MODEL

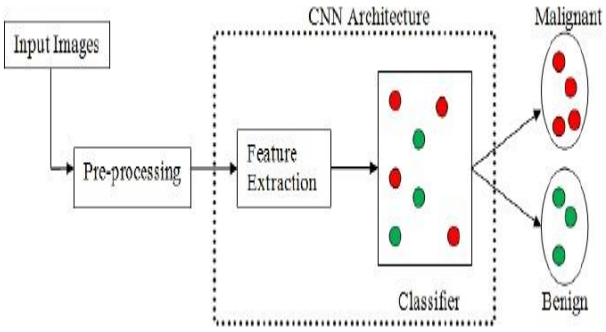
Derived from the models of biological visual systems, a Convolutional Neural Network (CNN) represents a breed of feed-forward neural networks. It is characterized by an arrangement of individual neurons that react to overlapping regions within their receptive field, aligning harmoniously with contemporary insights into image system structure. The utilization of neurons with identical parameters across diverse locations, when applied to overlapping areas of the preceding layer, engenders a phenomenon known as translational invariance. This attribute ensures consistency and reliability in adherence to the modern perception of image system organization.



CNNs possess a distinctive capability that allows them to recognize objects within their receptive field while remaining insensitive to variations in attributes like size, location, and orientation. This property is facilitated by the utilization of convolutional layers, which apply filters across the input data to detect features irrespective of their exact position. Moreover, the controlled connectivity in CNNs significantly reduces the computational requirements for training compared to fully connected neural networks.

The architectural structure of a Convolutional Neural Network, depicted in Figure 1, takes the form of a multi-layered feed-forward neural network. This arrangement comprises multiple hidden layers stacked in sequence, providing CNNs with the ability to acquire hierarchical features. These hidden layers typically consist of

convolutional layers followed by activation layers, and some may be succeeded by pooling layers. The sequential organization of these layers allows CNNs to progressively learn and represent hierarchical features. Starting from lower layers that capture simple features like edges and textures, the network gradually combines these to represent more complex concepts, culminating in object

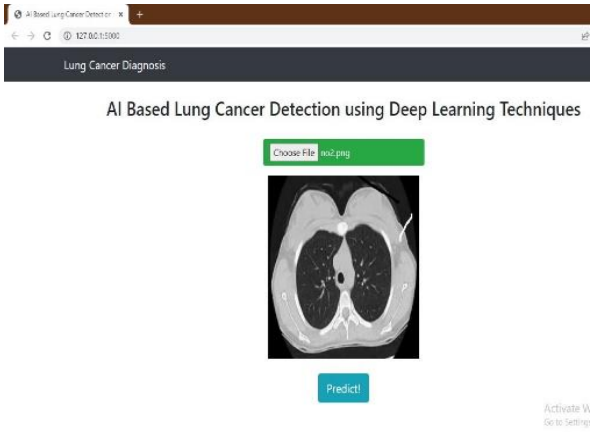


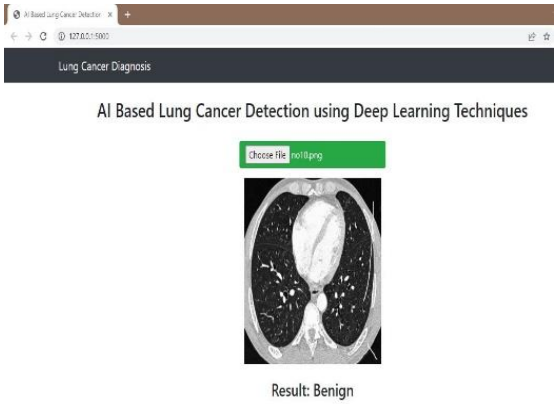
X. RESULT AND DISCUSSION

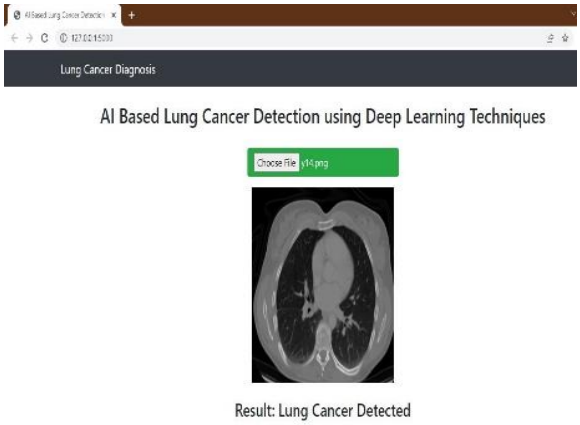
Lung cancer staging is determined by the size of lung nodules, which range from 5 mm to 25 mm, with nodules larger than 25mm considered atypical in images. Staging involves evaluating cancer size, tissue infiltration, and the presence of metastasis in lymph nodes or other organs. The staging is categorized into four levels based on severity:

1. Stage I: Cancer remains localized within the lung.
2. Stage II and III: Cancer remains restricted within the chest.
3. Stage IV: Cancer has extended beyond the chest, affecting other body parts.

For this study, a dataset comprising 15,419 lung images is employed. These images undergo processing via the proposed method. Diagnosis rules are established from these images and fed into a CNN classifier for learning. Subsequently, the learning process enables the CNN classifier to analyse lung images. Upon passing a lung image through the proposed method, the image undergoes a series of processing steps. Ultimately, the proposed method determines whether the examined lung image exhibits signs of malignancy.







XI. CONCLUSION

Lung cancer stands among the most perilous afflictions globally, necessitating accurate diagnosis and early identification to enhance survival rates. Current diagnostic strategies encompass X-ray, CT scan, MRI, and PET image analysis. Expert medical professionals discern cancer stages through their seasoned expertise. The treatment journey encompasses surgical intervention, chemotherapy, radiation therapy, and targeted therapy. However, this treatment path is prolonged, costly, and often distressing. To address these challenges, an innovative approach employing image

processing techniques, specifically molecule-based detection through CT scan images, is explored.

Various datasets supply CT scan images for analysis, notable for their brevity compared to X-ray and MRI images. A technique for image enhancement has been devised, serving the dual purpose of detection and early disease stage identification. An essential aspect is the incorporation of a time factor to promptly unveil abnormalities in the target image. Processing of CT scan images is undertaken, accurately isolating regions of interest like tumours from the original images. The pre-processing phase yields optimal outcomes through the utilization of Gabor Filters and Marker-Controlled Watershed segmentation.

Subsequent to isolating regions of interest, three key attributes are extracted: area, measurement, and warmth. This triad of properties serves as the foundation for lung cancer level detection. The findings highlight varying tumour stages, allowing for an accurate identification of lung cancer severity. The proposed methodology permits the measurement of cancer stages based on nodule size, enabling medical professionals to pinpoint the precise stage. In addition, the classification aspect is addressed through support vector machine modelling, offering an attractive avenue for classification purposes. These models amalgamate generalization controls with strategies designed to counter dimension-related challenges.

The utilization of kernel mapping facilitates a streamlined comparison of commonly employed model architectures, culminating in a compact framework. The generalization control aspect hinges on maximizing the margin, in harmony with the weight vector's composition within a canonical structure. This approach effectively addresses classification quandaries while aligning with the weight-based generalization mechanism. The proposed method displays promising potential in early-stage lung cancer detection, revolutionizing the landscape of disease identification and stage determination.

REFERENCES

[1] Staff, “Cancer Facts & Figures 2018,” Atlanta: American Cancer Society, Cancer, pp. 19-20, 2021.

[2] S. Avinash, K. Manjunath and S. Kumar, “An improved image processing analysis for the detection of lung cancer using Gabor filters and watershed segmentation technique”, 2019 International Conference on Inventive Computation Technologies (ICICT), 2019. Available: 10.1109/inventive.2016.7830084 [Accessed 24 January 2019].

[3] Q. Song, L. Zhao, X. Luo and X. Dou, “Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images”, Journal of Healthcare Engineering, vol. 2019, pp. 1-7, 2019. W.-K. Chen, Linear Networks and Systems. Belmont, CA: Wadsworth, 2019, pp. 123–135. P. Reid and S. Walter, “Yield from a fast track referral system for radiologists suspecting lung cancer”, Lung Cancer, vol. 115, p. S16, 2020.

[4] T. Jones, D. Duquette, M. Underhill, C. Ming, K. Mendelsohn-Victor, B. Anderson, K. Milliron, G. Copeland, N. Janz, L. Northouse, S. Duffy, S. Merajver and M. Katpadi, “Surveillance for cancer recurrence in long-term young breast cancer survivors randomly selected from a state wide cancer registry”, Breast Cancer Research and Treatment, vol. 169, no. 1, pp. 141-152, 2020.

[5] L. Sali, S. Delsanto, D. Sacchetto, L. Correale, M. Falchini, A. Ferraris, G. Gandini, G. Grazzini, F. Iafrate, G. Iussich, L. Morra, A. Laghi, M. Mascali and D. Regge, “Computer based self-training for CT colonography with and without CAD”, European Radiology, 2020.

[6] F. TjuJohore, M. Jony and P. Khatun, “A New Strategy to Detect Lung Cancer on CT Images”, International Research Journal of Engineering and Technology (IRJET), vol. 05, no.12, pp. 27-32, 2020.

[7] Valente, P. Cortez, E. Neto, J. Soares, V. de Albuquerque and J. Tavares, “Automatic 3D pulmonary nodule detection in CT images: A survey”, Computer Methods and Programs in Biomedicine, vol. 124, pp. 91-107, 2020.

[8] S.Marcham“Relationship between Slice Thickness to Artery Coronary Diagnostic Information on the Reconstruction of Maximum Intensity Protection (MIP)”, Journal of Medical Science and clinical Research, vol. 05, no. 06, pp. 23140-23145, 2019.

[9] F. Prior et al., "The public cancer radiology imaging collections of The Cancer Imaging Archive", Scientific Data, vol. 4, p. 170124, 2019. Available: 10.1038/sdata.2017.124.

[10] H. Chien and D. Mackay, “How much complexity is needed to simulate watershed streamflow and water quality? A test combining time series and hydrological models”, Hydrological Processes, vol. 28, no. 22, pp. 5624-5636, 2013.

[11] R. Lung CT, "The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki", Wiki.cancerimagingarchive.net, 2020. [Online]. Available: Lung+CT#4097ffced7f64f32a18885ceffdf7b86. [Accessed: 23- Mar- 2019].

[12] S. Makaju, P. Prasad, A. Alsadoon, A. Singh and A. Elchouemi, "Lung Cancer Detection using CT Scan

Images", Procedia Computer Science, vol. 125, pp. 107-114, 2020. Available: 10.1016/j.procs.2017.12.016.

[13] M. Firmino, G. Angelo, H. Morais, M. Dantas and R. Valentim, "Computer-aided detection (Cade) and diagnosis (Cadex) system for lung cancer with likelihood of malignancy", Biomedical Engineering Online, vol. 15, no. 1, 2019.

[14] H. Rutika r, "Automated detection and diagnosis from lungs CT scan images", International Journal of Emerging Technologies and Innovative Research (www.jetir.org UGC and sins Approved), ISSN:2349-5162, vol. 2, no. 3, pp. pp785-787, 2019.

[15] H. Ren, Y. Zhou and M. Zhu, "Tree Image Segmentation Based on an Improved Two-Dimensional Otsu Algorithm", International Journal of Hybrid Information Technology, vol.9, no. 9, pp. 199-210, 2019.

[16] Y. Zhang, H. Guo, F. Chen and H. Yang, "Weighted kernel mapping model with spring simulation based watershed transformation for level set image segmentation", Neurocomputing, vol. 249, pp. 1-18, 2019.

[17] Rana HK., Azam MS. and Akhtar MR. "Iris Recognition System Using PCA Based on DWT", SM Journal of Biometrics & Bio-statistics, vol. 2, no. 3, 2019.

[18] Rana HK, Azam MS, Akhtar MR, Quinn JMW, Moni MA.," A fast iris recognition system through optimum feature extraction", Peer Preprints7:e27363v2 https://doi.org/10.7287/peerj.preprints.27363v2, 2019.

[19] Y. Yuan, J. Wang, B. Li and M. Meng, "Saliency Based Ulcer Detection for Wireless Capsule Endoscopy Diagnosis", IEEE Transactions on Medical Imaging, vol. 34, no. 10, pp. 2046-2057, 2019.

[20] X. Jia and M. Meng, "A deep convolutional neural network for bleeding detection in Wireless Capsule Endoscopy images", 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019.

[21] X. Jia and M. Meng, "A deep convolutional neural network for bleeding detection in Wireless Capsule Endoscopy images", 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019.S. Suman, F. Husin, A. Malik, K. pogorelov

[22] M. Riegler, S. Ho, I. Hilmi and K. Goh, "Detection and Classification of Bleeding Region in WCE Images using Color Feature", Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing - CBMI '17, 2019.Suzuki K., "False-positive Reduction in Computer-aided Diagnostic Scheme for Detecting Nodules in Chest Radiographs", Academic Radiology, Volume 13, Number 10, pp.10-15, February 2019.

[23] G. Aviram and M. Revel, "Misclassification of Lymph Nodes in Lung Cancer Staging", Chest, vol. 151, no. 4, pp. 733-734, 2019.