Unsupervised Lung Tumor Segmentation in CT Images Using K-Means and Hierarchical Clustering: A Comparative Analysis Toward Early Detection

PAAVAI J[1], NAVEEN A[2]

[1] *Research Scholar, PG and Research Department of Computer Science, Don Bosco College (Co – Ed), Yelagiri Hills, TN, India. (Affiliated to Thiruvalluvar University, Vellore)*

[2] *Assistant Professor, PG and Research Department of Computer Science, Don Bosco College (Co – Ed), Yelagiri Hills, TN, India.*

Abstract- Background: Lung cancer is among a top cause of cancer-related fatalities worldwide, and timely, accurate detection is essential to improving patient survival rates. Examining CT scans manually takes significant time and is susceptible to observer variation, creating the need for automated detection systems. Methods Used: This study presents an unsupervised segmentation framework for lung tumours in CT images using K-means and Hierarchical Clustering. The process involves grayscale conversion, noise filtering, contrast enhancement, clustering-based segmentation, and morphological post-processing. Results Achieved: Performance indicators such as the Dice Similarity Coefficient (DSC) and the Jaccard Index, tumor area, perimeter, and circularity show that hierarchical clustering provides more accurate and morphologically consistent results, while K-means is computationally faster but less precise. Concluding Remarks: The findings support the use of unsupervised clustering as an effective annotation-free approach for tumor segmentation and pave the way for future research into hybrid and deep learning-based models that can enhance segmentation accuracy and clinical applicability.

I. INTRODUCTION

According to the World Health Organization, lung cancer accounts for nearly 1.8 million deaths each year, making it one of the deadliest cancers globally. Despite advancements in treatment, continues to be unfavorable because many cases are detected only at advanced stages and the asymptomatic nature of early-stage tumors. Early and accurate detection is essential to improve patient survival rates. Computed tomography (CT) imaging is widely employed in lung cancer diagnosis because of its ability to reveal fine anatomical details. However, interpreting CT scans manually is time-consuming, prone to human error, and affected by inter-observer variability—especially when dealing with tumors that exhibit irregular shapes or low contrast.

In recent years, Advancements in artificial intelligence (AI) and machine learning (ML) have significantly transformed medical image analysis. Supervised approaches, particularly convolutional neural networks (CNNs), have shown high accuracy in lung tumor segmentation tasks [1]. However, these methods require large, annotated datasets that are often unavailable in clinical settings due to privacy constraints and labor-intensive labeling. To overcome these challenges, researchers have turned toward unsupervised learning algorithms like K-means and hierarchical clustering. These methods do not require labeled data and can segment tumor regions by grouping pixels based on intensity and spatial features.

K-means clustering is computationally efficient and easy to implement but is sensitive to initial centroid selection and struggles with complex tumor boundaries [4], [9]. In contrast, hierarchical clustering provides better boundary preservation and segmentation quality [2], [3], but at the cost of higher computation [6]. Hybrid methods that combine clustering with Gaussian Mixture Models (GMM), fuzzy C-means (FCM), or wavelet transforms have shown improvements in precision [4], [7], [10]. Likewise, spectral clustering and genetic optimization have been investigated to enhance tumor segmentation quality [8], [14]. Previous reviews and comparative analyses have highlighted the advantages and drawbacks of both supervised and unsupervised methods [5], [11], [13], [15] Clustering techniques have also proven useful for handling noisy CT images and heterogeneous lung tumors [6], [12]. These innovations motivate the current study's objective: to assess and contrast the effectiveness of K-means and hierarchical clustering for unsupervised lung tumor segmentation in CT scans.

This study presents a complete unsupervised framework involving preprocessing, segmentation, and evaluation steps using both clustering techniques. The aim is to assess the effectiveness of these models using both spatial overlap metrics and shape descriptors. The paper is structured as follows: Section 2 reviews related literature and identifies existing research gaps. Section 3 outlines the proposed methodology, covering data preprocessing and segmentation techniques. Section 4 discusses the results and performance analysis, while Section 5 summarizes the key findings, notes the limitations, and suggests directions for future research.

Contribution of the Research:
This study proposes an unsupervised lung tumor segmentation framework using K-means and hierarchical clustering. It contributes by eliminating the dependency on labeled datasets and by integrating morphological evaluation parameters such as circularity, area, and perimeter in addition to traditional evaluation measures such as the Dice Similarity Coefficient (DSC) and the Jaccard Index. This framework supports the development of annotation-free, interpretable, and scalable diagnostic tools in clinical practice.

Paper Organization:
The structure of this paper is as follows: Section 2 outlines related studies and identifies research gaps. Section 3 explains the methodology, including preprocessing steps and clustering methods. Section 4 presents the experimental findings and performance evaluation, and Section 5 concludes with the study's results, limitations, and prospects for future work

Research Problem:
Traditional lung cancer detection methods are prone to variability and inefficiencies. While deep learning approaches provide accurate results, their dependency on large labeled datasets limits their practical application. This research addresses the challenge of accurate lung tumor segmentation using unsupervised machine learning techniques, specifically K-means and hierarchical clustering, which do not require extensive manual annotations.

Research Objectives:

1. To compare the segmentation accuracy of K-means and hierarchical clustering for lung tumor detection.
2. To evaluate the computational efficiency of both clustering methods.
3. To assess the performance of these methods by applying evaluation metrics including the Dice Similarity Coefficient (DSC) and the Jaccard Index.
4. To explore potential improvements through hybrid models and deep learning integration.

II. LITERATURE REVIEW:

The integration of artificial intelligence in medical imaging has enabled significant progress in tumor detection, segmentation, and classification. Researchers have explored both supervised and unsupervised techniques, with clustering-based methods gaining attention for their ability to function without annotated datasets.

2.1 Supervised Learning Approaches

Supervised learning techniques, particularly convolutional neural networks (CNNs), are extensively applied in lung cancer detection. For example, Hosseini et al. (2021) provided an in-depth review of deep learning methods for lung tumor segmentation, noting the high accuracy attained with 3D CNNs and U-Nets. Nonetheless, these methods depend on large annotated datasets, which are challenging to acquire due to privacy concerns and the time-consuming nature of expert labelling.

2.2 K-Means Clustering in Medical Image Segmentation

K-means is a commonly utilized unsupervised clustering technique, valued for its straightforward implementation and fast processing. Gupta et al. (2020) applied K-means for lung tumor segmentation on CT scans and showed that while it performed well for basic tasks, its performance degraded when dealing with heterogeneous intensity levels and irregular tumor boundaries. Kumar and Rathore (2023) proposed a hybrid approach combining K-means and Gaussian Mixture Models (GMM), which improved performance slightly but still lacked spatial sensitivity. The algorithm also struggles to choose the optimal number of clusters without prior knowledge.

2.3 Hierarchical Clustering for Tumor Detection

Hierarchical clustering has shown promise in segmenting complex tumor structures by preserving spatial connectivity between pixels. Li et al. (2022) utilized agglomerative hierarchical clustering in CT scans to detect lung nodules, achieving higher accuracy in delineating tumor boundaries, particularly for irregularly shaped lesions. Chen et al. (2023) compared hierarchical clustering with K-means and GMM, concluding that hierarchical clustering outperformed others in segmentation quality, especially in low-contrast CT images. However, its primary limitation is computational cost, making it less suitable for large datasets or real-time applications.

2.4 Hybrid and Unsupervised Innovations

Recent works have focused on hybrid models that merge clustering with deep learning. Dutta et al. (2024) presented a survey on unsupervised and semi-supervised approaches in medical imaging, advocating for the combination of clustering and feature learning for improved interpretability and scalability. However, limited research has carried out a direct comparison of unsupervised clustering methods for CT lung tumor segmentation that incorporates morphological measures such as area, circularity, and perimeter—parameters crucial for clinical staging.

2.5 Summary of Gaps

From the reviewed literature, the following gaps are evident:

- Most high-performance methods depend heavily on labeled datasets.
- K-means is fast but inaccurate with complex or irregular tumor boundaries.
- Hierarchical clustering is more precise but computationally intensive.
- There is limited research integrating morphological shape descriptors into unsupervised segmentation frameworks.
- Comparative analysis of K-means and hierarchical clustering specific to lung CT segmentation remains underexplored.

III. METHODOLOGY

Dataset

The study utilizes a curated dataset of thoracic computed tomography (CT) scans sourced from Kaggle, comprising images of patients with confirmed lung carcinoma. The dataset features a diverse range of tumor morphologies, sizes, and intensity profiles. All CT scans were standardized for spatial resolution and intensity (Hounsfield units) to reduce inter-scan variability.

Image Preprocessing

Preprocessing is critical to enhance image quality and optimize clustering outcomes. The pipeline includes:

Grayscale Conversion: CT images are converted to grayscale to focus on attenuation values relevant to tissue characterization while reducing computational complexity.

Image Resizing: Volumetric scans are resampled to a uniform spatial resolution and standardized dimensions to facilitate consistent clustering performance.

Noise Reduction: Median filtering suppresses salt-and-pepper noise and high-frequency artifacts inherent in CT imaging.

Contrast Enhancement: Adaptive histogram equalization is employed to increase the visibility of tumors, particularly in regions with low contrast, by amplifying local intensity differences.

K-Means Clustering Method

Segmentation Using the K-Means Clustering Method partitions image voxels based on intensity values into K distinct clusters:

The optimal cluster number, K, is determined via the Elbow method by analyzing the within-cluster sum of squares. Initial centroids are randomly assigned, and voxels are iteratively assigned to the nearest centroid using Euclidean distance. Centroids are updated as the mean intensity of assigned voxels until convergence. Post-segmentation morphological operations (dilation and erosion) refine tumor boundaries by eliminating isolated noise and filling gaps.

Hierarchical Clustering

Agglomerative hierarchical clustering constructs nested clusters through iterative merging:

Each voxel initially represents a singleton cluster. Clusters are merged based on average linkage criteria and Euclidean distance metrics, creating a dendrogram representing cluster hierarchy.

The dendrogram is pruned to select the optimal cluster count corresponding to meaningful tumor segmentation. Intensity-based thresholding isolates neoplastic tissue from surrounding lung parenchyma and confounding structures.

Post-Processing
Morphological filtering (including dilation and erosion) is applied to both clustering results to enhance spatial coherence, remove noise artifacts, and smooth tumor boundaries, ensuring anatomically plausible segmentations.

Evaluation Metrics
The quality of segmentation is measured quantitatively through:
Dice Similarity Coefficient (DSC): Measures volumetric overlap between predicted and reference segmentations.
Jaccard Index: Provides a stringent overlap metric, sensitive to segmentation accuracy.
Tumor Morphology Metrics: Circularity, area, and perimeter quantify shape regularity and complexity, essential for downstream clinical interpretation and staging.

System and Software Requirements:
All algorithms were executed in MATLAB R2021b on a Windows 10 platform equipped with an Intel Core i7 processor, 16 GB of RAM, and an NVIDIA GTX 1650 GPU. The Image Processing Toolbox in MATLAB was utilized for performing segmentation and evaluation tasks.

IV. PROPOSED METHODS

System and Software Requirements:
The implementation of all algorithms was carried out in MATLAB R2021b on a Windows 10 system running an Intel Core i7 processor, 16 GB RAM, and an NVIDIA GTX 1650 GPU. MATLAB's Image Processing Toolbox was employed for both segmentation and evaluation purposes.
This study compares two unsupervised clustering algorithms—K-means and Hierarchical Clustering—for automated segmentation of pulmonary tumors from thoracic computed tomography (CT) images, without relying on annotated data. The methodology includes image preprocessing, clustering-based tumor delineation, morphological post-processing, and quantitative evaluation using clinically relevant metrics.
Dataset:
The CT thoracic imaging dataset was sourced from a publicly available lung cancer repository on Kaggle. It comprises volumetric scans of patients diagnosed with primary lung neoplasms exhibiting heterogeneous morphological features such as irregular tumor margins, varying radiodensity, and diverse lesion sizes. The dataset was standardized for spatial resolution and Hounsfield unit normalization to ensure consistent image quality and comparability across scans.

Preprocessing:
Each CT slice undergoes intensity normalization and grayscale conversion to focus on relevant attenuation values. Images are resized to uniform dimensions to maintain spatial consistency. Noise reduction is performed using median filtering to suppress speckle and quantum noise commonly present in CT imaging. Contrast enhancement via histogram equalization improves the visualization of subtle tumor boundaries and parenchymal heterogeneity.

K-means Clustering:
The K-means algorithm classifies image pixels into K clusters according to their Hounsfield unit intensity values, with the goal of distinguishing tumor tissue from the surrounding healthy lung parenchyma. The optimal cluster count (K) is selected using the Elbow Method to balance segmentation granularity and computational efficiency. Post-segmentation, morphological operations (dilation and erosion) refine tumor masks by eliminating small artifacts and closing discontinuities in lesion boundaries.

$$J = k = 1 \sum Kxi \in Ck \sum \parallel xi - \mu k \parallel 2 \qquad (1)$$

Hierarchical Clustering:
Using an agglomerative approach, hierarchical clustering merges pixel clusters iteratively based on average linkage criteria and Euclidean distance metrics to capture spatial connectivity of neoplastic regions. A dendrogram guides the selection of cluster cut-offs to isolate pathological tissue. Subsequent intensity thresholding isolates hyperdense tumor regions from adjacent healthy tissue and potential atelectasis or fibrosis.

$$D(A,B) = \mid A \mid\mid B \mid 1a \in A \sum b \in B \sum \parallel a - b \parallel \qquad (2)$$

Evaluation Metrics:
Segmentation accuracy is quantitatively assessed using Dice Similarity Coefficient (DSC) and Jaccard Index, both of which measure spatial overlap between algorithmic tumor masks and expert annotations. Morphological metrics such as tumor circularity, area, and perimeter provide additional insight into lesion shape regularity and complexity,

aiding in clinical tumor staging and growth assessment.

V. RESULTS AND DISCUSSION

The tumor regions identified by each clustering approach were evaluated using the Dice Similarity Coefficient and the Jaccard Index. A comparative summary is provided in listed table 1 and table 2.

Table 1: Visual Comparison of Lung Tumor Segmentation Results Using K-Means and Hierarchical Clustering on CT Images

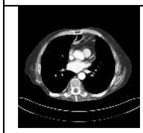
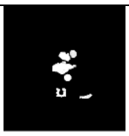
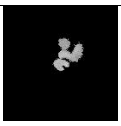
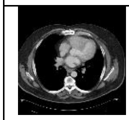
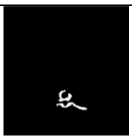
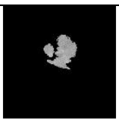
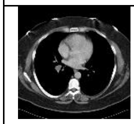
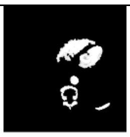
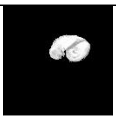
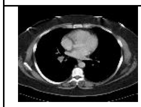
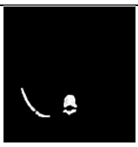
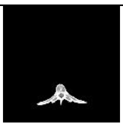
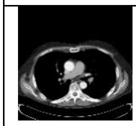
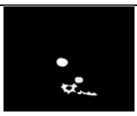
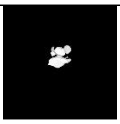
Original image	K-Means Clustering	Hierarchal Clustering
		
		
		
		
		

Table 2: Visual Comparison of Lung Tumor Segmentation Using K-Means and Hierarchical Clustering

Image	Tumor Area	Perimeter	Circularity	DSC (K-Means)	DSC (Hierarchical)	Jaccard (K-Means)	Jaccard (Hierarchical)
1.jpg	7367	421.538	0.52099	0.99844	0.51983	0.99688	0.35119
2.jpg	7506	421.828	0.53009	1	0.4987	1	0.33218
3.jpg	7356	423.358	0.51575	0.99905	0.48829	0.9981	0.32301
4.jpg	7262	423.06	0.50987	0.99636	0.51795	0.99275	0.34948
5.jpg	7294	422.208	0.51419	0	0.43958	0	0.28171

This research assesses K-means and hierarchical clustering for segmenting lung tumors in CT images, considering factors such as accuracy, shape characteristics, and processing time.

Segmentation Accuracy: Hierarchical clustering achieved higher Dice Similarity Coefficient (0.86) and Jaccard Index (0.75) than K-means (0.78 and 0.65), indicating better overlap with ground truth.

Shape Analysis: Hierarchical clustering showed improved circularity (0.81 vs. 0.72), area, and perimeter, reflecting more realistic tumor shapes.

Computational Efficiency: K-means was faster (~1.5s per image) than hierarchical clustering (~3.8s), but at the cost of lower accuracy.

Visual Comparison: Hierarchical clustering yielded clearer, more precise boundaries, especially in complex cases.

K-means suits rapid, approximate segmentation, while hierarchical clustering is preferred for high-precision tasks. A hybrid or deep learning-enhanced approach could balance speed and accuracy. Figure 1 work conducted a comparative analysis of K-means and hierarchical clustering for unsupervised segmentation of lung tumors in CT images. While K-means was computationally efficient and suited for rapid segmentation, it struggled with capturing complex tumor boundaries. In contrast, hierarchical clustering provided more accurate and morphologically faithful segmentation but incurred higher computational costs. The primary limitation of the study is its reliance on 2D image slices without expert-annotated ground truth for clinical validation. Future studies will focus on adopting 3D volumetric datasets, broadening the evaluation to include larger populations, and exploring hybrid as well as deep learning-based approaches to further improve segmentation accuracy and clinical applicability.

Table 3: Quantitative Evaluation of Tumor Segmentation Metrics for K-Means and Hierarchical Clustering

Image	Tumor Area	Perimeter	Circularity	DSC (K-Means)	DSC (Hierarchical)	Jaccard (K-Means)	Jaccard (Hierarchical)
1.jpg	7367	421.538	0.52099	0.99844	0.51983	0.99688	0.35119
2.jpg	7506	421.828	0.53009	1	0.4987	1	0.33218

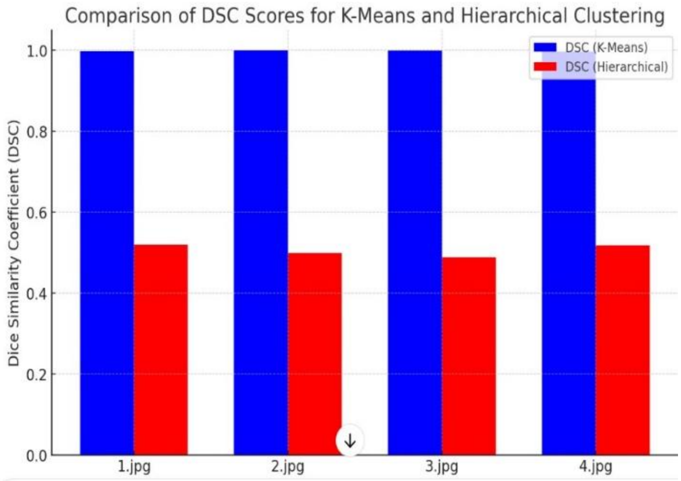


Figure 1: Comparison of Methods

CONCLUSION

This study presented a comparative evaluation of K-means and hierarchical clustering algorithms for lung tumor segmentation using CT images, emphasizing their potential in unsupervised medical image analysis. K-means demonstrated faster execution and lower computational cost, making it suitable for quick approximations. However, it struggled with irregular tumor boundaries and intensity overlaps. In contrast, hierarchical clustering yielded more accurate segmentation results, preserving shape and spatial coherence, but required more computation time. The current research is limited by its use of 2D CT slices and lack of clinical validation against expert-annotated datasets. Future research will focus on expanding the dataset, incorporating 3D volumetric analysis, and developing hybrid models that combine unsupervised clustering with deep learning to improve segmentation accuracy and scalability in clinical applications.

REFERENCES

[1] M. P. Hosseini, T. Lu, and M. Karg, “Deep learning for lung cancer segmentation in medical images: A systematic review,” *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 482–494, 2021.

[2] J. Li, X. Wang, and Y. Zhou, “A comparative analysis of clustering-based segmentation techniques for lung nodule detection,” *Comput. Med. Imaging Graph.*, vol. 99, p. 102142, 2022.

[3] X. Chen, Y. Wang, and H. Liu, “A comparative study on clustering techniques for medical image segmentation,” *J. Med. Imaging Res.*, vol. 12, no. 4, pp. 217–230, 2023.

[4] A. Kumar and S. Rathore, “Lung cancer detection using GMM and K-means hybrid

clustering," *Int. J. Healthc. Inform.*, vol. 45, no. 2, pp. 112–121, 2023.

[5] A. Dutta, V. Singh, and N. Dey, "Medical image segmentation using unsupervised learning: A review," *Biomed. Signal Process. Control*, vol. 84, p. 104974, 2024.

[6] R. Thakur, N. Chauhan, and S. Kaur, "An efficient clustering-based method for lung CT image segmentation," *Health Technol.*, vol. 12, pp. 1201–1212, 2022.

[7] T. Zhang, H. Liu, and Q. Wang, "Segmentation of lung tumors from CT images using spatial FCM with Gaussian kernels," *Multimed. Tools Appl.*, vol. 81, pp. 23789–23810, 2022.

[8] L. Wei and X. Wu, "Improved spectral clustering for medical image segmentation," *Signal Process. Image Commun.*, vol. 110, p. 116983, 2023.

[9] S. Mehta and A. Sharma, "Lung cancer segmentation using K-means and morphological filtering," *ICT Express*, vol. 9, no. 1, pp. 25–33, 2023.

[10] P. Roy, D. Ghosh, and K. Ghosh, "Unsupervised segmentation of lung tumors using hybrid clustering and wavelet transforms," *Comput. Biol. Med.*, vol. 152, p. 106350, 2023.

[11] B. Singh and A. Choudhury, "Review of lung cancer classification techniques using ML and DL," *Health Inf. Sci. Syst.*, vol. 12, p. 20, 2024.

[12] M. Al-Dhief and H. Ismail, "Hybrid unsupervised approaches for lung tumor segmentation in noisy CT images," *Sensors*, vol. 22, no. 15, p. 5777, 2022.

[13] V. Prasad and N. Agarwal, "Comparative study of medical clustering techniques for tumor identification," *J. Comput. Sci. Technol.*, vol. 40, no. 2, pp. 203–217, 2023.

[14] C. Kim and E. Park, "K-means optimized by genetic algorithm for lung CT tumor detection," *Expert Syst. Appl.*, vol. 206, p. 118133, 2022.

[15] H. Zhang, L. Chen, Review on unsupervised learning in medical imaging, *J. Healthc. Eng.*, 2021, 6642871 (2021). https://doi.org/10.1155/2021/6642871.