REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 22, No. 1 (2021)
http://www.veterinaria.org
Article Received: 10/04/2021 Revised: 22/04/2021 Accepted: 03/05/2021

Evaluating Lung Cancer Classification Performance Using Multiple Feature Extraction Methods with SVM and KNN Classifiers

Halaswamy B. M.[1*], Mamatha M. M.[2]

[1*]Senior Scale Lecturer, Department of Electronics and Communication Engineering, Government Polytechnic Hiriyur
[2] Lecturer, Department of Electronics and Communication Engineering, Government polytechnic Immadihalli

***Corresponding author:**
*Email: halaswamybm26@gmail.com

ABSTRACT
Lung cancer is one of the most prevalent causes of mortality worldwide, making early detection essential for improving patient survival rates. Computed tomography (CT) imaging serves as a crucial diagnostic tool; however, the large volume of generated images poses challenges in precise interpretation by radiologists. This study evaluates the effectiveness of lung cancer classification by utilizing various feature extraction techniques in combination with support vector machine (SVM) and k-nearest neighbours (KNN) classifiers. By analysing different feature sets, the research aims to identify the most effective combination for enhanced classification accuracy. The findings indicate notable improvements in classification performance, facilitating more reliable lung cancer detection.

Keywords: Lung Cancer, Feature Extraction, SVM, KNN, CT scan, Image Processing, GLCM, LBP, HOG, Image Segmentation.

Introduction

Several respiratory diseases, including chronic bronchitis, chronic obstructive pulmonary disease (COPD), acute respiratory distress syndrome (ARDS), emphysema, and lung cancer, significantly impact global health. Reports from the World Cancer organization in 2014 indicate that lung cancer is a leading cause of cancer-related deaths in both men and women. Additionally, statistics from 2012 suggest that lung cancer was responsible for over 1.56 million fatalities worldwide. The primary contributing factors to lung diseases include smoking, inhalation of drugs, exposure to allergens, and air pollutants. The severity of lung diseases can be effectively assessed using computed tomography (CT) imaging, which provides detailed visualization of soft tissues, facilitating accurate diagnosis. CT imaging plays a crucial role in lung cancer detection. Lung cancer is generally classified into
Two primary categories: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). NSCLC is further divided into three subtypes:
Squamous cell carcinoma, adenocarcinoma, and large cell carcinoma [1].
Lung cancer remains one of the deadliest diseases worldwide sample figure-1a. Early detection significantly enhances survival rates. Over the years, CT imaging has become an essential diagnostic tool for lung disease identification. Advances in image processing techniques have led to major improvements in detecting abnormalities with greater precision. However, relying solely on radiologists for diagnosis can sometimes lead to errors. To improve detection accuracy, this study proposes a custom segmentation approach combined with a Support Vector Machine (SVM) classifier. This method enhances both lung cancer identification and classification accuracy.
Compared to other imaging modalities, CT scans are particularly effective in identifying small lung nodules due to their ability to rapidly acquire high-resolution images while minimizing artefacts. However, challenges exist in interpreting CT scans, as human perception is limited, and radiologists may experience fatigue or distractions, increasing the likelihood of misdiagnosis. To address these limitations, computer-aided detection (CAD) systems assist radiologists in improving diagnostic accuracy.

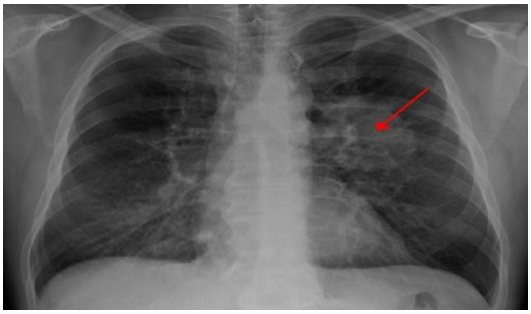

Figure 1a: Lung Cancer Chest X-Ray Image

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 22, No. 1 (2021)
http://www.veterinaria.org
Article Received: 10/04/2025 Revised: 22/04/2025 Accepted: 03/05/2025

Lung cancer progresses rapidly, making early detection critical for effective treatment. According to World Health Organization (WHO) reports, lung diseases claim over 7.6 million lives annually. Additionally, lung cancer-related deaths are projected to rise significantly, reaching approximately 17 million by 2030 [2]. In recent years, image processing techniques have been widely adopted in medical applications to enhance early disease detection, particularly in cancers such as lung and breast cancer [3].

The timely identification of tumors in the brain, chest, and lungs is essential for initiating early treatment. Image processing techniques play a key role in this process, enabling the early detection of malignant growths. The image processing workflow consists of four major steps: pre-processing, feature extraction, segmentation, and classification [4]. This study integrates machine learning approaches with CT imaging, utilizing a custom segmentation technique in combination with an SVM classifier for detecting early-stage lung cancer. Our objective is to refine lung cancer detection methods by incorporating advanced segmentation strategies to improve accuracy.

Literature Review

Lung cancer originates in the lung tissues and remains one of the most common malignancies affecting both men and women. Similar to other diseases, it develops due to repeated genetic mutations in cells. Tobacco use is the leading cause of lung cancer, accounting for approximately 85% of related fatalities [5-6]. This disease can develop in a single lung or both, within the classified air passage regions, and is characterized by uncontrolled growth of abnormal cells [7].

Recent studies indicate that the overall survival rate for lung cancer is approximately 18%, with most cases being identified within 5-6 years. However, survival rates vary among individuals, particularly those with diabetes. Research suggests that elevated insulin levels increase the likelihood of developing lung cancer [8]. Furthermore, studies have shown that diabetes mellitus may influence the survival outcomes of lung cancer patients [9]. Nicotine has been found to affect insulin secretion in individuals with Type 2 Diabetes (T2D) and alter insulin activity in smokers [10]. Smoking cessation is, therefore, a crucial step in managing diabetes and preventing related complications [11]. Several clinical and pre-clinical studies have demonstrated that nicotine in cigarettes impacts body composition, insulin sensitivity, and pancreatic β-cell function [12].

In recent years, machine learning (ML) has gained prominence in biomedical research, particularly in developing multidimensional models for data analysis. Machine learning is broadly classified into two categories: supervised and unsupervised learning. In supervised learning, labelled training data is used, comprising both input features and desired outcomes. On the other hand, unsupervised learning identifies patterns and clusters data without predefined labels. In clinical practice, electronic health records (EHRs) are extensively utilized by physicians for research and patient care management. The integration of ML techniques with EHR data enables personalized patient care and enhances hospital performance monitoring [14-15].

Support Vector Machines (SVMs) represent a cutting-edge ML approach widely applied in cancer diagnostics. SVM functions by mapping input vectors into a high-dimensional feature space and identifying an optimal hyper plane that separates data points into distinct classes. The goal is to maximize the margin between the closest data points and the decision boundary, thereby improving classification accuracy. SVMs are particularly effective in ensuring consistent and reliable classification outcomes [16]. The SVM model constructs an optimal hyper plane for binary classification by maximizing the minimum margin between two datasets. Kernel functions such as linear, sigmoid, and radial basis functions (RBF) are employed to handle non-linearly separable data.

Another widely used classification approach is the C4.5 decision tree algorithm, developed by Ross Quinlan. C4.5 is an extension of the earlier ID3 algorithm and is recognized for its simplicity and efficiency. An advanced version of this algorithm, the J48 classifier, utilizes entropy to construct decision trees. The algorithm selects the most relevant attribute as the root node and subsequently develops branches for different attribute values. This technique effectively classifies data into various categories [17].

Naïve Bayes (NB) is another classification method commonly applied in predictive analytics. This probabilistic approach calculates the likelihood of a given class based on the distribution of attributes within the dataset. During training, the model determines conditional probabilities for each class, enabling it to predict class labels for new instances [18].

A study by [19] introduced an SVM-based classification model utilizing CT scan images to categorize lung cancer into benign, malignant, and normal classes. Another study [20] explored segmentation techniques for detecting lung abnormalities in CT scans. The segmentation approach relied on selecting specific regions to improve diagnostic accuracy. Additionally, a radionics-based analysis was conducted on benign and malignant mediastina tissues for feature extraction [21]. This research emphasized the significance of mediastina lymph node metastases in non-small cell lung cancer (NSCLC) and their impact on treatment decisions.

ArtiPatle, Neeraj Sirohi, and Tanushree Sinha Roy [22] conducted a study on lung image classification, differentiating between cancerous and non-cancerous cases. Their method utilized CT scan images, where key regions of interest were extracted using an active shape model. Feature extraction was followed by classification using a fuzzy inference system, which combined fuzzy logic and neural networks. Their approach achieved an accuracy of 94.12%.

Another study by Raj Kumar Sagar, Ankit Shankhadhar, and Ritika Agarwal [23] explored the concept of content-based image retrieval in clinical applications. They reviewed existing methodologies for detecting lung nodules at early stages

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 22, No. 1 (2021)
http://www.veterinaria.org
Article Received: 10/04/2025 Revised: 22/04/2025 Accepted: 03/05/2025

and proposed a Computer-Aided Diagnosis (CAD) system to enhance detection accuracy. The CAD system was designed to improve the precision of lung nodule identification by incorporating multiple processing levels.
Pradyut Kumar Biswal and Moumita Mukherjee [24] introduced an automated CAD system for lung cancer detection using CT scan images. Their approach employed an iterative Thresholding technique combined with conventional histogram-based Thresholding for image segmentation. The proposed method involved extracting lung regions and applying rule-based separation to identify nodules. A decision-making algorithm was integrated to streamline the classification process and minimize complexity.
Despite the advancements in image processing techniques for lung cancer detection, there remains room for improvement in classification accuracy. While CAD systems have shown promise in identifying early-stage lung malignancies, challenges persist in achieving optimal sensitivity and specificity. Enhancing classifier performance through feature selection and hybrid models can further improve diagnostic precision. By refining classification techniques and integrating advanced machine learning models, CAD systems can be optimized for better accuracy and reliability in lung cancer diagnosis.

Proposed Methodology

Lung cancer detection and classification play a crucial role in improving early diagnosis and treatment effectiveness. In this study, multiple feature extraction techniques, including **Gray Level Co-Occurrence Matrix (GLCM), Local Binary Pattern (LBP), and Histogram of Oriented Gradients (HOG)**, are utilized alongside **Support Vector Machine (SVM) and k-Nearest Neighbours (KNN)** classifiers to enhance classification accuracy. The objective is to determine the optimal combination of features and classifiers that yield the best performance in lung cancer classification.
The proposed method shown in figure-1b follows a structured approach to classify lung cancer using computed tomography (CT) scan images. The process involves the following key steps:

1. **Pre-processing**: The input image undergoes noise filtering to enhance quality and remove unwanted artefacts.
2. **Feature Extraction**: Three prominent texture-based feature extraction methods—**GLCM, LBP, and HOG**—are applied to extract meaningful patterns from the images.
3. **Training and Classification**: The extracted features are fed into **SVM and KNN classifiers,** either individually or in different combinations, to analyse their impact on classification accuracy.
4. **Performance Evaluation**: The classification results are compared to determine the best feature-classifier combination for achieving higher accuracy.

This approach ensures a systematic evaluation of various features and classification techniques, allowing for better performance optimization in lung cancer detection.

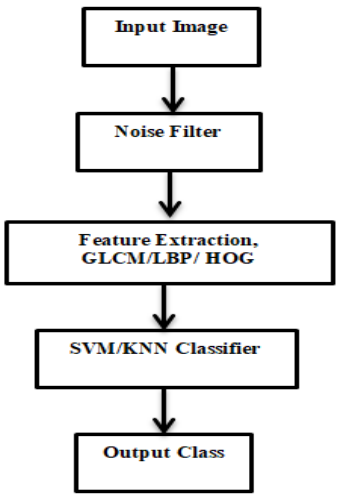

Figure 1b: Proposed Flow Diagram

Feature Extraction Techniques

1. The **GLCM** is a second-order statistical method used for texture analysis in image processing. It quantifies how often pairs of pixel intensities occur in an image at a given spatial relationship. This method helps extract texture features such as contrast, correlation, energy, and homogeneity, which are essential for distinguishing between healthy and cancerous lung tissues.

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 22, No. 1 (2021)
http://www.veterinaria.org
Article Received: 10/04/2025 Revised: 22/04/2025 Accepted: 03/05/2025

2. The **LBP** is a simple yet powerful feature descriptor used for texture classification. It operates by comparing each pixel in an image with its neighbouring pixels and assigning a binary value based on intensity differences. The LBP method has gained significant attention in medical imaging due to its robustness in detecting fine-grained textures and distinguishing between normal and abnormal tissue structures.
3. The **HOG** feature descriptor is widely used in object detection and computer vision applications. It works by calculating the distribution of gradient orientations in localized image regions, making it highly effective for capturing shape and edge information. In lung cancer classification, HOG features help identify tumour structures based on gradient patterns within CT images.

Classifier Selection and Performance Analysis

The **Support Vector Machine (SVM)**and **k-Nearest Neighbours (KNN)** classifiers are employed for the classification task:

1. **SVM**: A supervised learning model that identifies an optimal hyper plane to separate different classes with the maximum margin, making it well-suited for binary classification tasks such as distinguishing between malignant and benign tumours.
2. **KNN**: A non-parametric classifier that assigns a class label based on the majority voting of the nearest neighbouring samples in the feature space. KNN is effective in pattern recognition, particularly when the dataset has a well-defined structure.

Results

Figure 2a–2d illustrates the complete simulation process of the proposed method. The figures below depict the sequential steps involved in the method, along with the corresponding results and observations.

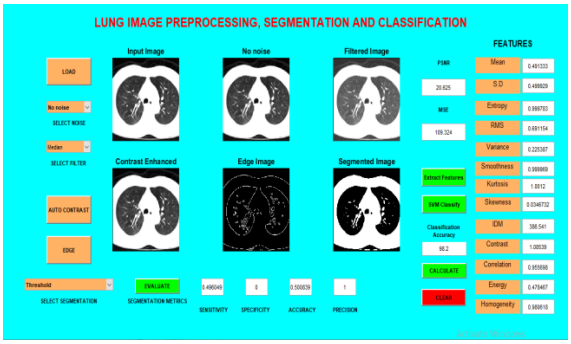

Figure 2a: Simulation of proposed work

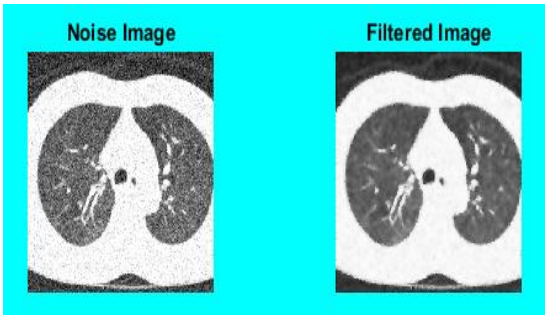

Figure 2b: Lung image with Noise and Filtered

Table 1. The accuracies are observed as below in the table:

Features	Classifier	Accuracy
GLCM	SVM	92.32
	KNN	93.20
GLCM+LBP	SVM	94.67
	KNN	93.64
GLCM+HOG	SVM	95.14
	KNN	94.22
GLCM+LBP+HOG	SVM	96.42
	KNN	95.13

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 22, No. 1 (2021)
http://www.veterinaria.org
Article Received: 10/04/2025 Revised: 22/04/2025 Accepted: 03/05/2025

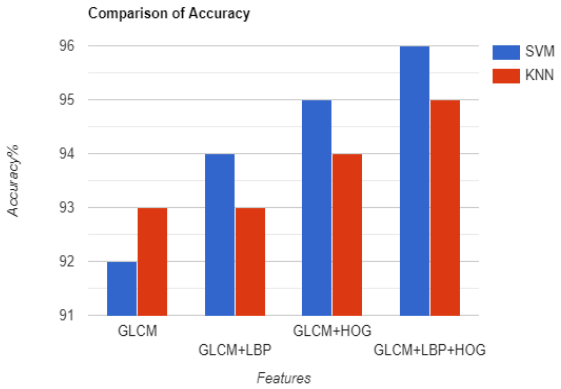

Figure 2c: Accuracy Comparison Bar Chart

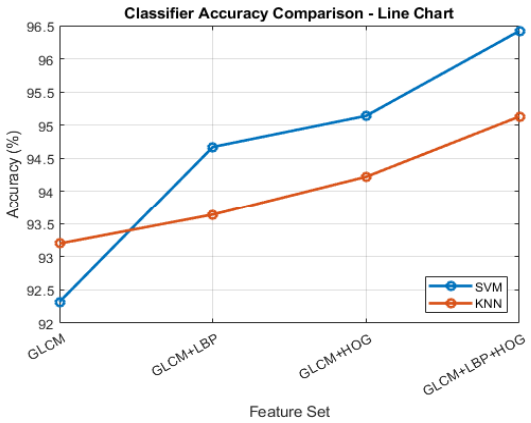

Figure 2d: Accuracy Comparison Line Chart

Conclusion

This study presents a performance analysis of lung cancer classification using multiple feature extraction techniques and various machine learning classifiers to identify the most effective combination for improved classification accuracy. By systematically evaluating different feature-classifier pairings, the proposed methodology achieves a maximum classification accuracy of 96%, demonstrating significant advancements in lung cancer detection.The proposed approach not only enhances the precision of lung malignancy identification but also improves the overall reliability of classification outcomes. The integration of multiple feature extraction techniques, such as GLCM, LBP, and HOG, with classifiers like SVM and KNN, enables a more robust and optimized classification framework.Additionally, this research highlights the combined advantages of multiple feature sets in machine learning for biomedical image analysis, paving the way for more sophisticated computer-aided diagnosis (CAD) systems in future medical imaging applications. The findings contribute to improving early detection and diagnosis, ultimately assisting healthcare professionals in making more accurate and timely clinical decisions.

References

1. V. Krishnaiah, G. Narsimha, C. Subhash.(2013), Diagnosis of lung cancer prediction system using data mining classification techniques, *In International Journal of Computer Science and Information Technologies*, 4(1): 39-45.
2. J. J. Dignam, L. Huang, L. Ries, M. Reichman, A. Mariotto, E. Feuer. (2009), Estimating cancer statistic and other-cause mortality in clinical trial and population-based cancer registry cohorts. *Wiley InterScience* [Online].
3. Disha Sharma, Gagandeep Jindal (2011), Identifying Lung Cancer Using Image Processing Techniques, *International Conference on Computational Techniques and Artificial Intelligence*.
4. B.N, Mithuna&Ravikumar, Pushpa& C.N, Arpitha (2018), A Quantitative Approach for Determining Lung Cancer Using CT scan Images, *1786-1790. 10.1109/ICECA.2018.8474670.*
5. Fan, Jianqing; Han, Fang; Liu, Han (2014), Challenges of Big Data analysis, *National Science Review*. 1 (2): 293–314. ISSN 2095- 5138. PMC 4236847. PMID 25419469. doi:10.1093/nsr/nwt032.
6. Tina M. St. John M.D. (2005), With Every Breath: A Lung Cancer Guidebook, (1):75-82. ISBN 0-9760450-2- 8, www.lungcancerguidebook.org.

7. B. Sobolev, A. Levy, and S. Goring, Eds (2016), Health Services Data: Big Data Analytics for Deriving Predictive Healthcare Insights, *in Data and Measures in Health Services Research*, Springer US, 1(1)1–17.
8. http://www.news-medical.net/news/20120530/Insulinuse-linked-to-lung-cancer-risk-in-diabetes.aspx (2012)
9. Dutkowska, Adam Antczak (2016), Comorbidities in lung cancer , *AgataEwa, Pneumonologiai Alergologia Polska*, 84 (3): 186–192.
10. Xie X, Liu Q, Wu J, Wakui M.(2009), Impact of cigarette smoking in type 2 diabetes development, Cta PharmacologicaSinica,30 (6):784- 787.
11. Chang SA (2012), Smoking and Type 2 Diabetes Mellitus. Diabetes & Metabolism, Journal,36(6):399-403
12. Maddatu, Judith et al.(2017), Smokingand the risk of type 2 diabetes, *Translational Research:*184(1),101-107.
13. Appari A, Eric Johnson M, Anthony DL.(2013), Meaningful use of electronic health record systems and process quality of care: evidence from a panel data analysis of U.S. acute-care hospitals, *Health ServRes*,48(1):354–75.
14. Fitzhenry F, Murff HJ, Matheny ME, et al.(2013), Exploring the frontier of electronic health record surveillance: the case of postoperative complications, Med Care51:509–16.
15. J.R. Quinlan.(1994), C4.5 programs for machine learning, Morgan Kaufmann Publishers,(16):235-240.
16. Vapnik,V.(1995), Support-vector networks, .Machine Learning. 20 (3): 273–297.
17. G. Dimitoglou, J. A. Adams, and C. M. Jim.(2012), Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability, CoRR, 4(8):1–9.
18. Hamid KarimKhani Z and et.al.(2015), A comparative survey on data mining techniques for breast cancer diagnosis and prediction Survey, Indian Journal of Fundamental and Applied Life Sciences.5 (S1): 4330- 4339..
19. S. G. Armato, III and W.F. Sensakovic(2004), Automated lung segmentation for thoracic CT: Impact on computer-aided diagnosis, Acad. Radiol., 11(9): 1011-1021.
20. I. sluimer, M. Prokop, and B. van Ginneken (2005), Towards automated segmentation of the pathological lung in CT, *IEEE Trans. Medical Image*, 24(8):1025-1038.
21. Y. Xu, M. Sonka, G. McLennan, J. Guo. And E.A. Hoffman (2006), MDCT-based 3-D texture Classification of emphysema and early smoking related lung anthologies, *IEEE Trans. Med. Imag.*, 25(4): 464-475.
22. Tanushree Sinha Roy, NeerajSirohi, ArtiPatle (2015), Classification of Lung Image and Nodule Detection Using Fuzzy Inference System, *International Conference on Computing, Communication and Automation (ICCCA2015).*
23. Ritika Agarwal, AnkitShankhadhar, Raj Kumar Sagar (2015), Detection of Lung Cancer Using Content Based Medical Image Retrieval, *Fifth International Conference On Advanced Computing And Communication Technologies.*
24. M. Mukherjee and P. K. Biswal (2018), Segmentation of lungs nodules by iterative thresholding method and classification with Reduced Features, *Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Coimbatore: 450-455.