International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Machine Learning Based Lung Cancer Classification in Histopathological Images- A Survey

Kiruba. B [1], J. Vijaykumar [2], Nisha. P [3]

[1]*Department of Electronics and Instrumentation, Bharathiar University, Coimbatore, Tamilnadu, India.*
[2]*Department of Electronics and Instrumentation, Bharathiar University, Coimbatore, Tamilnadu, India.*
[3]*Department of Electronics and Instrumentation, Bharathiar University, Coimbatore, Tamilnadu, India.*

Abstract - Examining histopathology slides visually is a key technique utilized by pathologists to determine the stage, type, and subtype of lung tumors. The most common subtypes of lung cancer are adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC), and distinguishing between them necessitates careful visual evaluation by a skilled pathologist. The deep learning and machine learning techniques used to categorize histopathological lung cancer images are thoroughly examined in this work to aid in diagnosing lung cancer. Digital tissue pathology improves diagnosis accuracy and gives the pathologist more detail and better image quality with multiple viewing options and team annotations. Histopathological images are very useful for examining the conditions of different biological structures and diagnosing diseases like cancer.

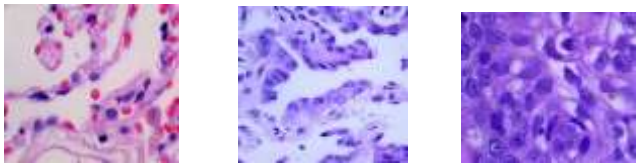
***Key Words*:** Lung cancer, Machine Learning, Histopathological Image, Deep Learning.

1.INTRODUCTION

The biggest cause of death is lung cancer, with a mortality rate of 18%. Different cancers have different death rates: intestinal cancer is 7.7% to 6.8%, liver cancer is 8.3%, and colon cancer is 9.4%. Lung cancer patients have varying degrees of disease and response to treatment. According to the US Organization, WHO, and EMA, carcinoma of the lungs is now the leading cause of death in the US and Europe, surpassing even heart disease. A correct diagnosis is necessary for each patient with lung cancer to obtain the right treatment. Treatment for lung cancer requires a combination of medicines. Cancer prognoses vary from 4% to 17%. Early lung cancer can be cured with surgery. Depending on the stage, patients with non-small cell lung cancer (NSCLC) who had their cancer removed had a 75% to 100% 5-year survival rate. I have been diagnosed with stage IIIA non-small lung cancer. AT this stage, survival is 25%. (Rahman et al.2025). In addition to accurately predicting the potential outcomes of a particular form of cancer, machine learning algorithms can identify and find intriguing patterns in complex datasets. It helps developers build models that associate a condition with several variables. Presently, image classifiers are used to identify anomalies and the existence of lung cancer hot spots using Computer Tomography (CT) images. The two main categories of image classification methods are supervised and unsupervised algorithms. According to our data, the SVM algorithm performs remarkably well, reaching hitherto unheard-of levels of training accuracy (99.99%), testing accuracy (96.60%), precision (96.39%), F1 score (96.39%), and Cohen Kappa score (94.67%). Furthermore, we examine the interpretability of the ML models by elucidating the underlying decision-making processes through feature importance analysis and visualization tools. (R. I. Sumon et al., 2024).

1.1 Lung Cancer Histopathological Image

A collection of histopathological pictures used to diagnose lung and colon cancer is known as the Dataset. Each image's initial resolution was some pixels, which was standardized by cropping it to pixels. This dataset is separated into three different classifications (adenocarcinoma, benign, and squamous) of pictures. In our investigation, we concentrated on the lung cancer subpopulation [9].



A) Benign B) Adenocarcinoma C) Squamous carcinoma

Fig.1 Lung cancer in a histopathological image

2. RELATED WORKS ON LUNG CANCER DETECTION AND CLASSIFICATION

As the leading cause of mortality and a deadly disease that affects people all over the world, lung cancer is growing at an alarming rate. As a result, it is essential to create a precise system that will help medical professionals identify and treat this illness. In the realm of lung cancer, numerous researchers have created a variety of models that are applied to the dataset.

2.1 Data Preprocessing

Dataset pre-processing was performed, and images were subjected to resizing, data enhancement, and data normalization operations. All images were resized to 256×256 pixels. Five types of enhancement techniques (horizontal flip, rotation, scaling, height shift, shift, and width shift) were applied to the dataset, where the rotation was set, and the height shift and width shift were set for the image size. The training set was expanded for training the MIM, and no data enhancement operation was taken for the test set because these enhancement methods change the relative position of cancer in histopathological images. The tumor tissue was removed and sent to the pathology department, so each doctor may use different stains when creating tumor sections due to

personal preferences [1]. Noise Ninja-GF-Based Noise Removal to remove noise from the CT images, we use the Noise-Gaussian Filtering (GF) technique in this step. This method uses weighted averages based on Gaussian functions, like a cunning ninja, to subtly remove noise and improve image clarity like a silent night watchman. A key component of the SCMO-MLL2C approach, GF-based noise obliteration coordinates the effort to identify and classify lung cancer from CT images. In this field, Gaussian Filtration (GF) stands out as the artistic brushstroke, lovingly entrusted with enhancing CT scan quality by eliminating distracting noise [2]. Throughout the processing process, pictures with an equivalent magnification of x 20(0.5 μm per pixel) were used to preserve both local features and the global perspective. The pathologists manually annotated the TTF-format WSIs using the ASAP platform, designating distinct regions of colored irregular polygons that corresponded to distinct histological lung tissue types [3]. Image preparation techniques, such as subsampling or wavelet transform, can be used to lower the computing cost of processing such images with CAD systems. The regions of interest can then be tentatively located by analysing the low-resolution images produced in this manner: only these regions proceed to the higher resolution processing step. Other preprocessing methods, including picture smoothing, denoising, and enhancement, may be used for image restoration in the event of low-quality input, such as extreme noise, low contrast with weak edges, and intensity inhomogeneity. Image smoothing typically refers to spatial filtering, such as Gaussian filtering and bilateral filtering, which remove tiny features and picture noise to highlight the main image structure. To eliminate the noise created during the image acquisition, filtering, compression, and reconstruction processes, image denoising techniques are employed. Wavelet thresholding, variational techniques, robust statistics, and partial differential equation (PDE)-based anisotropic diffusion are the main types of image denoising techniques [4]. The pipeline for pre-processing is Contrast Limited Adaptive Histogram Equalization (CLAHE). The sample images from the lung cancer data subset are labelled as the Dataset. The process of image data pre-treatment. The resized images are not displayed to scale. This technique improves contrast by operating on separate regions or tiles of the image. The contrast in each tile is enhanced to align the histogram of the entire output area closely with the histogram defined by the specified distribution [5]. To do the pre-processing steps, the image was enlarged, converted to BGR2RGB, and then converted to a NumPy array. Feature scaling is the next phase, which entails using the generalization technique on the picture. Labeling is then carried out by assigning a label (lung n or lung SCC) to each image. The image is then subjected to feature scaling using the generalization approach, in which the values of an easy image array are divided by 255, the greatest value that can be obtained (the image's maximum intensity Value) [6].

Table 1. Comparison of data preprocessing-based techniques

Ref	Cancer	Techniques	Result
Arunchalam et. al., (2024)	Benign, malignant, normal	Noise Ninja-Gaussian Filtering	classification
Yang et al. (2021)	Lung cancer subtypes	Extract ROI	classification
He et al. (2012)	Carcinoma detection	spatial filtering, bilateral filtering	Disease classification
Nayak et. al., (2024)	Binary colon lung cancer	CLAHE	classification
Hamed et. al. (2023)	Lung tissue	Bgr2rgb	Disease classification

2.2 Image segmentation techniques

The WSI of lung cancer is the region obtained by the threshold segmentation algorithm from the tissue. It has a sliding window with a size of 224x224 from tissue. The pixel has a step size of 128. [12].The transSegNet segmentation model is proposed with an input layer, an encoder, patch-embedded the feature –level information is forwarded to the output layer. This segmentation is proposed to accurately calculate and process the image to block further processing [13].A novel unsupervised segmentation method for pathology images. Therefore, the segmentation method is for invasive and non-invasive methods of pathological images. We extract the images as w x w pixel patches in this segmentation. It is also applicable to 2D and 3D medical image applications. A large NMI value has the best result segmentation. [14]. This method proposes the quality segmentation applied to images from two cancer types in the WHO Grade (LGG) and Lung adenocarcinoma(LUAD). This segmentation algorithm discriminates between background tissue and target nuclei through a threshold parameter. The choice of threshold parameter values leads to under-segmentation or over-segmentation of an image [15]. A dual encoder network of tissue semantic segmentation of a histopathological image called DETisSeg. The properties of cancer cells and TME are closely associated with the progression of cancer. Different types of tissues in the TME qualification. Therefore, automatic cancer tissue region segmentation is done by modern computer vision algorithms and HE-stained histopathological images. They have a large diversity of size, low contrast, texture features, small gaps, and inconsistent staining [he].The system architecture of lung cancer segmentation and classification of histopathological images. Before segmentation and classification to identify the disease in an image, input image data must be pre-processed and split into training and testing datasets. Deep learning algorithms have been used to automate feature detection in medical images [16].

Table 2. Comparison of segmentation-based techniques

Ref	Cancer	Techniques	Result
Rui et al. (2022)	Lung cancer	Threshold segmentation	Classification
Taliba et al (2024)	Healthy/abnormal cell	TransSegNet segmentation	Accuracy
Moriya et al (2018)	Lung adenocarcinoma	K-means segmentation and the multithreshold OTUs method	NMI score 0.626 compared to 0.168 and 0.167
Wen et al (2017)	LGG and LUAD	Nuclei segmentation quality	LGG and 75.43% for LUAD.
He et al (2024)	TME and TILs	Tissue semantic segmentation	Precision
Kanakaradi et al. (2024)	SCLC, NSCLC, SQCC,	U Net model, ResNet-50, EfficintNetB5, VGG-16	Accuracy 99%

2.3 Image feature extraction

The general structure of feature extraction using VGG16.The DF extraction VGG16 framework is basic work. This features into various machine learning models along with LBP features; we applied ensemble learning with this technique and evaluated the result.[17]. Image feature extraction of the CAD system identifies the disease signatures with their image features. The CAD system analyses the Histopathological image. Traditional features include morphometric with object size and shape, graph-based and minimum spanning, intensity and colour features, and texture features in the CAD system [4]. Radionics feature extraction was performed using the pixel spacing in three dimensions. The wavelet, Laplacian of Gaussian (LOG), square, square root, Logarithm, Exponential Gradient, and LBP2D filters were applied to the original MR images. The radionics features included 396 first-order features, 14 shape-based features, and 1496 texture features. T1 mapping has a total of 1906 radiomics features extracted from the original early detection of lung and colon cancer via the Ensemble DL model. The HIELCC-EDL technique utilizes histopathological images to identify lung and colon cancer. This technique uses the Wiener filtering method of noise reduction. They also use the channel of Residual Network for feature patterns. The LCC detection using three classifiers has an extreme learning machine, a competitive neural network, and long short-term memory. The HIELCC_EDL model has been applied to a benchmark dataset. [19].

Table 3. Comparison of image feature extraction techniques

Ref and year	Cancer	Techniques	Result
Singh et.al (2023)	Lung and colon	VGG16	Precision
He et al. (2012)	Lung carcinoma	CAD	Classification
Jiang et al. (2024)	Lesion	GLCM, GLZSM, GLRLM, GLDM	Classification
Alotaibi et.al. (2024)	LCC	HIELCC-EDL	Accuracy 99.60%

2.4 Image classifier techniques

The classification of lung cancer is the most important objective of this research. Machine learning algorithms are used for efficient classification, Random Forest, Decision the, and support vector machine classifier [20]. The enhancement of the hyperparameter on the combination of GWO-IWO-Decision Tree classifier for RAdam outperforms all other classifiers has achieving an accuracy of 91.57% in classifying Benign and adenocarcinoma classes [21]. We examine the variant efficient model for different image resolutions to classify lung and colon histopathology images into five categories: colon adenocarcinoma, benign tissue, lung adenocarcinoma, lung benign tissue, and lung squamous cell carcinomas [22]. We have used a deep learning method of classification on the model. The modified Alex Net's ability to accurately categorize lung cancer. The deep learning architecture DenseNet121, used for lung cancer classification, has a remarkable accuracy performance of 99.6%. It detects Lung Adenocarcinoma. It reliably identifies benign instances of lung tissue, highlighting its excellent performance in this task [9]. In this case, DFs from Alex Net are coupled with high-level features that were retrieved using discrete wavelet transforms (DWT). The output-categorized images are then produced by feeding the linear SVM with the combination of high-level fused features for classification. Hyperspectral data can be found in the histopathology image. It's spectral and (SVMs) [23]. CNN can only process a limited number of image sizes; image patches must be created from scanned images. The 60%, 20%, and 20% separation for the training, validation, and test sets was arbitrary, and there isn't a gold standard in place at the moment. A more robust model would be produced by a larger percentage of instances in the training set, but the data in the validation and test cohorts might not be representative. However, as information from the training set moves into the validation set during hyperparameter tuning, separation into the three sets is required. Several CNN architectures and variations have been developed in the past, and the ImageNet dataset demonstrates that some of them have high classification accuracy [24].

Table 4. Comparison of image classifier-based techniques

Ref	Cancer	Technique	Result
Karthika et.al. (2024)	Early cancer detection	RF, DCT, SVM	Classification 92.26%
Shanmugam et.al. (2023)	Benign, adenocarcinoma	GWO-IWO-Decision Tree	Classification 91.57%
Anjum et.al. (2023)	Lung and colon image	EfficientNetB2	Cclassification 97.24%
Rahman et.al (2025)	Benign and lung tissue, adeno	Alex Net, DenseNet121	Classification 100%
Sethy PK et.al. (2023)	Lung nodule sample	SVM	Classification 99.3%
Kriegsmann et.al. (2020)	ADC, SQCC, SCLC	CNN	Accuracy 100%

4. Conclusion

When it comes to analysing pathology photos, deep learning has many advantages over shallow learning techniques. These include the ability to perceive complicated things, the ability to clarify feature definitions, efficiency through equal calculation, and reasonableness for transfer learning. To classify photos of lung and colon cancer histology. This study in all three features-depth, width, and resolution-according to the resources that are the CNN model and EfficientNet model in a principled manner. This is the first study to use the pre-trained EfficientNet model for lung colon image categorization. Starting at 224x224 in the B0 model, all variations with varying resolutions rose to 600x600 in the B7 model. The last layer of each model is tuned for optimal performance, and several dropouts guard against overfitting. Accuracy of EfficientNetB7 model 0-20-40-60-80-100 epochs train valid Effective B7 model epoch train 50 40 30 20 10 0 20 40 60 80 100 Plots showing accuracy and loss for training and validating the EfficientNetB7 model. First, the combination of VGG-16 and EfficientNetB7 is not exclusive. Increases not just the classification finding accuracy but also their stability and consistency, as seen by the similarly high precision, recall, and F1-score values. Second, this study employs cutting-edge methods for image augmentation and dataset balance, which enhance the model's generalization. Third, our method shows that it can overcome the difficulties in classifying histopathological images, which frequently have a lot of visual differences and are very complex.

5. Survey on Performance Analysis

The performance analysis of the reviewed paper is compared based on its accuracy. The variables required to evaluate the accuracy include True positive (TP), True negative (TN), False positive (FP), and False negative (FN). For these values, TN and FN should be the lowest.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Presents a comparison between the articles reviewed based on their accuracy. Among the articles reviewed, the segmentation, feature extraction based on a classifier, has the highest accuracy of 100%. The most common classification of SVM, CNN, GLCM, and CAD is more efficient in lung cancer detection of Histopathological images.

REFERENCES

[1]Liu, M., Li, L., Wang, H., Guo, X., Liu, Y., Li, Y., & Luan, L. (2023). A multilayer perceptron-based model applied to histopathology image classification of lung adenocarcinoma subtypes. *Frontiers in Oncology*, *13*, 1172234

[2]Arunachalam, P., Geetha, S., Jose, N. N., & Vivekanandan, G. (2024, July). A Feline-Inspired Optimizer Enhanced through Self-Improvement, Coupled with Machine Learning, for the Identification of Lung Cancer in CT Scans. In *2024 Second International Conference on Advances in Information Technology (ICAIT)* (Vol. 1, pp. 1-6). IEEE.

[3]Yang, H., Chen, L., Cheng, Z., Yang, M., Wang, J., Lin, C., ... & Li, W. (2021). Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: a retrospective study. *BMC Medicine*, *19*, 1-14.

[4]He, L., Long, L. R., Antani, S., & Thoma, G. R. (2012). Histology image analysis for carcinoma detection and grading. *Computer methods and programs in biomedicine*, *107*(3), 538-556.

[5]Nayak, T., Gokulkrishnan, N., Chadaga, K., Sampathila, N., Mayrose, H., & KS, S. (2024). Automated histopathological detection and classification of lung cancer with an image pre-processing pipeline and spatial attention with deep neural networks. *Cogent Engineering*, *11*(1), 2357182.

[6]Hamed, E. A. R., Salem, M. A. M., Badr, N. L., & Tolba, M. F. (2023). An efficient combination of a convolutional neural network and the LightGBM algorithm for lung cancer histopathology classification. *Diagnostics*, *13*(15), 2469.

[7]Coudray, N., Ocampo, P.S., Sakellaropoulos, T. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat Med* **24**, 1559–1567 (2018). https://doi.org/10.1038/s41591-018-0177-5

[8]Sumon, R. I., Mazumdar, M. A. I., Uddin, S. M. I., & Kim, H. C. (2024, July). Exploring Deep Learning and Machine Learning Techniques for Histopathological Image Classification in Lung Cancer Diagnosis. In *2024 International Conference on Electrical, Computer and Energy Technologies (ICECET* (pp. 1-6). IEEE

[9]Rahman, Mahfujur. (2025). An Evaluation of Various Techniques for Classifying Histopathological Images of Lung Cancer. 7. 60-89. 10.5281/zenodo.14603781.

[10]M. Magdy Amin, A. S. Ismail, and M. E. Shaheen, "Multimodal Non-Small Cell Lung Cancer Classification Using Convolutional Neural Networks," in *IEEE Access*, vol. 12, pp. 134770-134778, 2024.

[11]Dad, I., He, J., & Baloch, Z. (2024). Graph-Based Analysis of Histopathological Images for Lung Cancer Classification Using GLCM Features and DeepWalk Embeddings.

[12]Rui, Xu & Wang, Zhizhen & Liu, Zhenbing & Han, Chu & Yan, Lixu & Lin, Huan & Xu, Zeyan & Feng, Zhengyun & Liang, Changhong & Pan, Xipeng & Liu, Zaiyi. (2022). Histopathological Tissue Segmentation of Lung Cancer with Bilinear CNN and Soft Attention. BioMed Research International. 2022. 10.1155/2022/7966553.

[13]Talib, L. F., Amin, J., Sharif, M., & Raza, M. (2024). Transformer-based semantic segmentation and CNN network for detection of histopathological lung cancer. *Biomedical Signal Processing and Control*, *92*, 106106.

[14]Moriya, T., Roth, H. R., Nakamura, S., Oda, H., Nagara, K., Oda, M., & Mori, K. (2018, March). Unsupervised pathology image segmentation using representation learning with spherical k-means. In *Medical Imaging 2018: Digital Pathology* (Vol. 10581, pp. 278-284). SPIE.

[15]Wen, S., Kurc, T. M., Gao, Y., Zhao, T., Saltz, J. H., & Zhu, W. (2017). A methodology for texture feature-based quality assessment in nucleus segmentation of histopathology images. *Journal of Pathology Informatics*, *8*(1), 38.

[16]He, P., Qu, A., Xiao, S., & Ding, M. (2024). Detisseg: A dual-encoder network for tissue semantic segmentation of histopathology images. *Biomedical Signal Processing and Control*, *87*, 105544.

[17]Kanakaraddi, S. G., Handur, V. S., Jalannavar, A., Chikaraddi, A., & Giraddi, S. (2024). Segmentation and Classification of Lung Cancer using Deep Learning Techniques. *Procedia Computer Science*, *235*, 3226-3235.

[18]Jiang, J., Xiao, Y., Liu, J., Cui, L., Shao, W., Hao, S., ... & Hu, C. (2024). T1 mapping-based radiomics in the identification of histological types of lung cancer: a reproducibility and feasibility study. *BMC Medical Imaging*, *24*(1), 308.

[19]Alotaibi, M., Alshardan, A., Maashi, M., Asiri, M. M., Alotaibi, S. R., Yafoz, A., ... & Khadidos, A. O. (2024). Exploiting histopathological imaging for early detection of lung and colon cancer via an ensemble deep learning model. *Scientific Reports*, *14*(1), 20434

[20]Karthika, M. S., Rajaguru, H., & Nair, A. R. (2024). Performance enhancement of classifiers through bio-inspired feature selection methods for early detection of lung cancer from microarray genes. *Heliyons*, *10*(16).

[21]Shanmugam, K., & Rajaguru, H. (2023). Exploration and enhancement of classifiers in the detection of lung cancer from histopathological images. *Diagnostics*, *13*(20), 3289.

[22]Anjum, S., Ahmed, I., Asif, M., Aljuaid, H., Alturise, F., Ghadi, Y. Y., & Elhabob, R. (2023). Lung Cancer Classification in Histopathology Images Using Multiresolution Efficient Nets. *Computational Intelligence and Neuroscience*, *2023*(1), 7282944.

[23]Sethy PK, Geetha Devi A, Padhan B, Behera SK, Sreedhar S, Das K. Lung cancer histopathological image classification using wavelets and AlexNet. *Journal of X-Ray Science and Technology*. 2023;31(1):211-221. doi:10.3233/XST-221301

[24]Kriegsmann, et. Al., M., Haag, C., Weis, C. A., Steinbuss, G., Warth, A., Zgorzelski, C., ... & Kriegsmann, K. (2020). Deep learning for the classification of small-cell and non-small-cell lung cancer. *Cancers*, *12*(6), 1604.

[25]yahia Ibrahim, N., & Talaat, A. S. (2022). An enhancement technique to diagnose colon and lung cancer by using double CLAHE and deep learning. *International Journal of Advanced Computer Science and Applications*, *13*(8)

[26]Singh, O., Singh, K.K. An approach to classify lung and colon cancer of histopathology images using deep feature extraction and an ensemble method. *Int. j. inf. tecnol.* 15, 4149–4160 (2023). https://doi.org/10.1007/s41870-023-01487-1.