



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IX **Month of publication:** September 2026

DOI: <https://doi.org/10.22214/ijraset.2026.84891>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

AI-Powered Lung Cancer Classification Using Hybrid Histopathological Image Analysis

D. Suguna Seleena Kennedy¹, Dr. R. Venkata Lakshmi²

¹M. Tech, Department of Computer Science and Engineering, Bapatla Engineering College, Bapatla, AP, INDIA

²Asst Professor, Department of Computer Science and Engineering, Bapatla Engineering College, Bapatla, AP, INDIA

Abstract: Lung cancer is a major cause of cancer-related mortality, and accurate classification of histopathological tissue patterns can support the analysis of lung cancer subtypes. Manual examination of histopathological images is time-consuming and requires careful assessment of cellular and tissue-level morphological patterns. To assist automated histopathological image analysis, this study presents a hybrid deep learning and machine learning framework for three-class classification of lung histopathological images. The proposed system employs an ImageNet-pretrained DenseNet201 convolutional neural network for deep feature extraction through transfer learning, followed by frozen-base training and fine-tuning of selected network layers. A learned 256-dimensional feature representation is subsequently extracted from the trained network and provided to a CatBoost classifier for final classification. The system categorizes images into lung adenocarcinoma, lung squamous cell carcinoma, and Benign Lung Tissue. On a validation set of 3,000 images, the proposed hybrid framework achieved an overall classification accuracy of 99.90%, with 2,997 images correctly classified. The framework combines deep convolutional feature learning with gradient-boosted machine learning classification to provide an automated and modular approach for lung histopathological image classification.

Keywords: Lung Histopathology, Histopathological Image Classification, Lung Cancer Classification, DenseNet201, Convolutional Neural Network, Transfer Learning, Deep Feature Extraction, CatBoost, Lung Adenocarcinoma, Lung Squamous Cell Carcinoma, Benign Lung Tissue.

I. INTRODUCTION

Lung cancer is one of the major causes of cancer-related mortality worldwide. Histopathological examination of tissue samples plays an important role in identifying abnormal tissue patterns and distinguishing major lung cancer subtypes, including adenocarcinoma and squamous cell carcinoma [9]. However, the examination of histopathological images involves the assessment of complex cellular and tissue structures, making automated image analysis an important area of research.

The growing availability of digitized histopathological datasets has provided opportunities for developing computer-aided methods for tissue classification. One such resource is the Lung and Colon Cancer Histopathological Image Dataset (LC25000), which contains labelled histopathological images from different tissue categories and has been widely used in studies involving automated image classification [1]. The lung portion of this dataset provides images representing adenocarcinoma, squamous cell carcinoma, and benign lung tissue, making it suitable for investigating automated classification of these tissue types.

Deep learning, particularly convolutional neural networks (CNNs), has been increasingly applied to histopathological image analysis because CNNs can learn visual patterns directly from image data. Instead of relying entirely on manually designed characteristics, these networks can develop hierarchical representations of tissue structures during training [6], [9]. DenseNet, introduced by Huang et al., uses connections between multiple layers to encourage feature reuse and improve the flow of information through the network [2]. These characteristics make DenseNet-based architectures useful for learning image representations in medical image analysis.

Although a CNN can be trained to perform classification directly, separating feature learning from the final classification stage provides another approach for developing a modular classification system. In such a framework, features learned by a CNN can be extracted and subsequently provided to a conventional machine-learning classifier. Previous research has explored this type of hybrid strategy for lung histopathological image classification. Hamed et al., for example, combined CNN-based feature extraction with LightGBM and demonstrated strong classification performance on lung histopathological images [8]. This provides motivation for investigating alternative machine-learning classifiers with deep features.

In this study, a hybrid DenseNet201–CatBoost framework is developed for the classification of lung histopathological images into three categories: lung adenocarcinoma, benign lung tissue, and lung squamous cell carcinoma.

An ImageNet-pretrained DenseNet201 network is employed to learn visual representations using transfer learning. The network is first trained with its convolutional base frozen and is subsequently fine-tuned by allowing selected layers to learn from the target dataset. A 256-dimensional feature representation obtained from the trained network is then extracted and supplied to a CatBoost classifier for the final classification stage [3].

The main objective of this work is to investigate whether the combination of DenseNet201-based feature extraction and CatBoost classification can provide an effective and modular approach for lung histopathological image classification. The performance of the proposed framework is evaluated using classification accuracy, precision, recall, F1-score, and confusion-matrix analysis.

A. Objectives of the System

The primary objective of this work is to develop an automated system for classifying lung histopathological images into Benign Lung Tissue, lung adenocarcinoma, and lung squamous cell carcinoma using a hybrid deep learning and machine learning approach. The specific objectives are:

- 1) To develop an automated system for classifying lung histopathological images.
- 2) To preprocess the images by resizing, normalization, and training-time data augmentation.
- 3) To use DenseNet201 for extracting learned visual features from histopathological images [2].
- 4) To apply transfer learning using ImageNet-pretrained DenseNet201 weights, followed by frozen-base training and fine-tuning of selected layers.
- 5) To generate a learned 256-dimensional feature representation from the trained network.
- 6) To use CatBoost for classification of the extracted features into the three target categories [3].
- 7) To evaluate the classification performance using accuracy, precision, recall, F1-score, and confusion-matrix analysis.
- 8) To provide an automated prediction interface that displays the predicted class and corresponding probability score.

II. LITERATURE SURVEY

Recent research has increasingly explored deep learning for automated analysis of histopathological images. The Lung and Colon Cancer Histopathological Image Dataset (LC25000) introduced by Borkowski *et al.* provides 25,000 images across five tissue categories, including lung adenocarcinoma, lung squamous cell carcinoma, and benign lung tissue, and has become a widely used dataset for developing image-classification methods [1].

Deep convolutional neural networks have demonstrated the ability to learn discriminative visual representations directly from histopathological images. Echle *et al.* discussed the growing role of deep learning in computational cancer pathology, while Davri *et al.* reviewed deep learning applications for lung cancer diagnosis and prediction using histological and cytological images [6], [9]. These studies demonstrate the potential of automated feature learning for assisting histopathological image analysis.

Huang *et al.* introduced DenseNet, a convolutional architecture in which each layer receives feature maps from preceding layers through dense connections. This promotes feature reuse and improves information propagation within the network [2]. The ability of DenseNet architectures to learn rich hierarchical representations makes them suitable for medical image feature extraction. In the present work, DenseNet201 is employed as the deep feature extraction component using transfer learning.

Hybrid approaches that combine deep feature extraction with conventional machine learning classifiers have also been investigated. Hamed *et al.* proposed a CNN–LightGBM framework for lung histopathology classification using LC25000 and reported an accuracy of 99.6% under their experimental protocol [8]. This demonstrates the potential of separating feature extraction and classification into different learning stages.

CatBoost, introduced by Prokhorenkova *et al.*, is a gradient-boosting framework that can perform multiclass classification using numerical feature representations [3]. In contrast to approaches that directly use the CNN classification layer, the present study combines DenseNet201-based feature extraction with CatBoost for the final classification stage.

Additional datasets and recent studies have further expanded research in computational pulmonary pathology. Diosdado *et al.* introduced LungHist700 as a resource for deep learning in pulmonary pathology [7], while Ahmed *et al.* investigated deep learning for histopathological slide analysis in lung cancer [11]. Systematic research has also examined the broader use of deep learning for extracting information from lung histopathology beyond conventional image classification tasks [10].

Based on these developments, this study proposes a DenseNet201–CatBoost hybrid framework for three-class classification of lung histopathological images. DenseNet201 is used to learn a 256-dimensional feature representation, which is subsequently classified using CatBoost. The approach therefore combines deep representation learning with gradient-boosted machine learning while maintaining a modular classification pipeline.

III. METHODOLOGY

A. Dataset

The proposed framework was developed using the Lung and Colon Cancer Histopathological Image Dataset (LC25000) [1]. The lung subset used in this study contains 15,001 images belonging to three classes: lung adenocarcinoma (lung_aca), Benign Lung Tissue (lung_n), and lung squamous cell carcinoma (lung_scc). The classes contain 5,000, 5,000, and 5,001 images, respectively. The dataset was divided into training and validation subsets using an 80:20 image-level split implemented through the training and validation subsets of the image generator.

B. Image Preprocessing

All images were resized to 224×224 pixels with three color channels to match the DenseNet201 input dimensions. Pixel values were rescaled to the range [0,1]. During training, data augmentation including rotation, image shifting, horizontal and vertical flipping, zooming, and shearing was applied to improve model generalization. No augmentation was applied to the validation images.

C. DenseNet201 and Transfer Learning

DenseNet201 was selected as the deep feature-learning network. DenseNet uses dense connections between successive layers to promote feature reuse and information propagation [2]. An ImageNet-pretrained DenseNet201 backbone was used for transfer learning, with a custom classification head containing the learned feature layer used for subsequent feature extraction.

Training followed a two-phase transfer-learning strategy. In Phase 1, the DenseNet201 backbone was frozen and the newly added head layers were trained using the Adam optimizer [4] with an initial learning rate of 0.001 for a maximum of 20 epochs. The classification head also incorporated dropout regularization to reduce overfitting [5]. The best validation accuracy during this phase reached 100.00% at epoch 20.

D. Fine-Tuning

In Phase 2, the last 50 layers of the DenseNet201 backbone were unfrozen along with the seven trainable layers of the classification head, resulting in 57 trainable layers. Fine-tuning was performed using a reduced learning rate of 1×10^{-5} . ReduceLROnPlateau and early stopping were applied to control further training. The best validation accuracy during fine-tuning was 99.90%, and the corresponding model weights were restored for subsequent feature extraction.

E. Deep Feature Extraction

The output of the custom 256-dimensional dense_features layer was used as the learned representation of each image. The trained DenseNet201 network converted each input image into a 256-dimensional numerical feature vector. Feature extraction from the 12,001 training images produced a $12,001 \times 256$ feature matrix, which was subsequently used to train the CatBoost classifier.

F. CatBoost Classification

The extracted 256-dimensional feature vectors were supplied to a CatBoost classifier for final multiclass classification [3]. The classifier was trained to distinguish lung adenocarcinoma, Benign Lung Tissue, and lung squamous cell carcinoma. CatBoost was configured with 500 boosting iterations, a learning rate of 0.05, and a tree depth of 8.

Thus, the proposed hybrid pipeline can be summarized as:



Fig. 1. Pipeline Summary

G. Performance Evaluation

The final hybrid framework was evaluated on the 3,000-image validation set, containing 1,000 images from each class. Performance was assessed using accuracy, precision, recall, F1-score, and confusion-matrix analysis. The DenseNet201–CatBoost framework achieved an overall validation accuracy of 99.90%.

IV. IMPLEMENTATION

A. System Architecture

The developed **PulmoScan** system implements a modular pipeline for automated lung histopathological image classification. The system accepts an input histopathological image, performs preprocessing, extracts deep visual features using the trained DenseNet201 network, and passes the resulting 256-dimensional feature vector to the CatBoost classifier for final classification [2], [3].

The overall implementation follows:

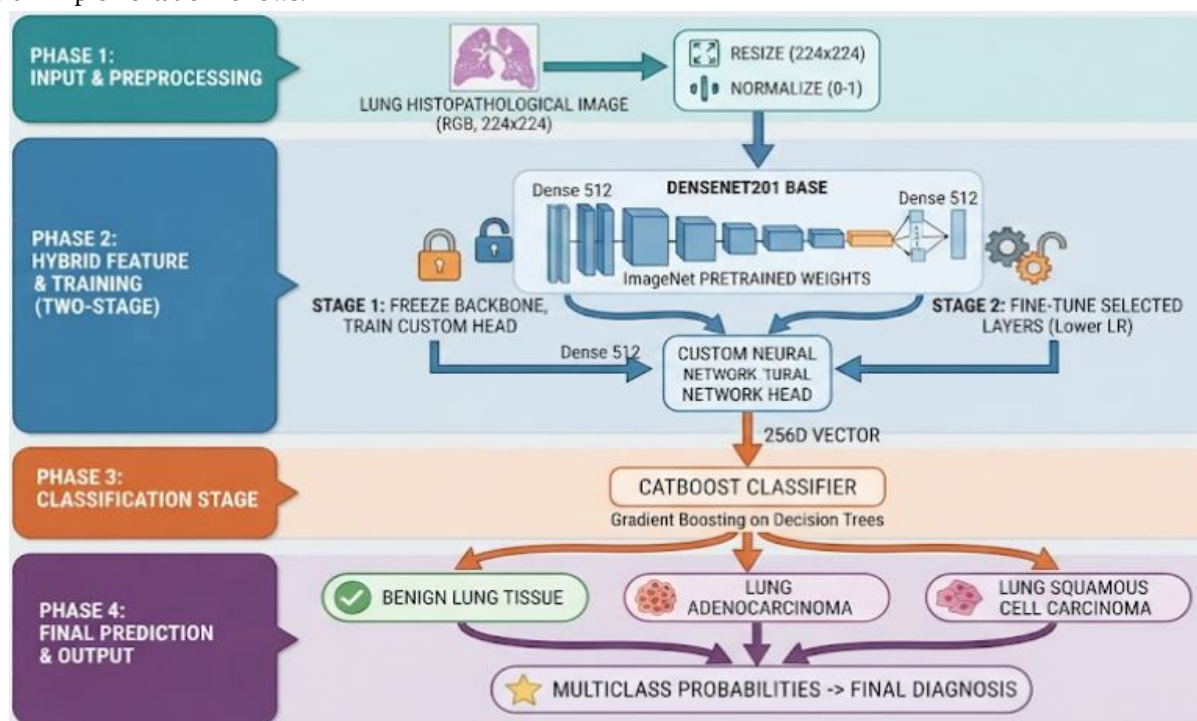


Fig. 2. Overall System Architecture

B. Deep Learning Model Implementation

The CNN component was implemented using DenseNet201 with ImageNet-pretrained weights. DenseNet architectures use dense connections between layers to promote feature reuse and information propagation [2]. Input images are resized to 224×224 pixels before being provided to the network. A custom classification head was added to the pretrained backbone, including the 256-dimensional dense_features layer used for subsequent feature extraction.

The trained DenseNet201 network was used as a feature extractor rather than as the final classifier. The output of the dense_features layer provides the intermediate representation used by the subsequent machine-learning classifier.

C. Hybrid Classification Implementation

The extracted 256-dimensional feature vectors were supplied to a CatBoost classifier for final multiclass classification [3]. CatBoost performs gradient-boosted decision-tree classification on the learned numerical feature representations. During implementation, features extracted from the training images were used to train the classifier, while features extracted from the validation images were subsequently used to generate the final predictions. The classifier produces one of three output categories: lung adenocarcinoma, Benign Lung Tissue, or lung squamous cell carcinoma.

D. PulmoScan Prediction Interface

A prediction interface was developed as part of the PulmoScan application to demonstrate the practical use of the trained model. The interface allows a histopathological image to be uploaded and displays the predicted tissue category and corresponding confidence score. Patient and slide information can also be associated with the uploaded sample.

For the demonstrated sample, the system predicted lung squamous cell carcinoma (lung_scc) with a probability of 99.9914%. The corresponding probabilities for lung adenocarcinoma and Benign Lung Tissue were 0.0064% and 0.0022%, respectively

E. Model Analysis and Report Generation

The application provides additional prediction information through its model-analysis and reporting components. The interface displays the probability distribution of the three classes and provides prediction-related information for the uploaded image. A downloadable PDF report is also generated containing the sample information, predicted classification, confidence value, and class-probability breakdown. The generated report includes a disclaimer stating that the model output is intended for research and academic purposes only and should not be considered a substitute for professional pathological diagnosis.

V. RESULTS

A. Classification Performance

The proposed DenseNet201–CatBoost framework was evaluated on a validation set of 3,000 images, comprising 1,000 images from each class. The system achieved an overall accuracy of 99.90%, with 2,997 images correctly classified. Lung adenocarcinoma achieved 100.00% precision, 99.70% recall, and 99.85% F1-score, while Benign Lung Tissue achieved 100.00% for all three metrics. Lung squamous cell carcinoma achieved 99.70% precision, 100.00% recall, and 99.85% F1-score. The macro and weighted averages were 99.90% for precision, recall, and F1-score.

TABLE I
CLASSIFICATION PERFORMANCE ON THE VALIDATION SET.

Class	Precision	Recall	F1-Score	Support
Lung Adenocarcinoma	100.00%	99.70%	99.85%	1000
Benign Lung Tissue	100.00%	100.00%	100.00%	1000
Lung Squamous Cell Carcinoma	99.70%	100.00%	99.85%	1000
Macro Average	99.90%	99.90%	99.90%	3000
Weighted Average	99.90%	99.90%	99.90%	3000

B. Confusion Matrix Analysis

The confusion matrix further illustrates the classification performance. Out of the 3,000 validation images, **2,997 were correctly classified**. Three lung adenocarcinoma images were incorrectly classified as lung squamous cell carcinoma, while all Benign Lung Tissue and lung squamous cell carcinoma images were correctly classified.

TABLE II
CONFUSION MATRIX OF THE VALIDATION SET

Actual \ Predicted	Lung Adenocarcinoma	Benign Lung	Lung Squamous Cell Carcinoma
Lung Adenocarcinoma	997	0	3
Benign Lung	0	1000	0
Lung Squamous Cell Carcinoma	0	0	1000

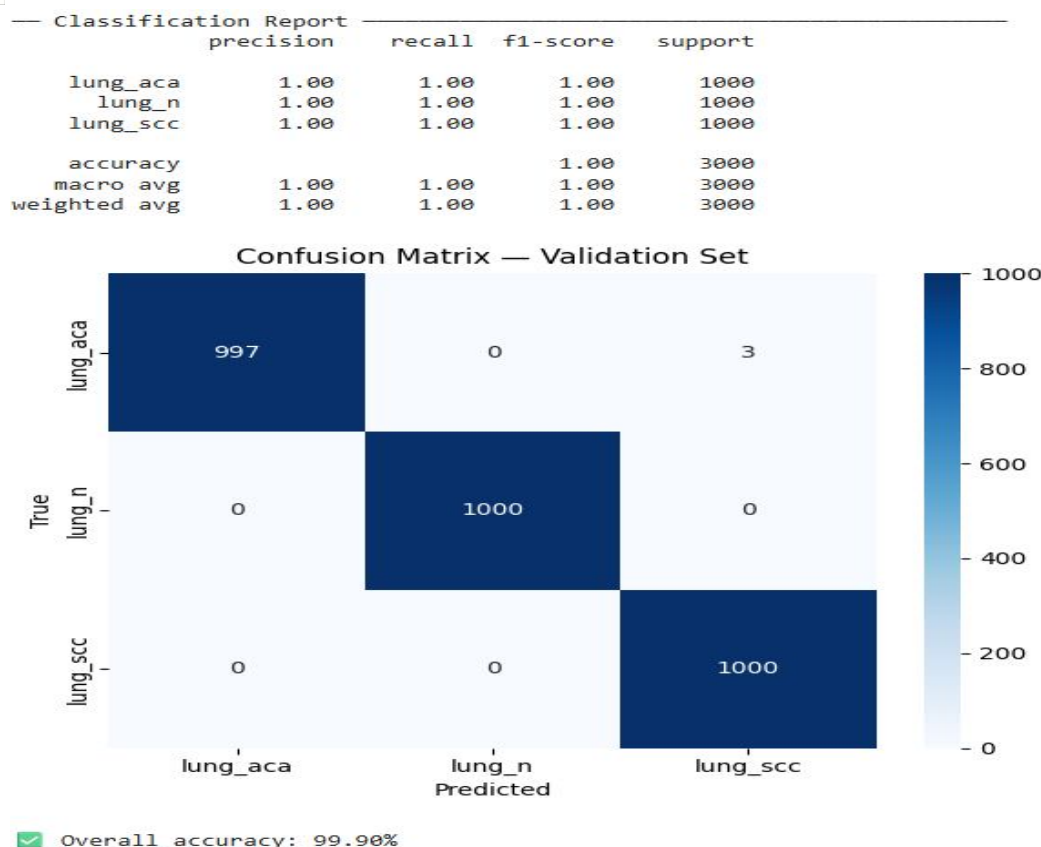


Fig. 3. Confusion matrix of the proposed DenseNet201– CatBoost framework

All three observed misclassifications occurred between the two malignant classes, with three adenocarcinoma images classified as squamous cell carcinoma.

C. Sample Prediction

The developed PulmoScan application was also evaluated using an individual histopathological image. For the demonstrated sample, the system correctly predicted lung squamous cell carcinoma (lung_scc).

The model assigned a predicted probability of 99.9914% to lung squamous cell carcinoma, while lung adenocarcinoma and Benign Lung Tissue received probabilities of **0.0064%** and **0.0022%**, respectively.

TABLE III
SAMPLE PREDICTION PROBABILITY

Class	Predicted Probability
Lung Squamous Cell Carcinoma	99.9914%
Lung Adenocarcinoma	0.0064%
Benign Lung Tissue	0.0022%

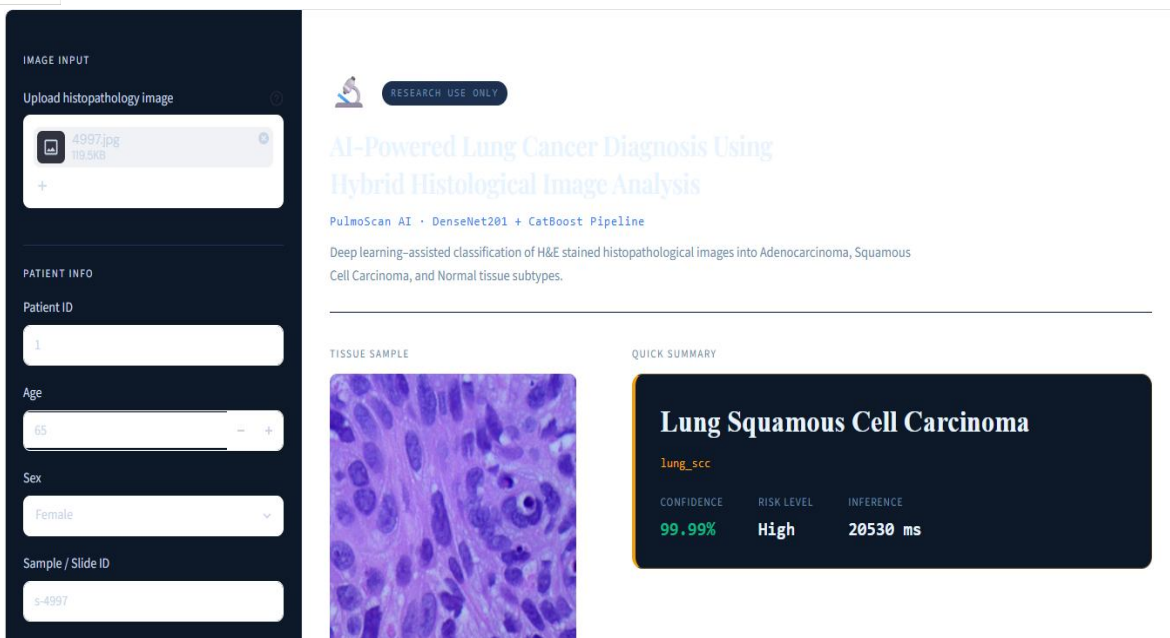


Fig. 4. PulmoScan prediction interface showing the sample image and predicted class.

D. Feature Representation Analysis

The trained DenseNet201 network generated a 256-dimensional feature vector for each input image. For the demonstrated sample, the extracted vector contained 60 non-zero activations and 196 zero activations. The feature values ranged from 0.000 to 9.678, with a mean activation of 0.8531.

The feature-vector visualization demonstrates that the CNN produces a numerical representation of the input image before classification by CatBoost. This intermediate representation provides an interpretable view of the features supplied to the machine-learning classifier.

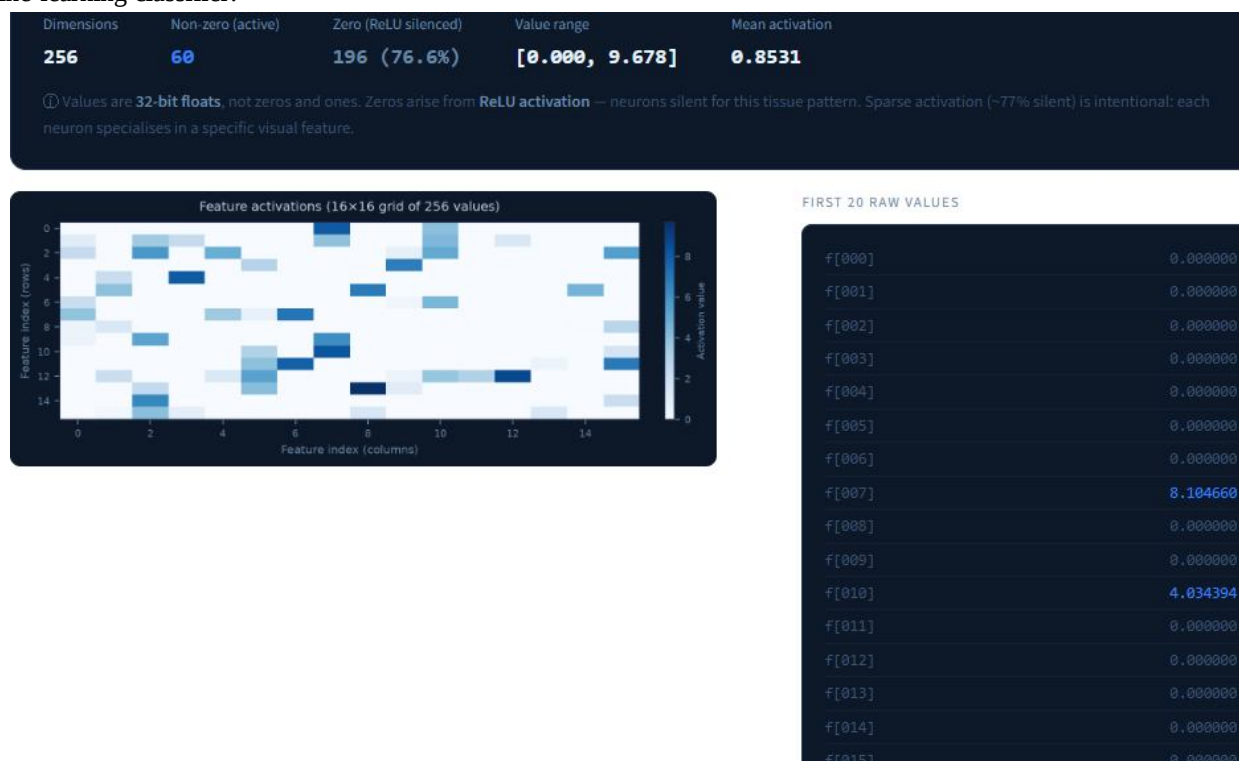


Fig. 5. Visualization of the extracted 256-dimensional feature representation.

E. Application Output and Report Generation

The PulmoScan interface provides the predicted class, probability distribution, patient/sample information, and additional prediction details. The system also generates a downloadable PDF report containing the input information, predicted classification, confidence value, and class-probability breakdown.

For the demonstrated sample, the application displayed a **99.99% confidence** for the predicted lung squamous cell carcinoma class. The generated report also includes a disclaimer stating that the output is intended for research and academic purposes and should not replace professional pathological diagnosis.

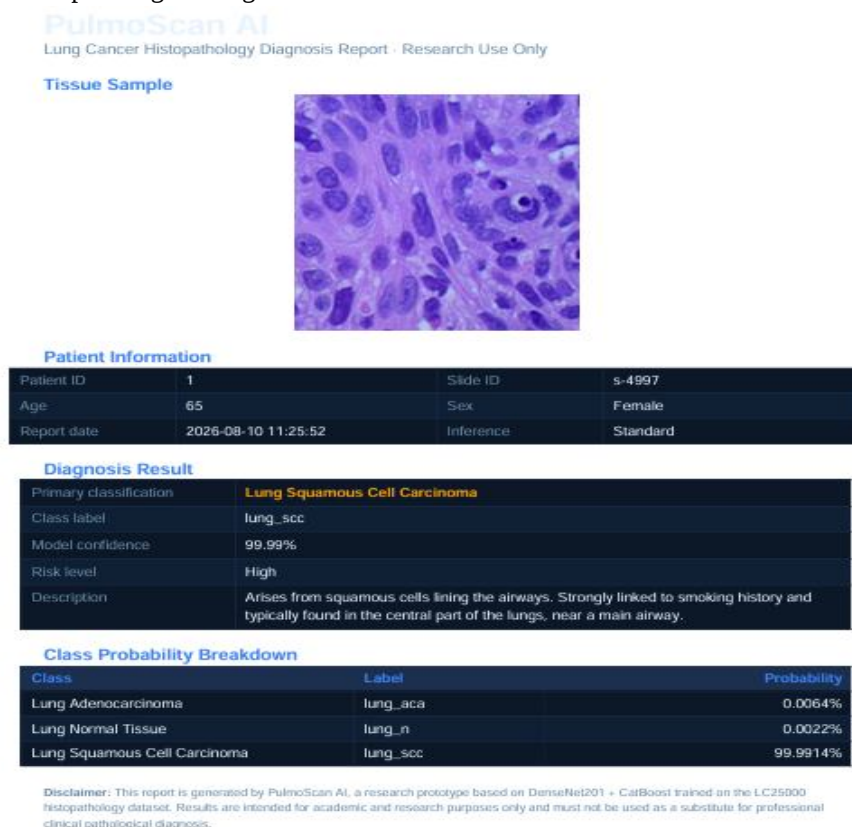


Fig. 6. Generated PDF prediction report from the PulmoScan application

F. Comparison with Previous Work

Previous research has demonstrated strong performance on LC25000 using hybrid deep-learning and machine-learning approaches. Hamed et al. [8], for example, reported 99.6% accuracy using a CNN–LightGBM framework on lung histopathological images from LC25000.

In comparison, the proposed DenseNet201–CatBoost framework achieved 99.90% validation accuracy under the experimental configuration used in this study. However, direct numerical comparison should be interpreted cautiously because differences in data partitioning, preprocessing, architectures, and evaluation protocols can affect reported performance.

TABLE IV
COMPARISON WITH PREVIOUS WORK

Method	Dataset	Reported Accuracy
CNN–LightGBM [8]	LC25000	99.6%
Proposed DenseNet201–CatBoost	LC25000 lung subset	99.90%

The results demonstrate that the proposed hybrid architecture provides strong classification performance while separating deep feature extraction from the final machine-learning classification stage.

VI. CONCLUSION

This study presented a hybrid deep learning and machine learning framework for automated classification of lung histopathological images. The proposed approach employs an ImageNet-pretrained DenseNet201 for deep feature learning and extracts a 256-dimensional feature representation from the trained network. These features are subsequently classified using CatBoost into three categories: lung adenocarcinoma, benign lung tissue, and lung squamous cell carcinoma.

The proposed DenseNet201–CatBoost framework achieved 99.90% accuracy on the 3,000-image validation set, demonstrating strong classification performance within the experimental setting. The system also provides automated predictions with class probabilities through the developed application interface. The separation of deep feature extraction and machine-learning classification provides a modular framework for lung histopathological image analysis. However, the results are based on the LC25000 dataset and the defined validation protocol; therefore, further evaluation on independent and clinically representative datasets is required to assess the generalizability of the proposed approach. The system is intended for research and academic purposes and should not be considered a substitute for professional pathological diagnosis.

VII. FUTURE SCOPE

The proposed framework can be further improved by evaluating its performance on larger, independent, and clinically representative histopathological datasets to assess its generalizability across different imaging conditions and patient populations. Future work may also investigate advanced preprocessing and explainable artificial intelligence (XAI) techniques to improve model robustness and provide greater insight into the visual features influencing classification. Alternative deep learning architectures and machine-learning classifiers can also be explored for further performance comparison and optimization. The system can additionally be extended to support patch-based and whole-slide image analysis, enabling the processing of larger histopathological specimens. Integration of relevant clinical information could further support more comprehensive analysis. Deployment on cloud-based or clinical research platforms may improve accessibility, scalability, and practical usability. These extensions could help advance the framework toward more comprehensive computer-aided histopathological analysis while maintaining its intended role as a research-support tool rather than a replacement for professional pathological diagnosis.

REFERENCES

- [1] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides, "Lung and Colon Cancer Histopathological Image Dataset (LC25000)," arXiv preprint arXiv:1912.12142, 2019.
- [2] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708, 2017.
- [3] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased Boosting with Categorical Features," in Advances in Neural Information Processing Systems (NeurIPS), vol. 31, pp. 6638–6648, 2018.
- [4] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2015.
- [5] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," Journal of Machine Learning Research, vol. 15, no. 56, pp. 1929–1958, 2014.
- [6] A. Echle, N. T. Rindtorff, T. J. Brinker, T. Luedde, A. T. Pearson, and J. N. Kather, "Deep Learning in Cancer Pathology: A New Generation of Clinical Biomarkers," British Journal of Cancer, vol. 124, pp. 686–696, 2021.
- [7] J. Diosdado, P. Gilabert, S. Seguí, and H. Borrego, "LungHist700: A Dataset of Histological Images for Deep Learning in Pulmonary Pathology," Scientific Data, vol. 11, article 1088, 2024.
- [8] E. A.-R. Hamed, M. A.-M. Salem, N. L. Badr, and M. F. Tolba, "An Efficient Combination of Convolutional Neural Network and LightGBM Algorithm for Lung Cancer Histopathology Classification," Diagnostics, vol. 13, no. 15, article 2469, 2023.
- [9] A. Davri, E. Birbas, T. Kanavos, N. Ntitsos, N. Giannakeas, A. T. Tzallas, and A. Batistatou, "Deep Learning for Lung Cancer Diagnosis, Prognosis and Prediction Using Histological and Cytological Images: A Systematic Review," Cancers, vol. 15, no. 15, article 3981, 2023.
- [10] R. Parra-Medina, G. Guerron-Gomez, D. Mendivelso-González, J. H. Gil-Gómez, J. P. Alzate, M. Gomez-Suarez, J. F. Polo, J. J. Sprockel, and A. Mosquera-Zamudio, "Deep Learning in Histopathology Images for Prediction of Oncogenic Driver Molecular Alterations in Lung Cancer: A Systematic Review and Meta-Analysis," Translational Lung Cancer Research, vol. 14, no. 5, pp. 1756–1769, 2025.
- [11] A. A. Ahmed, M. Fawi, A. Brychcy, M. Abouzid, M. Witt, and E. Kaczmarek, "Development and Validation of a Deep Learning Model for Histopathological Slide Analysis in Lung Cancer Diagnosis," Cancers, vol. 16, no. 8, article 1506, 2024.



D. Suguna Seleena Kennedy

M. Tech, Department of Computer Science and Engineering, Bapatla Engineering College,
Bapatla, AP, INDIA, sugunadaggumalli@gmail.com



Dr. R. Venkata Lakshmi

Asst. Professor, Department of Computer Science and Engineering, Bapatla Engineering College,
Bapatla, AP, INDIA, Lsubramanyam671@gmail.com



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)