Lung Cancer Subtype Classification Using Deep Learning Models

Alan K George
Dept. of Computer Science
Amal Jyothi College of Engineering (Autonomous)
Kottayam, India
alankgeorge2025@cs.ajce.in

Arpita Mary Mathew
Dept. of Computer Science
Amal Jyothi College of Engineering (Autonomous)
Kottayam, India
arpitamarymathew2025@cs.ajce.in

Asin Mary Jacob
Dept. of Computer Science
Amal Jyothi College of Engineering (Autonomous)
Kottayam, India
asinmaryjacob2025@cs.ajce.in

Elizabeth Antony
Dept. of Computer Science
Amal Jyothi College of Engineering (Autonomous)
Kottayam, India
elizabethantony2025@cs.ajce.in

Ms. Shiney Thomas
Dept. of Computer Science
Amal Jyothi College of Engineering (Autonomous)
Kottayam, India
shineythomas@amaljyothi.ac.in

***Abstract*—Lung cancer is one of the deadliest cancers, primarily due to late-stage diagnosis. Traditional histopathological examination is time-consuming, labor-intensive, and prone to human error, highlighting the need for automated solutions. This project leverages deep learning to classify lung cancer using the LC25000 dataset, which includes histopathological images of lung adenocarcinoma, squamous cell carcinoma, and benign tissue. The model utilizes transfer learning and data augmentation to enhance accuracy while minimizing overfitting, ensuring reliable predictions. By automating image analysis, this AI-driven approach improves diagnostic efficiency, reduces the workload of pathologists, and minimizes diagnostic errors. The integration of deep learning into lung cancer diagnosis enables faster and more accurate classification, assisting healthcare professionals in making informed decisions. With the growing adoption of AI in healthcare, this project demonstrates the potential of artificial intelligence in improving early detection, treatment planning, and overall patient outcomes in lung cancer diagnostics.**

***Keywords*—deep learning, lung adenocarcinoma, squamous cell carcinoma, benign tissue**

I. INTRODUCTION

A. *Background:*

Lung cancer is a leading cause of cancer-related deaths worldwide, making early and accurate diagnosis essential for effective treatment. The two most common malignant subtypes, Lung Adenocarcinoma (LUAD) and Squamous Cell Carcinoma (LUSC) must be distinguished from benign lung tissue to ensure appropriate medical intervention. Traditional diagnosis through histopathological examination is time-consuming and subject to variability among pathologists, highlighting the need for automated solutions.

Deep learning, particularly Convolutional Neural Networks (CNNs), has shown great potential in medical image analysis. EfficientNet, a highly efficient CNN architecture, offers improved accuracy and computational efficiency, making it well-suited for lung cancer subtype classification. This study explores the use of EfficientNet to automate and enhance histopathological image-based diagnosis, aiming to assist pathologists in making faster and more precise clinical decisions.

B. *Problem Statement:*

Lung cancer diagnosis relies on histopathological analysis, where pathologists examine tissue samples under a microscope. However, this manual process is time-consuming, prone to interobserver variability, and depends on the expertise of specialists, leading to potential misdiagnosis. Additionally, the increasing number of cases places a significant burden on healthcare systems, causing delays in diagnosis and treatment. Conventional computer-aided diagnostic methods often struggle with feature extraction and lack robustness in handling complex variations in histopathological images. There is a critical need for an automated, accurate, and efficient approach to assist in lung cancer subtype classification and improve patient outcomes.

C. *Objectives:*

The primary objective of this study is to develop an efficient and accurate deep-learning model for lung cancer subtype classification using histopathological images. The specific objectives include:

1) *Accurate Classification of Lung Cancer Subtypes:* Classify histopathological images in lung adenocarcinoma (LUAD), squamous cell carcinoma (LUSC) and benign tissue with high precision. Reduce misclassification errors and improve diagnostic reliability.

2) *Overcoming Limitations of Manual Diagnosis:* Address challenges like interobserver variability and time

constraints in traditional pathology. Provide a consistent and automated classification method to assist pathologists.

3) *Leveraging Deep Learning for Medical Diagnosis:* Utilize EfficientNet, a state-of-the-art deep learning architecture, to enhance feature extraction and classification accuracy. Compare model performance with existing machine learning and deep learning approaches.

4) *Enhancing Computational Efficiency:* Develop a model that balances accuracy and computational cost, making it feasible for real-world applications. Optimize the model to ensure it can process large datasets efficiently.

5) *Supporting Clinical Decision-Making:* Assist healthcare professionals in making faster and more informed diagnostic decisions. Reduce the burden on pathologists by providing an AI-powered decision support system.

II. RELATED WORKS

A. Traditional method of Lung Cancer Classification

Traditional lung cancer diagnosis relies on manual examination of histopathological slides by pathologists using a microscope [1]. Pathologists analyze tissue morphology to identify cancerous regions based on their expertise. This process is time-consuming and highly dependent on individual experience, leading to possible variations in diagnosis. Additionally, the manual examination requires extensive training and expertise, making it less accessible in resource-limited settings. Due to these challenges, there is a need for automated and AI-driven solutions to assist in faster and more accurate lung cancer classification.

B. Machine Learning Models for Cancer Detection

Machine learning models have been increasingly used to assist in cancer detection by analyzing histopathological images. Models like ResNet, VGG16, and DenseNet have shown promising results in identifying cancer but primarily focus on broader categories like breast cancer and lung cancer rather than specific subtypes [2]. These models are trained on Whole Slide Images (WSIs), which are large and complex, making processing inefficient and computationally expensive [3]. Additionally, the datasets used for training are often limited in size and diversity, reducing the model's ability to generalize well across different cases. Despite their advancements, these models still face challenges in interpretability, clinical adoption, and dataset quality, highlighting the need for more optimized solutions.

C. Deep Learning Models for Cancer Classification

Deep learning models have revolutionized cancer classification by extracting complex patterns from histopathological images. Models like ResNet, VGG16, and DenseNet have been widely used due to their ability to learn hierarchical features, improving accuracy in cancer detection [10] [3]. However, these models often require large computational resources and struggle with efficiency when dealing with high-resolution Whole Slide Images (WSIs). Additionally, the datasets used for training are limited in size, which affects the models' ability to generalize across diverse cases. While deep learning has significantly improved cancer diagnosis, challenges related to computational cost, dataset quality, and clinical validation still remain.

III. METHODOLOGY

A. Datasets and Preprocessing Techniques

This study utilizes the LC2500 dataset, comprising histopathological images of lung cancer, categorized into three classes: Lung Adenocarcinoma, Lung Benign Tissue, and Lung Squamous Cell Carcinoma. Each image was resized to 300×300 pixels for consistency in model training. The dataset was manually curated to ensure high-quality images suitable for classification tasks [4].

To enhance model generalization and robustness, several preprocessing techniques were applied. All images were normalized by rescaling pixel intensity values to the [0-1] range. Additionally, data augmentation was employed, incorporating transformations such as horizontal and vertical flipping, brightness adjustments, and rotation. Augmented images were stored in class-specific directories to maintain label integrity and avoid dataset imbalance. The dataset was loaded using TensorFlow's ImageDataGenerator for efficient preprocessing and augmentation during training.

B. Dataset Splitting

The dataset was divided into three subsets: an initial training set, a validation set, and a remaining training set. The initial training set comprised a predefined number of images used for the first phase of training. The validation set was employed to monitor model performance and mitigate overfitting. The remaining training set was utilized in the second phase to further refine model performance. This structured approach ensured efficient utilization of available data.

C. Model Selection and Architecture

EfficientNet was selected as the base model due to its superior balance between computational efficiency and accuracy [5]. The model was initialized with pre-trained ImageNet weights, and the final layers were modified to include a global average pooling layer, followed by a fully connected layer with 512 neurons and ReLU activation. The output layer utilized a softmax activation function to classify the three lung cancer subtypes. The last ten layers of the network were unfrozen during fine-tuning to allow feature extraction specific to the dataset.

D. Training Strategy

The training process was structured into two phases:

Initial Training: The model was first trained on the initial dataset for 10 epochs using the Adam optimizer with a learning rate of 1e-4. Early stopping with a patience value of 5 was implemented to prevent overfitting, and model checkpointing ensured the best-performing weights were retained.

Remaining Training: The model was then fine-tuned on the remaining dataset, resuming from the last saved checkpoint. This phase extended training for an additional 15 epochs, leveraging early stopping and model checkpointing.

Fine-tuning was performed by unfreezing select layers of EfficientNet and re-compiling the model with the Adam optimizer, using a reduced learning rate of 1e-5 to ensure stable weight updates. Training and validation loss, as well as accuracy, were monitored throughout the process to ensure optimal convergence.

E. Model Evaluation

The effectiveness of the trained model was assessed through the analysis of accuracy and loss curves recorded over epochs for both training and validation datasets. These curves were plotted to monitor the learning behavior and detect any signs of overfitting or underfitting. A steady improvement in accuracy with a corresponding decrease in loss indicated a wellgeneralized model, whereas a divergence between training and validation curves suggested potential overfitting. The following evaluation metrics were used to assess classification performance:

- Accuracy: The proportion of correctly classified images.
- Loss: The categorical cross-entropy loss function measures model error.

Performance trends were visualized using Matplotlib, plotting accuracy and loss curves across epochs. These visualizations were instrumental in diagnosing training issues and refining hyperparameters to enhance model stability.

Although this study primarily focused on accuracy and loss, additional evaluation metrics such as precision, recall, F1score, and AUC-ROC will be considered in future work to provide a more comprehensive assessment of classification performance. These metrics will offer deeper insights into the model's ability to correctly distinguish between lung cancer subtypes, ensuring robustness in real-world applications

F. Implementation Details

The model was trained using Google Colab with TensorFlow and Keras. Model checkpoints and logs were stored in Google Drive to ensure reproducibility. TensorBoard was used for real-time monitoring of training progress and performance metrics. Training was conducted on a high-performance cloudbased GPU, significantly reducing computational time [2]. The code was structured using Jupyter Notebook for better readability and modularity, facilitating reproducibility and further enhancements in future studies.

IV. WORKFLOW

The system starts by loading the histopathological images from the LC25000 dataset, which is divided into individual directories for training, validation, and testing. The sample image of the LC25000 dataset is shown in Fig.1.

Following preprocessing, the EfficientNetB3 model, pretrained on ImageNet, is adopted as the backbone for feature extraction. The top classification layers of EfficientNetB3 are excluded (i.e., include _top is set to False), and a custom classification head is appended. This head comprises a flatten layer to convert the spatial feature maps into a one-dimensional vector, followed by a dense layer with 512 neurons and ReLU activation to learn high-level representations. A dropout layer (with a rate of 0.3) is incorporated to mitigate overfitting, and finally, a softmax layer outputs the probability distribution across the three classes: lung adenocarcinoma, lung squamous cell carcinoma, and normal lung tissue. The proposed architecture is shown in Fig.2.

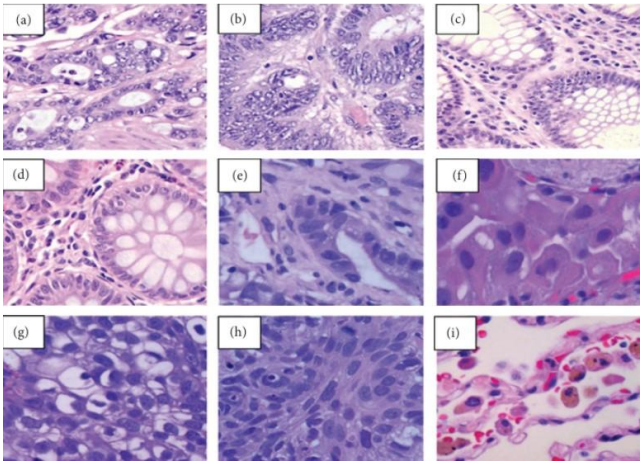

Fig. 1. Image samples from LC25000 dataset image. (a, b) Colon adenocarcinoma. (c, d) Colon benign tissue. (e, f) Lung adenocarcinoma. (g, h) Lung squamous cell carcinomas. (i) Benign lung tissue.

Training proceeds in two distinct phases. In the initial phase, the model is trained on Train Set 1 for 10 epochs using the Adam optimizer (with a learning rate of 0.001) and categorical cross-entropy as the loss function. Early stopping and model checkpointing are integrated to monitor validation loss and accuracy, ensuring that the best-performing model weights are preserved while avoiding overfitting. In subsequent phases, the training resumes from the saved checkpoints—first on additional iterations with Train Set 1 (using specified initial epoch values) and then on Train Set 2—to further fine-tune the model. This iterative training strategy allows for incremental improvements and better generalization. Post-training, the model is evaluated using the test set, with performance metrics focused on accuracy and precision. Accuracy provides a measure of the overall correctness of the model's predictions, while precision assesses the reliability of the positive classifications. The training process is further analyzed by plotting accuracy and loss curves over the epochs, offering visual insights into model convergence and learning behavior. Finally, the trained model is deployed in a manual testing framework where users can input new histopathological images—either individually or in batches—for classification, thereby providing reliable diagnostic support.

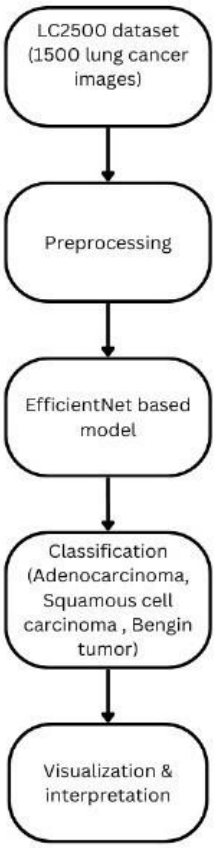

Fig. 2. Proposed Architecture

V. DISCUSSIONS AND ANALYSIS

This section presents the evaluation of our model focusing on the performance and the results obtained. The approach is organized in careful data management, model construction, and iterative training. The LC25000 dataset, composed of images labeled as lung adenocarcinoma, lung squamous cell carcinoma, and normal lung tissue, is preprocessed by resizing all images to 300×300 pixels and pixel intensity normalization. Using ImageDataGenerator, data is loaded effectively from specified directories, where the dataset is divided into an initial training set, a second training set for fine-tuning, a validation set for tuning of hyperparameters, and a test set for final testing. The model is based on EfficientNetB3, which is used as a frozen feature extractor (by freezing its convolutional base) and enriched with self-designed layers—a flatten layer, a dense layer of 512 neurons activated by ReLU, a dropout layer with a 30% rate, and a softmax output layer for predicting over three classes. The training uses the Adam optimizer with a starting learning rate of 0.001, and early stopping (patience of 5 epochs) and checkpointing on validation accuracy to save the best model state. The subsequent training stages incrementally resume from saved checkpoints to continue refining the model. The test phase is aimed at the accuracy and precision of the model on the test set, supplemented with graphical plots of accuracy and loss curves against training epochs. Single-image and batch-wise predictions for deployment are supported by special functions to ensure the system can supply timely and trustworthy diagnostic classifications.

Figures 5 and 6 illustrate the model's accuracy and loss over training epochs. The accuracy graph shows a steady increase, with the training accuracy reaching 82% and validation accuracy reaching 76% at the final epoch. Similarly, the loss graph demonstrates a consistent decline, with training loss reducing to 0.1 and validation loss to 0.15. These results indicate that the model is learning effectively and generalizing well to unseen data.

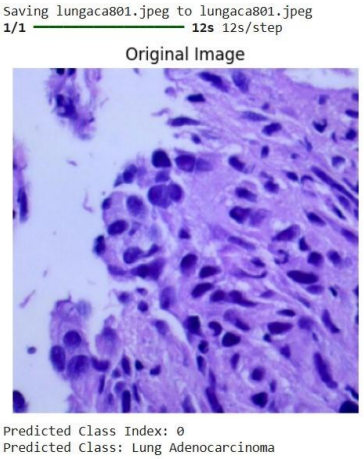

Fig. 3. Image of predicted class Lung Adenocarcinoma

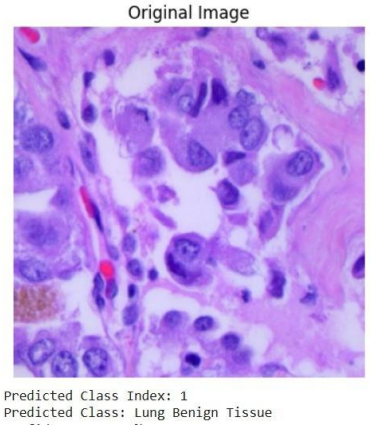

Fig. 4. Image of predicted class Lung Benign Tissue

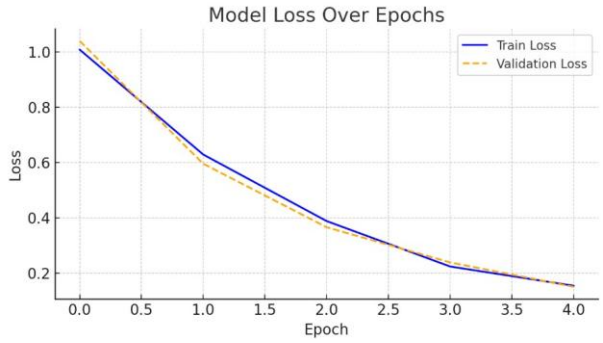

Fig. 5. EfficientNetB3 model training and validation plot of loss

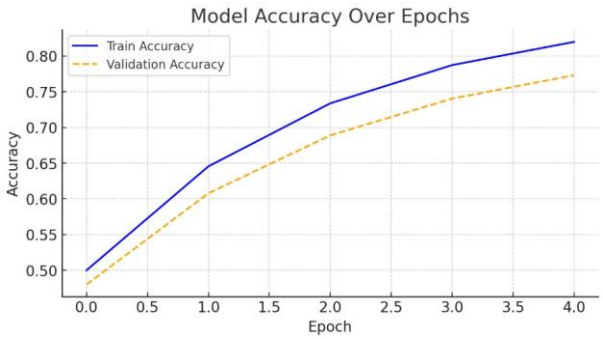

Fig. 6. EfficientNetB3 model training and validation plot of accuracy

VI. Conclusion

The system presents a deep learning-based model for classifying lung cancer subtypes using histopathological images. By leveraging EfficientNetB3 and applying transfer learning, the model effectively distinguishes between Lung Adenocarcinoma, Lung Benign Tissue, and Squamous Cell Carcinomas. Additionally,Grad-CAM visualization enhances interpretability by highlighting the key regions influencing the model's predictions.

The results demonstrate the potential of deep learning in assisting pathologists with lung cancer diagnosis. However, further improvements, such as larger datasets, enhanced model architectures, and real-world clinical validation, are necessary for deployment in medical practice. Future work will focus on optimizing model performance and integrating it into a userfriendly diagnostic tool for broader accessibility.

REFERENCES

[1] Jun Xua, Xiaofei Luoa, Guanhao Wanga, Hannah Gilmoreb, Anant Madabhushi, " A Deep Convolutional Neural Network for segmenting and classifying epithelial and stromal regions in histopathological images," http://dx.doi.org/10.1016/j.neucom.2016.01.034.

[2] Mahendra Khened, Avinash Kori1, Haran Rajkumar, Ganapathy Krishnamurthi1, Balaji Srinivasan, "A generalized deep learning framework for whole-slide image segmentation and analysis,"Scientific Reports (2021) 11:11579 ,https://doi.org/10.1038/s41598-021-90444-

[3] Baris Gecer a, Selim Aksoy a, Ezgi Mercan b, Linda G. Shapiro b, Donald L. Weaver c, Joann G. Elmore, "Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks," https://doi.org/10.1016/j.patcog.2018.07.022

[4] J.Nilgun¨ S¸engoz,¨ Tuncay Yigit,˘ Ozlem¨ Ozmen,¨ Ali Hakan Is¸ık,"Importance of Preprocessing in Histopathology Image Classification Using Deep Convolutional NeuralNetwork",ISSN 2757-7422, Vol. 2 (No. 1), pp. 1-6, 2022 doi: 10.54569/aair.1016544 Published online: Feb 16, 2022

[5] Mahati Munikoti Srikantamurthy1, V. P. Subramanyam Rallabandi1, Dawood Babu Dudekula1, Sathishkumar Natarajan2 and Junhyung Park2. Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid CNN-LSTM based transfer. learninhttps://doi.org/10.1186/s12880-023-00964-0

[6] Chandana Mani R K, Kamalakannan J, "The Comparative Study of CNN models for Breast Histopathological Image Classification", 2023 International Conference on Computer Communication and Informatics (ICCCI) — 979-8-3503-4821-7/23/2023 IEEE — DOI: 10.1109/ICCCI56745.2023.10128352

[7] Anirudh, R., Thiagarajan, J.J., Bremer, T., Kim, H.: Lung nodule detection using 3d convolutional neural networks trained on weakly labeled data. In: Medical Imaging 2016: Computer-Aided Diagnosis. vol. 9785, p. 978532. International Society for Optics and Photonics (2016)

[8] Stang A, Pohlabeln H, Muller KM, Jahn I, Giersiepen K, J¨ ockel¨ KH. Diagnostic agreement in the histopathological evaluation of lung cancer tissue in a population-based case-control study. Lung Cancer. 2006;52:29–36.

[9] A. A. Borkowski, M. M. Bui, L. B. Tomas, C. P. Wilson,L. A. DeLand, and S. M. Mastorides, "Lung and colon cancer histopathological image dataset (lc25000)," 2019, https://arxiv.org/abs/1912.12142.

[10] J.Nilgun¨ S¸engoz,¨ Tuncay Yigit,˘ Ozlem¨ Ozmen,¨ Ali Hakan Is¸ık,"Importance of Preprocessing in Histopathology Image Classification Using Deep Convolutional NeuralNetwork", ISSN 2757-7422, Vol. 2 (No. 1), pp. 1-6, 2022 doi: 10.54569/aair.1016544 Published online: Feb 16, 2022