International Journal on Emerging Research Areas (IJERA) *ISSN : 2230-9993*

DOI: 10.5281/zenodo.14651088

Classification of Lung Cancer Subtypes Using Deep Learning Model

Alan K George
Department of Computer Science and Engineering
Amal Jyothi College of Engineering
Kanjirapally, Kerala, India
alankgeorge2025@cs.ajce.in

Arpita Mary Mathew
Department of Computer Science and Engineering
Amal Jyothi College of Engineering
Kanjirapally, Kerala, India
arpitamarymathew2025@cs.ajce.in

Asin Mary Jacob
Department of Computer Science and Engineering
Amal Jyothi College of Engineering
Kanjirapally, Kerala, India
asinmaryjacob2025@cs.ajce.in

Elizabeth Antony
Department of Computer Science and Engineering
Amal Jyothi College of Engineering
Kanjirapally, Kerala, India
elizabethantony2025@cs.ajce.in

Shiney Thomas
Department of Computer Science and Engineering
Amal Jyothi College of Engineering
Kanjirapally, Kerala, India
shineythomas@amaljyothi.ac.in

***Abstract*—Cancer is a leading cause of death worldwide, affecting millions of people each year. There is an urgent need for improved cancer detection, diagnosis, and treatment methods. Histopathological examination, involving the microscopic analysis of tissue samples, is the gold standard for cancer diagnosis. However, this process can be time-consuming and subjective, relying heavily on pathologists' expertise. Deep learning models, particularly convolutional neural networks (CNNs), excel at image analysis and pattern recognition. CNNs can be trained on large datasets of histopathological images to learn the complex features associated with different cancer types. Once trained, these models can automate cancer detection, classify cancer subtypes, segment tumor regions and predict treatment response. Deep learning models, particularly convolutional neural networks (CNNs), have successfully classified various cancer subtypes. For instance, studies have shown the effectiveness of CNN, CNN Gradient Descent, VGG-16, VGG-19, Inception V3, and Resnet-50 in accurately classifying lung cancer subtypes from histopathological images. Transfer learning, a technique that adapts pre-trained CNN models to new tasks, has further enhanced classification accuracy, especially when working with limited medical image datasets. The ability to accurately classify cancer subtypes using deep learning can aid pathologists in making more informed diagnoses and guide treatment strategies. Continued research and development in this field promise to revolutionize cancer diagnosis and prognosis, leading to more personalized and effective treatment strategies.**

***Index Terms*—Deep Learning, Convolutional Neural Networks (CNNs), EfficientNet, Histopathological Images**

I. INTRODUCTION

A. General Background

Cancer stands as a formidable global health challenge, impacting countless lives and demanding continuous innovation in diagnosis and treatment. Traditional histopathological examination, while considered the gold standard, suffers from limitations such as subjectivity and time-consuming processes.

There is a profound impact of cancer on a global scale. It is a leading cause of death worldwide, with high morbidity and mortality rates associated with various forms of the disease. This stark reality underscores the urgency for advancements in cancer care, particularly in early and accurate detection. The current reliance on histopathological examination, which involves the microscopic analysis of tissue samples, has inherent drawbacks. It is a labor-intensive process, demanding significant expertise from pathologists, and can be prone to interobserver variability. Moreover, manual analysis of the vast and intricate details present in tissue samples can be time-consuming, delaying diagnosis and treatment decisions.

To address this, deep learning, a subset of artificial intelligencecan be used as a transformative technology in cancer diagnostics. Deep learning models, especially convolutional neural networks (CNNs), are exceptionally well-suited for image analysis and pattern recognition tasks. These models can be trained on massive datasets of histopathological images to learn the intricate features that distinguish cancerous tissues from healthy ones, and to differentiate between various cancer subtypes.

DOI: 10.5281/zenodo.14651088

B. Objectives

To improve cancer detection and classification using deep learning techniques and emphasize the significant role of deep learning in analyzing histopathological images, aiming to overcome the limitations of traditional methods that rely on manual feature extraction and are prone to human error. Deep learning, with its ability to automatically discern intri- cate patterns and features from raw image data, presents a transformative approach to cancer diagnosis. This approach is particularly significant given the challenges associated with accurately and efficiently analyzing the vast amounts of data inherent in histopathological images. The aim is to explore the application of diverse deep learning architectures, including convolutional neural networks (CNNs), which are particularly adept at processing image data. The focus is on developing robust and reliable algorithms that can achieve high accuracy in detecting and classifying various cancer types, including breast cancer, lung cancer, and colon cancer. The overarching goal is to create powerful tools that augment pathologists' capabilities, enabling them to make more informed decisions and ultimately contributing to enhanced patient care and outcomes.

C. Scope

The goal is to develop tools that can assist pathologists in making more informed decisions, ultimately leading to better patient outcomes and to explore the application of various deep learning architectures, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid models like CNN-LSTM.

II. Related Works

The paper presents a substantial body of work focusing on utilizing deep learning for cancer detection and classification, leveraging histopathological images. A significant portion of this research centers around employing convolutional neural networks (CNNs) for breast cancer tissue classification. [1] Transfer learning, a technique where models pre-trained on massive datasets like ImageNet are adapted for the specific task, has proven highly effective in this context. Researchers have also explored data augmentation techniques, which artificially expand the training dataset by generating variations of existing images, to further enhance model performance. The relative performance of different CNN architectures, such as VGG-16, ResNet-50, and AlexNet, has been a subject of investigation in several studies. [2]

Beyond simply using these architectures as black boxes, researchers have explored how different deep learning method- ologies can be applied to various tasks in histopathology image analysis. The automatic localization of diagnostically relevant regions of interest (ROIs) in WSIs has been addressed in several studies. [3] [4] These frameworks frequently utilize multiple fully convolutional networks (FCNs), trained to emulate the decision-making processes of pathologists viewing images at varying magnifications. For distinguishing between subtypes of carcinoma, techniques such as Gaussian-based hierarchical voting and repulsive balloon models have been employed to delineate cells within the tissue. [5]

The choice of features used to represent the histopathology images significantly impacts the performance of classification models. The sources highlight the evolution of feature extrac- tion techniques, from handcrafted features to those learned by deep neural networks. [6] [7]Earlier methods relied heavily on manually engineered, low-level image features, such as color, texture, and local binary patterns (LBP). These features were typically used in conjunction with traditional machine learning classifiers like support vector machines (SVMs). [6] However, the advent of deep learning has enabled the automatic extraction of more intricate and informative features directly from the raw pixel intensity values of the images. This has led to significant improvements in classification accuracy. For example, one study showed that a deep CNN model outperformed three methods based on handcrafted features in the task of classifying epithelial and stromal regions in both breast and colorectal cancer images. [7]

The sources also acknowledge the unique challenges inherent in analyzing WSIs, particularly their massive size. To handle these computationally demanding images, researchers have de- veloped efficient processing strategies, often involving dividing the WSIs into smaller, overlapping patches. [9] Another critical issue is the potential for class imbalance within the datasets, where one type of tissue might be significantly overrepresented compared to others. To address this, techniques like ensemble segmentation models have been proposed, which combine the predictions of multiple models trained on different subsets of the data. The sources also stress the importance of incorporating uncertainty estimation frameworks into the analysis pipeline. These frameworks provide a measure of confidence in the model's predictions, which is crucial for practical applications. [9] [10]

III. Proposed System

A. General Background

In this study, deep learning-based systems are proposed to analyze histopathological images, particularly in the context of cancer detection and classification. These systems often leverage Convolutional Neural Networks (CNNs), sometimes combined with other techniques like Long Short-Term Mem- ory (LSTM) networks or transfer learning, to extract features from images and perform classification. The proposed systems aim to assist pathologists in tasks such as identifying tumor regions, classifying cancer subtypes, and assessing the severity of cancer. The systems are trained and validated on datasets of histopathological images, and their performance is evaluated using metrics such as accuracy, sensitivity, and specificity. The ultimate goal of these systems is to improve the efficiency and accuracy of cancer diagnosis and contribute to better patient outcomes.

DOI: 10.5281/zenodo.14651088

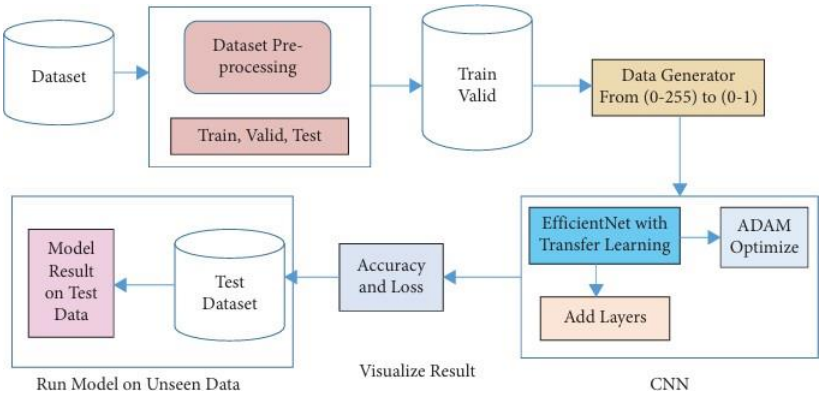

Fig. 1. Block diagram of proposed work

B. Images Dataset Acquisition

The dataset used in this research is the LC25000 Lung and Colon Cancer Histopathological Image dataset, which contains a total of 25,000 JPEG images, each sized at 768x768 pixels. This dataset is organized into two main folders: one for colon cancer, with 10,000 images, and one for lung cancer, with 15,000 images. The lung cancer folder is further divided into three subfolders, representing two types of lung cancer and benign lung tissue images. [11]

In histopathology images, cancerous and non-cancerous tissues are identified as follows: 1.Malignant tissue that are characterized by a darker appearance and abnormal nuclear growth compared to normal tissue. 2.Benign tissue that shows normal tissue growth and appears lighter in color. To expand the dataset, three augmentation techniques are applied to the images: Random rotation by up to 25% in both directions, Ran- dom noise addition to simulate variability, Horizontal flipping of the images.

These augmentations increase the lung cancer dataset to 15,000 images, with equal representation of five classes: lung adenocarcinoma, lung squamous cell carcinoma, benign lung tissue, colon adenocarcinoma, and benign colon tissue, each with 5,000 images. This results in a balanced dataset of 10,000 colon images (5,000 per class) and 15,000 lung images.

C. Preprocessing

In the preprocessing phase, it's essential to standardize the image size for optimal performance of the CNN model. Initially, all images are sized at 768×768 pixels. Depending on the EfficientNet model variant, the image resolution is adjusted accordingly. [12] The entire dataset is split into training, validation, and testing sets. To minimize overfitting—which can occur if there is insufficient data for the model to accurately learn class distinctions—we prioritize training data. The training set contains 25,000 images, evenly distributed across five classes: colon adenocarcinoma, benign colon tissue, lung adenocarcinoma, benign lung tissue, and lung squamous cell carcinoma. Additionally, 5,000 images are allocated for validation and another 5,000 for testing, with equal class distribution in each. In the final step of preprocessing, it is ensured that all images are labeled consistently within their respective subfolders. For example,lung images are labeled as "lungaca1" for adenocarcinoma, "lungn1" for benign, and "lungscc1" for squamous cell carcinoma. This systematic labeling aids the ImageDataGenerator in the training phase, allowing it to accurately assign labels for model learning.

D. Proposed CNN model

Convolutional neural networks (CNNs) are a promising method for classifying and analyzing histopathology images. CNNs are a type of deep learning algorithm that can learn features from data in a data-driven fashion without the need for handcrafted feature extraction. This makes them well-suited for analyzing complex and information-rich histopathology images. We employ transfer learning with the ImageNet dataset, fine-tuning the final layers of each model variant to enhance accuracy and performance. The dataset is split into three parts: 80% for training, 10% for validation, and 10% for testing.For EfficientNet B0, images are resized to 224×224 pixels, with similar resizing applied to other models. After preprocessing and training, model performance is then evaluated.

E. EfficientNet

EfficientNet models utilize a simple yet powerful compound scaling strategy. Developed by Google in 2019, these neural network architectures are optimized to maximize accuracy while minimizing computational cost, making them particularly effective for classification tasks. EfficientNets outperform other architectures, such as DenseNet, Inception, and ResNet, on the ImageNet benchmark, and they also run more efficiently. This approach allows for scaling up a ConvNet model to meet any target resource constraints while preserving model efficiency, a benefit that is especially useful in transfer learning scenarios. EfficientNet balances three dimensions—network width (number of filters per layer), depth (number of layers), and resolution (image height and width). This balance is achieved by proportionally scaling each dimension with fixed scaling coefficients,

DOI: 10.5281/zenodo.14651088

unlike traditional models that scale only one dimension at a time. This compound scaling approach results in a streamlined and effective model structure.

IV. Result Analysis

In evaluating the proposed CNN model for lung and colon cancer classification, the study uses multiple EfficientNet variants (B0 to B7) to identify the optimal balance of accuracy and computational efficiency. Each variant was tested for accuracy and loss during training, validation, and testing phases using a dataset of 25,000 histopathology images across five classes: lung adenocarcinoma, lung squamous cell carcinoma, benign lung tissue, colon adenocarcinoma, and benign colon tissue. The training was conducted for 100 epochs, and images were resized according to each model's requirements, with EfficientNet B0 using 224x224 pixels and B7 using 600x600 pixels. Among the variants, EfficientNet B2 achieved the highest classification accuracy at 97% with an image resolution of 260x260 pixels, demonstrating effective accuracy with a minimal loss rate of 0.07. While higher variants like B6 and B7 also performed well in accuracy, they required longer training times and higher computational resources, resulting in diminishing returns on efficiency compared to B2. The training times varied significantly across models, with lower variants (B0-B2) completing within minutes, while the higher models (B3 and above) required hours on Google Colab's GPU setup, indicating that resource allocation impacts model selection. The study also addressed the issue of class imbalance by ensuring each class contained 5,000 images, thus maintaining a balanced dataset that reduced overfit- ting risks. Techniques like batch normalization and dropout were applied to optimize training stability, and softmax activation was employed for multi-class classification.

Overall, the results validate the proposed method's robustness in histopathological image classification, with EfficientNet B2 being the recommended variant for its balance of speed, computational efficiency, and accuracy. The study highlights the potential for further improvement by expanding the dataset and increasing the number of classes, which could further enhance model performance and application to broader diagnostic scenarios.

V. Future Scope

Future research on lung and colon cancer classification in histopathology images could focus on several key areas to enhance model performance and applicability. Expanding the dataset with a broader variety of histopathological images, including other cancer types and additional benign tissues, would improve the model's generalization. Additionally, ex- ploring higher image resolutions or adapting multi-scale ap- proaches may reveal more detailed features, potentially increas- ing classification accuracy. It also involves applying advanced augmentation techniques and using generative models to synthetically increase dataset size may help alleviate overfitting and class imbalance issues further. Investigating other deep

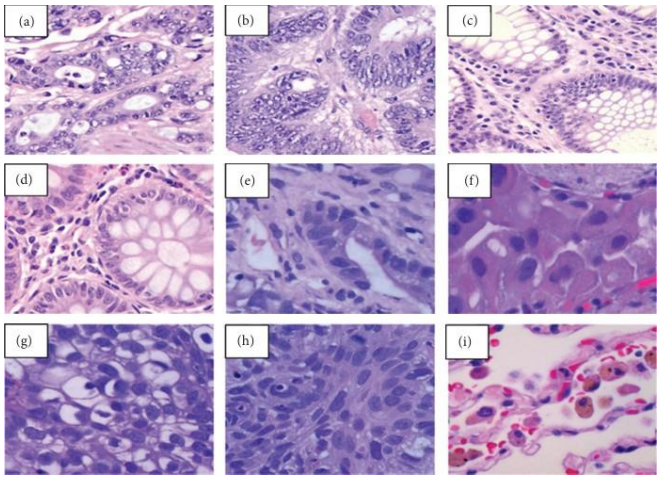

Fig. 2. Image samples from LC25000 dataset image. (a, b) Colon adenocarcinoma. (c, d) Colon benign tissue. (e, f) Lung adenocarcinoma. (g, h) Lung squamous cell carcinomas. (i) Benign lung tissue.

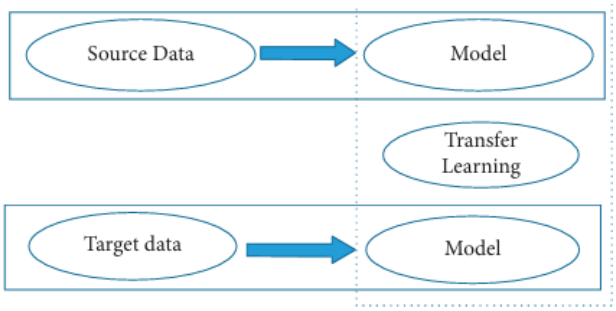

Fig. 3. Transfer learning schematic diagram

learning architectures, such as Transformer models or hybrid CNN-transformer networks, could provide new insights into handling complex histopathological patterns. Furthermore, studies like that of Lis Jose et al. [13], which demonstrated the effectiveness of hybrid machine learning models for lung disease detection using chest X-ray images, could inspire novel methods for combining imaging modalities to enhance diagnostic accuracy.

VI. Conclusion

Deep learning is transforming cancer histopathology image analysis, moving from traditional methods to a more computational approach. Deep learning models, particularly Convolutional Neural Networks (CNNs), are proving to be highly effective for classifying benign and malignant cancer subtypes using histopathology images. CNNs automatically learn and extract complex features from raw pixel data, requiring minimal preprocessing. They learn the entire process from input image to output classification, leading to greater accuracy and efficiency in diagnosis. The use of transfer learning, which leverages pretrained models on large datasets, further enhances their accuracy and ability to generalize to unseen data. Researchers are employing techniques like ensemble learning, combining predictions

DOI: 10.5281/zenodo.14651088

from multiple CNNs to improve overall performance. Deep learning models have achieved remarkable results in accurately identifying and classifying tumors, even those that are small or exhibit subtle visual differences, demonstrating their potential to aid pathologists and improve patient care.

REFERENCES

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

[2] Chandana Mani R K, Kamalakannan J, "The Comparative Study of CNN models for Breast Histopathological Image Classification", 2023 International Conference on Computer Communication and Informat- ics (ICCCI) — 979-8-3503-4821-7/23/2023 IEEE — DOI: 10.1109/IC-CCI56745.2023.10128352

[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] Jun Xua, Xiaofei Luoa, Guanhao Wanga, Hannah Gilmoreb, Anant Mad-abhushi, " 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.

[5] Mahendra Khened, Avinash Kori1, Haran Rajkumar, Ganapathy Krishna-murthi1, 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-

[6] Vani Rajasekar, V. Rangaraaj, M.P. Vaishnnave, S. Premkumar, Vellian-giri Sarveshwaran,"Lung cancer disease prediction with CT scan and histopathological images feature analysis using deep learning techniques ", https://doi.org/10.1016/j.rineng.2023.101111

[7] Stang A, Pohlabeln H, Mu¨ller KM, Jahn I, Giersiepen K, Jo¨ckel KH. Diag-nostic agreement in the histopathological evaluation of lung cancer tissue in a population-based case-control study. Lung Cancer. 2006;52:29–36.

[8] J.Nilgu¨n S¸ engo¨z, Tuncay Yig˘it, O¨ zlem O¨ zmen, Ali Hakan Is¸ık,"Importance of Preprocessing in Histopathology Image Classification Using Deep Convolutional Neural Network", ISSN 2757-7422, Vol. 2 (No. 1), pp. 1-6, 2022 doi: 10.54569/aair.1016544 Published online: Feb 16, 2022

[9] 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)

[10] Min Li, Ziwei Yan, Xiaojian Ma, Chenchen, Chengchen, Yushuai Yuan, Shuailei Zhang, Fangfang Chen, Yujie Bai, Panyunzhou, Xiaoyi Lv And Mingrui Ma, "Research on the Auxiliary Classification and Diagnosis of Lung Cancer Subtypes Based on Histopathological Images"Received March 14, 2021, accepted March 25, 2021, date of publication April 5, 2021, date of current version April 13, 2021.Digital Object Identifier 10.1109/ACCESS.2021.3071057

[11] 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.

[12] Q. Dong, S. Gong, and X. Zhu, "Imbalanced deep learning by minority class incremental rectifcation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 6, pp. 1367–1381, 2019.\

[13] Lis Jose, Akhil Lorence, Akhil Manohar, Amal Jose Chacko, Arjun J, "Lung Disease Detection from Chest X-ray Images Using Hybrid Machine Learning Model", International Journal on Emerging Research Areas (ISSN:2230-9993), vol.04, issue 01, 2024 doi: 10.5281 /zenodo.125 25268

DOI: 10.5281/zenodo.14651088