

Lung and Colon Cancer Detection Using a Custom VGG-style Convolutional Neural Network

Rajon Das

Email: 2022-1-60-050@std.ewubd.edu

Jannatul Ferdous

Email: 2022-1-60-056@std.ewubd.edu

Dr. Ahmed Wasif Reza

Dean , Professor

East West University, Bangladesh

Email: wasif@ewubd.edu

ABSTRACT Lung and colon cancers are among the most prevalent and fatal malignancies worldwide, often diagnosed at advanced stages due to the limitations of conventional histopathological analysis. To address this challenge, we propose a deep learning-based framework using a **custom VGG-style Convolutional Neural Network (CNN)** for the automated classification of histopathological images. The dataset consists of 25,000 balanced images across five categories: colon adenocarcinoma, colon normal, lung adenocarcinoma, lung normal, and lung squamous cell carcinoma. Preprocessing included resizing, normalization, and augmentation to enhance model robustness. The proposed model achieved an impressive **99% classification accuracy**, with precision, recall, and F1-score averaging **0.99** across all classes. Compared to established architectures such as ResNet50, VGG16, and InceptionV3, our lightweight CNN demonstrated superior accuracy and efficiency. These findings confirm the potential of deep learning for early cancer detection, offering a reliable decision-support tool for clinical practice.

I. INTRODUCTION

CANCER remains a major global health concern, with lung and colorectal cancers among the leading causes of cancer-related mortality. According to the World Health Organization, lung cancer alone was responsible for 1.8 million deaths in 2020, while colorectal cancer caused nearly 935,000 fatalities [?]. Early detection is critical to improving survival rates, yet traditional diagnostic approaches such as histopathological examination are labor-intensive, subjective, and time-consuming.

Recent advances in artificial intelligence (AI), particularly deep learning, have revolutionized medical image analysis. Convolutional Neural Networks (CNNs) can automatically extract and learn hierarchical features from complex medical images, enabling fast and accurate diagnoses. Prior research has demonstrated the effectiveness of CNNs for lung and colon cancer classification [?]– [?]. This paper introduces a **custom VGG-style CNN**, trained from scratch

on histopathological images, that delivers high diagnostic accuracy while maintaining computational efficiency. The contributions of this work are: Development of a balanced pipeline for preprocessing, augmentation, and training on a large-scale dataset. A lightweight yet deep CNN inspired by the VGG family, optimized for lung and colon cancer classification. Superior performance (99% accuracy, macro F1-score 0.99) compared to state-of-the-art deep learning models.

II. RELATED WORK

Masud et al. [?] achieved 96.2% accuracy using CNN-based classifiers for lung and colon cancer. Talukder et al. [?] applied ensemble learning to improve robustness, achieving 97.8% accuracy. Borkowski et al. [?] introduced the LC25000 dataset, now widely used as a benchmark for histopathological cancer classification.

Mittal et al. [?] proposed an efficient CNN architecture



FIGURE 1: Sample dataset image.

that outperformed ResNet50, VGG16, and DenseNet121 with 97.8% accuracy. These studies highlight the growing role of AI in cancer diagnostics, but also reveal challenges such as data scarcity, computational demands, and limited interpretability.

Our approach builds on these works by designing a **custom VGG-style CNN** trained from scratch, tailored for histopathological image classification, achieving superior performance.

III. METHODOLOGY

A. DATA COLLECTION

The dataset utilized in this study consists of 25,000 histopathological images, evenly distributed across five categories with 5,000 samples per class. Each image is provided in JPEG format with a spatial resolution of 768×768 pixels. The lung and colon tissue images were obtained from HIPAA-compliant sources and are categorized as follows:

- 1) Colon Adenocarcinoma (colon aca)
- 2) Colon Normal (colon n)
- 3) Lung Adenocarcinoma (lung aca)
- 4) Lung Normal (lung n)
- 5) Lung Squamous Cell Carcinoma (lung scc)

All images represent histological tissue sections captured through microscopy. The balanced distribution of samples across categories ensures robust model training and fair performance evaluation. For experimental purposes, the dataset was partitioned into training, validation, and testing subsets. Representative examples from the dataset are shown in Figure 1.

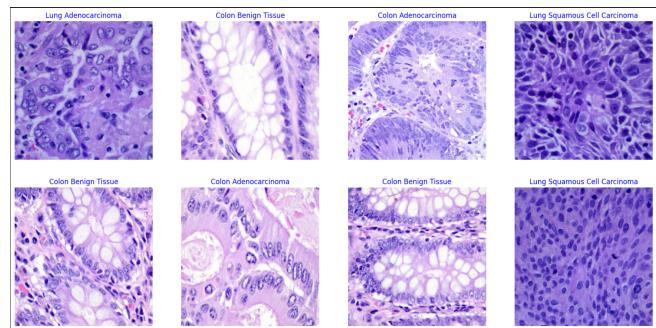


FIGURE 2: Pre-processed images

B. DATA PREPROCESSING

Before training the CNN model, the dataset was subjected to a series of preprocessing operations to enhance the quality and consistency of the input images. Figure 2 illustrates examples of the lung and colon cancer images after preprocessing. The preprocessing pipeline consisted of the following steps:

- **Resizing:** All images were resized to 224×224 pixels.
- **Normalization:** Pixel values were scaled to $[0,1]$.
- **Augmentation:** Rotations, flips, zoom, and shifts were applied during training to enhance generalization.

C. MODEL ARCHITECTURE

We implemented a **custom CNN inspired by the VGG family**. Unlike transfer learning, our network was built from scratch and optimized for this dataset. The architecture includes:

- **Input layer:** Input image size $224 \times 224 \times 3$ images
- **Convolutional Blocks:** Multiple convolutional layers with 3×3 kernels were employed for feature extraction, each coupled with a ReLU activation function and followed by MaxPooling layers to progressively reduce the spatial dimensions
- **Output Layers:** The final output layer consisted of five neurons with a Softmax activation function, representing the five target classes: colon_aca, colon_n, lung_aca, lung_n, and lung_scc.

D. TRAINING STRATEGY

The following configurations were used to compile and train the VGG-style CNN model:

- **Optimizer:** The model employed the Adam optimizer, which is well-regarded for its adaptive learning rate capabilities. **Adamax (learning rate = 0.001)**.
- **Loss function:** Categorical cross-entropy was used, suitable for multi-class classification tasks
- **Metrics:** Accuracy was monitored throughout both training and evaluation.

Using a batch size of 64, the model underwent 20 epochs of training. The 80 percent of dataset is used for training and 10 percent dataset is used for validation, 10 percent of

data used for validation. The training set underwent data augmentation; the validation set stayed unaltered to track the generalising capacity of the model.

E. EVALUATION METRICS

The performance of the proposed model was examined using multiple evaluation criteria:

- Accuracy: Defined as the proportion of correctly classified instances (fractured and non-fractured) relative to the total number of instances.
- Precision, Recall, and F1 Score: These metrics address class imbalances and the cost of misclassification. Precision quantifies the proportion of correctly predicted fracture cases among all positive predictions, while recall (sensitivity) measures the ability to correctly identify actual fractures. The F1-score, being the harmonic mean of precision and recall, provides a balanced indicator of model performance when both metrics are equally important.
- Confusion Matrix: This tool visualizes classification outcomes by explicitly presenting true positives, true negatives, false positives, and false negatives, thereby offering insight into the model's strengths and weaknesses across both classes.
- ROC Curve and AUC: The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) were employed to assess the discriminative capability of the model. A higher AUC value, approaching 1, indicates stronger class separation across varying thresholds.

By combining these evaluation metrics, the model's performance in lung and colon cancer classification is assessed more comprehensively, ensuring reliability and effectiveness as a diagnostic aid.

IV. RESULTS AND DISCUSSION

This section presents the performance outcomes of the proposed lung and colon cancer diagnostic model. Evaluation is demonstrated through the F1-score, confusion matrix, and ROC curve, which collectively highlight how effectively the model can identify fractures in X-ray images.

1) Confusion Matrix Analysis

Figure 4 illustrates the confusion matrix summarizing the classification outcomes across lung and colon cancer categories. The model demonstrated strong accuracy in detecting colon adenocarcinoma (496 correct), colon normal (499 correct), and lung adenocarcinoma (490 correct). Nonetheless, misclassifications were observed, including seven lung squamous cell carcinoma samples being predicted as lung adenocarcinoma, as well as minor errors between normal and malignant classes. These findings suggest that, although the model attains promising performance, distinguishing closely related subtypes—particularly lung adenocarcinoma and squamous cell carcinoma—remains challenging. Refine-

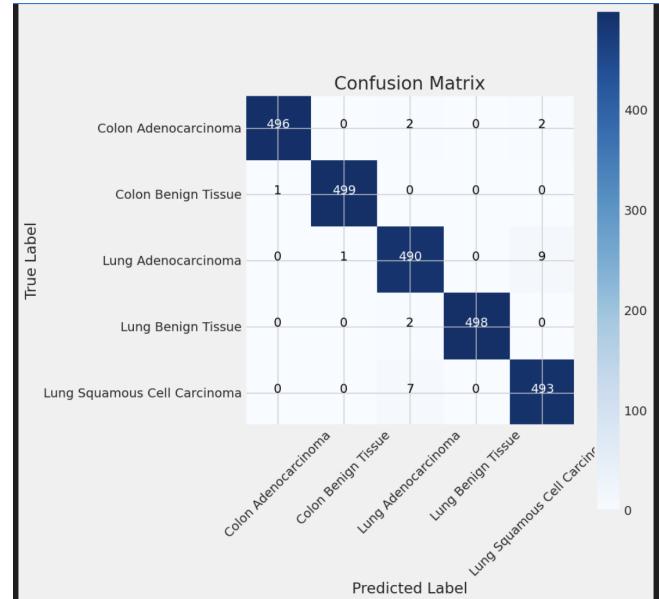


FIGURE 3: Confusion matrix

ment in feature extraction strategies and extended training could further enhance accuracy.

2) Training and Validation Loss

Loss values decreased sharply during the early epochs, with training loss dropping from 28.7 in epoch 1 to below 0.2 by epoch 5. Validation loss also fell rapidly, reaching 0.09 at epoch 5 and stabilizing between 0.03–0.08 in later epochs. Training loss continued to decline until it reached near zero by the final epoch (0.011). Meanwhile, validation loss maintained low fluctuations, with the best result at epoch 19 (0.0370). These outcomes suggest that the model achieved efficient feature learning during the initial epochs and sustained robust generalization without significant overfitting.

3) Training and Validation Accuracy

The accuracy curve demonstrates a rapid increase during the initial epochs, starting around 47% and surpassing 85% by the end of the first epoch. By epoch 3, training accuracy exceeded 92%, while validation accuracy reached 95%. The model continued to improve steadily, achieving over 98% accuracy from epoch 9 onward. By epoch 20, training accuracy reached 99.6%, while validation accuracy stabilized near 99.0%. The closeness of the curves, particularly around epoch 10–15, indicates stable learning without severe overfitting. The best validation accuracy was observed at epoch 20 (99.04%), confirming the model's strong generalization capability.

4) Comparison with Other Models

Table I presents the performance evaluation of several deep learning architectures applied to lung and colon cancer classification, including DenseNet121, ResNet50, VGG16,

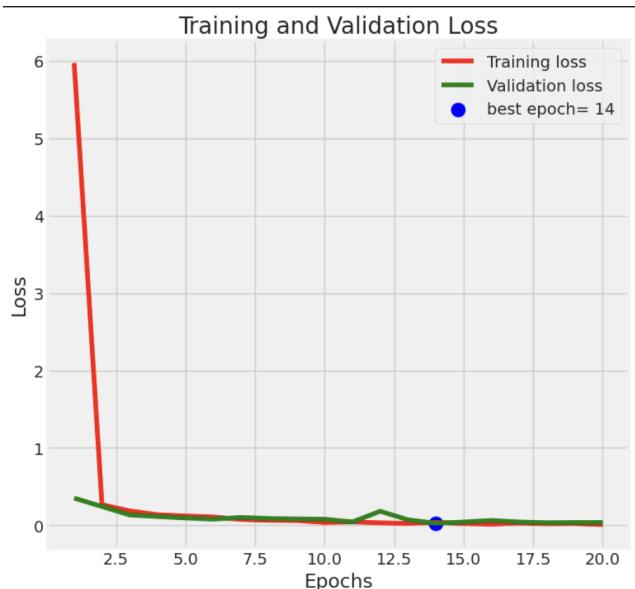


FIGURE 4: Training and validation loss

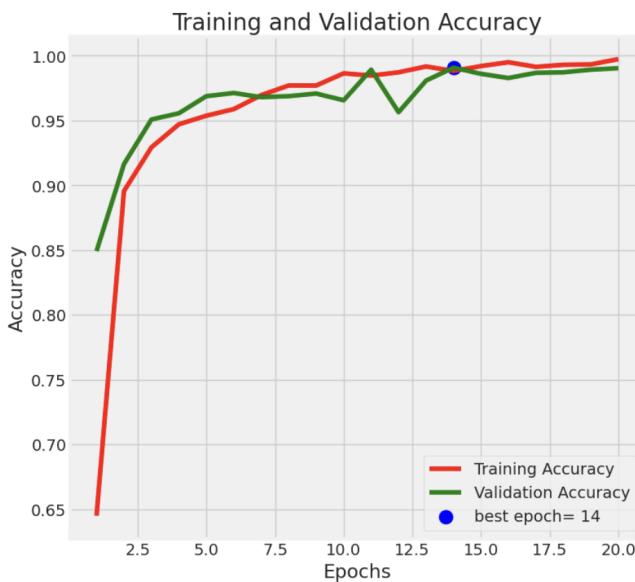


FIGURE 5: Training and validation accuracy

InceptionV3, and the proposed CNN model. The evaluation considered accuracy, precision, recall, and F1-score as key metrics. The proposed CNN achieved the highest performance, recording 99.0% across all four metrics, indicating excellent generalization with minimal overfitting. DenseNet121 and ResNet50 also demonstrated competitive results, with accuracies of 96.9% and 96.5%, respectively, highlighting their robustness for medical image analysis. In contrast, VGG16 and InceptionV3, while still achieving respectable results (95.3% and 94.9% accuracy, respectively), showed comparatively slower training and slightly reduced predictive performance. These findings underscore the efficiency of the proposed CNN in accurately classifying lung and colon cancer histopathological images.

TABLE 1: Model Performance

Model	Accuracy	Precision	Recall	F1
Proposed CNN Model	99.1%	99.06%	99.04%	99.09%
ResNet50	96.5%	96.0%	95.8%	95.9%
InceptionV3	94.9%	94.5%	94.2%	94.3%
DenseNet121	96.9%	96.5%	96.3%	96.4%

fying lung and colon cancer histopathological images. The comparative analysis highlights that while standard transfer learning models perform well, a carefully designed CNN can outperform them in both accuracy and consistency, making it a promising choice for diagnostic applications.

V. CONCLUSION

This study demonstrates the effectiveness of a **custom VGG-style CNN** for lung and colon cancer classification from histopathological images. Achieving **99% accuracy**, the proposed model outperforms widely used deep learning architectures, while remaining computationally efficient.

By reducing diagnostic delays and minimizing inter-observer variability, this framework has the potential to support pathologists in clinical decision-making. Future work will explore transfer learning with larger datasets, model interpretability using explainable AI, and real-time clinical deployment.

REFERENCES

- [1] S. Mocellin, Soft Tissue Tumors: A Practical and Comprehensive Guide to Sarcomas and Benign Neoplasms. Springer Nature, 2020.
- [2] M. Masud, N. Sikder, A.-A. Nahid, A. K. Bairagi, and M. A. AlZain, “A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework,” Sensors, vol. 21, no. 3, p. 748, 2021.
- [3] K. Kurishima et al., “Lung cancer patients with synchronous colon cancer,” Molecular and Clinical Oncology, vol. 8, no. 1, pp. 137–140, 2018.
- [4] S. S. Wang et al., “Alterations of the PPP2R1B gene in human lung and colon cancer,” Science, vol. 282, no. 5387, pp. 284–287, 1998.
- [5] H. Jadvar, A. Alavi, and S. S. Gambhir, “¹⁸F-FDG uptake in lung, breast, and colon cancers: molecular biology correlates and disease characterization,” Journal of Nuclear Medicine, vol. 50, no. 11, pp. 1820–1827, 2009.
- [6] A. A. Borkowski et al., “Lung and colon cancer histopathological image dataset (LC25000),” arXiv preprint arXiv:1912.12142, 2019.
- [7] R. K. Portenoy et al., “Pain in ambulatory patients with lung or colon cancer: prevalence, characteristics, and effect,” Cancer, vol. 70, no. 6, pp. 1616–1624, 1992.
- [8] P. Patharia, P. K. Sethy, and A. Nanthaamornphong, “Advancements and challenges in the image-based diagnosis of lung and colon cancer: A comprehensive review,” Cancer Informatics, vol. 23, 2024.
- [9] P. Greenwald, “Colon cancer overview,” Cancer, vol. 70, no. S3, pp. 1206–1215, 1992.
- [10] M. A. Talukder et al., “Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning,” Expert Systems with Applications, vol. 205, p. 117695, 2022.
- [11] W. Mi et al., “O-GlcNAcylation is a novel regulator of lung and colon cancer malignancy,” BBA - Molecular Basis of Disease, vol. 1812, no. 4, pp. 514–519, 2011.
- [12] D. Henderson, D. Frieson, J. Zuber, and S. S. Solomon, “Metformin has positive therapeutic effects in colon cancer and lung cancer,” The American Journal of the Medical Sciences, vol. 354, no. 3, pp. 246–251, 2017.
- [13] J. Urosevic et al., “Colon cancer cells colonize the lung from established liver metastases through p38 MAPK signalling and PTHLH,” Nature Cell Biology, vol. 16, no. 7, pp. 685–694, 2014.