**Lung and Colon Cancer Detection using Histopathological Images**

Seanmark Paz, Yaakov Sternberg, Jabali Shah

AAI 521: Applied Computer Vision for AI

Prof. Siamak Aram

December 12th, 2024

# 

[**Introduction 3**](#_x4jo0p64gw2o)

[**EDA and Pre-Processing 4**](#_kx4dicwj8vvc)

[Exploratory Data Analysis 4](#_5tadyh2wurbp)

[Pre-Processing Steps 4](#_duxixfmom677)

[**Modeling Methods 5**](#_wsdtr6vgfwb0)

[Model Architecture 5](#_kyoxc7xkbk6m)

[Alternatives Considered 5](#_3pkzjn8a4v5f)

[Training and Validation 6](#_5t6sj68raq3)

[**Validation, Performance Metrics, and Results 7**](#_cls03huco9i)

[Validation Techniques and Metrics 7](#_6ur2zbwkptv3)

[Results and Findings 7](#_5nkx1lolj9oi)

[Challenges 9](#_nhxklmsnie05)

[**Conclusion 10**](#_ppfp9mfar2i3)

[Key Takeaways 10](#_d4wzsjomhnys)

[**References 11**](#_5p6nqclc9pda)

# 

# Introduction

Cancer detection through histopathological image analysis presents complex diagnostic challenges, particularly for lung and colon malignancies. While computational approaches have shown promising potential, significant gaps remain in automated tissue classification methodologies. This project develops a machine learning model designed to classify histopathological images across five critical tissue categories: benign lung tissue, benign colon tissue, lung adenocarcinoma, lung squamous cell carcinoma and colon adenocarcinoma.

The motivation stems from ongoing limitations in current diagnostic practices. Manual examination of tissue slides remains time-intensive and susceptible to inter-observer variability, creating substantial clinical bottlenecks. By using advanced computer vision, this project aims to streamline cancer detection by building the model that seeks to provide a supportive diagnostic tool for two of the most common cancer types across the world, potentially saving lives and enhancing diagnostic accuracy.

The dataset has 25,000 histopathological images, each 768 x 768 pixels in JPEG format, distributed equally across five classes (Bilgin et al., 2022). These images were augmented from an original set of 750 lung tissue and 500 colon tissue samples using the Augmentor package (Bloice et al., 2017). The dataset is HIPAA compliant and validated by experts, ensuring its quality for medical research (Borkowski et al., 2019).

# 

# EDA and Pre-Processing

## Exploratory Data Analysis

Exploratory analysis confirmed a balanced dataset with 5,000 images per class, this is a methodological safeguard against the class imbalances that often plagues medical imaging datasets. Visual inspections revealed distinct differences between benign and cancerous tissues, benign tissues presented orderly cellular arrangements, while adenocarcinoma samples told a different story—irregular cell boundaries, chaotic architectural patterns that hinted at underlying pathological processes.

## Pre-Processing Steps

The dataset was preprocessed using PyTorch transforms (Paszke et al., 2019) to transform raw image data into a robust, learnable format . This included resizing images to 224 x 224 pixels, random horizontal flipping, random rotation up to 15°, and normalization with mean and standard deviation values of [0.5, 0.5, 0.5]. This approach enhances model generalizability. Each transformation helped with standardizing input, preventing overfitting, and simulating the natural variability of biological imaging.

The dataset was split into training (70%), validation (15%), and testing (15%) sets using the PyTorch random\_split function. This ensured reproducibility and balanced allocation across all classes. Dataloaders were then created to handle batching and shuffling during training and evaluation. These steps ensured that the dataset was well-prepared for deep learning without requiring manual feature extraction.

# 

# Modeling Methods

## Model Architecture

The project utilized a pre-trained MobileNetV2 architecture (Sandler, 2018), strategically modified for five-class cancer tissue classification. The final fully connected layer was replaced with a dropout layer (p=0.5) and a dense layer with softmax activation (Srivastava et al., 2014). This lightweight model was chosen for its efficiency and strong performance on smaller datasets. Cross-entropy loss was used as the criterion, and the AdamW optimizer was applied with a learning rate of 1e-4 and weight decay for regularization.

The model training was conducted on a GPU for faster computation, with early stopping implemented to prevent overfitting. Early stopping was triggered if validation accuracy did not improve for two consecutive epochs.

## Alternatives Considered

Multiple pre-trained architectures were evaluated during the initial exploration. ResNet and EfficientNet presented compelling capabilities in medical imaging (He et al., 2016; Tan & Le, 2019). While these models offer significant advantages in terms of feature extraction and transfer learning, their computational overhead made them less attractive for this specific research context. The MobileNetV2's lightweight design strikes a balance between accuracy and efficiency, making it ideal for resource-constrained environments ultimately emerging as the preferred solution for the specific research landscape.

Ensemble learning methods were explored conceptually as a means to combine predictions from multiple models to enhance overall performance. However, this approach was deferred for future work due to time constraints and the remarkable accuracy achieved with the single MobileNetV2 model. Future iterations could also involve using specialized segmentation-focused architectures like U-Net, opening new avenues for refined classification.

## Training and Validation

The dataset was split into training (70%), validation (15%), and testing (15%) sets using PyTorch’s random\_split function to ensure balanced class representation. The model was trained using the AdamW optimizer with a learning rate of 1e-4 and weight decay for regularization. A learning rate scheduler was implemented to reduce the learning rate dynamically after plateaus in validation accuracy. Training ran for a maximum of 20 epochs, with early stopping triggered if validation accuracy did not improve for two consecutive epochs. This ensured efficient utilization of resources and prevented overfitting.

During validation, Grad-CAM visualizations (Selvaraju et al., 2017) provided crucial insights into the model's decision-making process, revealing the biological relevant regions in histopathological images most critical to classification. The validation loss and accuracy were tracked after each epoch, with both stabilizing to indicate convergence by the fifth epoch. These validations went beyond the traditional metrics and ensured the robustness of the training data and reliable generalization to unseen data.

# 

# Validation, Performance Metrics, and Results

## Validation Techniques and Metrics

Validation of the model was conducted using a 70/15/15 split for training, validation, and testing datasets. Early stopping was implemented, halting training after two consecutive epochs without improvement in validation accuracy. Metrics such as accuracy, precision, recall, and F1-score were calculated to assess performance, along with a confusion matrix and Grad-CAM visualizations to interpret the model's decision-making process.

## Results and Findings

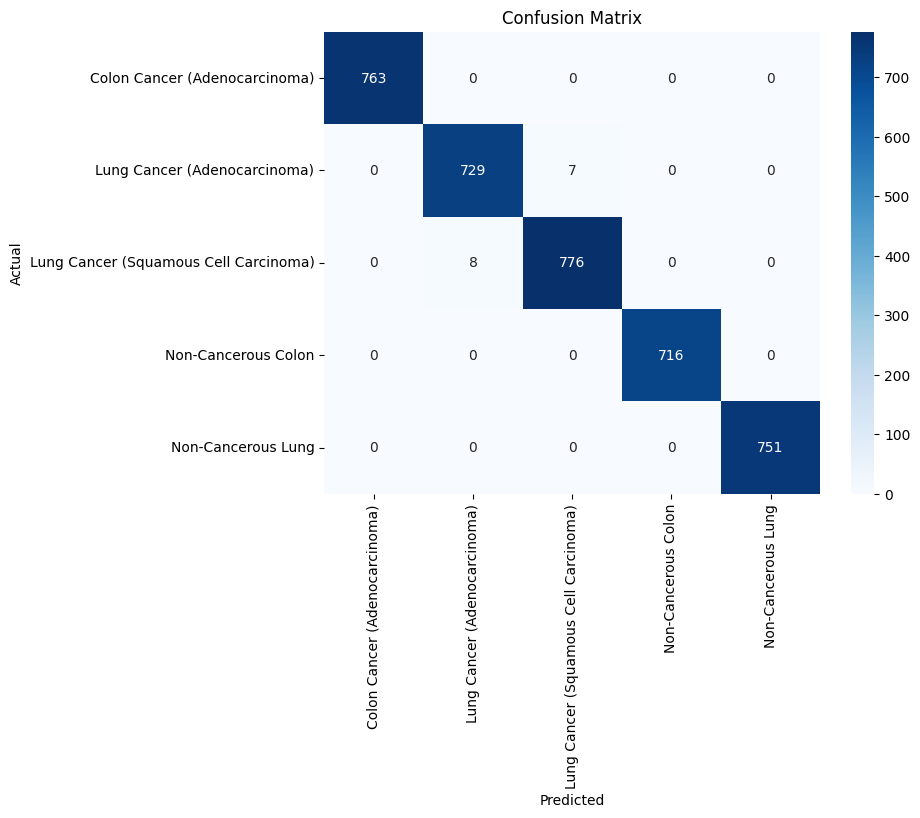
The model achieved remarkable result with an overall test accuracy of 99.87%, detailed class metrics as follows:

Classification Report:

|  | precision | recall | f1-score | support |
| --- | --- | --- | --- | --- |
| Lung Benign | 1.00 | 1.00 | 1.00 | 763 |
| Lung Adeno | 0.99 | 0.99 | 0.99 | 736 |
| Lung Squamous | 0.99 | 0.99 | 0.99 | 784 |
| Colon Adeno | 1.00 | 1.00 | 1.00 | 716 |
| Colon Benign | 1.00 | 1.00 | 1.00 | 751 |
| **Accuracy** |  |  | 1.00 | 3750 |
| **Macro Avg** | 1.00 | 1.00 | 1.00 | 3750 |
| **Weighted Avg** | 1.00 | 1.00 | 1.00 | 3750 |

The confusion matrix revealed minimal misclassifications, with high precision and recall across all classes. Grad-CAM visualizations highlighted the regions of images that contributed most to the model's predictions, providing interpretability and confidence in the model's outputs.

Training and validation accuracy curves demonstrated stable convergence, reaching over 99% accuracy within five epochs. Loss values for both training and validation stabilized quickly, reflecting the efficiency of the MobileNetV2 model and the effectiveness of the preprocessing pipeline. Challenges included the computational demand of training on a large dataset, which was addressed through GPU acceleration and hyperparameter optimization.



## Challenges

This project presented several notable challenges that required thoughtful resolution. First, the computational demand of training a convolutional neural network (CNN) on such a large dataset was significant. Addressing this challenge involved leveraging GPU acceleration, which effectively reduced computation time by approximately 80%, enabling the efficient training of the model.

Managing the complexity of the MobileNetV2 model required careful tuning of hyperparameters, including dropout rates and learning rates, to balance training stability and performance. Overfitting was mitigated by implementing early stopping and data augmentation.

Neural networks mostly operate as black boxes which makes it difficult to understand their decision-making processes. Grad-CAM visualizations helped to address this issue, providing meaningful insights into the model’s decision-making process by highlighting biologically relevant regions within the images.

# Conclusion

This project demonstrates the potential of deep learning methodologies in histopathological image classification for lung and colon cancer detection. By carefully applying a pre-trained MobileNetV2 model to lung and colon histopathological images, the project demonstrates a promising approach to automated tissue classification. The project's high accuracy and precision are encouraging, but it is a step and not a definitive solution. Grad-CAM visualizations added interpretability, enhancing the trustworthiness of the model in clinical use. This transformative potential of AI in diagnostic processes will supplement the ability to assist pathologists, not replace them, remains the fundamental aspiration.

## Key Takeaways

1. **Data Augmentation and Regularization**: These techniques were instrumental in improving the generalization capabilities of the model. Augmentations like rotations and flips introduced variability, while dropout layers mitigated overfitting.
2. **MobileNetV2 Effectiveness:** The lightweight yet powerful MobileNetV2 architecture proved highly efficient, achieving near-perfect accuracy with reduced computational requirements.
3. **Interpretability:** Grad-CAM visualizations offered critical insights into the model's black-box decision-making, increasing its reliability and acceptance in clinical settings.
4. **Future Directions:** Incorporating pre-trained models like ResNet or EfficientNet and exploring ensemble learning techniques could further enhance performance. Integrating real-time diagnosis feedback loop would make this constantly evolving and beneficial in healthcare environments.

# References

Bilgin, C., Yang, J., Wernick, M. N., Kanmaz, P., & Maier, A. (2022). LC25000: A dataset of 25,000 histopathological images for lung and colon cancer classification. Kaggle.<https://doi.org/10.34740/kaggle/dsv/4264051>

Bloice, M. D., Stocker, C., & Holzinger, A. (2017). Augmentor: An image augmentation library for machine learning. *The Journal of Open Source Software, 2*(19), 432.<https://doi.org/10.21105/joss.00432>

Borkowski, A. A., Bui, M. M., Thomas, L. B., Wilson, C. P., DeLand, L. A., & Mastorides, S. M. (2019). Lung and colon cancer histopathological image dataset (LC25000). *arXiv Preprint*, arXiv:1912.12142.<https://arxiv.org/abs/1912.12142>

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N., & Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine, 24*(10), 1559–1567.<https://doi.org/10.1038/s41591-018-0177-5>

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). IEEE.<https://doi.org/10.1109/CVPR.2016.90>

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32, pp. 8024–8035). Curran Associates, Inc.<https://papers.nips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf>

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4510–4520). IEEE.<https://doi.org/10.1109/CVPR.2018.00474>

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626). IEEE.<https://doi.org/10.1109/ICCV.2017.74>

Sirinukunwattana, K., Raza, S. E. A., Tsang, Y. W., Snead, D. R., Cree, I. A., & Rajpoot, N. M. (2016). Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging, 35*(5), 1196–1206.<https://doi.org/10.1109/TMI.2016.2525803>

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(1), 1929–1958.<http://jmlr.org/papers/v15/srivastava14a.html>

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 6105–6114). PMLR.<http://proceedings.mlr.press/v97/tan19a.html>