# Computational Pathology for Colon Cancer: Combining Deep Feature Extraction, PCA Clustering, and Nuclear Segmentation

Noah Kaleekal

Irvington High School

02 September 2025

**Abstract**

Colon adenocarcinoma is among the most prevalent cancers worldwide, and an accurate histopathological diagnosis is critical for guiding treatment. However, conventional diagnoses can be limited by observer variability and the potential for false negatives. In this study, a computational framework is developed, that integrates image classification, deep feature visualization, and morphometric analysis to distinguish Hematoxylin and Eosin (H&E)-stained slides between Colon adenocarcinoma (COAD) and normal colon tissue. A ResNet18-based classifier was then trained on a balanced dataset of COAD and normal tissue samples from The Cancer Genome Atlas. The model achieved a macro-averaged F1-score of 0.9330, with perfect recall for benign samples. Deep features extracted from the network were further analyzed with Principal Component Analysis, revealing clear clustering between cancerous and normal tissues. To capture morphological differences, nuclei segmentation was applied with Cellpose. Quantitative analysis showed that COAD tissue exhibited significantly larger nuclei area fractions compared to normal tissue, consistent with the well-established phenomenon of nuclear enlargement in malignancy. Representative examples confirmed that COAD nuclei were more irregular and heterogeneous, while nuclei of benign cells remained uniform. Taken together, these results demonstrate that deep learning methods can reliably capture visual and morphometric features, distinguishing malignant from benign colon tissues. Beyond high classification accuracy, the integration of feature visualization and segmentation provides biologically interpretable insights, aligning with pathological hallmarks of cancer. This work highlights the potential for automated image analysis to complement diagnostic workflows and reduce misclassification in colon cancer pathology.

## Introduction

Early diagnoses are the key towards reducing cancer mortality rates. Hematoxylin and Eosin staining (H&E) has long been regarded as the gold standard for pathological analysis to inform diagnoses [18]. To obtain such a sample, patients' tumors are excised, sectioned, and stained. Hematoxylin is a basic, purple dye that stains acidic structures, such as the Nucleus or Mitochondria, which contain nucleic acids and proton pump machinery, respectively. Conversely, Eosin, in pink, stains basic structures, including the cytoplasm and Extracellular matrix (ECM) [5].

Pathologists analyze these stained tissues for a wealth of features, which provide information on the benignity or malignancy of the tumor. Cellular organization patterns, like the remodeling of tissue architecture into tube-like structures, appearance of cellular nests within tissues, or significant ECM deposition, often provide clues as to whether a tissue is cancerous [17] [15] [1]. Furthermore, analysis of cell features, including size, Nucleus-to-Cytoplasm ratio, and more, can provide insight into the replicative properties of a cell. Importantly, infinite replicative potential is a hallmark of cancer [20] [2].

Recent advances in deep learning have enabled the extraction of highly discriminative image features from H&E slides. Feature embedding methods, such as those derived from Convolutional neural networks (CNNs), allow for the separation of benign and malignant samples in lower-dimensional space, often revealing structure invisible under traditional microscopy [10] [4]. In addition, dimensionality reduction methods such as Principal Component Analysis (PCA) help visualize how these extracted features naturally cluster, further validating the distinction between normal and cancerous tissues [7].

Beyond classification, segmentation-based approaches provide an opportunity to quantify nuclear morphology at scale [3]. By computing metrics such as area fraction or distribution of nuclear sizes, one can capture differences that reflect underlying biological mechanisms. For instance, larger and more variable nuclear size is strongly associated with aggressive tumor phenotypes, while normal tissues typically display more uniform morphology [14]. These computationally extracted morphometric features can therefore complement classification models, providing interpretable biological insights.

Altogether, combining classification, feature extraction, and morphometric analysis offers a powerful pipeline for histopathology. Such an approach could not only improve diagnostic accuracy, but also provide pathologists with quantitative evidence to support clinical decisions. In this study, such methods are applied for H&E slides to compare cancerous and benign colon tissues, ultimately demonstrating their diagnostic potential.

# Methods

## Dataset

Histopathology slides were obtained from The Cancer Genome Atlas (TCGA) repository [11]. The dataset consisted of Colon adenocarcinoma (COAD) and normal colon tissue samples. To ensure balance across classes, 30 images per group were randomly selected for analysis. Images were stored in PNG format and processed at a resolution of $224 \times 224$ pixels unless otherwise noted.

## Preprocessing

All images were normalized to have pixel intensity values in the range [0,1]. To reduce computational complexity and ensure uniformity across samples, color channels were standardized by subtracting the dataset mean and dividing by the standard deviation. Augmentations, including randomized flips and rotations, were applied during training to improve model generalization.

## Classification Model

A supervised image classification model was trained to distinguish between COAD and benign colon tissue. Specifically, a ResNet18 backbone was employed and pre-trained on ImageNet [6], with the final fully connected layer modified to output two classes. Training was performed using cross-entropy loss and the Adam optimizer, with a learning rate of 0.001. Data were split into 80% training and 20% test partitions. Performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrices.

## Feature Extraction and PCA

Deep features were extracted from the penultimate layer of ResNet18 for all images. To visualize separability between classes, PCA [9] was applied to reduce the 512-dimensional feature space to two principal components. PCA projections were plotted to illustrate separate clustering of COAD and normal tissue samples.

### Segmentation and Morphometric Analysis

To assess morphological differences between nuclei, segmentation was performed using the Cellpose v3.0.8 model [16]. For each image, the total area of segmented nuclei was divided by the total image area to obtain a nuclei area fraction. Box-and-Whisker were generated to compare distributions of nuclei area fraction between COAD and benign tissues. In addition, representative images were overlaid with segmentation masks to provide qualitative visualization of nuclear morphology.

### Statistical Analysis

Differences in nuclei area fraction between COAD and benign tissues were assessed using a two-sample t-test. Results were considered statistically significant at $p < 0.05$. All analyses were conducted in Python 3.9 using `PyTorch` [12], `scikit-learn` [13], `matplotlib` [8], and `scipy` [19].

## Results

### Classification Performance

This supervised model achieved strong performance in distinguishing colon adenocarcinoma (COAD) from normal colon tissue. The classification report is summarized in Table 1. The model achieved a macro-averaged F1-score of 0.9330, with perfect precision for COAD and perfect recall for benign tissue. The confusion matrix (Table 2) demonstrates that only two COAD samples were misclassified as benign, while no benign samples were misclassified.

Table 1: Classification report for colon histology images.

| | **Precision** | **Recall** | **F1-score** |
|---|---|---|---|
| COAD (adenocarcinoma) | 1.0000 | 0.8667 | 0.9286 |
| Normal colon tissue | 0.8824 | 1.0000 | 0.9375 |
| Macro average | 0.9412 | 0.9333 | 0.9330 |

Table 2: Confusion matrix showing predicted vs. true labels.

| | **Predicted** | |
|---|---|---|
| **True** | COAD | Normal |
| COAD | 13 | 2 |
| Normal tissue | 0 | 15 |

### Feature Representation with PCA

To assess the separability of classes in feature space, 512-dimensional features from ResNet18 were extracted and applied to PCA. As shown in Figure 1, COAD and normal tissues form distinct clusters, with minimal overlap. This demonstrates that deep features capture biologically relevant structure, capable of differentiating cancerous from normal tissue.

### Nuclei Segmentation and Morphometric Analysis

To investigate morphological differences, nuclei were segmented using Cellpose. The fraction of image area occupied by nuclei was calculated for each sample. As shown in Figure 2, COAD tissue had significantly larger nuclei area fractions compared to normal tissue ($p < 0.05$). This is consistent with the previously reported phenomenon of nuclear enlargement and polyploidy in malignant cells.

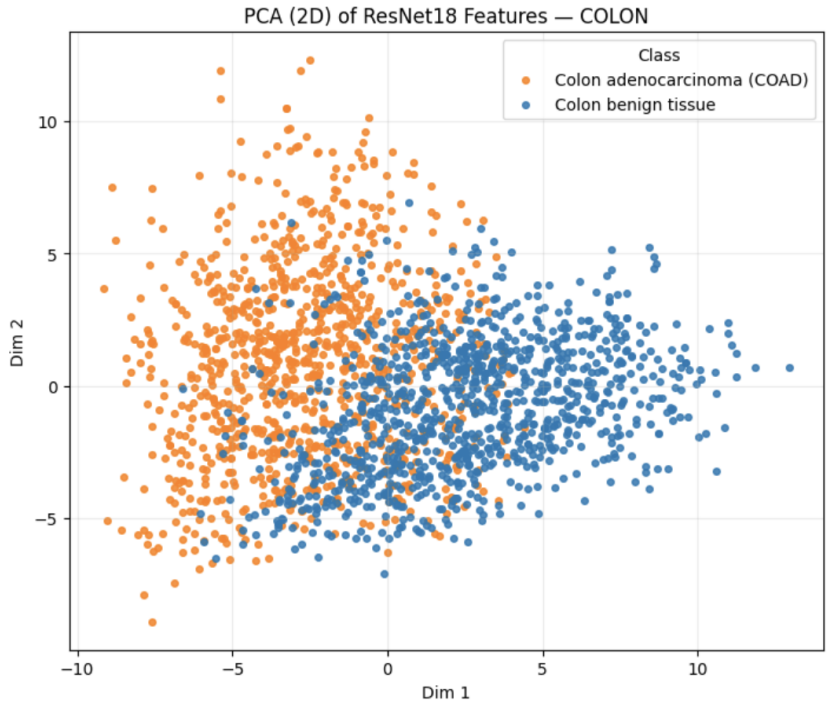


Figure 1: PCA projection of ResNet18 features. Orange: COAD samples, Blue: normal tissue.

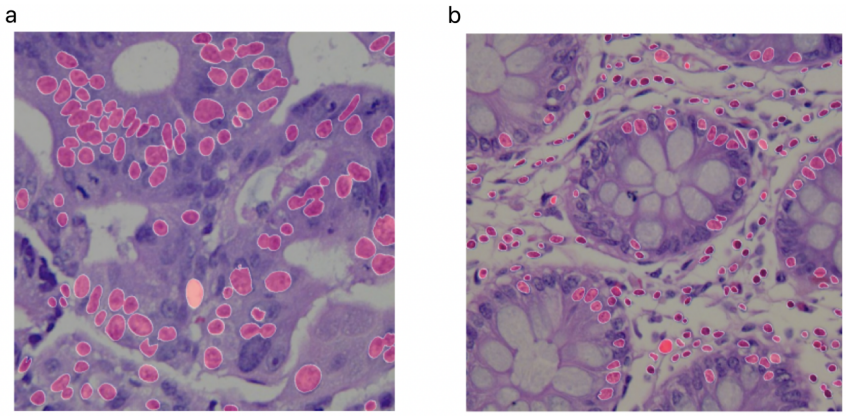


Figure 2: Comparison of nuclei area fraction between COAD and normal tissue.

### Qualitative Visualization

Representative segmentation overlays are shown in Figure 3. In COAD tissue (panel a), nuclei appear enlarged and irregular, while in normal tissue (panel b), nuclei are smaller and more uniform in shape and size.

## Conclusion

In this work, a machine learning framework is presented to distinguish colon adenocarcinoma (COAD) from normal colon tissue using Hematoxylin and Eosin (H&E)- stained histopathology slides. This classifier achieved high accuracy, with strong precision and recall across both classes of tissue. Feature visualization with PCA confirmed that deep representations captured biologically

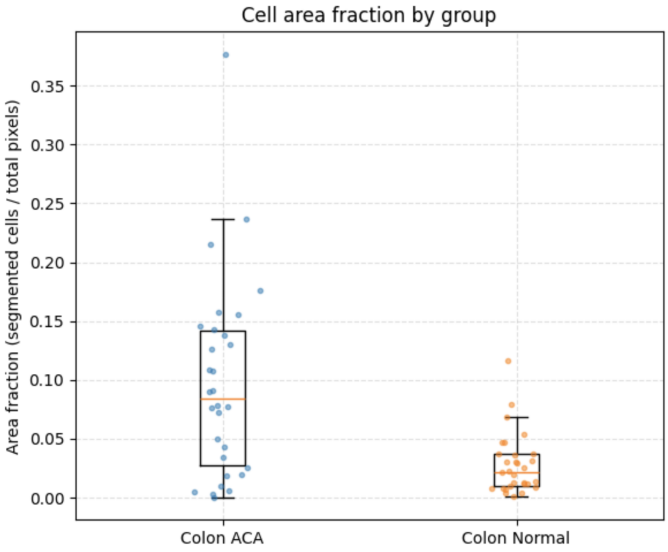


Figure 3: (a) Colon adenocarcinoma (COAD). (b) Normal colon tissue.

relevant differences, as the model was capable of separating malignant from benign samples. In addition, segmentation-based morphometric analysis highlighted enlarged and irregular nuclei in COAD tissue, consistent with established pathological hallmarks.

These findings demonstrate that automated analysis can go beyond simple classification by providing interpretable biological insights. The combination of deep learning and image segmentation not only improves diagnostic accuracy but also has the potential to reduce observer variability and false negatives in pathology. Future work will extend this approach to larger, multi-institutional datasets and integrate genomic or clinical data to further support personalized treatment decisions.

These fundings underscore the value of computational pathology as a complementary tool for cancer diagnosis, capable of assisting clinicians in making faster and more consistent decisions.

# References

[1] Sana Ahuja, Sufian Zaheer, and Sunil Ranga. "Histomorphological Evaluation of Desmoplastic Tumor Stroma in Malignant Ovarian Surface Epithelial Tumors, 2023". In: *Journal of Mid-Life Health* 14.2 (2023), pp. 107–111.

[2] Shruthi Balachandra, Sharanya Sarkar, and Amanda A Amodeo. "The Nuclear-to-Cytoplasmic Ratio: Coupling DNA Content to Cell Size, Cell Cycle, and Biosynthetic Capacity, 2023". In: *Annual Reviews Genetics* 56 (2022), pp. 165–185.

[3] Haoran Chen and Robert F Murphy. "Evaluation of cell segmentation methods without reference segmentations, 2023". In: *Molecular Biology of the Cell* 34.6 (2023).

[4] Srinidhi L Chetan, Ozan Ciga, and Anne L Martel. "Deep neural network models for computational histopathology: A survey, 2021". In: *Journal of Mid-Life Health* 67 (2021).

[5] Andrew H Fischer et al. "Hematoxylin and eosin staining of tissue and cell sections, 2008". In: *Cold Spring Harbor Protocols* (2008).

[6] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[7] Ying-Lin Hsu, Po-Yu Huang, and Dung-Tsa Chen. "Sparse principal component analysis in cancer research, 2015". In: *Translational Cancer Research* 3.3 (2014), pp. 182–190.

[8] John D Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95.

[9] Ian T. Jolliffe and Jorge Cadima. *Principal component analysis*. Springer, 2016.

[10] Jian Lu et al. "Functional and embedding feature analysis for pan-cancer classification, 2022". In: *Frontiers in Oncology* 12 (2022).

[11] Cancer Genome Atlas Network. "Comprehensive molecular characterization of human colon and rectal cancer". In: *Nature* 487.7407 (2012), pp. 330–337.

[12] Adam Paszke, Sam Gross, Francisco Massa, et al. "PyTorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019), pp. 8024–8035.

[13] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. "Scikit-learn: Machine learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[14] Ishitha Singh and Tanmay P Lele. "Nuclear morphological abnormalities in cancer – a search for unifying mechanisms, 2022". In: *Results and Problems in Cell Differentiation* 70 (2022), pp. 443–467.

[15] Louisa M Solis et al. "Histologic patterns and molecular characteristics of lung adenocarcinoma associated with clinical outcome, 2012". In: *Cancer* 118.11 (2012), pp. 2889–2899.

[16] Carsen Stringer et al. "Cellpose: a generalist algorithm for cellular segmentation". In: *Nature Methods* 18.1 (2021), pp. 100–106.

[17] Mehran Taherian, Hua Wang, and Huamin Wang. "Pancreatic Ductal Adenocarcinoma: Molecular Pathology and Predictive Biomarkers, 2023". In: *Cells* 11.19 (2022), p. 3068.

[18] Liang-Jun Tseng, Arata Matsuyama, and Valerie MacDonald-Dickinson. "Histology: The gold standard for diagnosis, 2023". In: *The Canadian Veterinary Journal* 64.4 (2023), pp. 389–391.

[19] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python". In: *Nature Methods* 17.3 (2020), pp. 261–272.

[20] Min Zhou, Mei Zhou, and Yang Jin. "Tumour Cell Size Control and Its Impact on Tumour Cell Function, 2025". In: *Cell Proliferation* (2025).