A Dual Approach: Using DINO and CLIP for Histopathological Image Analysis in Lung and Colon Cancer

Hananeh Shirzadnia , Viraat Saaran , Remziye Maral Demirsecen[1]

Abstract

Lung and colon cancers are leading causes of cancer-related mortality worldwide and are typically diagnosed through histopathological analysis. This paper presents a novel approach utilizing self-supervised learning (SSL), specifically the DINO model, for the automatic analysis of histopathology images related to lung and colon cancers, alongside the Contrastive Language-Image Pretraining (CLIP) model for image classification. SSL enables models to learn from unlabeled data, which improves scalability and efficiency for large datasets. High-dimensional embeddings were extracted using DINO and CLIP on a dataset of 25,000 histopathology images of cancerous and healthy tissues. These embeddings were then processed through clustering, t-SNE visualization for dimensionality reduction, and classification with a Random Forest classifier. Our approach achieved a classification accuracy of 92% with the DINO model and 97% with the CLIP model on the test set. Clustering revealed distinct groupings of tissue types, demonstrating DINO's potential for feature extraction. t-SNE visualization further highlighted the clear separation between cancerous and healthy tissues. In our study, while CLIP achieved higher accuracy than DINO, the DINO model, as an SSL-based approach, demonstrates strong potential as an alternative to traditional supervised methods in histopathological image analysis.

1. Introduction

Cancer detection has always been one of the major challenges in pathology and among physicians in general, particularly during diagnosis and treatment planning. In general, manual identification of cancer from microscopic histopathological images is subjective and can vary from expert to expert based on their expertise and many other factors. Since it varies, it is further improved because there is a lack of precise and objective quantitative methods that can classify these histopathology images. Nowadays Digital pathology [1] is becoming more popular where tissue samples are scanned into high-resolution images.

Deep learning techniques, especially convolution neural networks, have been widely used in medical image analysis, yielding impressive results in cancer detection and tissue classification tasks. Despite their great performance, most of these models depend on huge amounts of labeled data, which is generally expensive and difficult to obtain, especially in the medical domain, where expert annotations are needed [2].

Thus, self-supervised learning offers an alternative by which models can learn useful representations from unlabeled data. DINO is a self-supervised framework based on the teacher-student architecture that learns image representations without labels. In this paper, we will apply DINO to a set of histopathology images and extract the resultant embeddings for clustering, visualization, and classification.

Additionally, CLIP is utilized as another technique to enhance the classification of previously unseen images by utilizing large-scale image-text pairs for training and zero-shot inference. The CLIP approach is particularly appealing in medical image classification where acquiring enough annotated training data is often difficult and time-consuming. Nevertheless, it was originally trained on natural image-text pairs, which limits CLIP's performance when applied to histopathology images with distinct semantic and visual characteristics [7]. We illustrate the potential of self-supervised and contrastive learning in analyzing histopathology images to distinguish between cancerous and normal tissues of the lungs and colon.

2. Related Work

It has been realized that pre-training with self-supervised learning can considerably improve the performance of foundation models for histopathological image analysis. The representation learning from large, unlabeled data is much needed in medical imaging because the labeled dataset is always limited. For processing histopathological images, Vision Transformers are strong architectures with their ability to model both local and global features within complex tissue structures.

TransPath: Transformer-Based Self-supervised Learning for Histopathological Image Classification by Wang et al. explores the application of self-supervised learning using vision transformers for histopathological image analysis. In their work, the authors used a vision transformer architecture pre-trained on large, unlabeled histopathological image datasets. This approach of using the ViT architecture allows the model to learn feature embeddings from tissue images without using the labeled data. The study shows that using self-supervised learning with transformers could give accurate representations, which can be useful in various histopathological classification tasks. Although the paper does not focus on the specific type of cancer, such as lung or colon cancer, it shows the need to use self-supervised learning in histopathology and its ability to improve performance in these various histopathological tasks [3].

Similarly, Vorontsov et al. (2022) presented the Virchow Foundation model, which relies on pre-training large-scale tissue samples to perform pan-cancer detection. This model is very effective at detecting common and rare cancers, such as cervical and bone cancer, for which other models have limitations. The Virchow model had shown strong generalization across several cancer types, indicating its potential for clinical-grade cancer detection even in cases presenting with rare and difficult-to-diagnose cancers. Besides, this robustness against out-of-distribution data inherently makes the model valuable in real-world clinical scenarios, as clinical data may differ across different institutions. This paper presents a unique work that uses the foundation models coupled with self-supervised learning techniques to extract useful features in histopathological images, resulting in better detection and diagnosis in clinical settings [4].

3. Methodology

3.1. Dataset

For this study, we have employed the open-access *Lung and Colon Cancer Histopathological Image Dataset (LC25000)*, proposed by Borkowski et al [5]. There are 25,000 color histopathological images in this dataset, evenly distributed across five different classes, each containing 5,000 images. All the images are 768 × 768 pixels in size and are JPEG formatted. The dataset is organized into two primary subfolders:

- **colon_image_sets**: Contains two subfolders, *colon_aca* and *colon_n*, representing 5,000 images of colon adenocarcinomas and 5,000 images of benign colonic tissues, respectively.

- **lung_image_sets**: Contains three subfolders, *lung_aca*, *lung_scc*, and *lung_n*, representing 5,000 images of lung adenocarcinomas, 5,000 images of lung squamous cell carcinomas, and 5,000 images of benign lung tissues, respectively.

This dataset provides a rather balanced distribution of both cancerous and benign samples that would be ideal for conducting robust model training and testing in various classification tasks in lung and colon cancer detection. Because of its size and diversity, it is more suitable for applications that involve foundation models.

3.2. Models

3.2.1. DINO

DINO stands for Self-Distillation with No Labels and represents the new frontier in self-supervised learning applied mainly to Vision Transformers. Contrary to standard supervised learning that relies on labeled data, in DINO, models can learn visual representations directly from images without labels. The core of DINO is its teacher-student architecture, where both networks share the same structure, but the teacher network is updated using an exponential moving average of the student's weights. This structure allows the model to stabilize learning over time without relying on labeled input. This flexibility in architecture means DINO can be implemented with either Vision Transformers or traditional convolutional neural networks, such as ResNet-50.

The DINO training process begins by giving the teacher and student models different cropped views of the same image, after which the student model learns to match the teacher's output and in the following the output of each model is normalized by a softmax function. The main objective is to minimize the difference between the output of the teacher and the output of the student using a cross-entropy loss function. During the learning phase of the student model, the teacher model also gets updated slowly using a momentum-based technique instead of using direct backpropagation. This is known as the exponential moving average; it provides a controlled learning environment for the teacher who learns from students and eventually refines the learned representation over time [6].

The self-supervised learning approach of DINO allows us to generate clusters without using the labeled data. This similar feature clustering is very successful in tasks like ImageNet classification. Not only does DINO perform well on linear classification, but it also performs well in k-nearest neighbor classification, outperforming other methods in benchmark tests. What makes DINO stand out is its ability to work with large, unlabeled datasets while still giving great results. This flexibility of using unlabeled data makes it suitable for real-world applications like clinical imaging and histopathology, where labeled data is often challenging to obtain.

3.2.2. CLIP

CLIP (Contrastive Language-Image Pre-training) is a state-of-the-art approach that uses natural language to learn visual representations. CLIP was developed by OpenAI to bridge the gap between the two modalities. The core principles of CLIP are joint embedding space, contrastive learning, and semantic alignment.

Unlike traditional image classifiers that rely on labeled datasets, CLIP learns a shared representation for images and textual descriptions causing it to understand domain-specific semantic relationships between images and text. This allows robust generalization across various domains, making it a powerful model for tasks requiring visual and contextual understanding.

CLIP employs a Vision Transformer (ViT) as its image encoder to process visual inputs, while a text transformer interprets textual descriptions. Both modalities are transformed into feature vectors, and mapped into a shared feature space. This shared representation allows a direct comparison of similarities between the image and text embedding space.

In the shared space, CLIP employs contrastive learning to align image and text representations, by bringing matching pairs—such as a tissue image and its corresponding description—closer together while pushing non-matching pairs apart. The contrastive learning optimizes a contrastive loss based on cross-entropy, which calculates similarity scores across all image-text pairs. During training, the model evaluates these similarities and updates the loss accordingly. If an image-text pair is dissimilar, the loss increases, reinforcing the separation between their embeddings. This process enables the model to establish a semantic relationship between images and text, facilitating more accurate cross-modal understanding.

One of the fundamental advantages of CLIP is its ability to generalize well to unseen data, due to its semantic understanding. It is able to find meaningful relationships between images and text, as a result of its training on a large and diverse corpus of datasets. However, fine-tuning CLIP for specific tasks can be challenging due to the scale of the data that it was initially trained on.

4. Experiments and Results

4.1. Training Details : DINO

This model has been developed based on the DINO self-supervised learning framework for lung and colon cancer histopathological images using the Vit-small variant of the ViT architecture. The model was pre-trained for 10 epochs overall with a batch size of 32 using the AdamW optimizer. Notably, this model has been pre-trained and fine-tuned on this dataset without labeled annotations, relying on self-supervised learning to extract meaningful and discriminative features from the images.

Throughout training, the model indicated remarkably consistent improvement in its loss function. At the beginning of the training (epoch 0) the training loss was 10.93. Over 10 epochs, this loss went down smoothly to 10.29 by epoch 9. The loss reduction after training is a notable indication that the model learned well to identify the most relevant features within the histopathological images to fine-tune with limited labeled data.

The learning rate was varied during the training process, starting at $1.87e^{-5}$ and linearly growing to $3.5e^{-4}$ in the final epoch. This increase in learning rate coupled with weight decay provided convergence for the model and helped prevent overfitting. Also during training, by controlling the learning rate over each epoch, the model's performance improved.

This indicates the strength of the self-supervised learning framework of DINO. The core concept behind this framework is self-distillation in the teacher model, through updates of momentum, refines the learning of the student model without the use of labeled data. This is particularly useful in domains like medical image analysis, where usually only a limited amount of labeled data is available. The ability of DINO to extract valuable features from unlabeled data makes it especially suitable for applications where data annotation is expensive or scarce, such as histopathology or other clinical imaging tasks. Through this framework, DINO achieves high feature representations, which are necessary for accurate analysis, despite the inconvenience of missing labels.

4.2. Training Details: CLIP

We fine-tuned the CLIP model on our dataset, splitting the data into 80% for training, 10% for testing, and 10% for validation. As with DINO, the model was pre-trained for 10 epochs with a batch size of 32 using the AdamW optimizer. The AdamW optimizer with a learning rate of $5e^{-5}$ and weight decay of $1e^{-4}$ was applied to update the model's weights effectively.

The base model for CLIP is ”openai/clip-vit-base-patch32”; it uses a ViT-based architecture to provide image embeddings given input images in the form of 32x32 pixel patches. While its exact training dataset is not officially published, it is believed that the model was indeed trained on an extensive dataset (around 400 million image-text pairs scraped from the internet) equal to web-scale datasets.

To improve the model's generalization and reduce the chance of overfitting, various data augmentation techniques were applied during training. These techniques included random resized cropping, random horizontal and vertical flipping, and color jittering (adjusting brightness, contrast, saturation, and hue). The transformations exposed the model to a wider range of image variations, improving generalization.

For each batch of images during training, the following steps were performed:

Feature Extraction: Image features were extracted using the vision encoder of the CLIP model. The encoder mapped each input image to a high-dimensional embedding space.

Text Embedding Generation: Text embeddings were generated from the descriptions of the classes (e.g., ”Lung Normal Histopathology Image”) using the text encoder of the CLIP model.

Feature Normalization: Both the image and text embeddings were normalized to unit vectors. This normalization ensures that the embeddings are on the same scale for accurate calculation of cosine similarity.

Similarity Computation: The cosine similarity between the normalized image and text embeddings was calculated. This similarity score represents the alignment between the image and its corresponding textual description.

Loss Optimization: Cross-entropy loss was applied to the similarity scores to optimize the model. This loss function encourages high similarity for matching image-text pairs and low similarity for non-matching pairs.

4.3. Clustering

We first pre-trained the DINO model to extract embeddings from our dataset of histopathological images; then, we studied the learned feature representations using clustering. The main objective was to divide the images by similarity in their embeddings and explore the hidden structure of the data. The DINO embeddings were originally shaped (25000, 1, 192), where 25,000 is the number of images that have been processed, 1 is an extra batch dimension to maintain consistency, and 192 is the length of the embedding vector for each image. We removed the batch dimension to make the embedding easier (for example, clustering and visualization) using np.squeeze(). The final result is a 2D array of shapes (25000, 192).

With the resulting 192-dimensional embeddings, we applied a suitable clustering algorithm, such as K-means, that grouped the 25,000 images based on their embedded vectors. The different sizes of the five following clusters result from said clustering. Cluster 2 had the maximum number of images, pointing to a dominating pattern or visual feature common across most of the dataset. In contrast, the other clusters had fewer images, maybe representing more specific or distinctive patterns. Whose final distribution of sizes, in turn, was the following: Cluster 0 had 3568 images, Cluster 1 had 2444 images, Cluster 2 held 10,374 images, Cluster 3 comprised 4684 images, and Cluster 4 included 3930 images.

These final values indicate that the model identified certain dominant features across the image dataset over some features that appeared less frequently or had a smaller representation in the overall pattern. As a result, the model has successfully recognized key visual patterns in the whole dataset, which enhances its ability for prediction and classification tasks. The large size of Cluster 2 indicates that it could be an important pattern and thus may be of further interest. Smaller clusters might contain some rare or unique features, which can also be interesting to consider for further analysis. Also, the bar chart which shows the number of images in each cluster (Figure 1) was generated visually to show the imbalanced distribution of images. This chart helps us in understanding the variation in the dataset.

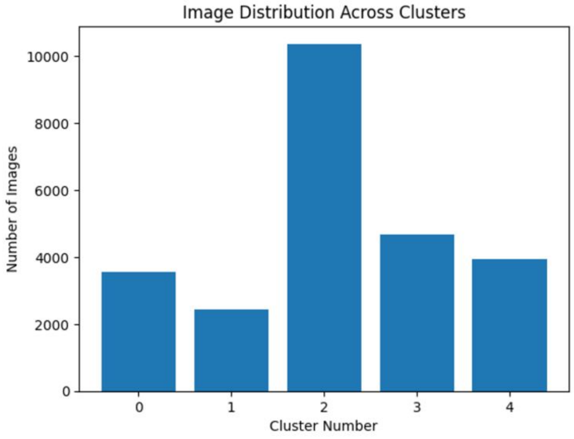


Figure 1. Image distribution across five clusters.

4.4. t-SNE Visualization

In an attempt to get a deeper understanding of the structure of the feature space learned by the DINO model, t-SNE was applied for the visualization of high-dimensional image embeddings. These are 192-dimensional vector embeddings that represent the learned features of histopathological images; we used t-SNE to project them into a 2D space to make relationships between the images more interpretable.

t-SNE is a nonlinear technique of dimensionality reduction that tries to preserve the local structure of the data. In this context, images similar in the original feature space are positioned closer together in the 2D projection, while dissimilar images are placed farther apart. This approach allows us to explore the natural groupings and distinctions between images which is learned by the model visually.

The t-SNE plot shown in Figure 2 indicates clear separation for similar images with similar features. Each point corresponds to a single image. Clustered points show images with common features. This suggests that the model is successful in distinguishing between normal and cancerous tissues or even between different cancer subtypes, such as adenocarcinoma or squamous cell carcinoma.

Another clear observation in the t-SNE plot is that clusters are very well-defined, which suggests that DINO has learned to separate well-distinguished groups of images. These might correspond to major tissue types or categories of cancer reflected earlier in clustering analysis; for instance, one cluster predominantly consists of images of lung cancer, and the other might correspond to normal colon tissue.

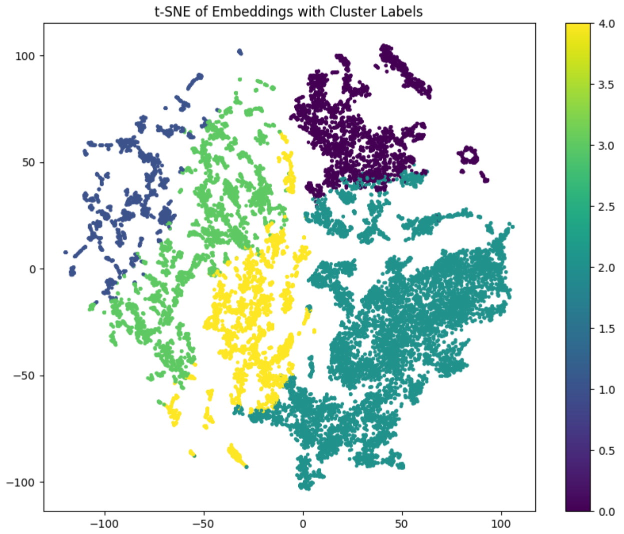


Figure 2. t-SNE visualization of the embeddings extracted from our dataset after applying the DINO model.

4.5. Image Similarity Analysis

We performed a search using cosine similarity between image embeddings to identify the most similar images in the dataset to a given query image, which provides useful insights into the relationships between different histopathological images.

As shown in Figure 3, the query image which is shown far left with the top five most similar images lined up in descending order from most similar to least similar. The similarity scores are located above each image, and measured by the cosine similarity of the query image embedding to other image embeddings. These scores demonstrate just how well each image corresponds to the query. A higher score, closer to 1, means the images are more similar, while values approaching 0.9997 reflect very slight differences in the image features. In this example, the query image is from the category ”Colon Adenocarcinoma.” The five most similar examples include both images from the same category and some from other categories, such as ”Colon Normal.” This shows the model’s ability to learn visual similarities between distinct tissue types and cancer classifications. While images of the same category tend to cluster together, similarity search shows that some from various categories, such as cancerous and normal tissue, may still share visual features the model picks up in its embeddings.

The similarity scores provide a clear way to measure how closely related each image is to the query. The small range of similarity values, from 1 to 0.9997, underlines how precisely the model ranks images based on their visual similarity. This precision shows the model’s ability to distinguish detailed visual features that might not always align directly with medical classifications, such as differentiating between tissues or cancer types. This visualization of similarity search results will give a deeper understanding of how the model clusters similar histopathological images based on learned features. It has potential applications in image retrieval, diagnostic support, and finding similar cases in large datasets. It also allows us to evaluate the generalization capability of the model across various tissue types and cancer classifications and understand the relations between the images and the effectiveness of the feature learning process by the model.

4.6. Model Training and Evaluation

4.6.1. DINO

We performed a search using cosine similarity between images The 192-dimensional embeddings extracted from the histopathological images were used as input for a random forest classifier, which was trained to predict the cancer type, colon, or lung, based on these feature embeddings. Then the dataset was split into three parts: 60% was allocated to the training set for training the model, 20% to a validation set to perform some fine-tuning on its parameters, and the remaining 20% as the test set used in the evaluation of its final performance.

The Random Forest classifier achieved 91% accuracy on the test set, which has proven to be quite powerful in correctly classifying most of the images. Performance metrics for the model, such as precision, recall, and F1-score, are all consistent across different classes of cancer. Both the macro and weighted averages of precision, recall, and F1-score were 0.92, indicating that the performance of the model is well-balanced across classes.

Further, the confusion matrix in Figure 4 gives a clear visualization of the classifier’s performance; it provides infor-

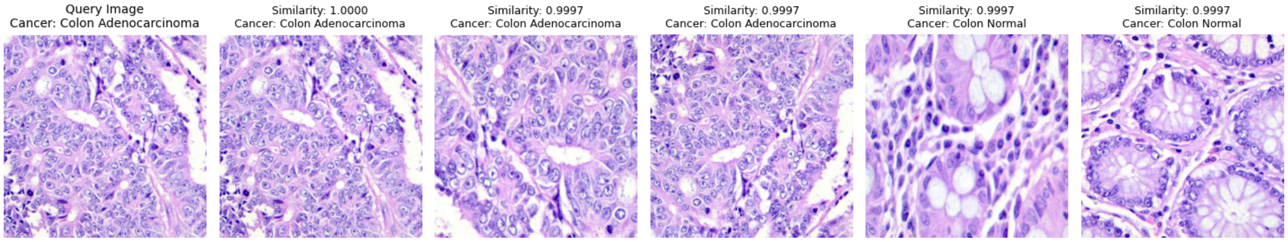


Figure 3. Similarity search results are based on cosine similarity between the query image embedding and other image embeddings.DINO Found the five closest matches in terms of visual features.

mation on correct and incorrect predictions for every type of cancer. Results seem to indicate that feature embeddings have captured all important insights about histopathological images, and this embedding lets the random forest classifier discern among tissue types that are cancerous versus non-cancerous.

These findings support the fact that the extracted embeddings can be useful in medical image classification. The successful performance of the classification tasks indicates that the embeddings are appropriate for key visual feature identification, and future improvements may be carried out by fine-tuning the classifier or trying more advanced models to extract even more discriminative features.

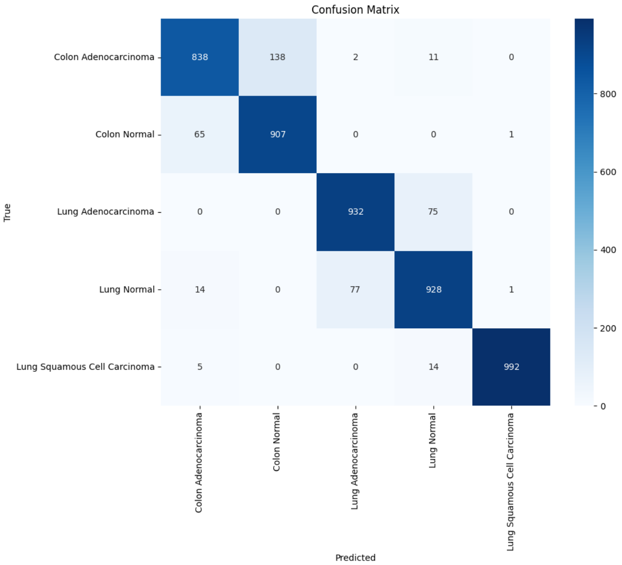


Figure 4. Confusion matrix showing the performance of a Random Forest classifier in classifying histopathology images using features extracted with DINO.

4.6.2. CLIP

The Confusion Matrix in Figure 6 provides a clear visualization of the model's performance, illustrating how well the model performed across each class. It offers valuable insights into correct and incorrect predictions for each type of cancer, highlighting areas of strength and potential for improvement.

Based on the figure, Colon Adenocarcinoma, Colon Normal, and Lung Normal have excellent classification results. Colon Adenocarcinoma has 490 correct predictions out of 500 samples, while Colon Normal misclassified only 2 images and Lung Normal misclassified only 3 images.

There are some misclassifications between Lung Adenocarcinoma and Lung Squamous Cell Carcinoma indicating there is a small challenge in distinguishing these cancer types. However, the model still maintains an accuracy of over 90%, highlighting its overall robustness.

In Figure 5, there is a detailed Classification Report on the CLIP model. Similar to the DINO model, in the Classification Report in Figure 5, performance metrics for the model, precision, recall, and F1-score, were consistently high across all classes of cancer, exceeding 91%.

Specifically, for Colon Adenocarcinoma, Colon Normal, and Lung Normal, the F1-score, precision, and recall impressively exceed 98%. However, a slight decrease is observed for Lung Adenocarcinoma and Lung Squamous Cell Carcinoma, where the F1-score is 0.93, suggesting potential areas for improvement in distinguishing these two subtypes, as reflected in the Confusion Matrix.

Additionally, Both the macro and weighted averages of precision, recall, and F1-score were 0.97, indicating that the performance of the model is very well-balanced across classes, underscoring the model's reliability and overall strength in cancer classification.

High precision indicates that CLIP's predictions are highly reliable, reducing the likelihood of unnecessary follow-up tests. High recall ensures that the model effectively identifies most instances within the dataset, thus minimizing the risk of undiagnosed cases. F1-score, which provides a harmonic balance between precision and recall, demonstrates strong performance across all classes. These results are particularly significant in the healthcare field, where accurate classification and early detection are crucial for effective diagnosis and treatment.

Overall, CLIP achieves an impressive classification accuracy of 97%, outperforming the DINO model. It exhibits high precision and recall, signifying a low rate of false positives and false negatives. This combination of characteristics makes the model suitable for deployment in clinical settings.

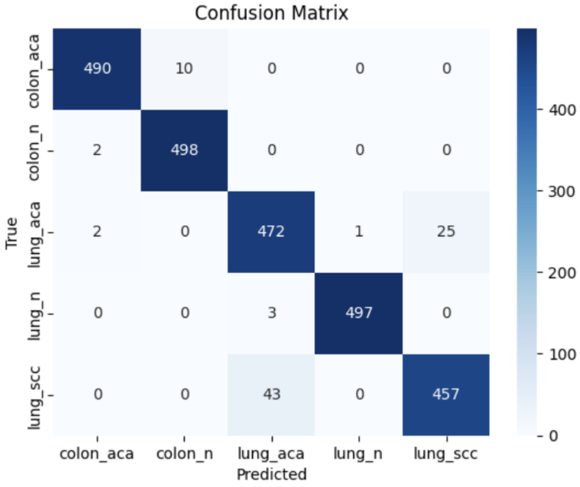
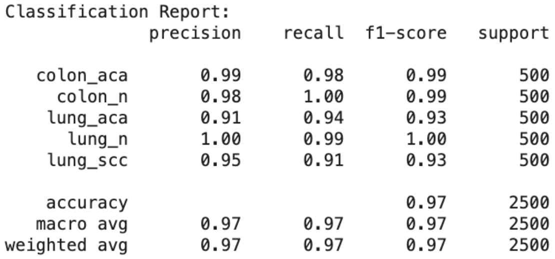


Figure 5. Confusion matrix showing the performance of CLIP model in classifying histopathology images



```
Classification Report:
              precision    recall  f1-score   support

   colon_aca       0.99      0.98      0.99       500
     colon_n       0.98      1.00      0.99       500
    lung_aca       0.91      0.94      0.93       500
      lung_n       1.00      0.99      1.00       500
    lung_scc       0.95      0.91      0.93       500

    accuracy                           0.97      2500
   macro avg       0.97      0.97      0.97      2500
weighted avg       0.97      0.97      0.97      2500
```

Figure 6. Classification report for the CLIP model showing the performance details

5. Challenges

The most important challenge with the pre-training of the DINO and CLIP models was time-consuming training for fine-tuning CLIP models, which are very computationally intensive. For example, training for 10 epochs with a batch size of 32 takes about 4 hours. The next significant obstacle was related to the availability of GPU because most of the fine-tuning models require higher computational resources than those available.

Additionally, constraints such as the limited availability of high-quality medical Q&A datasets for histopathology images during training presented a challenge to fully optimizing our model.

6. Conclusion

In this work, we have shown the potential of DINO and CLIP in lung and Colon cancer detection on histopathology images, achieving performance that are comparable to the best convolutional networks specifically designed for this task. CLIP, as a contrastive model, reached an accuracy of 97% on our dataset, while the DINO model achieved 91% classification accuracy with a Random Forest classifier. Our experiments indicated that even with limited labeled data, self-supervised and contrastive learning models have a strong ability to learn dominant features from histopathology images.

Additionally, these models are capable of generalizing well across unseen data which shows their robustness in medical image classification tasks. Leveraging such models could significantly aid pathologists in early and accurate cancer detection. Furthermore, by applying self-supervised learning techniques, we saw the possibility of using large, unlabeled datasets to address the challenges of acquiring large labeled data in medical images.

7. References

1. R. Kumar, R. Srivastava, and S. Srivastava, ”Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features,” BioMed Research International, vol. 2015, Art. no. 457906, Aug. 2015.

2. L. Deininger, B. Stimpel, A. Yuce, S. Abbasi-Sureshjani, S. Schönenberger, P. Ocampo, K. Korski, and F. Gaire, ”A comparative study between vision transformers and CNNs in digital pathology,” F. Hoffmann-La Roche AG, Basel, Switzerland, 2023.

3. X. Wang, S. Yang, J. Zhang, M. Wang, and J. Zhang, ”TransPath: Transformer-based Self-supervised Learning for Histopathological Image Classification,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

4.E. Vorontsov, A. Bozkurt, A. Casson, G. Shaikovski, M. Zelechowski, K. Severson, E. Zimmermann, J. Hall, N.

Tenenholtz, N. Fusi, E. Yang, P. Mathieu, A. van Eck, D. Lee, J. Viret, E. Robert, Y. K. Wang, J. D. Kunz, M. C. H. Lee, J. H. Bernhard, R. A. Godrich, G. Oakley, E. Millar, M. Hanna, and T. J., ”A foundation model for clinical-grade computational pathology and rare cancers detection,” arXiv preprint arXiv:2309.07778, 2023.

5. A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides, ”Lung and Colon Cancer Histopathological Image Dataset (LC25000),” arXiv preprint arXiv:1912.12142v1 [eess.IV], Dec. 2019.

6. M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, ”Emerging Properties in Self-Supervised Vision Transformers,” arXiv preprint arXiv:2104.14294, 2021.

7. Sun, X., Zou, X., Wu, Y., Wang, G., Zhang, S. (2025). ”Fairness Analysis of CLIP-Based Foundation Models for X-Ray Image Classification,” arXiv preprint arXiv:2501.19086, 2025.