Histopathology Classification via Fine-tuned Vision-Language Models ML Singapore!

Group 20: Abhishek Balaji, Kale Aprup, Misra Aditya, Vaishnav Muralidharan

A0246018J, A0265847M, A0266884J, A0235268Y

Abstract

Histopathology plays a central role in disease diagnosis, particularly cancer, but is challenged by visual complexity and the need for expert interpretation. While Vision-Language Models (VLMs) like CLIP have shown promise (especially when adapted to medical domains via contrastive learning on datasets such as Quilt-1M) their effectiveness for clinical tasks like multi-label classification remains underexplored. This project proposes a fine-tuning approach for CLIP (ViT-L/14) tailored to histopathology image-text pairs. We selectively unfreeze the final layers of both encoders, fuse their embeddings, and train a dedicated MLP head using Binary Cross-Entropy loss. Evaluated on a curated 43k-pair subset of Quilt-1M, our model achieves a 3-label accuracy of 68%, demonstrating strong potential for identifying multiple co-occurring findings. Framed within the needs of Singapore's healthcare system, this system aims to support faster, more consistent diagnoses and improve trust in AI-assisted pathology workflows.

Introduction

Histopathology (the microscopic examination of tissue samples) remains the gold standard for diagnosing diseases, particularly cancer. In Singapore, where an aging population places growing demands on the healthcare system, timely and accurate pathology services are increasingly critical. Yet, interpreting histopathology slides is a complex, time-intensive process that requires significant expertise and is prone to inter-observer variability (van den Bent 2010). These challenges motivate the development of machine learning (ML) tools to assist pathologists in improving diagnostic accuracy, efficiency, and consistency.

Recent advances in self-supervised learning have introduced powerful Vision-Language Models (VLMs) such as CLIP (Radford et al. 2021), which learn joint image-text representations from web-scale data. Their extension to medical domains is gaining momentum. The Quilt-1M dataset and resulting QuiltNet model (Ikezogwo et al. 2023) demonstrated the value of domain-specific VLM fine-tuning, achieving state-of-the-art results on histopathology tasks via contrastive learning over one million image-text pairs.

However, QuiltNet primarily supports zero-shot classification and retrieval by aligning embeddings, rather than optimizing for clinical tasks like multi-label classification—where multiple co-occurring findings must be identified from a single image-report pair. While linear probing on pre-trained embeddings is feasible, end-to-end (or partial) fine-tuning may yield richer and more task-specific representations. Prior deep learning work has already demonstrated expert-level performance in topics such as detecting metastases on lymph node slides, achieving AUCs close to 0.99 (Ehteshami Bejnordi et al. 2017), highlighting the potential of supervised learning for diagnostic accuracy.

This work builds on these insights by proposing a fine-tuned CLIP-based architecture tailored for multi-label histopathology classification from both images and associated textual descriptions.

Problem Statement and Proposed Application

This project addresses the need for automated multi-label classification in histopathology by jointly leveraging visual information from tissue images and semantic cues from accompanying textual reports.

In real-world clinical settings, access to large expert-annotated datasets is often limited, and inputs (ranging from high-resolution slides to free-text notes) can vary significantly in format and quality. CLIP's dual-encoder architecture and strong generalization from large-scale pretraining make it particularly well-suited for such multi-modal, low-resource environments.

Real-World Scenario (Singapore): A pathologist at the National University Hospital (NUH) reviews a digitized breast biopsy slide. Alongside the image, they input a transcribed note: *"Lobular carcinoma in situ infiltrating in Indian file pattern has worse prognosis than infiltrating ductal carcinoma."* Within seconds, our model processes both modalities and returns a ranked list of relevant findings:

Breast pathology	Soft Tissue	Cytopathology

This second read helps the pathologist focus attention on subtle but significant features, reducing diagnostic oversight and accelerating downstream decisions for multidisciplinary teams (e.g., oncology, surgery). Over time, integrating such

a tool could enhance reporting consistency, improve efficiency in high-volume labs, and provide educational support for junior pathology staff.

Methodology

Dataset

We used a curated subset of 43,866 image-text pairs from the Quilt-1M dataset (Ikezogwo et al. 2023), a large-scale vision-language dataset for histopathology. Each pair is labeled with up to three subpathology classes selected from 60 possible labels. The full dataset comprises one million pairs generated through a multi-stage pipeline involving video mining, frame extraction, ASR transcription, and large language model (LLM)-based classification.

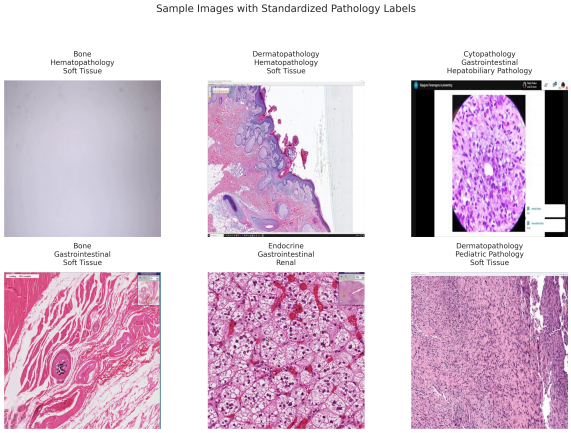

Figure 1: Sample image-text pairs from the Quilt-1M dataset.

Source: The dataset was built from 1,087 hours of clinician-narrated histopathology videos on YouTube. Relevant frames were extracted using an ensemble classifier that filtered out non-histology scenes, duplicates, and low-quality frames.

Text Processing: Audio narrations were transcribed using ASR and refined with LLMs, UMLS-based ontologies, and heuristic rules to generate informative captions.

Labeling: Captions were processed by an LLM classifier to assign up to three subpathology labels across 18 broader categories (e.g., gastrointestinal, renal, dermatopathology).

Data Cleaning and Preprocessing

Each subset includes a folder of image patches and a CSV file with image paths, captions, top-3 labels, and diagnostic descriptions.

Key Preprocessing Steps:

- **Missing Data Removal:** Dropped rows with `NaN` values in `caption`, `image_path`, or `corrected_text`.
- **Label Standardization:** Merged semantically similar labels (e.g., *Immunology* vs. *Immunopathology*) using a custom mapping to ensure consistency and improve F1-score reliability.

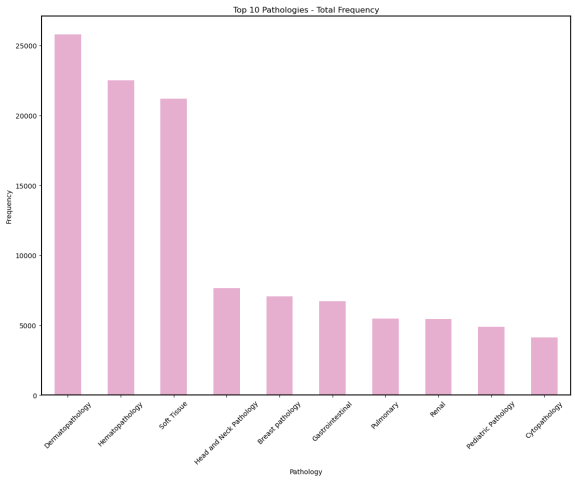

Figure 2: Frequency distribution of subpathology classes in our Quilt-1M subset.

Model Architectures Explored

We evaluated several modeling approaches for multi-label histopathology classification, first isolating the contributions of individual modalities before exploring multimodal fusion.

Baseline: Zero-Shot CLIP As a baseline, we evaluated OpenAI's CLIP model (`ViT-L/14`) in a zero-shot setting to benchmark the capabilities of large vision-language models without task-specific training. CLIP embeds both images and text into a shared embedding space. We computed image embeddings and compared them against text embeddings generated from class-specific prompts (e.g., "A pathology image of [class name]") using cosine similarity. The top-3 classes with the highest similarity scores were selected as predictions.

This approach required no fine-tuning and relied entirely on pretrained representations. We used the `clip` library for inference.

Image-Only: Fine-tuned ResNet-50 + MLP To benchmark a strong unimodal visual baseline, we finetuned a ResNet-50 backbone (He et al. 2016) pre-trained on ImageNet. The final fully connected layer of the ResNet-50 was replaced with an identity mapping (`nn.Identity()`) and a 4-layer Multi-Layer Perceptron (MLP) classification head was trained on top of the resulting 2048-dimensional features:

```
Linear(2048, 512) -> ReLU -> Linear(512,
    512) -> ReLU -> Linear(512, 512) ->
ReLU -> Linear(512, num_classes)
```

To enable task-specific adaptation, we unfroze the last ResNet layer (`layer4`), and trained it alongside the MLP head. The model was trained for 30 epochs using the `AdamW` optimizer with a learning rate of 1×10^{-5}, binary cross-entropy loss (`nn.BCEWithLogitsLoss`), gradient norm clipping (max norm 1.0), and mixed precision. Batch size was set to 32.

Text-Only: Fine-tuned BioBERT + MLP To evaluate the textual modality (captions and report snippets), we employed BioBERT (`dmis-lab/biobert-v1.1`) (Lee et al. 2020), a BERT-based model pre-trained on large biomedical corpora. We extracted the pooled output of the `[CLS]` token (768-dimensional) as the text representation, which was then passed through a three-layer MLP classifier:

```
Linear(768 -> 512) -> ReLU -> Linear(512
    -> 512) -> ReLU -> Linear(512 ->
   num_classes)
```

For task-specific adaptation, the final two BioBERT transformer encoder blocks were unfrozen, and trained along with the MLP head. This selective tuning strategy preserves foundational biomedical language understanding while allowing the model to adapt to histopathology-specific text. Training was performed using the `AdamW` optimizer with differential learning rates: 1×10^{-5} for BioBERT and 1×10^{-4} for the MLP head. We used the same training setup as the ResNet model—mixed precision, BCEWithLogits loss, gradient clipping, and 30 epochs. Due to the higher memory footprint of transformer models, we used a reduced batch size of 16.

Rationale for Initial Approach Choices We selected these initial models to represent strong uni-modal baselines—vision-only, text-only—and a zero-shot VLM. This allowed us to isolate the contribution of each modality and establish performance bounds before pursuing multi-modal fusion.

Early experiments with naive fusion (e.g., concatenating embeddings from ResNet and generic BERT) yielded poor results, likely due to mismatched embedding spaces. This motivated the use of modality-specialized models (ResNet, BioBERT) and pretrained joint encoders like CLIP for downstream integration.

Experiments and Results

Experimental Setup of Initial Approach

Dataset Split: The curated 43,866 image-text pair subset from Quilt-1M was split into training and testing sets using an 80/20 ratio. The split was kept consistent across all experiments.

Models Tested: We evaluated the three architectures described in the previous section :

1. Zero-Shot CLIP (`ViT-L/14`)
2. Fine-tuned ResNet-50 + MLP (Image-Only)
3. Fine-tuned BioBERT + MLP (Text-Only)

Evaluation Metrics: Each sample is annotated with exactly three positive labels, making this a multi-label classification task. We used the following metrics:

- **Micro AUC**
- **F1-Score (Micro/Macro):** Calculated on top-3 binary predictions to evaluate both overall and per-class performance.
- **Hamming Loss (Top-3):** Measures the fraction of incorrect labels in top-3 predictions.
- **Top-k Accuracy (Top-3):** Reports the proportion of samples where at least $k \in \{1, 2, 3\}$ true labels are present among the top-3 predicted classes.

Note: Macro AUC was excluded due to instability—some classes had no positive predictions in the test set.

Evaluation of Initial Approach

The performance of the three initial models on the test set is summarized in Table 1. Predictions for F1-scores, Hamming Loss, and Custom Accuracies were derived by selecting the top-3 classes with the highest predicted probabilities for each sample.

Interpretation of Zero-Shot CLIP (`ViT-L/14`): These results indicate that while CLIP demonstrates strong general vision-language understanding, its pretrained embeddings lack the task-specific alignment needed for accurate multi-label histopathology classification. The model struggles to retrieve even a single correct label among its top predictions.

Interpretation of Fine-tuned ResNet-50 + MLP (Image-Only): Fine-tuning the final layers of a pre-trained ResNet-50 yields substantial predictive power (Avg Train Loss after 30 epochs: 0.0421). The high AUC-micro and 1-label accuracy indicate strong sensitivity to atleast one correct label. The considerable gap between F1-micro and F1-macro suggests uneven performance across classes, likely favoring more frequent labels. The drop from 1-label to 3-label accuracy reflects the challenge of identifying all true labels, though a 45% 3-label accuracy remains promising for a purely visual model.

Interpretation of Fine-tuned BioBERT + MLP (Text-Only): Fine-tuned BioBERT effectively leverages the diagnostic value in textual descriptions, outperforming zero-shot CLIP by a wide margin—highlighting the benefit of domain-specific pretraining. While it trails the ResNet model in micro F1 and accuracy, its higher F1-macro suggests better balance across classes, possibly capturing pathology nuances described in text. Nonetheless, compared to vision-based approaches, text alone remains a weaker signal for this task.

Overall Initial Findings: Both fine-tuned unimodal models—ResNet (image-only) and BioBERT (text-only)—substantially outperform the zero-shot CLIP baseline, demonstrating the importance of task-specific adaptation. ResNet achieves higher F1-micro and custom top-k accuracies, indicating stronger discriminative power from visual features. BioBERT, while slightly weaker overall, achieves a higher F1-macro, suggesting better performance on less frequent classes. These complementary strengths motivate our final approach: fusing both modalities using a fine-tuned CLIP model to capture information that neither modality fully encodes alone.

Experimental Setup of CLIP-based approaches

CLIP: Contrastive Language-Image Pretraining

OpenAI's CLIP model (Radford et al. 2021) is trained on 400 million image-text pairs to learn a shared embedding space where semantically related visual and textual inputs

Table 1: Performance Metrics for Initial Model Approaches on the Test Set.

Model Approach	AUC-μ	F1-μ (Top-3)	F1-M (Top-3)	Hamming (Top-3)	1-Label Acc	2-Label Acc	3-Label Acc
Zero-Shot CLIP (ViT-L/14)	0.6343	0.1317	0.0615	0.1140	0.4460	0.0263	0.0001
ResNet-50 + MLP (Image-Only)	0.9808	0.7422	0.2747	0.0258	0.9612	0.8136	0.4518
BioBERT + MLP (Text-Only)	0.9751	0.6842	0.3509	0.0316	0.9533	0.7633	0.3363

Note: AUC-μ = Micro AUC; F1-μ/M = Micro/Macro F1; Acc = Accuracy.

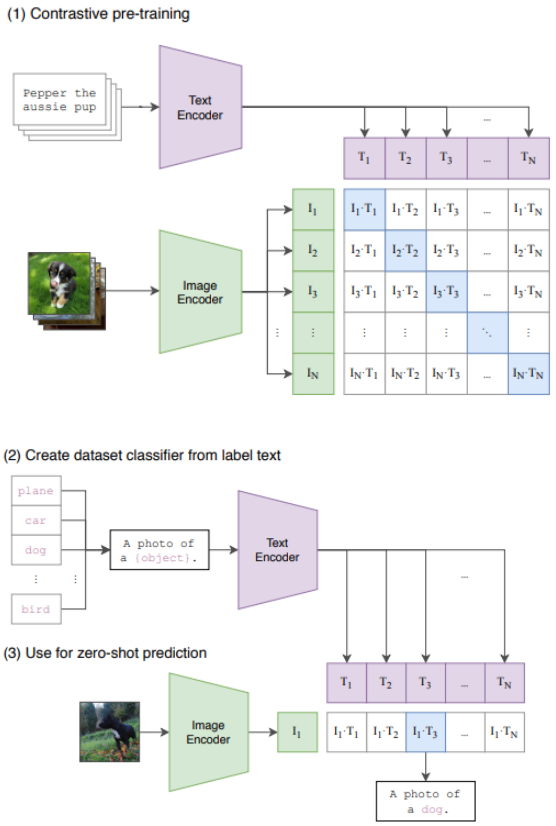

Figure 3: Summary of CLIP Training Approach

are aligned. It uses a dual-encoder architecture comprising a Vision Transformer (or ResNet) for images and a Transformer-based text encoder. Training is performed via contrastive learning, maximizing cosine similarity between embeddings of matched image-text pairs while minimizing it for mismatched ones.

CLIP Approach with Frozen Encoder Weights

Our initial approach with CLIP relied on the fusion of embeddings from a `ViT-B/32` image encoder and a CLIP text encoder. Then, these embeddings were passed through a 4-layer MLP. Two different fusion methods were explored:

- **Concatenation Fusion**
- **Sum Fusion**

Concatenation fusion outperformed element-wise summation (55% vs. 47% top-3 accuracy), suggesting that retaining separate image and text representations allows the MLP to better discriminate pathology classes. Consequently, concatenation was used in our final model.

While CLIP's image and text encoders produce aligned embeddings, they are not fine-tuned for histopathology classification. We therefore unfreeze the final layers of the image encoder to enable task-specific adaptation and improve the quality of learned representations.

Final Model Architecture

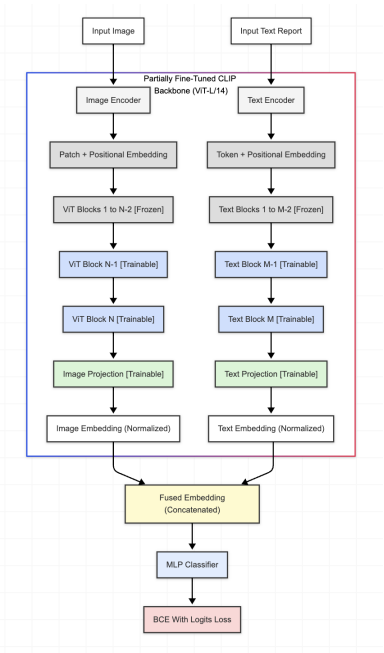

Figure 4: Final Model Architecture Diagram

To realize this application, we investigated and implemented a novel fine-tuning strategy for a state-of-the-art Vision Language Model (VLM). Our approach utilizes OpenAI's CLIP model with a Vision Transformer backbone `ViT-L/14` as the base, leveraging its powerful pre-trained representations. The core technical aspects are:

- **Selective Fine-tuning:** Instead of full fine-tuning or only training a linear probe, we unfreeze and train only the final two transformer blocks of both the image (ViT) and text encoders. This exploits the desirable property of leveraging robust, general features from deeper layers while adapting the most task-specific layers (Wortsman et al. 2022).
- **Multi-modal Fusion:** We extract embeddings for both the input image and the input text report using the par-

tially fine-tuned CLIP encoders. These embeddings are then concatenated to create a fused representation capturing information from both modalities.

- **Classification Head:** A simple Multi-Layer Perceptron (MLP) head is trained on top of the fused embeddings to perform the final multi-label classification into N predefined sub-pathology classes (Guo 2024).
- **Loss Function:** We employ `BCEWithLogitsLoss`, appropriate for multi-label classification tasks where multiple classes can be positive simultaneously.

This approach offers potential advantages over alternatives like using only image embeddings (ignoring text) or simple linear probing (less capacity for complex interactions). It directly optimizes for the downstream multi-label task using combined visual and textual cues. However, it requires careful tuning of learning rates for different model parts (CLIP vs. head) and managing the computational cost of fine-tuning parts of a large VLM.

Evaluation

Table 2: Performance Metrics for the Final Fine-tuned CLIP + MLP Model.

Metric	Value	Metric	Value
AUC-μ	0.9926	1-L Acc (Top-3)	0.9875
F1-μ (Top-3)	0.8628	2-L Acc (Top-3)	0.9216
F1-M (Top-3)	0.5820	3-L Acc (Top-3)	0.6796
Hamming (Top-3)	0.0137		

Note: Metrics based on top-3 predictions where applicable. AUC-μ=Micro AUC; F1-μ/M=Micro/Macro F1; Hamm.=Hamming Loss; Acc=Accuracy.

The performance of our final model, which employs selective fine-tuning of CLIP (`ViT-L/14`) and fuses image and text embeddings via concatenation into a 4-layer MLP head, is presented in Table 2.

Interpretation of Fine-tuned CLIP + MLP (Multi-modal): This multi-modal approach significantly outperforms all initial baseline and uni-modal models (compare Table 2 with Table 1). The AUC-micro reached 0.9926, demonstrating excellent overall discrimination capability. Critically, the F1-micro score (0.8628) based on top-3 predictions saw a substantial improvement of over 12 percentage points compared to the best uni-modal approach (ResNet+MLP, 0.7422). This strongly supports our hypothesis that fusing visual and textual information learned within CLIP's joint embedding space, combined with task-specific fine-tuning, leads to superior performance.

Furthermore, the custom accuracy metrics show substantial gains, particularly for identifying the complete set of labels. The 1-label accuracy remains very high (0.9875), but the 2-label accuracy (0.9216 vs 0.8136 for ResNet) and especially the 3-label accuracy (0.6796 vs 0.4518 for ResNet) demonstrate the model's enhanced ability to capture the multiple relevant findings concurrently.

These results validate our final model architecture and fine-tuning strategy. The selective unfreezing allows adaptation without catastrophic forgetting, while the fusion mechanism effectively combines complementary information from both modalities within the powerful representation space learned by CLIP and adapted via Quilt-1M-derived data.

Discussion

Baseline CNN Models

We found that the fine-tuned ResNet-18 marginally outperformed a vanilla CNN on accurately predicting all three labels, and both these models largely outperformed a fine-tuned ResNet-50. This can be attributed to the fact that ResNet-18 provides a compact hypothesis space for training. Given that the size of the Quilt dataset is small compared to the data that ResNet was pretrained with. On the other hand, the CNN and ResNet-50 had more parameters than the ResNet-18 model, leading to a more hypothesis space for relatively small amounts of data and hence overfitting.

Naive Concatenation of Image and Text Embeddings

We initially used two encoder models: a ResNet-18 CNN to retrieve image embeddings and a DistilBERT encoder to retrieve text embeddings. A naive concatenation of the text encodings and image encodings were then passed into a 4-layer MLP, with a final output layer whose size equaled the number of labels. We then applied a sigmoid function to obtain a probability distribution. However, in this case, the model performed worse than ResNet-18 without the text augmentation.

While concatenation is a well-known fusion method and has been used by (Guo 2024), the vector representation provided by the text and image encoder belonged to two different vector spaces; the vector provided by the image encoder would not be similar to that of the text encoder, even though they were describing the same three pathology labels.

Thus, the MLP did not combine the two representations accurately, probably due to a small representation space and lack of more data. As a result, we needed encoders that would both encode information to a common representation space.

Unfreezing the Image Encoder Weights

Given that unfreezing the last two transformer blocks of the image encoder yielded an improvement of over 10%, we can derive two possibilities: either the MLP does not present with an adequate hypothesis representation space to learn the concept of identifying pathology classes, or the image and text encoders do not provide with an adequate intermediate representation that can be learned by the MLP.

We can disprove the first conjecture, that the MLP is too small a representation of the hypothesis space, as increasing the number of layers in the MLP from 4 to 5 did not improve the accuracy in predicting the top-3 labels. Hence, we can conclude that the previously frozen text and image encoders led to a slight 'domain misalignment', as they were not providing representations suitable for discriminating between pathology classes.

Model Suitability and Limitations

Despite promising results, our proposed approach has several limitations that reflect broader challenges in applying large vision-language models to clinical domains:

- **Partial fine-tuning due to resource constraints:** While we selectively unfroze the final layers of the CLIP model to adapt it to the histopathology task, full fine-tuning was computationally prohibitive. This limits the model's ability to fully specialize to domain-specific patterns and may reduce performance compared to more thoroughly fine-tuned architectures.
- **Restricted dataset usage and class imbalance:** Due to storage and compute limitations, we used only a 1/20th subset of the Quilt-1M dataset. While this reflects real-world constraints (where large, balanced medical datasets are often unavailable) it also means the model was trained on a narrower and more imbalanced sample distribution. Some pathology classes were significantly underrepresented, likely affecting performance on rare conditions.
- **Limited interpretability of predictions:** The current system outputs only a top-3 list of predicted pathology classes without any supporting visual or textual explanations. In clinical settings where accountability and traceability are essential, this lack of interpretability may limit user trust and hinder integration into diagnostic workflows.

Future work could address these limitations by leveraging more efficient fine-tuning techniques, expanding access to larger and more balanced datasets, and incorporating explainability features to support clinical decision-making.

Ethics

Privacy and Ethical Considerations: In line with privacy guidelines, no personally identifiable patient information is present in QUILT-1M. Only public YouTube video IDs are released, and sensitive speech segments were filtered out during the ASR correction stage. The authors also ensured that only narrative-style videos were used to minimize incidental background conversations.

Conclusion

Our approach delivers a transformative leap in histopathology classification: by jointly fine-tuning image and text encoders within a single semantic space and training a lightweight MLP head, we consistently outperform both uni-modal and zero-shot baselines across every key metric—achieving substantially higher micro-AUC and F1 scores, dramatically lower Hamming loss, and superior top-k accuracy. Crucially, this unified embedding framework enables real-time, robust multi-label predictions that fit seamlessly into digital pathology workflows. In practice, integrating our model as a clinical decision-support tool can accelerate diagnostic turnaround, reduce inter-observer variability, and provide junior pathologists with expert-level second-reads—ultimately driving more consistent diagnoses and better patient outcomes. Our results thus underscore the power of fine-tuned vision–language models to redefine what is possible in computational pathology.

Future Work: Future work could explore full encoder fine-tuning, advanced fusion methods (e.g., cross-attention), and training at scale using all parts of the Quilt-1M dataset and other related medical datasets. Incorporating explainability (e.g., Grad-CAM), model distillation for deployment, and human-in-the-loop learning would further enhance clinical usability and adaptability.

References

Ehteshami Bejnordi, B.; Veta, M.; Johannes van Diest, P.; and et al. 2017. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22): 2199–2210.

Guo, Y. 2024. Multimodal Multilabel Classification by CLIP. arXiv:2406.16141.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Ikezogwo, I. E.; Arnold, S. J.; Taiwo, J. E.; Onu, V. E.; and Ebiegberi, J. O. 2023. Quilt-1M: A Large-Scale Histopathology Vision–Language Dataset from Narrated Microscopy Videos. *Medical Image Analysis*, 85: 102857.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

van den Bent, M. J. 2010. Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. *Acta Neuropathologica*, 120(3): 297–304. Epub 2010 Jul 20.

Wortsman, M.; Ilharco, G.; Kim, J. W.; Li, M.; Kornblith, S.; Roelofs, R.; Gontijo-Lopes, R.; Hajishirzi, H.; Farhadi, A.; Namkoong, H.; and Schmidt, L. 2022. Robust fine-tuning of zero-shot models. arXiv:2109.01903.

Team Roles

→ Abhishek: Exploratory data analysis and research on CLIP model. Wrote report sections on abstract, introduction, dataset, model architectures, experimental setup and evaluation for initial approach, evaluation of final approach.

→ Aprup: Baseline models training and evaluation. Exploratory data analysis and data preprocessing. Research on multilabel classification.

→ Aditya: CLIP models training and evaluation. Research on CLIP and multimodal multilabel classification.

→ Vaishnav: Baseline and CLIP models training and evaluation. Data preprocessing. Research on CLIP and multimodal multilabel classification.