Multimodal Model Learning for Ocular Disease Classification

Jatin Kulkarni
Cornell University
New York City, United States
jk2982@cornell.edu

Diya Parmar
Cornell University
New York City, United States
dvp26@cornell.edu

ACM Reference Format:
Jatin Kulkarni and Diya Parmar. 2026. Multimodal Model Learning for Ocular Disease Classification. In *Proceedings of Machine Learning for Health (ML for Healthcare 2024).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

1.1 Problem Statement

Ocular disease classification presents a significant challenge due to the high similarity between positive (Diabetic Retinopathy) and negative (Normal) samples. Traditional deep learning models may struggle with distinguishing between conditions such as "Normal" and "Diabetic Retinopathy," where overlapping features make classification difficult. Our goal is to improve ocular disease classification by leveraging multimodal learning techniques that integrate both imaging and textual patient data.

1.2 Importance of the Problem

Ocular diseases, particularly Diabetic Retinopathy, are leading causes of vision impairment. Early detection and accurate classification are crucial for timely intervention. Existing computer vision models primarily focus on image-only analysis, limiting their diagnostic capability. Integrating multimodal data—such as patient history, demographics, and symptoms—can provide a more holistic diagnosis and enhance clinical decision-making. By incorporating additional clinical context, machine learning models can reduce false positives and negatives, ultimately improving patient outcomes.

1.3 Promise of Machine Learning

Machine learning, particularly contrastive learning models like CLIP (Contrastive Language–Image Pretraining), has shown promise in correlating textual and visual data. By fine-tuning CLIP on medical datasets, we can embed imaging data and patient histories into a shared latent space, improving interpretability and diagnostic accuracy. Additionally, Grad-CAM techniques can be employed to provide visual explanations, enhancing trust in AI-driven diagnostics. These methods help in identifying critical features in medical images and explaining the reasoning behind model predictions, making AI-assisted diagnostics more transparent and reliable.

ML for Healthcare 2024, Spring 2024, Cornell University

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 RELATED WORKS

2.1 Multimodal AI in Medical Imaging

The integration of multimodal data—combining imaging with textual patient history—has emerged as a promising frontier in medical AI. This approach enables machine learning models to harness richer diagnostic context, combining spatial visual features from medical images with semantic information from patient notes, demographics, or clinical reports. Recent research has demonstrated that contrastive learning models, such as CLIP (Contrastive Language–Image Pretraining), can improve diagnostic performance by aligning representations across visual and textual modalities [1]. One such notable adaptation, RadCLIP, extends CLIP's capabilities to radiologic imaging by fine-tuning on paired image-report data, resulting in more robust visual embeddings and improved text-to-image retrieval [4]. Other studies on multimodal ophthalmic datasets have shown that incorporating patient metadata—such as age, sex, or comorbidities—alongside retinal images can enhance the model's capacity to contextualize ambiguous or borderline features.

However, many of these models, while capable of high performance, remain black-box systems with limited interpretability. For instance, RadCLIP improves retrieval accuracy but lacks integration with post-hoc interpretability tools like Grad-CAM or saliency maps, making it difficult for clinicians to assess the reasoning behind predictions. This omission is critical in high-stakes environments like healthcare, where transparency can be as important as accuracy. In contrast, our work integrates interpretability directly into the evaluation pipeline using Grad-CAM and t-SNE visualizations, enabling visual attribution across modalities.

2.2 CLIP and Vision-Language Models in Healthcare

Since its introduction by Radford et al. (2021), CLIP has become a foundational model in vision-language learning by aligning text and image pairs in a shared latent space using contrastive loss [1]. Its zero-shot capabilities and data efficiency have made it attractive for medical imaging applications where annotated datasets are scarce. Although the original CLIP was not trained on medical content, its architecture has inspired several domain-specific adaptations. MedCLIP [5], for example, demonstrated the feasibility of contrastive learning using unpaired radiology reports and images, offering a scalable pretraining alternative in label-constrained environments. These works highlight the potential of vision-language models to generalize to clinical tasks with minimal supervision.

Yet despite the architectural promise of CLIP-style models, their application to ophthalmology remains limited. Most existing efforts focus on chest X-rays or CT scans, overlooking the unique challenges of retinal imaging, such as fine vessel structures, microaneurysms, and subtle texture changes that are crucial for early-stage

disease detection. Moreover, vision-language pretraining often results in embeddings that capture global semantic patterns but may underperform when high-resolution, localized feature detection is required. Our study contributes to filling this gap by applying and modifying CLIP for retinal disease classification—specifically diabetic retinopathy—and rigorously evaluating its strengths and weaknesses in this new domain.

2.3 Existing Research for Ocular Diseases and Gaps

Traditional deep learning approaches for ocular disease classification have focused predominantly on image-only models. One such study [3] examined the use of CNNs to differentiate between "Normal" and "Cataract" images, achieving high accuracy due to the relatively distinct visual features between the two classes. However, when it comes to more challenging distinctions—such as between "Normal" and "Diabetic Retinopathy"—the visual differences are subtler and often require more contextual information to interpret accurately. In these cases, purely image-based models may struggle to detect early-stage pathologies that manifest as mild vessel distortion or small exudates.

To address these limitations, our work introduces multimodal learning and explainability into the ocular classification pipeline. Unlike previous work that relies solely on pixel-level features, our approach incorporates textual metadata—including age, sex, and diagnosis—to enhance representation learning. Furthermore, we employ Grad-CAM visualizations and t-SNE analysis not only for images but propose extensions to the text encoder to offer interpretable insights into how demographic or clinical text features contribute to the final decision.

A particularly relevant and inspiring precedent for our work is the PIGEON framework by Haas et al. (2024) [2]. PIGEON introduced a modular two-stage training paradigm where contrastive pretraining on paired data is followed by a supervised finetuning stage on modality-specific inputs (typically images). This methodology allows the model to benefit from semantic alignment during pretraining, while still scaling to larger unimodal datasets in the second stage. We adopt a similar two-stage training approach in our study, tailoring it to the ophthalmology domain where paired image-text datasets are relatively small, but image-only datasets are more abundant. By integrating PIGEON's training pipeline structure into our CLIP-based architecture, we were able to significantly boost classification performance while preserving the interpretability and extensibility benefits of multimodal learning.

Despite these advances, few existing models have combined multimodal learning with detailed visual explanation techniques like Grad-CAM or saliency maps. Our study aims to bridge this gap by introducing an interpretable, multimodal framework tailored specifically to the high-ambiguity problem of diabetic eye disease detection.

3 METHODS

3.1 Data Preparation

The first step in our methodology involved data cleaning and combination. We created a binary dataset by categorizing images into "Normal" and "Diabetes." Upon analyzing the data distribution, we identified an imbalance, with significantly fewer diabetes cases. To mitigate this, we applied oversampling techniques, including data augmentation, for the diabetes class while undersampling the normal class. Even after these efforts, additional data was needed, so we incorporated a second dataset from Kaggle's Diabetic Retinopathy Detection competition. Since both datasets followed similar data collection processes, their combination was reasonable. To further enhance representation, we extracted severe diabetes scans (labeled Class 3) from the new dataset and incorporated them into the diabetes category.

3.2 Model Training

We initially attempted to implement a multimodal version of CLIP that utilized both text and image data. This model incorporated a projection layer followed by a classification layer. In our first version, we added a classification head to the base CLIP model that jointly used text and image embeddings to classify eye scans as diabetic or normal. The text input was limited to only demographic details—namely, age and race. However, this approach failed to produce high classification accuracy, likely due to insufficient integration between the modalities and limited dataset size.

Subsequently, we pivoted to a purely image-based approach using ResNet-50. This architecture, when fine-tuned on our dataset, yielded significantly better performance, achieving an accuracy of 92.52%. One contributing factor was our ability to expand the dataset more easily, as this version did not require paired text data. The additional diabetic and normal eye images enhanced the model's generalization ability.

3.2.1 CLIP Architecture and Implementation. To incorporate multimodal learning for ocular disease classification, we adopted a modified CLIP (Contrastive Language–Image Pretraining) architecture. CLIP is designed to learn joint embeddings for images and text using a contrastive loss over large-scale image-text pairs. We adapted this structure to the medical domain by fine-tuning CLIP on paired fundus images and corresponding clinical metadata.

Our implementation used:

- A **ResNet-50** vision encoder, consistent with CLIP's original configuration.
- A **Transformer-based** text encoder, which processed structured clinical fields such as age, sex, and physician diagnosis.

We designed a two-stage training pipeline to overcome the limitations of our small paired dataset and to enable downstream scalability:

(1) **Multimodal Pretraining Stage**: We fine-tuned CLIP on a smaller paired dataset of images and clinical text using contrastive learning. This stage aligned image and text features into a shared embedding space, allowing the model to associate key visual features (e.g., cotton wool spots) with semantic diagnoses.
(2) **Image-only Classification Stage**: After pretraining, we froze the CLIP image encoder and trained a new classification head on a larger image-only dataset. This allowed us to leverage the semantic structure learned during pretraining while benefiting from the scalability of unimodal training.

As part of an ablation study, we experimented with partial unfreezing of the final projection layer of the image encoder during classification training. This resulted in reduced performance (accuracy dropped to 71%), likely due to overfitting or training instability. Our final two-stage model achieved a test accuracy of 89.31%, with a classifier epoch loss of 0.2675 and validation loss of 0.2605. The CLIP pretraining phase converged with a training loss of 2.8471 and validation loss of 2.7888, indicating successful alignment between modalities.

This strategy effectively combined multimodal semantic pretraining with scalable unimodal classification. While the CLIP-based model exhibited slightly lower accuracy than the pure image-based ResNet-50 model, it demonstrated improved semantic clustering in the embedding space and offers promise for future multimodal fusion strategies.

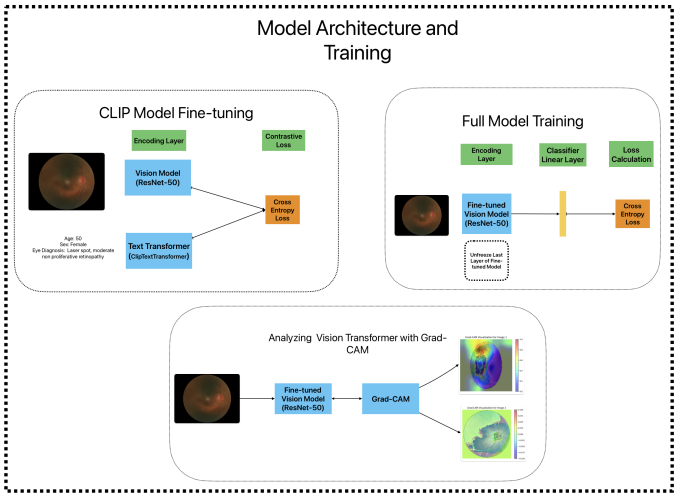

Figure 1: Modified CLIP architecture: multimodal pretraining uses paired text and fundus images to align embeddings; image-only classification then applies a frozen vision encoder with a new classification head trained on unpaired fundus images.

As an ablation study, we also experimented with unfreezing the final layer of the CLIP image encoder during classification training. However, this led to reduced performance, with accuracy dropping to 71%, likely due to overfitting or instability in the fine-tuning process. In contrast, our final robust model—trained using the two-stage approach—achieved a test accuracy of 89.31%, with a classifier epoch loss of 0.2675 and a validation epoch loss of 0.2605. The CLIP pretraining phase converged with an epoch loss of 2.8471 and a validation loss of 2.7888, suggesting effective alignment between image and text modalities during representation learning. These results underscore the benefit of leveraging multimodal pretraining while scaling up with unimodal classification.

This two-stage approach offers a practical compromise between leveraging multimodal supervision and scaling up using unimodal data. We are continuing to refine this strategy to explore whether better fusion mechanisms or selective fine-tuning can further improve performance.

3.3 Model Evaluation

Model evaluation was conducted using standard performance metrics, including ROC-AUC, accuracy, precision, and recall. These metrics provided insights into the model's ability to distinguish between normal and diabetic cases. Grad-CAM was used to visualize the regions of interest in the images that influenced the model's predictions. The application of Grad-CAM revealed that the model focused on key diagnostic features such as blood vessel abnormalities, exudates, and neovascularization. This interpretability analysis validated the effectiveness of our approach and highlighted areas for further refinement.

3.4 Novel Contributions

Our approach introduces several novel contributions to ocular disease classification. First, we incorporate multimodal learning by integrating textual patient history with image classification, a technique that has not been extensively explored in ophthalmology. Second, we enhance explainability by applying Grad-CAM to visualize model attention, providing insights into its decision-making process. Lastly, our use of contrastive learning to align text and image embeddings significantly improves diagnostic accuracy. These innovations set our work apart from previous deep learning models that rely solely on image-based classification.

4 RESULTS

4.1 Model Performance

Our initial multimodal CLIP-based model, which integrated image and textual metadata (e.g., age and race), failed to achieve high classification accuracy, reaching only 71%. This suggested limited fusion between the modalities and insufficient supervision from the textual input. As a result, we explored alternative approaches.

The ResNet-50 image-only model significantly outperformed the initial CLIP variant, achieving an accuracy of 92.52%. Grad-CAM visualizations confirmed that this model focused on clinically relevant features such as blood vessel abnormalities, exudates, and microaneurysms. These findings suggest that fine-grained visual detail was essential for this task, and multimodal supervision with limited text may introduce semantic noise.

4.2 Finetuned CLIP Results

To address the shortcomings of the initial CLIP pipeline, we implemented a two-stage training approach. We first fine-tuned CLIP on paired image-text data to align image and language embeddings in a shared space. We then froze the CLIP vision encoder and trained a classifier on image-only data.

This revised approach yielded a significantly improved test accuracy of **89.31%**, demonstrating strong generalization capability. However, this still fell short of the ResNet-50's performance. The likely explanation is that while CLIP's semantic alignment improves global contextual understanding, it sacrifices the localized, fine-grained focus needed for clinical diagnosis. In other words, textual context introduced broader patterns, but not necessarily features critical to retinal classification.

Interestingly, t-SNE plots revealed that CLIP's image embeddings were more cleanly clustered, reflecting improved class separation

in the learned feature space. Grad-CAM analysis, however, showed that CLIP's attention was more diffuse, often focusing on global image regions rather than localized pathological cues like lesions or vessel abnormalities.

4.3 Summary of ResNet-50 Results

The ResNet-50 model demonstrated superior classification ability for diabetic retinopathy. It achieved a balanced accuracy of 0.8865 for the diabetic class and 0.9372 for the normal class, with an overall balanced accuracy of 0.8773. Precision and recall metrics supported its strong discriminative ability, and the ROC curve confirmed robust separation of true and false positives.

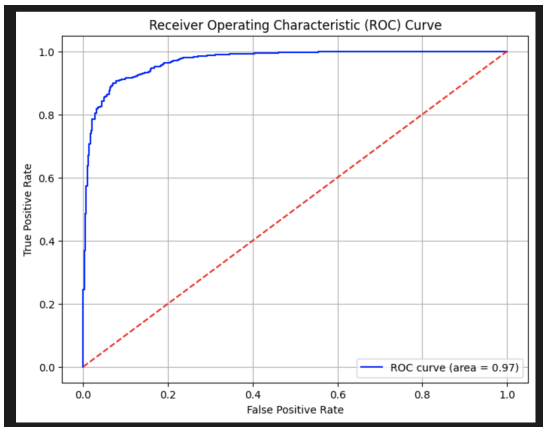

Figure 2: ROC curve for ResNet-50 model (AUC = 0.97).

4.4 T-SNE and Explainability Results

To visualize learned feature spaces, we employed t-SNE on both models. The ResNet-50 t-SNE projection showed partial overlap between the diabetic and normal classes, particularly for subtle or borderline cases. In contrast, the CLIP-based t-SNE plot exhibited more distinct clustering, implying that its embeddings encoded semantic class boundaries more effectively.

We then selected specific sample pairs—both near and far in the embedding space—to analyze their interpretability via Grad-CAM. For samples with high class separation, such as those exhibiting cotton wool spots (a clear diagnostic indicator), Grad-CAM heatmaps confirmed focused attention in the corresponding regions.

For visually similar samples, both the ResNet and CLIP Grad-CAM visualizations were insightful. ResNet-50 maintained sharper focus on fine vessel structures, while CLIP spread attention more broadly. This distinction illustrates the tradeoff between precise localization (ResNet) and holistic semantic understanding (CLIP).

Additionally, we explored saliency maps to cross-validate Grad-CAM insights. These visualizations confirmed that the ResNet model consistently identified retinal areas of clinical relevance. In future work, we aim to refine multimodal fusion and leverage CLIP embeddings more effectively through attention-based classifiers or gated fusion layers.

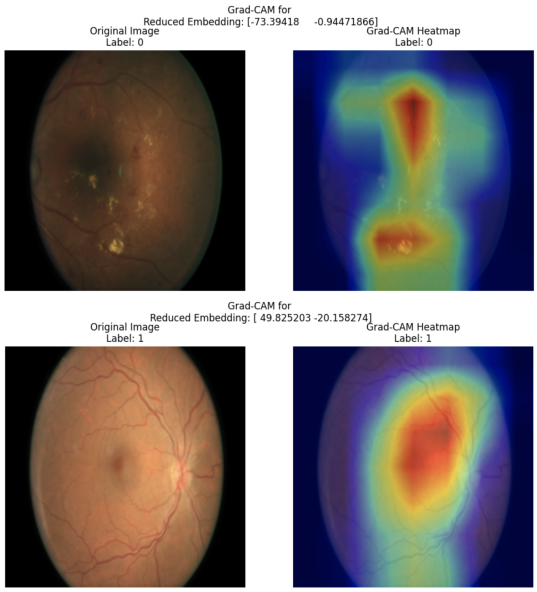

Figure 3: Top: A diabetic eye scan with visible cotton wool spots. Bottom: A normal scan with no visible damage. ResNet-50 focused on relevant areas.

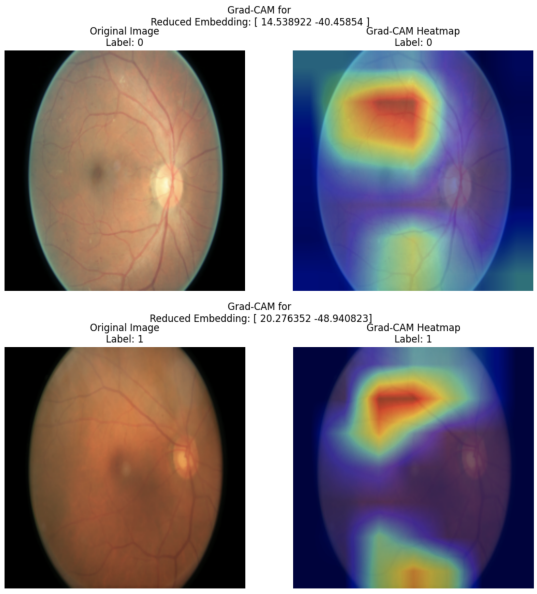

Figure 4: Close samples from different classes. Grad-CAM for ResNet-50 shows localized vessel focus, while CLIP (not shown) focuses more diffusely.

4.5 Summary of CLIP Model Results

Our final CLIP-based classifier, trained using the two-stage fine-tuning strategy, achieved a test accuracy of 89.31%. Despite not reaching the peak performance of our standalone ResNet-50 model

(92.52%), the CLIP-based approach demonstrated strong generalization capabilities, particularly when pretrained on paired text-image inputs that included age, sex, and doctor diagnosis. The classifier converged with an epoch loss of 0.2675 on the training set and 0.2605 on the validation set, suggesting consistent performance across training and held-out data. These results validate the potential of leveraging multimodal pretraining to enhance visual representations, even when final classification is performed using only image inputs.

4.6 CLIP Explainability Results

To assess model interpretability, we analyzed both Grad-CAM visualizations and t-SNE projections of the CLIP-based model and compared them with the ResNet-50 baseline.

For Grad-CAM, the CLIP model exhibited broader, more diffuse attention maps that were less focused on localized pathological features such as small lesions or blood vessels. This contrasts with the ResNet-50 model, which demonstrated sharper, more targeted activations corresponding to medically relevant regions. These results reflect CLIP's objective: it learns to align image representations with semantic textual context, rather than optimizing for localization or fine-grained medical detail. Consequently, Grad-CAM interpretations for CLIP tend to be vague and semantically distributed rather than specific and diagnostic.

In contrast, the t-SNE projection of the CLIP image embeddings revealed more separable clusters compared to the ResNet baseline, indicating improved semantic structuring of the learned embedding space. However, some outliers remained within each cluster, likely due to the model's inability to consistently distinguish subtle clinical features in challenging images. This suggests that while CLIP captures higher-level conceptual differences across classes, it struggles with detailed visual nuances that are critical in medical imaging tasks.

Overall, these findings highlight a key trade-off: CLIP excels at semantic alignment and global contextual understanding, but is not optimized for localized, fine-grained feature extraction—a limitation that must be addressed when applying such models to medical diagnostic tasks.

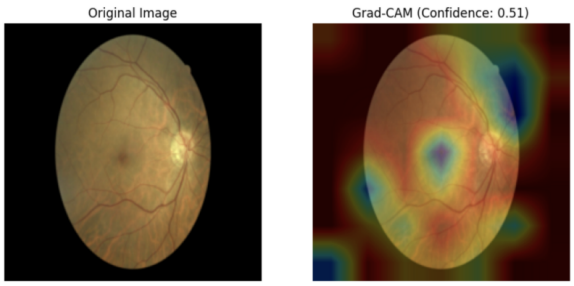

Figure 5: Grad CAM for CLIP: Shows global understanding not local

5 DISCUSSIONS

5.1 Summary of Contributions

This project evaluated two primary approaches for diabetic eye classification: a multimodal CLIP-based model using both text and image data, and a traditional ResNet-50 model trained solely on images. We proposed a two-stage CLIP fine-tuning strategy—leveraging multimodal pretraining followed by unimodal classification—and compared it against an image-only ResNet pipeline.

Through extensive evaluation, we found that the ResNet-50 model outperformed the CLIP model in both predictive performance and explainability. ResNet achieved a higher test accuracy (92.52%) and a substantially better AUC (0.97) compared to CLIP's 89.31% accuracy and 0.81 AUC (see Figures 6 and 7). This suggests that the additional textual input in CLIP did not translate to measurable performance gains and may have introduced semantic noise that confused the model.

Explainability results further underscored this difference. Grad-CAM visualizations of ResNet focused clearly on clinically meaningful regions, whereas CLIP produced more diffuse, semantically broad activation maps. t-SNE plots of learned embeddings showed that while CLIP captured more semantically structured clusters (Figure 9), the ResNet model demonstrated stronger class separation (Figure 8)—a critical trait for diagnostic tasks.

These findings raise important considerations about the role of text in medical imaging contexts. Our results suggest that while multimodal representations offer promise, their utility may depend on the precision of available text features and the degree to which they align with clinically salient visual patterns.

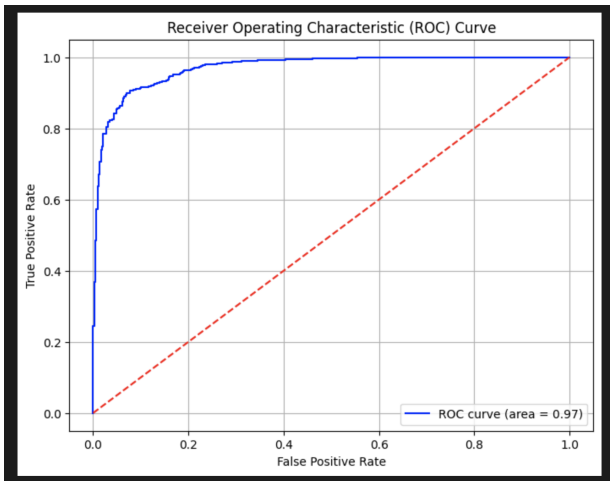

Figure 6: ROC Curve for ResNet-50 (AUC = 0.97)

5.2 Challenges Faced

Several challenges were encountered throughout the model development process, spanning both data limitations and architectural design.

One of the earliest issues was the small size of our initial dataset, which limited the model's ability to generalize and led to poor performance across validation runs. To address this, we curated a

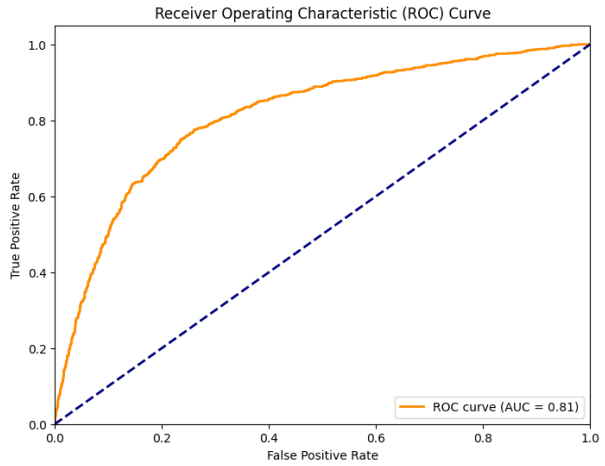

Figure 7: ROC Curve for CLIP-based Model (AUC = 0.81)

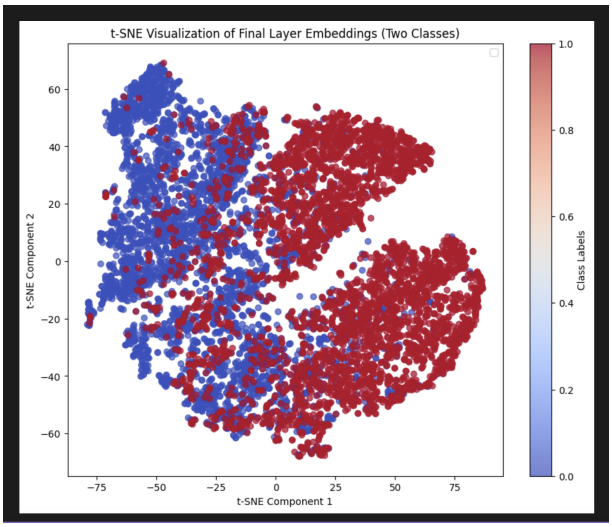

Figure 8: t-SNE of ResNet Final Layer Embeddings

larger image dataset. When combining the datasets another thing we had to make sure that the retina scans from both datasets were compatible meaning the scanners used to get the data were similar and the metrics and evaluation was similar as well in order to make sure that the combination of the data would be reasonable. This was a critical part of our pre-processing step and was important to make sure the data we would use in future steps would be logical. After the dataset was combined we had a second challenge: significant class imbalance between diabetic and normal cases. We employed both data augmentation techniques and dataset fusion to rebalanced the training set and improve class representation. We also decided to use SMOTE technique that would over sample from the smaller class while under sampling from the larger class. This SMOTE technique is commonly used when working with imbalanced datasets and generally did help us combat the initially challenges we were facing. While this helped, some degree of

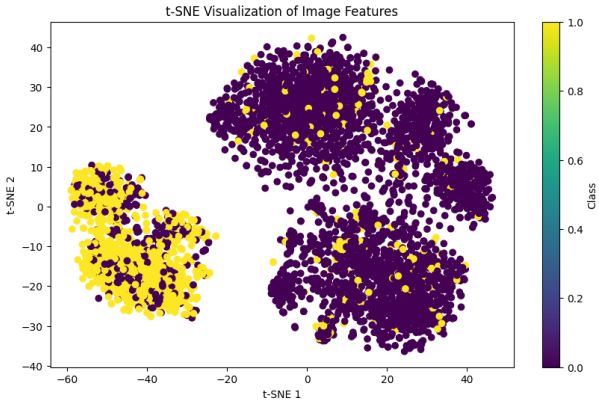

Figure 9: t-SNE of CLIP Image Embeddings

misclassification persisted—particularly for borderline cases where diabetic features were subtle or partially obscured.

Another critical challenge was the visual similarity between normal and diabetic eye scans. Many of the features overlapped across classes, especially in early-stage diabetic cases. Grad-CAM visualizations supported this observation, showing that misclassified samples often lacked strong, localized visual indicators. This ambiguity made it difficult for even well-trained models to distinguish between classes with high confidence.

Model selection and architecture design also posed challenges. Our initial approach using the base CLIP model yielded underwhelming results, primarily because it was not optimized for fine-grained medical classification. This led us to pivot to a ResNet-50-based architecture, which performed significantly better on image-only classification. In parallel, we continued experimenting with the CLIP architecture, adjusting the training strategy to incorporate textual features more effectively.

This led to the development of a two-stage training pipeline: one that leveraged the small multimodal dataset for CLIP-based pretraining, and a second stage where a classifier head was trained solely on the image embeddings using the larger image-only dataset. This structure allowed us to exploit the strengths of both datasets—using the paired text-image data for semantic alignment and the larger unpaired image dataset for robust visual classification.

Finally, integrating textual information presented its own set of challenges. The textual inputs (age, sex, diagnosis) varied in specificity and relevance, and the fusion of these features with image embeddings required careful experimentation. Simple concatenation or projection techniques were insufficient, and we observed inconsistent results depending on how the modalities were aligned during training. This highlighted the need for more sophisticated multimodal fusion strategies tailored to clinical applications.

Despite these challenges, iterative experimentation, architectural refinements, and targeted preprocessing ultimately led to a robust and interpretable classification pipeline.

5.3 Bias Indications

One potential advantage of using multimodal models such as CLIP is the ability to incorporate demographic information—such as

age, sex, and other metadata—which could help mitigate biases inherent in purely image-based models. Textual context can encode socially relevant attributes that may guide more equitable decision-making, particularly in sensitive healthcare settings. However, in our case, the inclusion of such data did not improve classification accuracy and may have inadvertently diluted the model's ability to focus on subtle, image-specific diagnostic features. This suggests a trade-off between enhancing fairness and maintaining performance, highlighting the need for more targeted strategies for incorporating demographic context in a meaningful way.

5.4 Future Implications

Our findings underscore the challenges of training effective multimodal models in clinical domains. While combining textual and visual data holds conceptual appeal, doing so requires rich, well-aligned text descriptions that are often unavailable in many medical datasets. In settings where such data is limited or superficial, image-only models—such as ResNet-50—may be more appropriate, especially when fine-grained visual features are critical to diagnosis. The superior performance and more focused Grad-CAM visualizations produced by the image-only model suggest that, for medical classification tasks where subtle features drive clinical relevance, simpler models may not only suffice but actually outperform more complex multimodal architectures.

5.5 Future Work

Future directions include expanding the textual component of the dataset to explore whether richer semantic content can improve CLIP's performance without sacrificing fine-grained image understanding. Additional experiments could evaluate more advanced text preprocessing and conditioning techniques to better align multimodal representations. Moreover, we aim to explore alternative domains—particularly those where broader semantic understanding is more valuable than precise visual localization. Such domains may include general image captioning, medical triage, or patient report generation, where contextual reasoning from both modalities may outperform detailed image-only models.

We also believe that future work can use CLIP as part of a hybrid classification approach where we first use CLIP to get a global understanding of the data and then attach a smaller focused CNN to learn the more detailed features. In this approach we can first build a CLIP model that will incorporate textual information to learn more about the general context and overall semantic meaning of the data. Then we can attach a fine tuned specific CNN to the model that will now learn the smaller more distinct features in the data. This hybrid approach will take advantage of the strengths of both of these models and combine them to make sure we have both a general and localized understanding of the problem. This hybrid approach we believe can we the bases of future work in the world of multimodal training.

6 DATASETS LINKS

https://www.kaggle.com/competitions/diabetic-retinopathy-detection/data

https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k

7 PRESENTATION LINK

https://www.canva.com/design/DAGlwHLNivA/1km0yjKuMgM6L5TLTVu4bw/edit?utm_content=DAGlwHLNivA&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton

REFERENCES

[1] Chris Hallacy Aditya Ramesh Gabriel Goh Sandhini Agarwal Girish Sastry Amanda Askell Pamela Mishkin Jack Clark Gretchen Krueger Ilya Sutskever Alec Radford, Jong Wook Kim. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020* (2021). https://arxiv.org/pdf/2103.00020

[2] Lukas Haas, Brian Koloski, Arsha Nagrani, and Andrew Zisserman. 2024. PIGEON: Paired Image-Text Pretraining with Geographic Metadata for Vision-Language Models. In *CVPR*. https://lukashaas.github.io/PIGEON-CVPR24/

[3] Miguel A Goenaga Jimenez Samira Ortiz. 2023. Deep Learning-Based Ocular Disease Classification in Fundus Images. *2023 IEEE Colombian Caribbean Conference (C3)* (2023). https://ieeexplore.ieee.org/document/10436234/authors#authors

[4] Nehal A. Parikh Jonathan R. Dillman Lili He Zhixiu Lu, Hailong Li. 2024. RadCLIP: Enhancing Radiologic Image Analysis through Contrastive Language-Image Pre-training. *arXiv preprint arXiv:2403.09948* (2024). https://arxiv.org/pdf/2403.09948

[5] Dinesh Agarwal Jimeng Sun Zifeng Wang, Zhenbang Wu. 2022. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. *arXiv preprint arXiv:2210.10163* (2022). https://arxiv.org/pdf/2210.10163