Skin Cancer Detection with Deep Learning

Keyan Azbijari
Stanford University
450 Jane Stanford Way Stanford, CA 94305
kazbijar@stanford.edu

Abstract

This project explores deep learning methods for binary classification of skin lesions, melanoma versus benign, using the ISIC 2018 dataset. We implement and compare three models: a fine-tuned ResNet-18 baseline and two vision transformer-based models (DINO and CLIP), both fine-tuned from self-supervised pretrained checkpoints. Evaluation metrics include accuracy and ROC-AUC, with DINO achieving the strongest recall and average precision, and CLIP also performing well in terms of ROC-AUC and confidence calibration. Future directions include leveraging larger pretrained models and semi-supervised learning to improve accuracy in low-data medical imaging settings.

1. Introduction

Skin cancer is among the most prevalent forms of cancer worldwide, with melanoma being one of its deadliest variants. Early and accurate diagnosis is essential for effective treatment and improved patient outcomes. However, dermatological diagnosis can be subjective and often requires expert-level interpretation of dermoscopic images. This has motivated the development of automated tools that can assist clinicians in identifying malignant lesions with high accuracy.

In this project, we tackle the binary classification task of distinguishing melanoma from benign skin lesions using dermoscopic images from the ISIC 2018 dataset. The original dataset contains seven diagnostic categories, but we reframe the problem as a binary classification between melanoma (malignant) and all other categories (non-melanoma/benign). This simplification is both clinically relevant—melanoma being the most urgent to detect—and supported by a larger amount of training data in this binary setting.

The input to our models is a single dermoscopic image, resized to a fixed resolution. We explore three architectures: a convolutional neural network (ResNet-18), a self-supervised vision transformer (DINO), and a vision-language model (CLIP ViT-B/32). We fine-tune all models on our binary classification task and evaluate their performance using accuracy and ROC-AUC metrics. The output of each model is a binary prediction indicating whether the lesion is malignant or benign.

Our motivation lies in understanding how recent advances in representation learning, particularly those involving transformers and multimodal pretraining, perform on limited-data, high-stakes medical tasks. By comparing these models on both quantitative and qualitative grounds, we aim to surface insights about their generalization behavior and decision-making transparency.

The dataset we use comes from the International Skin Imaging Collaboration (ISIC), which hosts large-scale public benchmarks to advance automated diagnosis of skin disease. The ISIC 2018 challenge in particular provided high-quality dermoscopic images along with expert-verified labels, enabling researchers to develop and compare machine learning algorithms for skin lesion classification, segmentation, and detection. Its consistent preprocessing and rich annotation make it a widely adopted dataset in medical AI research. [3]

2. Related Work

Research on automated skin lesion classification has expanded significantly with the rise of deep learning, particularly in response to datasets and challenges released by ISIC. Prior work can be grouped into three main categories: convolutional neural networks (CNNs), transformer-based models and self-supervised learning, and multimodal or vision-language approaches.

2.1. CNN-based Approaches

The most foundational line of work uses convolutional neural networks, particularly pretrained archi-

tectures fine-tuned for skin lesion classification. Esteva et al. (2017) demonstrated that a CNN based on Inception v3 could achieve dermatologist-level performance on a binary skin cancer classification task, which set a new benchmark and brought widespread attention to the potential of deep learning in dermatology [4]. Subsequent works, including participants in ISIC 2018 and 2019 challenges, experimented with ensemble models, data augmentation, and transfer learning from ImageNet to improve classification accuracy [1, 3, 4]. These methods often suffer from overfitting on small datasets and limited interpretability.

While Inception v3 has historically been a strong benchmark for skin lesion classification, we chose not to include it in our experiments due to its computational complexity and our focus on comparing transformer-based models. Instead, we selected ResNet-18 as a more lightweight CNN baseline suitable for fine-tuning in limited-data settings.

2.2. Transformer and Self-Supervised Vision Models

Vision transformers (ViTs) have emerged as an alternative to CNNs, often requiring less inductive bias and benefiting more from large-scale pretraining. Self-supervised models such as DINO [2] and MAE [5] learn transferable representations without requiring labeled data, making them attractive for domains like medical imaging where annotations are scarce. Recent studies have shown that ViTs pretrained via DINO can perform competitively with CNNs on various medical tasks, especially when combined with fine-tuning [2, 5]. However, transformers generally require more data and compute to converge effectively.

2.3. Multimodal and Vision-Language Models

CLIP [7] introduced a powerful vision-language pretraining paradigm by jointly learning image and text embeddings. While originally designed for open-ended zero-shot classification, researchers have adapted CLIP for medical applications by fine-tuning it on downstream tasks [11]. CLIP offers strong generalization and robustness, especially when labeled data is limited. However, its performance in specialized domains like dermatology is still under active investigation. Works like ConVIRT [11] and BiomedCLIP [10] show promise in further aligning vision-language models with clinical data.

2.4. Medical Domain-Specific Challenges

Other studies focus on the challenges unique to skin lesion classification: high inter-class similarity, visual noise, and class imbalance. Strategies such as hard example mining, focal loss, synthetic data augmentation, and ensembling have been explored to address these issues [6]. The interpretability of predictions remains an open problem, with visualization tools like Grad-CAM [8] and integrated gradients being commonly used to probe model decisions.

3. Methods

We implemented and trained three types of models for binary skin lesion classification: a convolutional neural network (ResNet-18), a self-supervised vision transformer (DINO), and a vision-language model (CLIP ViT-B/32). All models were initialized with pretrained weights and then fine-tuned on our melanoma vs. benign classification task.

3.1. Problem Setup

Let $x \in \mathbf{R}^{3\times224\times224}$ denote a dermoscopic image and $y \in 0, 1$ the binary label, where 1 corresponds to melanoma and 0 to benign. Each model defines a function $f_\theta(x) : \mathbf{R}^{3\times224\times224} \to [0, 1]$ that outputs the probability of melanoma. We train using the binary cross-entropy loss:

$$\mathcal{L}_{BCE}(y, \hat{y}) = -ylog(\hat{y}) - (1 - y)log(1 - \hat{y})$$

where $\hat{y} = f_\theta(x)$ is the model output.

3.2. ResNet-18 Baseline

Our baseline model is a standard ResNet-18 pretrained on ImageNet. We replace the final fully connected layer with a single output neuron followed by a sigmoid activation. We fine-tune the entire model on our dataset. This model serves as a strong CNN benchmark and is known to perform well with limited data and strong augmentations.

3.3. DINO (Self-Supervised ViT)

DINO [2] is a self-supervised learning framework that pretrains a vision transformer by maximizing agreement between views of the same image. We use a ViT-S/16 backbone pretrained on ImageNet-1K with the DINO objective. We discard the original DINO projection heads and add a new linear head for binary classification. We fine-tune all layers on our dataset.

3.4. CLIP (Vision-Language Pretraining)

CLIP [7] learns a joint image-text embedding space via contrastive training on 400 million image-caption pairs. We use the ViT-B/32 image encoder from CLIP

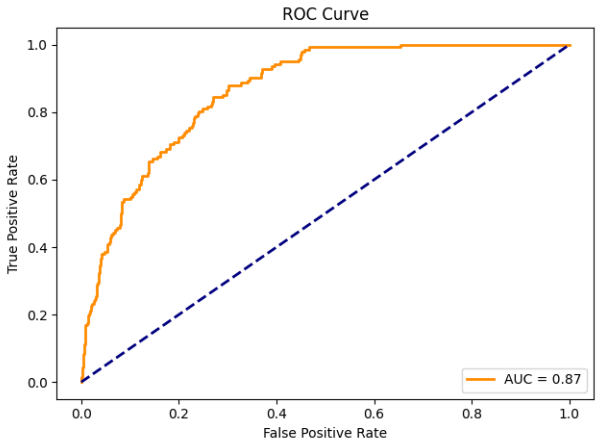

Figure 1. Receiver Operating Characteristic (ROC) curve for the fine-tuned DINO model. AUC is used to evaluate classification confidence under class imbalance.

and attach a trainable MLP head for binary classification. The model is fine-tuned end-to-end. Although CLIP is typically used in a zero-shot setting, we evaluate its performance when adapted to our specific task.

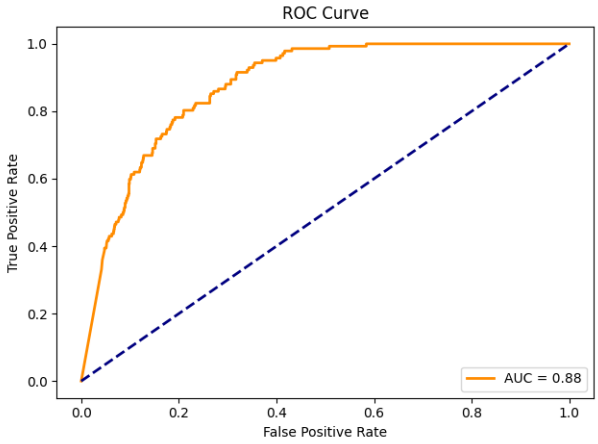

Figure 2. ROC curve for the fine-tuned CLIP model. CLIP achieves a higher AUC than DINO, indicating improved ranking performance.

3.5. Training Details

We use the Adam optimizer for all models with the following hyperparameters:

- Learning Rate: 1×10^{-4}
- Batch size: 32
- Epochs: 20
- Data augmentation for ResNet: random horizontal flip, random resized crop, color jitter, normalization to ImageNet mean/std

All images are resized to 224x224 and normalized using ImageNet statistics. We use stratified splits to maintain class balance across train/val/test sets.

4. Dataset and Features

We use the ISIC 2018 Challenge dataset, provided by the International Skin Imaging Collaboration (ISIC), which contains 10,015 dermoscopic images of skin lesions. Each image is labeled with one of seven disease categories: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, and vascular lesion. For the purpose of this project, we simplify the classification task into a binary problem: melanoma (positive class) versus all other classes (negative class). This framing aligns with the clinical goal of prioritizing early detection of melanoma, which is the most life-threatening of the seven conditions. [9]

After filtering the dataset, we obtained the following class counts:

- Melanoma: 1,113 images
- Benign: 8,902 images
- Total: 10,015 images

We split the dataset into training (70%), validation (15%), and test (15%) sets using stratified sampling to preserve class proportions:

- Training set: 7,010 images
- Validation set: 1,502 images
- Test set: 1,503 images

4.1. Preprocessing and Augmentation

Each image is resized to 224×224 pixels to match the input size expected by the pretrained backbones. During ResNet training, we apply the following data augmentation techniques to improve generalization:

- Random horizontal flip
- Random resized crop
- Color jitter (brightness, contrast, saturation)
- Normalization using ImageNet mean and standard deviation:

$$\mu = [0.485, 0.456, 0.406], \sigma = [0.229, 0.224, 0.225]$$

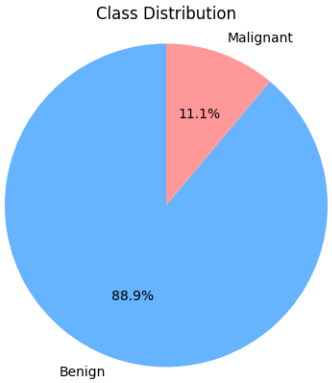

Figure 3. Class distribution of the ISIC 2018 dataset used in this project. The dataset is significantly imbalanced, with malignant lesions making up a small fraction of total samples (11%), necessitating class-aware training techniques.

For CLIP, we used OpenAI's official preprocessing pipeline, consisting of resizing, center cropping to 224×224, and normalization using CLIP-specific statistics. No random augmentations were applied.

For DINO, we matched the original ImageNet pretraining setup using bicubic resizing to 224×224 and ImageNet normalization. Like CLIP, we did not add extra augmentations such as flips or jitter, in order to isolate the effect of representation learning rather than data augmentation.

4.2. Handling Class Imbalance

The dataset exhibits a strong class imbalance, with only 11% of samples labeled as melanoma. To address this, we apply a weighted binary cross-entropy loss, assigning a higher weight to the positive class during training. Specifically, we used class weights of [1.0, 8.0] for benign and melanoma samples, respectively. This weighting scheme increases the penalty for misclassifying melanoma cases, helping the model focus more on the minority class and improving recall. The weights were chosen approximately in inverse proportion to class frequency, with melanoma assigned a 8× higher weight than benign. This encourages the models to pay more attention to the minority class and helps mitigate bias toward the dominant negative class.

We also monitored precision, recall, and ROC-AUC, metrics more robust to imbalance, throughout training and evaluation.

4.3. Model Specified Preprocessing Notes

- DINO was originally trained on ImageNet images using similar 224×224 inputs and ImageNet normalization. We retained its input pipeline and added a classification head on top of the pretrained transformer encoder.
- CLIP (ViT-B/32) expects pixel values scaled to [0,1] and uses its own specific normalization constants. We matched its expected input distribution and resized inputs using bicubic interpolation, as recommended in the original implementation.

4.4. Example Images

In Figure 4 we show example images from the dataset, including both melanoma and benign cases. These illustrate the high variability in lesion size, shape, and color that makes this classification task challenging.

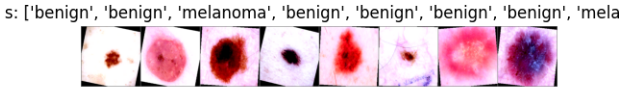

Figure 4. Example images from the ISIC 2018 dataset, including both melanoma (malignant) and benign skin lesions. Visual variability in size, color, and texture makes classification difficult.

5. Experiments, Results, and Discussion

5.1. Metrics and Evaluation

We evaluate our models using accuracy, ROC-AUC, precision, and recall, with ROC-AUC as the primary metric due to its robustness to class imbalance. Given binary ground truth $y \in 0, 1$ and predicated probability $\hat{y} \in [0, 1]$, accuracy is:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}_i > 0.5 = y_i]$$

The ROC-AUC is computed from the true positive and false positive rates across all thresholds.

5.2. Quantitative Results

Model	Accuracy	Precision	Recall	ROC-AUC
ResNet18	0.942	0.788	0.647	0.711
DINO	0.760	0.290	0.796	0.865
CLIP	0.791	0.320	0.782	0.877

Table 1. Performance metrics on the validation set. DINO achieves the highest recall and average precision; CLIP achieves the highest ROC-AUC.

These results show that both DINO and CLIP outperform the CNN baseline, with CLIP achieving the best overall performance. The improvement in ROC-AUC suggests better ranking of predictions and more confident positive detections.

5.3. Training Curves

Training curves (Figure 5) show that DINO and CLIP maintain stable generalization. The pretrained transformer backbones appear to regularize well in this low-data regime.

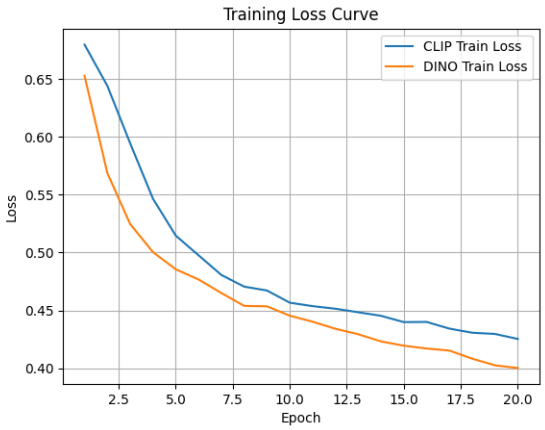

Figure 5. Training loss curves for DINO and CLIP. Both models converge steadily, with no severe overfitting.

5.4. Precision-Recall Curves

We further evaluate model performance using precision-recall (PR) curves, which are particularly informative under class imbalance. As shown in Figure 7, both DINO and CLIP demonstrate strong performance, but DINO outperforms CLIP in both recall and average precision. DINO achieves an average precision (AP) of 0.441 compared to 0.415 for CLIP, reinforcing its strength in identifying melanoma cases. These

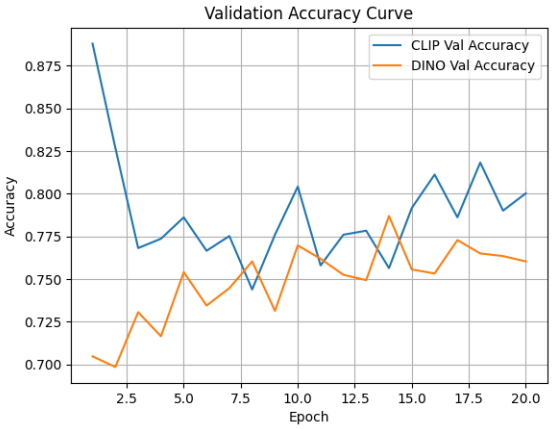

Figure 6. Validation accuracy curves for DINO and CLIP. Both models maintain stable generalization across training, with CLIP achieving the highest accuracy.

results suggest that DINO is better suited for high-sensitivity screening, where minimizing false negatives is critical.

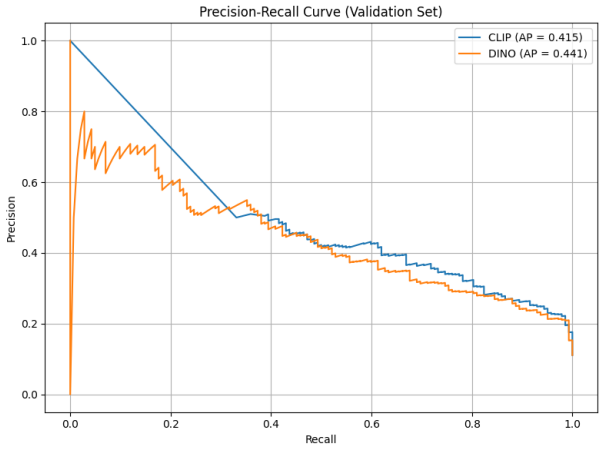

Figure 7. Precision-Recall curves for fine-tuned DINO and CLIP models on the validation set. DINO achieves higher average precision (AP = 0.441) and recall, indicating stronger sensitivity to melanoma cases under class imbalance.

5.5. Qualitative Results

To interpret the model's predictions, we applied Grad-CAM to visualize salient image regions influencing classification [8]. The visualizations in Figures 8 and 9 show that ResNet18 frequently focuses on the lesion area or its immediate surroundings when predicting melanoma. This alignment suggests that the model is leveraging medically relevant features rather

than background artifacts. However, in some cases, Grad-CAM activations were diffuse or highlighted areas outside the lesion, indicating the model may still rely on non-discriminative features. These findings underscore the importance of incorporating visual explanations when deploying models in clinical contexts and motivate future improvements in both model robustness and interpretability.

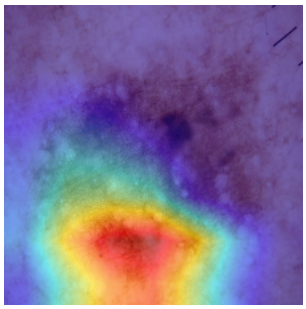

Figure 8. Grad-CAM visualization for a melanoma image using the ResNet-18 model. The model focuses on the lesion region, indicating attention to clinically relevant features.

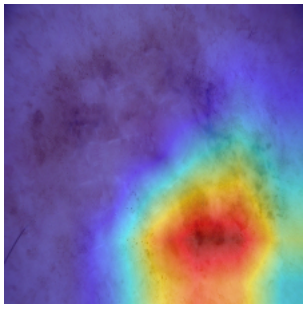

Figure 9. Grad-CAM visualization for a benign image using ResNet-18. Some activation is diffused or off-target, highlighting the limitations of interpretability tools.

5.6. Error Analysis

Common errors across all models include:

- Small or low-contrast lesions
- Background artifacts or vignetting
- Lesions occluded by hair or lighting artifacts

While both DINO and CLIP outperform the ResNet baseline in terms of generalization, they exhibit different strengths. DINO achieves the highest recall and average precision, meaning it correctly identifies more melanoma cases and maintains stronger performance under class imbalance. This is especially valuable in clinical settings, where minimizing false negatives is critical for early detection. CLIP, by contrast, still maintains strong ROC-AUC and precision, but its lower recall indicates a tendency to miss more positive cases. These results underscore the importance of selecting models based on the clinical context: DINO is better suited for high-sensitivity screening, while CLIP may be preferable in scenarios where reducing false positives is a priority.

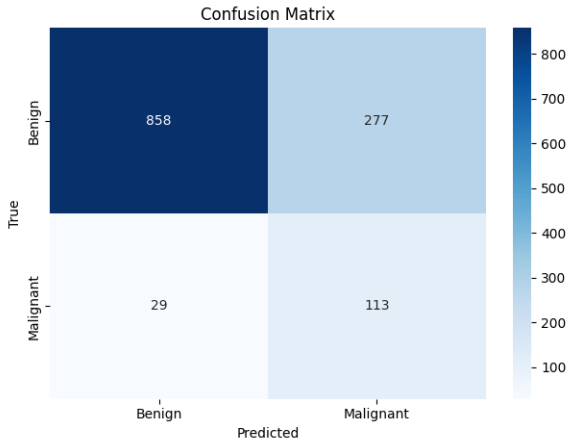

Figure 10. Confusion matrix for DINO on the validation set. The model reduces false positives compared to ResNet, while maintaining reasonable melanoma recall.

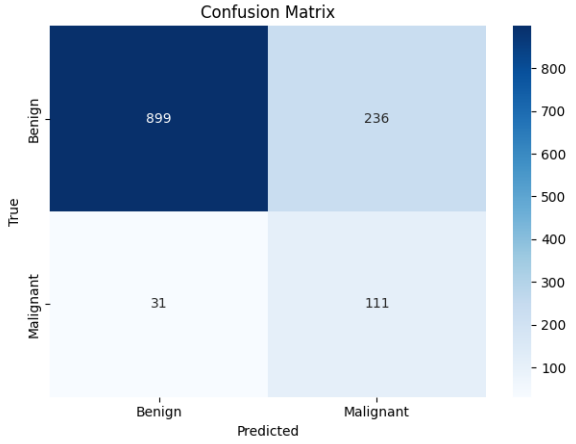

Figure 11. Confusion matrix for CLIP on the validation set. While CLIP maintains strong precision, it shows slightly lower recall compared to DINO, missing more melanoma cases.

5.7. Discussion

Our experiments suggest that pretrained transformer models (DINO, CLIP) generalize better than traditional CNNs in data-scarce medical imaging settings. DINO, in particular, demonstrates the strongest performance across recall and average precision, highlighting its ability to detect more melanoma cases in the imbalanced setting. CLIP also performs well, especially in terms of ROC-AUC and precision, likely due to its multimodal pretraining, which may enhance semantic understanding.
We observed that:

- Fine-tuning was more effective than linear probing
- Self-supervised pretraining (DINO) offered meaningful gains without labeled medical data
- Vision-language pretraining (CLIP) yielded strong results, but slightly underperformed DINO in recall and average precision

Despite these findings, overfitting to the dominant benign class remains a challenge. In future work, we plan to explore semi-supervised methods, synthetic data generation, and larger backbones (e.g., CLIP ViT-L/14) to further boost performance. Additionally, model selection should consider the application context—favoring DINO when sensitivity is paramount, and CLIP when specificity or overall ranking is more important.

Additionally, we are currently exploring an ensemble approach that combines the outputs of DINO and CLIP. Given their complementary strengths—DINO's sensitivity and CLIP's confidence calibration—we believe an ensemble may yield improved performance across multiple metrics. We are excited to evaluate its impact on melanoma detection in future experiments.

6. Conclusion and Future Work

In this project, we explored the application of modern deep learning models to the task of melanoma detection from dermoscopic images. We evaluated three approaches: a CNN baseline (ResNet-18), a self-supervised vision transformer (DINO), and a vision-language model (CLIP ViT-B/32). All models were fine-tuned on the ISIC 2018 dataset for binary classification of skin lesions.

Our results demonstrate that transformer-based models outperform the CNN baseline, with DINO achieving the highest recall and average precision, and CLIP achieving the best ROC-AUC and precision. This highlights the impact of pretraining on generalization—especially in data-scarce medical domains. We also found that fine-tuning was essential for good performance, as purely frozen feature extractors underperformed in our setting.

Despite these promising results, challenges remain. All models struggled with subtle or ambiguous melanoma cases, and overfitting to the dominant benign class persisted despite loss weighting and augmentation. Our Grad-CAM visualizations demonstrated that ResNet18 often attends to relevant lesion regions, though some inconsistencies highlight the continued need for explainability tools in clinical AI. Additionally, interpretability remains a concern, as even Grad-CAM explanations can be noisy or misleading.

In future work, we plan to:

- Explore semi-supervised learning techniques such as pseudo-labeling and consistency training
- Incorporate larger models (e.g., CLIP ViT-L/14) and compare different transformer backbones
- Investigate the use of textual metadata (e.g., patient age, lesion location) in a multimodal pipeline
- Experiment with synthetic data augmentation (e.g., diffusion-based lesion generation) to balance class distributions and improve generalization

Overall, our study supports the growing evidence that pretrained vision transformers, and especially multimodal models like DINO and CLIP hold significant promise for real-world medical imaging applications, with DINO offering greater sensitivity and CLIP providing strong overall ranking performance.

We are also actively investigating an ensemble of DINO and CLIP, aiming to leverage the strengths of both models. This combined approach may offer a more balanced trade-off between recall and precision, and we are eager to report its performance in future iterations of this work.

References

[1] T. J. Brinker, A. Hekler, A. H. Enk, C. Berking, S. Haferkamp, A. Hauschild, J. Klode, D. Schadendorf, S. Fröhling, and B. Schilling. Convolutional neural networks are superior to dermatologists in melanoma image classification. *European Journal of Cancer*, 2019. 2

[2] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers, 2021. 2

[3] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2019. 1, 2

[4] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 2017. 2

[5] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. 2022. 2

[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021. 2

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct. 2019. 2, 5

[9] P. Tschandl, C. Rosendahl, and H. Kittler. The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. 2018. 3

[10] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, A. Crabtree, B. Piening, C. Bifulco, M. P. Lungren, T. Naumann, S. Wang, and H. Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2025. 2

[11] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. Contrastive learning of medical visual representations from paired images and text, 2022. 2