Skin Cancer Detection from Smartphone Camera Images via Transfer Learning

Darren Chan
dchan04@stanford.edu

Yuxiang ”Jack” Zhang
jyxzhang@stanford.edu

Flora Yuan
floray@stanford.edu

Abstract

Skin cancer is one of the most common and deadly cancers worldwide, yet early detection significantly improves patient outcomes. While state-of-the-art computer vision models for skin lesion classification have shown strong performance, they are predominantly trained on dermoscopic images—captured by dermatologists using specialized equipment during clinical exams, limiting their accessibility in everyday settings. In this work, we aim to close the domain gap by investigating whether skin lesion models can be adapted to perform effectively on images captured with consumer-grade smartphone cameras, enabling accessible, at-home risk assessment. We propose a two-stage transfer learning pipeline: a classifier is first trained on clinical images from the ISIC archive (taken with DSLR cameras) and then fine-tuned on the MIDAS dataset, a real-world collection of iPhone-captured clinical images. Our final model achieves an accuracy of 0.70, ROC-AUC of 0.79, and recall of 0.86 when tested on smartphone images, demonstrating promising performance in this low-resource setting. In the process, we also develop a baseline classifier on ISIC clinical images alone, achieving 0.99 accuracy, 0.97 ROC-AUC, and 0.94 recall—competitive with leading approaches, despite operating without dermoscopic input.

1. Introduction

Skin cancer represents a major global health concern, with high incidence rates and fatal outcomes if not detected early, yet when caught early, it is highly treatable - making timely and accurate diagnosis critical to patient survival. In clinical practice, diagnosis typically depends on dermoscopic images captured using specialized equipment operated by trained dermatologists. These high-resolution images enable detailed inspection of skin lesions, but the equipment is costly and not widely available, limiting its use to well-resourced medical settings. As a result, there is a growing interest in computer vision models that can support or augment clinical diagnosis.

Recent advances in deep learning have achieved impressive performance in skin lesion classification, with convolutional neural networks (CNNs) trained on dermoscopic images reaching or even surpassing dermatologist-level accuracy. However, these models overwhelmingly rely on dermoscopy, namely high-quality, magnified images captured using specialized tools, limiting their usefulness in everyday, real-world applications. In contrast, clinical images, which broadly refers to images of skin lesions taken with a commercial camera or smartphone camera operated by any person, are far more accessible to the general population, but significantly harder for models to interpret due to many factors, including inconsistent lighting, varying angles, and poor image quality.

Our motivation is to bridge this gap, namely answering the question, ”can we build a model that performs accurate skin cancer classification on smartphone-grade images, thereby enabling accessible at-home risk assessment?” We aim to reduce the reliance on dermoscopic images, which are both expensive and require trained professionals to collect. We want to do this by adapting models to function well on widely available clinical images.

In this project, the input to our algorithm is a clinical image of a skin lesion, captured using a consumer smartphone at a standardized 6-inch distance. These images come from the MIDAS dataset and are preprocessed with center-cropping, resizing to 224×224 pixels, and normalization using ImageNet statistics. We do not use dermoscopic input during training or inference for this model. We use a pretrained ConvNeXt-Tiny as the CNN backbone, with a custom MLP head for classification. The model is trained in two stages: first on DSLR-quality clinical images from the ISIC dataset, then fine-tuned on real-world smartphone images from MIDAS to adapt it to noisy data. The output of our model is a binary prediction indicating whether the input lesion is malignant (label = 1) or benign (label = 0). To evaluate the model, we report metrics including accuracy, recall, ROC-AUC, precision, and F1 score—emphasizing recall to minimize false negatives in a medical context.

This work focuses on improving the robustness of skin cancer classification models in realistic, non-clinical settings. By training on high-quality clinical images and fine-tuning on smartphone-captured data, we aim to evaluate how well a state-of-the-art CNN architecture can generalize to lower-quality inputs. Our results contribute to understanding the limitations and potential of deep learning models when applied to accessible image modalities, and highlight important considerations for deploying such systems outside controlled clinical environments.

2. Related Works

Performing early-stage skin cancer detection with deep learning algorithms has been a popular topic of study. The vast majority of such studies use dermoscopic images and have achieved impressive results, often surpassing human physician in terms of accuracy. Typically, input images undergo preprocessing to normalize, denoise, and remove artifacts (e.g. hair, blood vessels) from the data before being passed into feature extraction, sequentialization, embedding, and classification networks.

2.1. Skin Cancer Detection Using Dermoscopic Images

Between 2018 and 2022, a total of 40 published studies applied AI-based approaches to skin cancer detection, and among them 37 were based on dermoscopic images. [9] Popular publicly available datasets used were the International Skin Imaging Collaboration Challenge (ISIC) – which has 5 datasets for each year the challenge was hosted from 2016 to 2020 – Human Against Machine with 10,000 images (HAM10000), and PH2 with 200 images. [10, 13, 5] Despite achieving significant success in studies, models trained on dermoscopic images still have problems to overcome before becoming ready for practical applications. One such problem is that public datasets are often small in size due to the high cost of obtaining dermoscopic images in practice. A number of algorithms have very high accuracy on small datasets ($n < 5000$), leading to concerns of overfitting. [6]

2.1.1 Convolution-Neural-Network-Based Models

Deep Convolution Neural Networks (DCNNs) and DCNN variants, such as EfficientNet and DenseNet, are the most popular architecture for this task and have consistently outperformed physicians and achieving over 95% accuracy on public datasets. [6]

Notably, Singh et al. achieved an average 99.02% accuracy on the ISIC datasets. First, images are pre-processed by removing artifacts (such as hair follicles and blood vessels) and enhanced using histogram equalization. Next, preprocessed images undergo segmentation, which was accomplished by a novel thresholding-based method along with a pentagonal neutrosophic structure to form a segmentation mask of the suspected skin lesion. Finally, the segmented images are passed to the classifier, which is a DCNN with custom architecture trained on the pre-processed and augmented dataset without segmentation. [11]

Other DCNN-based classifiers also achieved impressive results. Jaisakthi et al. leveraged EfficientNet models through transfer learning to achieve an AUC of 96.81% on the ISIC datasets. [3] Kaur et al. used a DCNN and achieved accuracy as high as 90.42% on the ISIC 2020 dataset. [4] Nawaz et al. proposed a hybrid framework that combines faster region-based convolutional neural networks (RCNNs) with fuzzy k-means clustering (FKM) for classification. Their classifier achieved average accuracy of 95.40%, 93.1%, and 95.6% on the ISIC 2016, ISIC 2017, and PH2 datasets. [8]

2.1.2 Transformer-Based Models

Studies have also explored transformers in skin cancer detection from dermoscopic images. Xin et al. proposed a vision transformer network named SkinTrans that achieved 94.1% accuracy on a private dataset of 1113 dermoscopic images. The authors first pre-trained their model on the HAM10000 dataset, then applied transfer learning to the private dataset. During preprocessing, z-score normalization where $z = \frac{x-\mu}{\sigma}$ and data augmentation, such as horizontal and vertical flip, random crop, random rotation, and color jitter, were applied. The images are serialized with a multi-scale sliding window, then embedded with patch embedding and fed into a Vision Transformer (ViT). [14]

2.2. Skin Cancer Detection Using Clinical Images

Skin cancer detection machine learning models that use clinical images exist, but they are comparatively fewer in quantity than studies using dermoscopic images. Clinical images are taken with day-to-day cameras, such as ones attached to smartphones, and can be taken far away from the target lesion. Clinical images in practice often include artifacts such as clothing and background objects, lack identification of target lesion, and be out of focus. These characteristics pose significant challenge for machine learning models to learn accurate skin cancer detection.

Popular datasets used by studies before 2020 are DermIS and Dermquest. Dermquest has been disabled by the time of this writing. DermIS is an European dermatology atlas for healthcare professionals and contains approximately 500 skin lesion images.

Nasr-Esfahani et al. conducted a melanoma detection study using clinical images in 2016. The authors proposed a custom CNN architecture where the final layer is a fully-connected layer that outputs binary classification. Preprocessing efforts included illumination correction, segmentation, and gaussian filter to smooth the skin area outside of the lesion. The group achieved 81% accuracy over a private dataset of 170 clinical images. [7]

Another study trained a CNN-based model on the HAM10000 dataset but tested its performance on clinical images, which performed on par with physicians in terms of sensitivity. [1]

3. Methods

3.1. KNN

The K-Nearest Neighbors algorithm is a non-parametric classification method that operates on the principle of similarity. Given a feature vector $x \in \mathbb{R}^d$ from our ResNet18 feature extractor, KNN classifies it by finding the k closest training examples in the feature space and assigning the most common class among these neighbors.

Mathematically, for a test point x, we:

1. Calculate the distance to all training points x_i in the feature space:

$$d(x, x_i) = \|x - x_i\|_2$$

2. Find the k nearest neighbors $\mathcal{N}_k(x)$ based on these distances.

3. Assign the class label y using majority voting:

$$y = \arg\max_c \sum_{i \in \mathcal{N}_k(x)} \mathbb{I}(y_i = c)$$

where $\mathbb{I}$ is the indicator function and c represents the possible classes.

3.2. ConvNeXt Feature Extractor

We use a pretrained ConvNeXt-Tiny model as the backbone of our feature extraction pipeline. ConvNeXt is a modern convolutional neural network architecture inspired by Vision Transformers but retains efficient convolutional designs. We remove the final classification layer and use the remaining network as a fixed feature extractor to obtain compact, semantically rich embeddings of skin lesion images.

For input and normalization, we see the following. Let the input image tensor be $X \in \mathbb{R}^{B\times 3\times 224\times 224}$, where B is the batch size, and 224×224 is the image resolution with 3 RGB channels. Each image is normalized using ImageNet statistics:

$$X_{\text{norm}} = \frac{X - \mu_{\text{ImageNet}}}{\sigma_{\text{ImageNet}}}$$

where $\mu_{\text{ImageNet}} = [0.485, 0.456, 0.406]$ and $\sigma_{\text{ImageNet}} = [0.229, 0.224, 0.225]$.

Moving on to feature extraction, the normalized image is passed through the ConvNeXt-Tiny encoder to obtain a feature map:

$$F = f_{\text{ConvNeXt}}(X_{\text{norm}}), \quad F \in \mathbb{R}^{B\times C\times H\times W}$$

where C is the number of output channels, and H, W are the spatial dimensions of the feature map.

To reduce the spatial dimensions, we apply global average pooling:

$$f = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} F[:, :, h, w], \quad f \in \mathbb{R}^{B\times C}$$

This produces a fixed-length feature vector f for each image. This algorithm leverages transfer learning by reusing pretrained weights from ImageNet that already encode meaningful low-level and mid-level visual features. The hierarchical structure of ConvNeXt allows it to extract multi-scale patterns from skin lesion images. The global average pooling step ensures spatial invariance while producing compact feature embeddings, which are effective for downstream tasks like binary classification.

3.3. MLP

A multi-layer perceptron (MLP) is a fundamental type of feedforward neural network composed of one or more layers of affine transformations followed by non-linear activation functions. Each layer performs a transformation of the form $f(x) = Wx + b$, and non-linearities such as ReLU are applied to enable the network to model complex, non-linear decision boundaries. When stacked, these layers form a universal function approximator, capable of learning a wide range of mappings from input to output spaces.

In our project, we use a two-layer MLP as a classification head on top of features extracted by a pretrained ConvNeXt-Tiny model. The input to the MLP is a feature vector $x \in \mathbb{R}^{512}$ generated from the global average pool layer of the backbone. The first linear layer maps this to a 256-dimensional hidden representation, followed by a ReLU activation: $h_1 = \text{ReLU}(W_1 x + b_1)$, where $W_1 \in \mathbb{R}^{256\times 512}$. To regularize training and prevent overfitting, we optionally apply dropout with probability 0.5 after the ReLU layer. The resulting representation is passed through a second linear layer producing a scalar logit $\hat{y} = W_2 h_1 + b_2$, where $W_2 \in \mathbb{R}^{1\times 256}$ To train the MLP, we use the binary cross-entropy loss with logits:

$$\mathcal{L}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\sigma(\hat{y}_i)) + (1 - y_i)\log(1 - \sigma(\hat{y}_i))\right],$$

where $y_i \in \{0, 1\}$ is the ground truth label and $\sigma(\cdot)$ denotes the sigmoid function. The model is optimized using Adam with a learning rate of 10^{-4} and weight decay $\lambda = 10^{-5}$. During inference, we apply a threshold of 0.5 to $\sigma(\hat{y})$ to obtain a binary classification of the lesion as benign or malignant.

3.4. ResNet

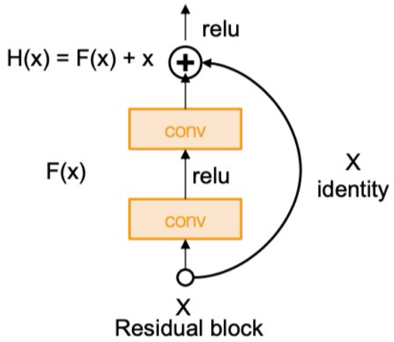

Figure 1. Illustration of the architecture of a residual block.

Residual Network (ResNet) is a convolutional neural network structure characterized by the depth of its architecture (ie. stacking many convolutional layers) and the use of skip connections (also called residual connections) to prevent vanishing gradients caused by the architecture's depth. ResNet learns a residual function $F(x)$ such that at each residual block (Figure 1) the model learns $H(x) = F(x) - x$, which allows the model to take a shortcut between convolutional layers in case of vanishing gradients, instead of learning a direct mapping $H(x)$. [2] In our implementation, we used a pretrained ResNet18 variant, which implies a network of 18 convolutional layers. We also customize the fully-connected layer into a 2-layer MLP network to perform classification.

3.5. EfficientNet

EfficientNet is a family of neural networks that scale model depth, width, and input resolution in a principled way using a compound scaling method. Introduced by Tan and Le [12], EfficientNet achieves strong performance across a range of computer vision benchmarks while using significantly fewer parameters and FLOPs compared to traditional architectures. The baseline model, EfficientNet-B0, is built around inverted bottleneck blocks and uses mobile-friendly depthwise separable convolutions, making it computationally efficient while retaining expressive power.

In our project, we use EfficientNet-B0 pretrained on ImageNet as an alternative backbone to ConvNeXt. We replace the original classification head, which originally maps to 1000 ImageNet classes, with a new binary classification head tailored to our task. Specifically, we substitute the final classifier with a dropout layer (with probability 0.2) followed by a fully connected layer that outputs a single logit. The model is trained end-to-end using the Adam optimizer with a learning rate of 10^{-4}, and binary cross-entropy loss with logits is used to supervise learning. We also apply a learning rate scheduler (ReduceLROnPlateau) that lowers the learning rate when validation accuracy plateaus. During training, we fine-tune all weights in the network, and the best-performing model on the validation set (based on accuracy) is saved for evaluation.

4. Dataset and Features

The data used for this analysis came from two datasets– the MIDAS dataset, which contains smartphone images in a variety of angles and lighting conditions, and the ISIC archive, which contains high resolution clinical images.

4.1. MIDAS Dataset

The MIDAS dataset is a prospectively gathered dataset with paired clinical and dermoscope images. Clinical images were taken on smartphone cameras from varying distances and under varied lighting conditions. After filtering out controls (which had no associated clinical metadata), dermoscopic images, and duplicate images per patient, we were left with a dataset of 634 clinical images (266 malignant, 368 benign).

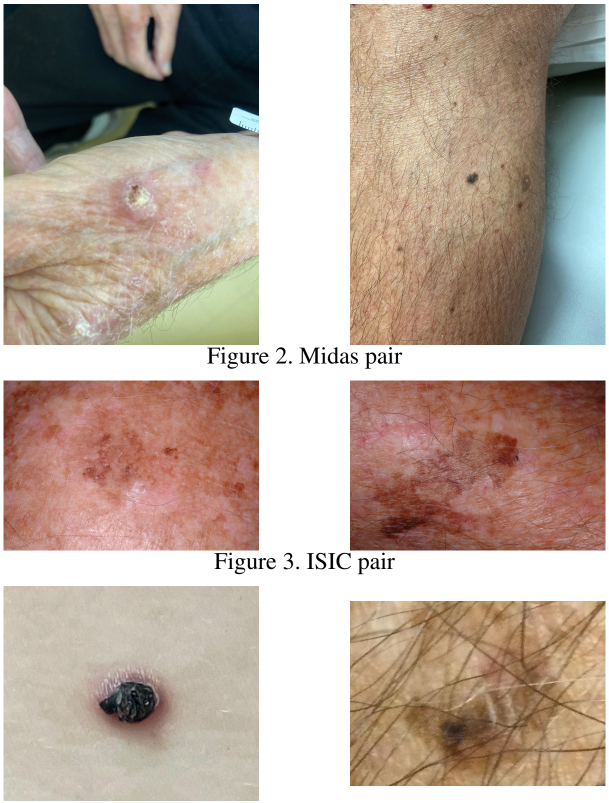

Figure 2. Midas pair

Figure 3. ISIC pair

Figure 4. Cropped Midas pair

Figure 5. Comparison of MIDAS and ISIC clinical images

In addition to using the raw MIDAS images, we also sought to determine whether cropping into the lesions would standardize the images, remove artifacts, and improve performance. Therefore, we chose a subsection of images that were taken at no greater than 6 inches away, manually cropped to the lesion, and removed low-quality images–images for which the lesion was out of focus or blurry, which left us with a dataset of 590 images (230 malignant and 360 benign). In total, we ended up with two MIDAS datasets: a large uncropped dataset and a small manually cropped dataset.

4.2. ISIC Dataset

Given that our dataset of cropped MIDAS images was quite small, we sought to develop a classifier on the ISIC archive. The ISIC archive is the largest archive of quality controlled skin cancer images. Though many images in the dataset are dermoscopic images, the ISIC archive does contain a collection of 3,597 clinical images taken on high resolution DSLR cameras. As a result, the ISIC dataset is both larger as well as more standardized than the MIDAS dataset, which makes it amenable to develop a classifier to fine tune for MIDAS. Notably, in addition to images labeled ”malignant” or ”benign,” the ISIC archive also contains images labeled as ”indeterminate,” meaning that the diagnosis is ambiguous or uncertain. We filtered out these images, as is common practice in similar papers using ISIC data, resulting in a dataset of 2,866 skin lesion images (1153 malignant and 1713 benign). Figure 5 shows some examples that highlight the differences between each dataset.

4.3. Dataset Preprocessing and Principal Component Analysis

All data were normalized to the ImageNet mean and standard deviation. Data augmentation was applied on the training sets within each dataset. For the ISIC images, we applied random cropping, horizontal flipping, and color jitter. For the MIDAS images, we applied random cropping and horizontal flipping–applying color jitter worsened our performance on our validation set so it was not applied.

To assess the feasibility of applying transfer learning from ISIC clinical images to MIDAS smartphone images, we conducted a series of visual analyses comparing the two domains. Specifically, we examined both the raw pixel-level image distributions and the learned feature representations extracted from a pretrained ConvNeXt encoder. To reduce the high-dimensional image and embedding data into a human-interpretable format, we applied Principal Component Analysis (PCA), enabling us to visualize potential differences or overlaps in the feature space between the two datasets. Interestingly, the visualizations revealed a substantial degree of overlap between the ISIC and MIDAS images, both in raw pixel space (Figure 6) and in the encoder feature space (Figure 7)—suggesting that despite differences in camera hardware, the underlying lesion characteristics are sufficiently similar to support effective transfer learning.

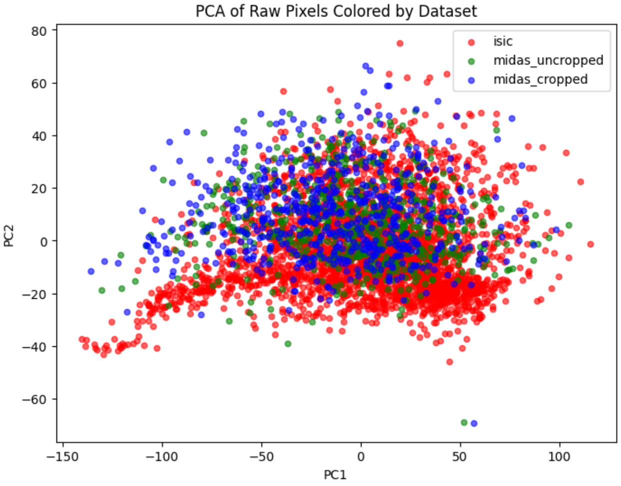

Figure 6. PCA plot of raw pixel values across the MIDAS (cropped and uncropped) and ISIC data

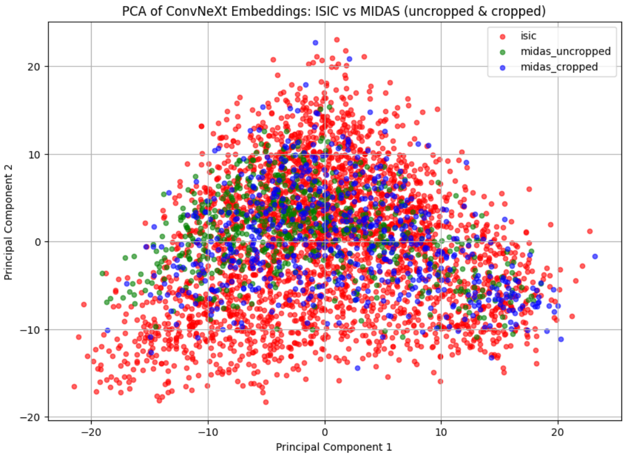

Figure 7. PCA plot of ConvNeXt encoded features across the MIDAS (cropped and uncropped) and ISIC data

5. Experiments

To develop an effective model for lesion classification on the MIDAS dataset, we began by conducting experiments using the raw, uncropped MIDAS images. Next, we leveraged the high-quality ISIC dataset to train a baseline model, and subsequently applied transfer learning by fine-tuning this model on the cropped MIDAS images. Cropping the MIDAS images helped align them more closely with the visual characteristics of the ISIC dataset, thereby making transfer learning more applicable and potentially more effective.

All models were evaluated using accuracy, precision, recall, and the area under the receiver operating characteristic curve (ROC-AUC). Accuracy is defined as the proportion of correctly classified samples among all samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. Precision measures the proportion of true positive predictions among all positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (also known as sensitivity) quantifies the proportion of actual positive samples correctly identified by the model:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Finally, the receiver operating characteristic area under the curve (ROC-AUC) summarizes the model's ability to discriminate between classes across all classification thresholds. The ROC curve plots the true positive rate (recall) against the false positive rate:

$$\text{False Positive Rate} = \frac{FP}{FP + TN}$$

and the ROC-AUC is the area under this curve, with values closer to 1 indicating better discriminatory performance.

To select the final models for evaluation on the test set, we prioritized a combination of receiver operating characteristic area under the curve (ROC-AUC) and recall metrics based on performance on the validation set. ROC-AUC was chosen as an assessment of overall classification ability–although ROC-AUC can be misleading in the case of class imbalance, our classes were well balanced meaning that ROC-AUC was a good measure of overall importance.

Recall was given special emphasis due to its critical clinical importance. In medical imaging applications—particularly in cancer detection—recall represents the true positive rate, indicating the proportion of malignant cases correctly identified by the model. Emphasizing recall helps reduce the incidence of false negatives, which is vital in diagnostic scenarios where failing to detect a malignant lesion could have severe consequences. Consequently, models demonstrating both high recall and competitive ROC-AUC were selected for final testing and downstream analyses.

5.1. Baseline MIDAS Data

In order to find the best model for our MIDAS data, we first ran our experiments on the raw (uncropped MIDAS images). Specifically, we first sought to evaluate which feature encoders worked best, evaluating using ResNet18, ResNet50, EfficientNet, and ConvNeXt, and passing the latent embeddings into our MLP. All models used the Adam optimizer with a learning rate of 1e-4 and weight decay of 1e-5. The Adam optimizer, which provides stable and consistent parameter updates due to its combination of AdaGrad and RMSProp was used because we observed oscillating and unstable training loss. To address potential overfitting, we also added dropout to our MLP head, with a dropout probability of 0.5. We trained the model on 10 epochs and then selected our best results given performance on the validation set. At this stage, we trained all models with the ConvNeXt backbone unfrozen to allow fine-tuning, given that ConvNeXt was pretrained on general images rather than clinical imaging data.

5.2. ISIC Dataset

Because we weren't seeing high performance among different feature extractors, we then decided to develop a classifier on the ISIC database to compare performance and as well as to fine tune our MIDAS model on. In order to allow for smooth gradient updates and minimize overfitting, we used the Adam optimizer, learning rate of 1e-4, weight decay penalty of 1e-5, dropout probability of 0.5. We additionally trained the model on 10 epochs with the ConvNeXt backbone unfrozen, selecting the best model based on validation set performance.

5.3. Transfer Learning on the Manually Cropped MIDAS Dataset

Our next experiment was to fine-tune the ISIC-trained model to our small cropped MIDAS dataset in order to evaluate if this transfer learning would result in improved performance. In order to allow our model to adapt to new data yet not overfit, we constructed a two stage training strategy. Our first stage was to freeze the backbone and train the model without weight decay penalties. We used a learning rate of 1e-4 and trained the model for two epochs. In the second stage, we unfroze the backbone and introduced dropout with probability 0.5 and weight decay with penalty 1e-5. We lowered the learning rate to 1e-5 and employed a learning rate scheduler to help the optimizer navigate smoother regions of the loss landscape. We hypothesize that stage one of the training regimen allows the model to learn and adapt to the new data, while stage two allows for fine-tuning without overfitting.

6. Results

The results of our experiments, presented in Table 2, demonstrate that transfer learning from the ISIC dataset to the cropped MIDAS dataset yields significant performance improvements compared to training on the raw MIDAS images alone. Notably, the recall—a critical metric for melanoma classification—increased substantially from

0.2857 to 0.8571. This improvement indicates that the transfer-learned model is more effective at correctly identifying malignant cases. Our results further show high performance on the ISIC dataset showing performance is comparable to dermoscope-trained models when the images are not captured using smartphones. In the following tables, Unc. represents ”uncropped” and Crop. represents ”cropped.”

Model	Accuracy	Precision	Recall	ROC-AUC
Midas Unc. (ConvNeXt)	0.6667	0.6000	0.5769	0.7040
Midas Unc. (ResNet)	0.6667	0.6429	0.4500	0.6625
Midas Unc. (EffNet-B0)	0.3750	0.3333	0.2500	0.4102
Midas Cropped (Final)	0.7363	0.6154	0.8823	0.8823
ISIC (Final)	0.9651	0.9293	0.9884	0.9941

Table 1. Validation results across models and datasets.

We report test set performance (Table 2) only for the final selected models, as these consistently outperformed earlier architectures on the validation set (Table 1). This evaluation strategy ensures that only the most promising models—based on validation accuracy and recall—are assessed on held-out test data, thereby reducing the risk of overfitting and preserving the integrity of the test set.

Model	Accuracy	Precision	Recall	ROC-AUC
Midas Unc. (Final)	0.5714	0.4667	0.2857	0.5840
Midas Crop. (Final)	0.7000	0.5769	0.8571	0.7927
ISIC (Final)	0.9916	0.8571	0.9364	0.9676

Table 2. Test performance of final models across datasets.

Across all models, early results indicated overfitting (achieving zero loss on the training data but poor generalization to the validation data), which motivated fine-tuning the weight decay, dropout, and data augmentation techniques described earlier.

Despite very strong performance on ISIC images, adapting our ISIC-trained model to perform well on the cropped MIDAS images was challenging, primarily because our model was actually underfitting to the new data. We observed that removing regularization techniques (weight decay and dropout) improved performance at first, but then led to overfitting. As a result, we tested a variety of training regimens that varied freezing vs unfreezing the backbone (which had now been pretrained on ISIC images), learning rates, and regularization. In the end, we were able to come to a two-stage training strategy described previously that mitigated the problem of underfitting appropriately.

Overall, our best model achieved recall of 0.8571 and ROC-AUC 0.7927. Figure 8 shows the confusion matrix results from our best model (determined via validation performance) evaluated on our dataset of cropped MIDAS smartphone camera images, and Figure 9 shows the ROC-AUC curve.

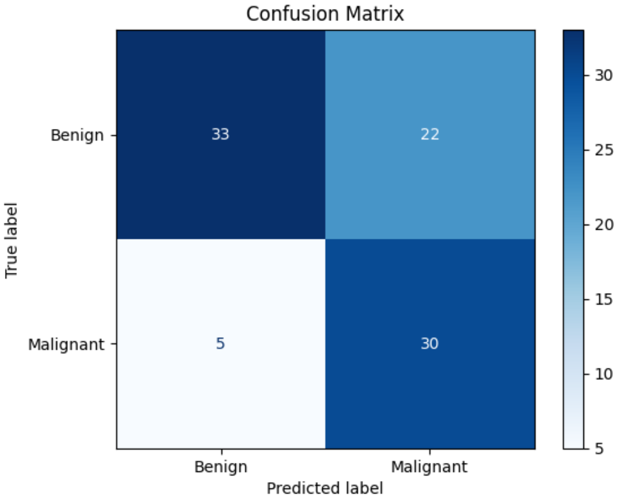

Figure 8. Confusion matrix results on the cropped MIDAS images test set.

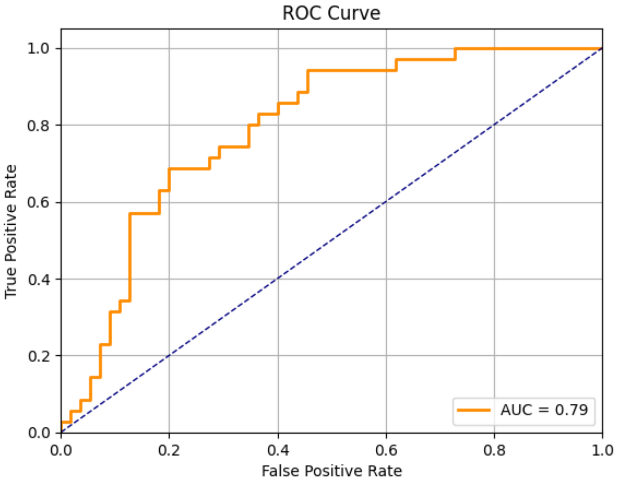

Figure 9. ROC-AUC curve for the cropped MIDAS images test set.

As indicated by the confusion matrix, because we optimized for recall, our model has high recall (true positive rate) at the cost of false positive rate. Figure 10 shows two examples of correctly predicted skin lesions and two examples of incorrectly predicted lesions, with more predictions shown in Figure 13. Though we sought to identify trends that could explain incorrect examples, we were unable to identify any meaningful patterns across incorrectly predicted images.

Further results for the ISIC model, including the confusion matrix, ROC-AUC curve, and sample image predictions can be found in the appendix.

7. Discussion

Overall, we observed that a ConvNeXt-based classifier trained on high-quality clinical images from the ISIC archive was able to improve performance compared to base-

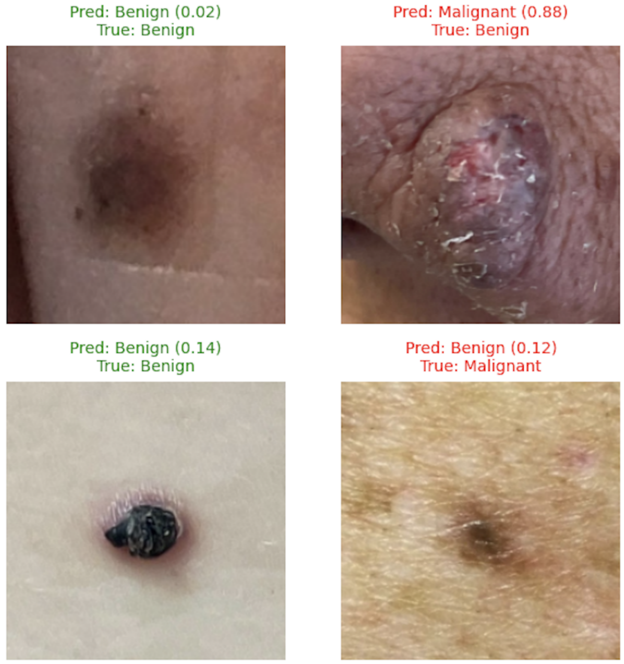

Figure 10. Images demonstrating correctly predicted and incorrectly predicted skin lesions

line when fine-tuned on the MIDAS dataset, which consists of more variable and lower-resolution iPhone images. Crucially, this performance was achieved while also manually cropping in so that the smartphone imaged matched the relative size of the ISIC images.

This supports the idea that models initially trained on higher-quality clinical images (e.g., from ISIC) can be effectively fine-tuned for deployment on more practical, real-world data sources like smartphone images. Our PCA analysis of the raw and encoded image embeddings further confirms that the domains overlap significantly, suggesting that the distribution shift is manageable with appropriate transfer learning strategies.

The fine-tuned MIDAS model attained promising results with an ROC-AUC of 0.79 and recall of 0.86, demonstrating clinical relevance for early detection. This methodology emphasizes the potential for at-home risk assessment, lowering barriers to early screening and empowering patients in under-resourced or remote settings.

However, our work is not without limitations. The cropped MIDAS dataset remains relatively small and non-standardized, with many images in the dataset taken in poor lighting conditions from a variety of angles. Further validation on diverse smartphone image collections is necessary to ensure broader generalizability.

Additionally, our ISIC-trained model outperforms the fine-tuned counterpart and achieves performance comparable to state-of-the-art models trained on dermoscopic images—surpassing the average diagnostic performance of dermatologists as reported in prior studies. This suggests that, with appropriate architectures, high-resolution clinical photographs can capture clinically meaningful features comparable to those seen in dermoscopic imaging. These results also point to the potential for achieving better performance on high-definition images captured via smartphone cameras, provided they are taken under standardized conditions (e.g., consistent distance and lighting with flash). This underscores the need for further evaluation and validation on real-world smartphone image datasets to better understand their clinical utility in early detection settings

8. Conclusion

Our results demonstrate the feasibility of using transfer learning to adapt high-performing skin lesion classification models to smartphone-captured clinical images. While existing models often rely on dermoscopic images taken in controlled clinical environments, our work achieves robust performance on images taken on smartphone cameras.

Future work should focus on image preprocessing techniques, such as applying lesion localizing and segmentation algorithms on clinical images to remove background objects, artifacts, and noise. A robust algorithm to semantically segment the skin lesion would significantly boost the model's ability to predict images with poor quality and thereby unlock the potential for it to be applied in day-to-day clinical uses and help millions of patients. Additionally, to evaluate performance on clinical datasets, future efforts should obtain performance metrics (accuracy, precision, recall, ROC-AUC) of human physicians' predictions on the test set. These metrics would contextualize the confidence of the model predictions and alleviate concerns of overfitting. In terms of architecture, ViT-based models are also worth exploring, as we believe the significant success they have exhibited with dermoscopic datasets can be transferred to clinical datasets.

References

[1] T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, B. Schilling, S. Haferkamp, D. Schadendorf, S. Fröhling, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer*, 111:148–154, 2019.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] S. M. Jaisakthi, P. Mirunalini, C. Aravindan, and R. Appavu. Classification of skin cancer from dermoscopic images using deep neural network architectures. *Multimedia Tools and Applications*, 82:15763–15778, 2023.

[4] R. Kaur, H. GholamHosseini, R. Sinha, and M. Lindén. Melanoma classification using a novel deep convolutional neural network with dermoscopic images. *Sensors*, 22(3), 2022.

[5] T. Mendonça, M. Celebi, T. Mendonca, and J. Marques. Ph2: A public database for the analysis of dermoscopic images. *Dermoscopy image analysis*, 2, 2015.

[6] H. Naseri and A. A. Safaei. Diagnosis and prognosis of melanoma from dermoscopy images using machine learning and deep learning: A systematic literature review. *BMC Cancer*, 25(1):75, 2025.

[7] E. Nasr-Esfahani, S. Samavi, N. Karimi, S. M. R. Soroushmehr, M. H. Jafari, K. Ward, and K. Najarian. Melanoma detection by analysis of clinical images using convolutional neural network. In *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 1373–1376. IEEE, 2016.

[8] M. Nawaz, Z. Mehmood, T. Nazir, R. A. Naqvi, A. Rehman, M. Iqbal, and T. Saba. Skin cancer detection from dermoscopic images using deep learning and fuzzy k-means clustering. *Microscopy Research and Technique*, 85(1):339–351, 2022.

[9] R. H. Patel, E. A. Foltz, A. Witkowski, and J. Ludzik. Analysis of artificial intelligence-based approaches applied to non-invasive imaging for early detection of melanoma: A systematic review. *Cancers*, 15(19), 2023.

[10] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, A. Halpern, B. Helba, H. Kittler, K. Kose, S. Langer, K. Lioprys, J. Malvehy, S. Musthaq, J. Nanda, O. Reiter, G. Shih, A. Stratigos, P. Tschandl, J. Weber, and P. Soyer. A patient-centric dataset of images and metadata for identifying melanomas using clinical context, 2021.

[11] S. K. Singh, V. Abolghasemi, and M. H. Anisi. Skin cancer diagnosis based on neutrosophic features with a deep neural network. *Sensors*, 22(16), 2022.

[12] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[13] P. Tschandl, C. Rosendahl, and H. Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[14] C. Xin, Z. Liu, K. Zhao, L. Miao, Y. Ma, X. Zhu, Q. Zhou, S. Wang, L. Li, F. Yang, S. Xu, and H. Chen. An improved transformer network for skin cancer classification. *Computers in Biology and Medicine*, 149:105939, 2022.

Appendices

A. Acknowledgments and Contributions

Darren focused on implementation and running experiments on the ISIC classifier and the uncropped MIDAS classifier, as well as the experiments, results, and discussion section. Jack focused on the related works, ResNet methods, conclusion, and appendices sections, and ran experiments on the ResNet-MLP architecture. Flora focused on the introduction, KNN, ConvNeXt Feature Extractor, MLP, and Efficient methods, and ran experiments on the EfficientNet architecture.

We would like to thank the entire teaching team and staff of the Spring 2025 CS231N course. Our sincerest gratitude extends to instructors Fei-Fei Li, Ehsan Adeli, Justin Johnson, and Zane Durante; course assistants (CAs) Chaitanya Patel (Head CA), Kyle Sargent, Tiange Xiang, Gabriela Aranguiz-Dias, Zhoujie Ding, Sabri Eyuboglu, Yunfan Jiang, Emily Jin, Matthew Jin, Jiaman Li, Ryan Li, Rahul Mysore Venkatesh, Fan-Yun Sun, Iris Xia, Serena Zhang, Minji Kang, Shutong Zhang, Stephen Tian, Aditesh Kumar, Wenlong Huang, and Cristóbal Eyzaguirre; and fantastic guest lecturers. In particular, we would like to thank Minji Kang, our CA mentor for this project, for providing helpful and insightful feedback on this project.

B. Additional Figures

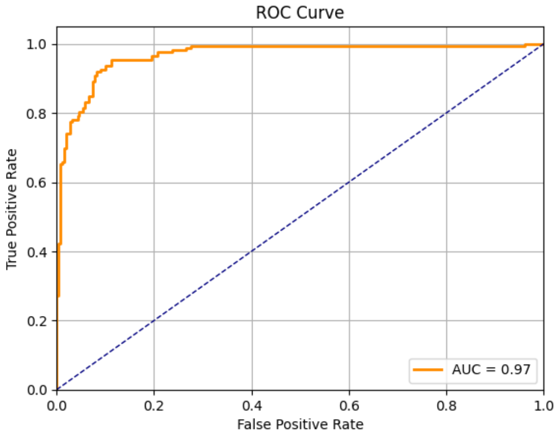

Figure 11. ROC-AUC curve for ISIC images test set.

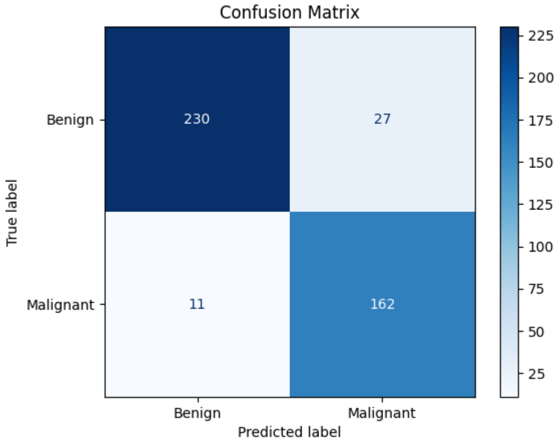

Figure 12. Confusion matrix results on ISIC images test set.

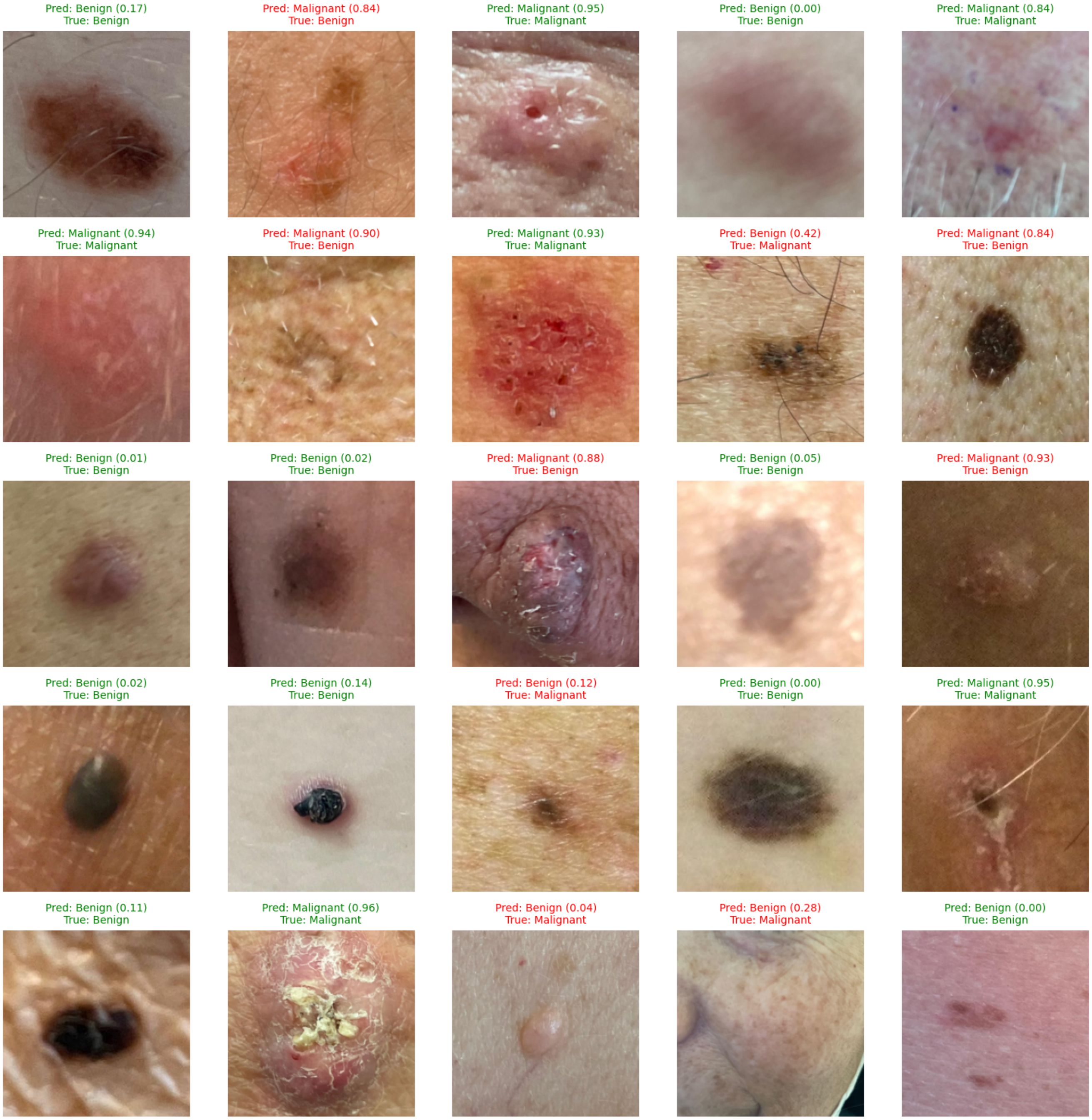

Figure 13. Examples of images from the Midas test set along with the model's prediction for each image.

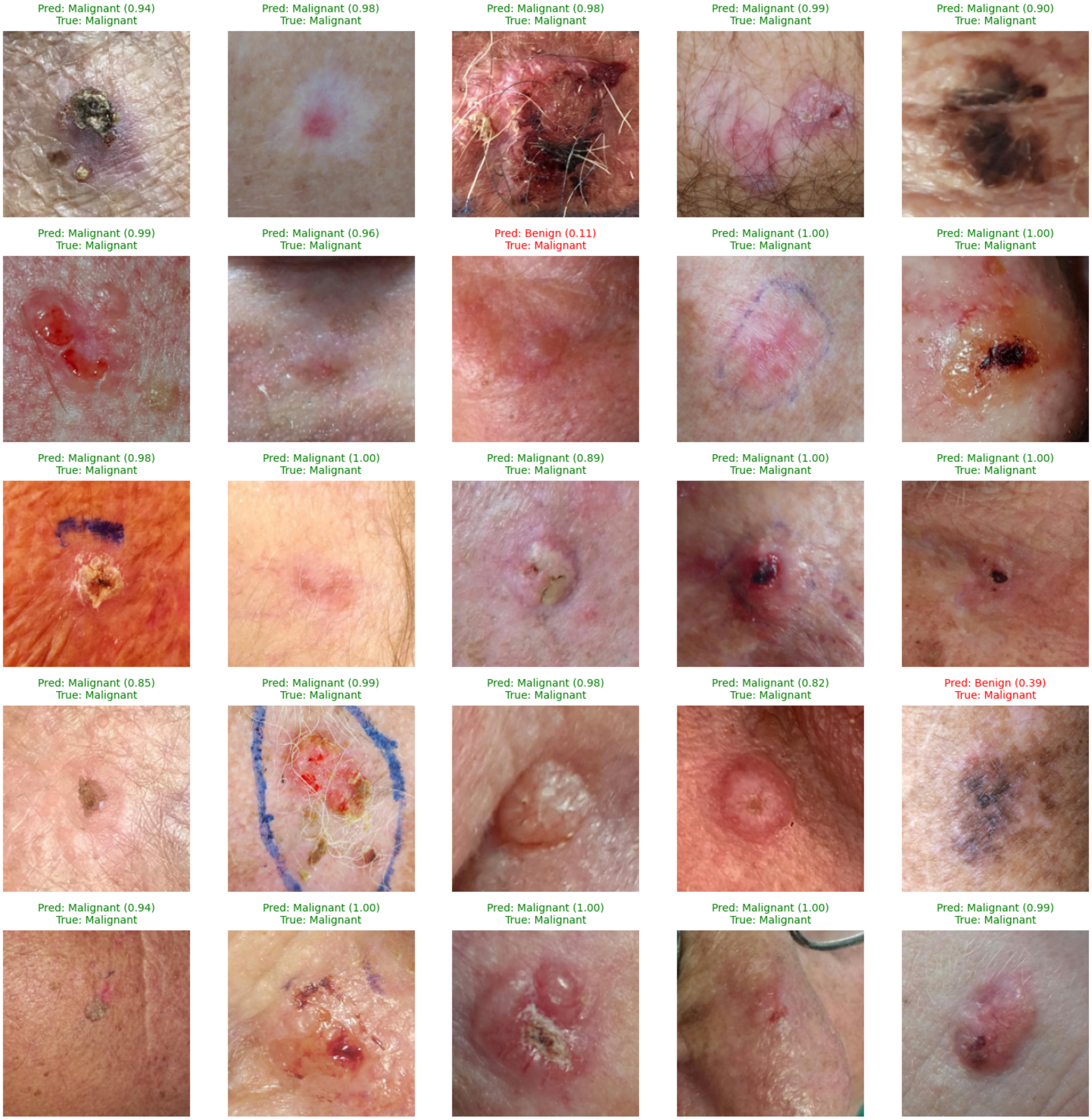

Figure 14. Examples of images from the ISIC test set along with the model's prediction for each image.