Skin Lesion Classification

1. Introduction:

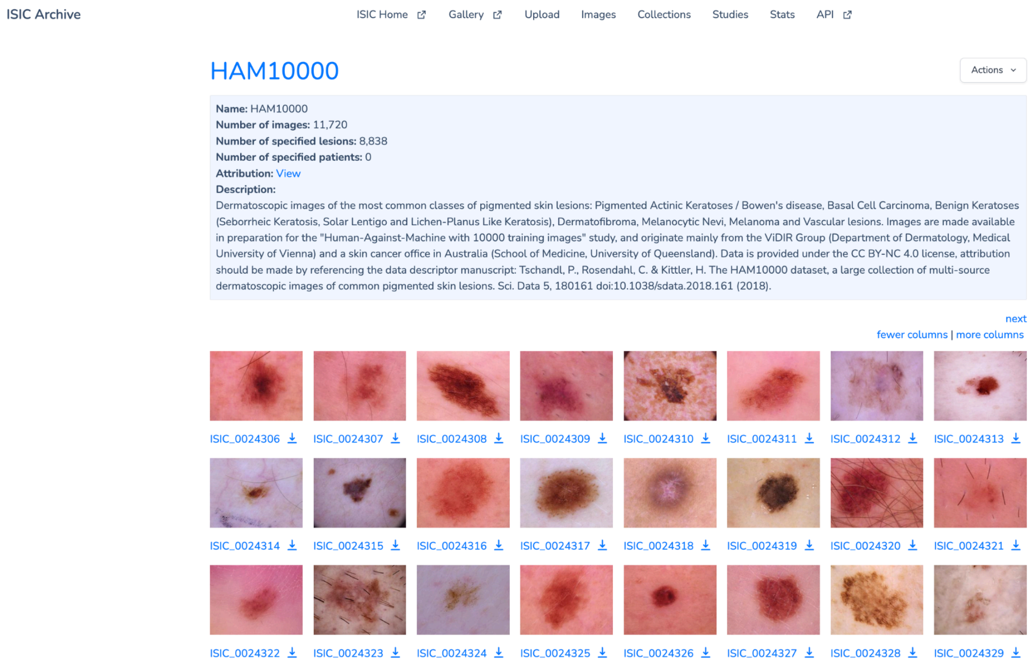
Melanoma is a type of skin cancer that develops from melanin-producing cells in the body. It typically occurs in the skin, but rarely occurs in other parts of the body. Visually inspecting the skin is the first step in diagnosing the condition. Moles that are irregular in color or shape are suspicious of melanoma. A few of the early signs of melanoma can be identified by using asymmetry, borders, colors, diameters. By leveraging computer vision and deep learning techniques, we can build systems that assist dermatologists in diagnosing skin lesions at an early stage.
The purpose of this project is to build a computer vision system that processes the images of skin lesions and classify them into different categories, such as benign, malignant, etc. The system will use deep learning models to recognize these patterns.

2. Discussion on data and chosen framework:

ISIC (International Skin Imaging Collaboration) Archive: ISIC has an open-source platform for the contribution of images of skin lesions under Creative Commons licenses. The images are associated with real diagnoses and other clinical metadata, and they are available for use in the public domain.

Link: https://api.isic-archive.com/collections/?pinned=true

The archive consists of several datasets and I am looking at the HAM10000 ("Human Against Machine with 10000 training images") dataset for modeling and testing.



HAM10000

Name: HAM10000
Number of images: 11,720
Number of specified lesions: 8,838
Number of specified patients: 0
Attribution: View
Description:
Dermatoscopic images of the most common classes of pigmented skin lesions: Pigmented Actinic Keratoses / Bowen's disease, Basal Cell Carcinoma, Benign Keratoses (Seborrheic Keratosis, Solar Lentigo and Lichen-Planus Like Keratosis), Dermatofibroma, Melanocytic Nevi, Melanoma and Vascular lesions. Images are made available in preparation for the "Human-Against-Machine with 10000 training images" study, and originate mainly from the ViDIR Group (Department of Dermatology, Medical University of Vienna) and a skin cancer office in Australia (School of Medicine, University of Queensland). Data is provided under the CC BY-NC 4.0 license, attribution should be made by referencing the data descriptor manuscript: Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5, 180161 doi:10.1038/sdata.2018.161 (2018).

The HAM10000 dataset contains dermatoscopic images from different populations, acquired and stored by different modalities. The final dataset consists of 11720 dermatoscopic images which can serve as a training set for our system. It is widely cited and provides high quality images, making it suitable for segmentation and classification. Also its moderate size and well-documented structure makes it a good choice for initial model prototyping and also suits the computational constraints of running in a laptop.

3. Tooling and Frameworks:

OpenCV: Open source computer vision library that we will be using for image preprocessing like hair removal, converting to grayscale, resizing, etc. It will also be used in postprocessing after segmentation, like finding contours, cropping lesion area.

Pytorch: Deep learning framework chosen for building, training and evaluating the neural network.

ResNet-18/ ResNet-34 / ResNet-50: ResNet is a model based on Convoluted Neural Networks architecture that uses residual connections to train deeper networks more effectively. We will start with RestNet-18 as it is the most lightweight model in the ResNet family and can be run in my Macbook with limited computation resources. Based on the performance we can take a call to run in ResNet-50.

Scikit-learn / Statsmodel / Numpy / Pandas: For evaluation metrics and analysis.

4. System Design:

The high-level system architecture follows a classic medical imaging pipeline:

1. Dataset preparation
2. Image preprocessing
3. Segmentation (classical morphology + U-Net attempt)
4. Augmentation and normalization
5. Model training (ResNet34)
6. Class imbalance handling
7. Post-training threshold tuning
8. Model evaluation and visualization

This modular design allows experimenting with variations in each stage.

5. Implementation Process and Improvements:

This section expands the midterm progress to include full experimentation cycles, failures, lessons learned, and final refinements.

5.1 Initial Setup and Baseline Model:

The initial approach used:

- ResNet18 pretrained on ImageNet
- Standard transforms (resize → tensor → normalize)

- CrossEntropy loss
- Adam optimizer

Baseline results were:

- High accuracy (>85%)
- Very low recall for malignant (~0.65–0.75)

This is a known pitfall in medical classification: accuracy hides poor performance on the minority class.

5.2 Handling Class Imbalance

Given the ~81/19 imbalance, the model tended to overpredict benign lesions.

5.2.1 Weighted BCE Loss

Replacing CrossEntropy with BCEWithLogitsLoss(pos_weight=X) helped penalize malignant misclassifications more heavily.

After tuning:

- Recall increased from ~0.75 → 0.93
- Precision dropped (expected trade-off)

This matched expectations: increasing recall typically increases false positives.
Weighted BCE became central to the system.

5.3 Image Preprocessing Improvements

5.3.1 Augmentation

To inject variety, augmentation included:

- Horizontal & vertical flips
- Rotations
- Color jitter
- RandomResizedCrop
- Gaussian noise

These augmentations reduced overfitting and increased robustness.

5.3.2 Segmentation Attempts

Two segmentation strategies were explored:

A. Classical Segmentation (skimage.morphology):

Steps:

- Convert to grayscale
- Otsu thresholding
- Morphological closing
- Small object removal
- Mask application

This successfully isolated lesions but sometimes removed lesion borders or created inconsistent masks, hurting precision. However, recall remained high.

B. U-Net Segmentation (attempt and revert):
A U-Net model was briefly tested. Although U-Net produced better masks, it significantly increased:

- Training time
- Memory usage
- System complexity

The improvement to classification accuracy was not significant enough to justify computational cost within project constraints. Therefore, classical segmentation was retained.

5.4 Threshold Tuning:
Even after weighted loss, the default probability threshold of 0.5 yielded:

Accuracy	0.8831
Precision	0.6749
Recall	0.7907
F1 Score	0.7282
ROC-AUC	0.9352

A medically useful classifier often prioritizes recall (not missing malignant cases), but extremely low precision is undesirable.

Using Precision-Recall curve analysis, thresholds were optimized for:

- Maximum F1 score
- Recall ≥ 90% (clinical priority)

Two tuned thresholds were studied:

(1) Best F1 Threshold

Accuracy	0.8963
Precision	0.7412
Recall	0.7326
F1 Score	0.7368
ROC-AUC	0.8347

(2) High-Recall Threshold (≥90%)

Accuracy	0.8209
Precision	0.5281
Recall	0.9012
F1 Score	0.6660
ROC-AUC	0.8511

6. Final System Performance:

This section analyzes the final model trained with:

- ResNet34
- Weighted BCE
- Segmentation
- Augmentation
- Threshold tuned for ≥90% recall

Accuracy	0.8209
Precision	0.5281
Recall	0.9012
F1 Score	0.6660
ROC-AUC	0.8511

6.1 Interpretation:

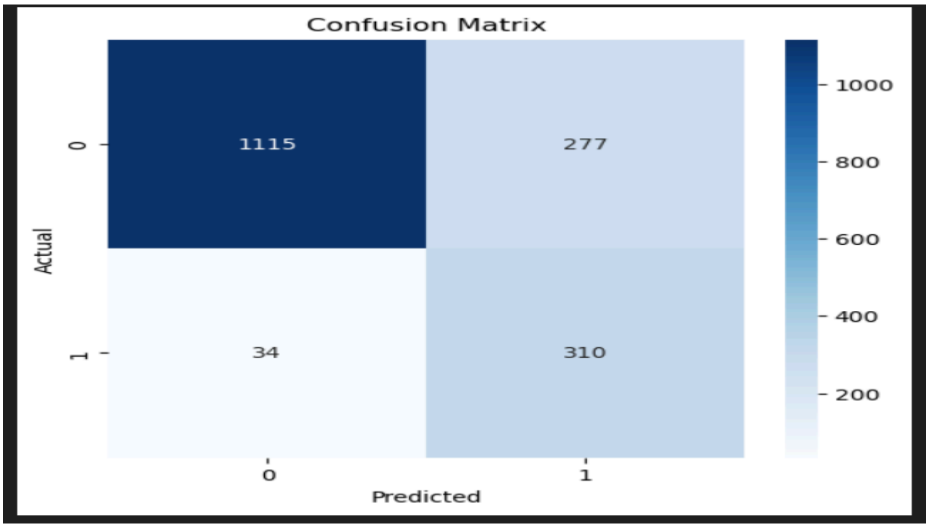
- High recall (0.90) means the model rarely misses malignant cases.
- Precision (0.53) indicates ~47% false positives, acceptable for a screening tool.
- F1 score shows a good balance between recall and precision.
- ROC-AUC (0.8511) indicates strong discriminative capability.

This is a clinically meaningful balance: prioritize catching malignant lesions even at the cost of false alarms.

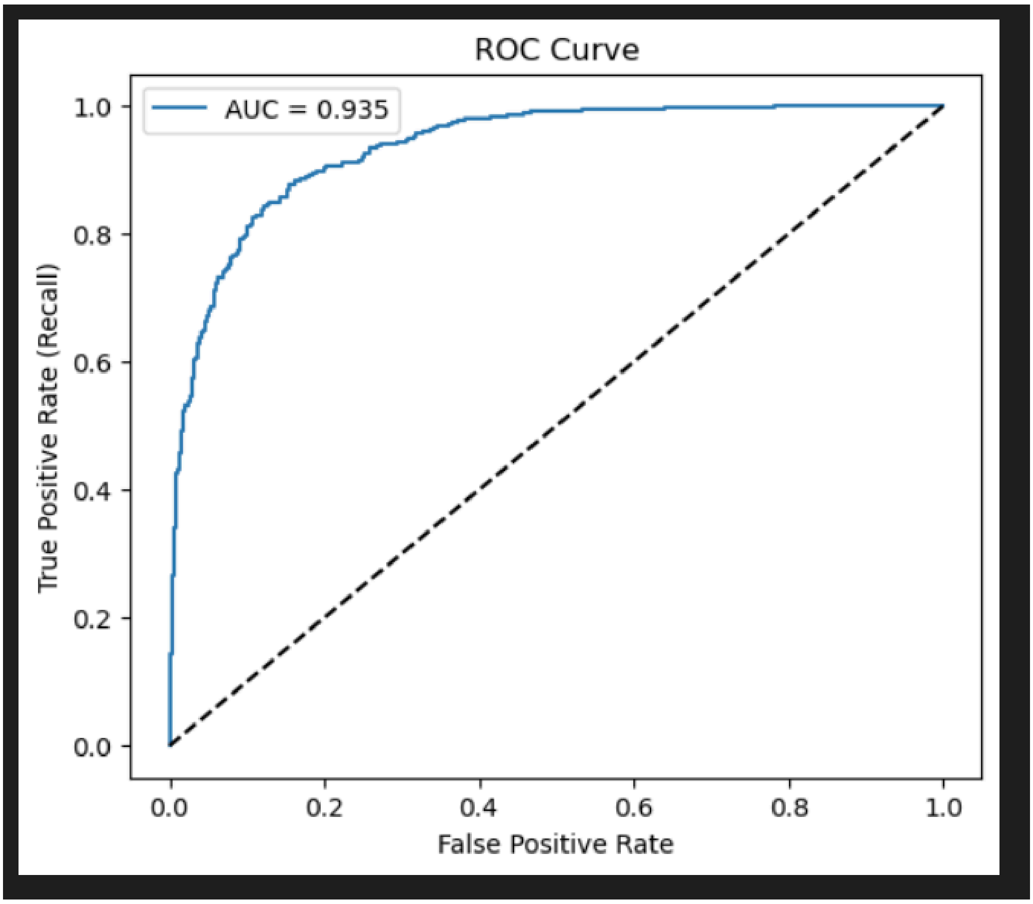
7. Visualizations

Three major visualizations strengthen the evaluation:

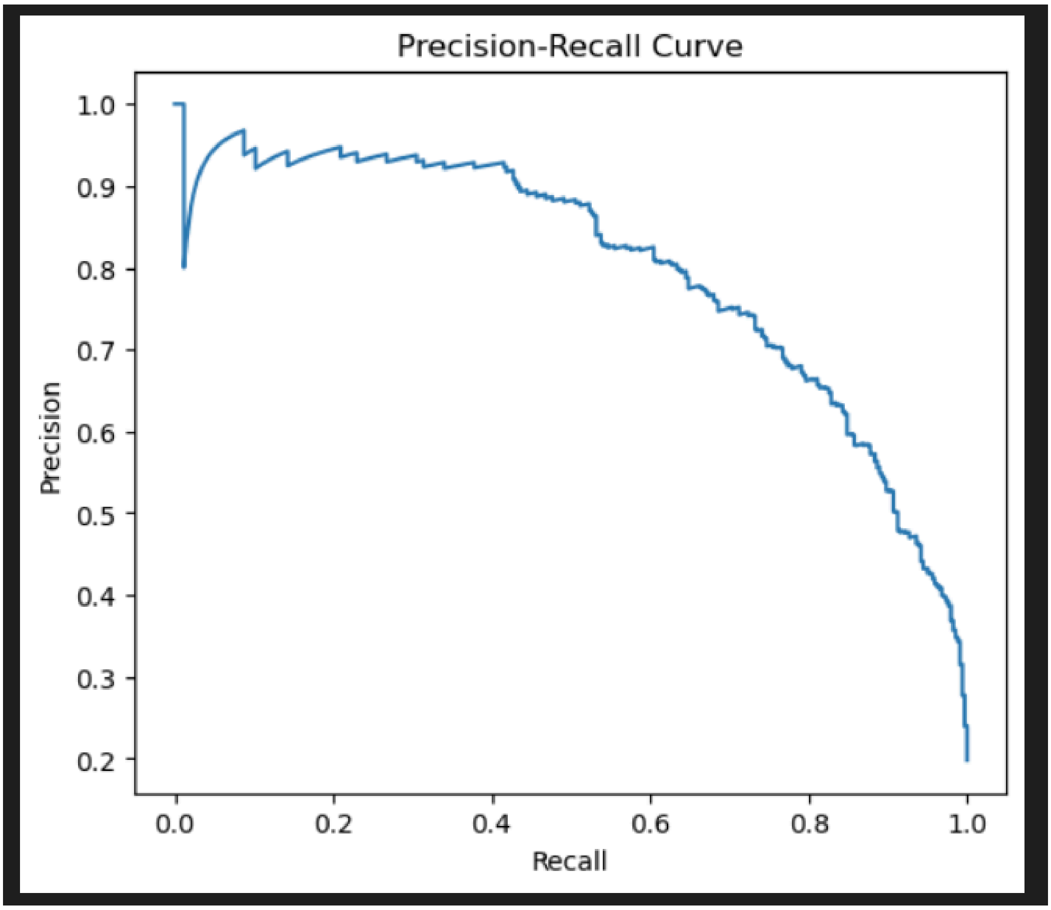
7.1 Confusion Matrix:



7.2 ROC Curve:



7.3 Precision–Recall Curve:



8. Strengths, Weaknesses, and Lessons Learned

Strengths

- High malignant recall (>90%)
- Strong generalization using augmentation
- Effective handling of class imbalance
- Useful interpretability via ROC and PR curves
- Segmentation helps localize lesion area

Weaknesses

- Precision remains moderate (≈53%)
- Segmentation via morphology can be inconsistent
- Limited compute constrained deeper architectures (EfficientNet, ResNet50)
- No testing on external datasets (ISIC 2024 removed)

Lessons Learned

- Weighted loss dramatically changes classifier behavior
- Segmentation helps but must be reliable

- Threshold tuning is essential in imbalanced medical problems
- More training data (especially malignant) would improve precision

9. Future Directions

If revisiting this system, improvements may include:

9.1 Advanced Segmentation

Return to U-Net, but simplify:

- Pre-trained U-Net models
- Light-weight UNet++ variants
- Use masks only during training for attention-guided learning

9.2 Using EfficientNet or ViT

These models often outperform ResNet in medical imaging tasks.

9.3 Multi-class Lesion Classification

Move beyond binary benign/malignant to classify lesion subtypes.

10. Conclusion

In this project, I learned how challenging and rewarding it is to build a system that can reliably classify skin lesions. Starting with basic preprocessing and a simple ResNet model, the results showed clear gaps, especially in detecting malignant cases. Each step that followed, handling class imbalance, applying data augmentation, experimenting with segmentation, and tuning the threshold helped me understand not just how to improve the model, but why those improvements mattered.

The final system still isn’t perfect, but it’s much better at catching malignant lesions while keeping overall performance stable. If I had more time, I would explore deeper architectures and more robust segmentation methods. Still, this project gave me a strong foundation in end-to-end model development and showed how small design decisions can meaningfully impact real-world performance.