Datathon @ UCI 2025 Malignant Melanoma Detector

Lucas Chin, Daanesh Bogale, Tiffany Lai, Brayden Weimholt

Why We Chose This Project

We chose this project because it brings together our passion for **digital image processing** and a mission with **real-world impact**: improving early skin cancer detection, and making it accessible for everyone.

Skin cancer is one of the most common cancers globally, yet many people lack access to dermatologists. The ISIC 2024 dataset mimics phone-quality images, which presents a realistic opportunity to build AI that supports clinical triage.

Rather than working with clean textbook data, we wanted to challenge ourselves by analyzing **real-world medical data** — where innovation truly happens.

What's in the Data

The dataset includes over 400,000 cropped lesion images and paired metadata.

- train-image
- train-image.hdf5
- **train-metadata.csv**
- test-image.hdf5
- test-metadata.csv
- sample_submission.csv

We focused on metadata exploration to understand malignancy patterns before building any predictive models.

Figure 1. Example of train-image

What's in the Data

- Train-metadata.csv (55 columns, 401059 rows)
  - isic_id: Unique case identifier, **primary key**
  - Patient info: age_approx, sex, attribution
  - Lesion info: size, shape, color irregularity, eccentricity (deviance from center of vein)
  - Location info: anatom_site_general, tbp_lv_location, tbp_lv_location_simple, x, y, z
  - Label: target
    - Binary
    - 0 = benign, 1 = malignant

How We Navigated the Data

Data Cleaning

- Remove diagnosis fields to prevent **data leakage**
  - e.g. iddx_full, iddx_1 to iddx_5, mel_thick_mm, mel_mitotic_index, lesion_id
- Dropped rows with missing sex, and missing anatom_site_general (1.5-2.5% of their sample)
- Inputted missing age_approx with the **median age(60)** – <1% of the data missing age

Exploratory Data Analysis

- **Validated** the CNN's training/classification processes with a logistic regression
  - Explored Correlations, Model Fitting
  - ROC Plot and Interpretation
- Compared sample sizes and **distribution of "target"** classifier on logarithmic scale

What We Found

Correlation Matrix

- Highly Correlated Subsets
  - Color elements: A, B, C, H, L levels inside and outside the lesion
    - ABC positively correlated with themselves and L; H is negatively correlated with A
  - Size elements: Area, Perimeter, Diameter highly correlated
- Reduce this multicollinearity prior to fitting logistic regression model
  - Drop one of each pair heavily correlated features
  - NumPy Correlation Matrix–**drop columns with r > 0.9**

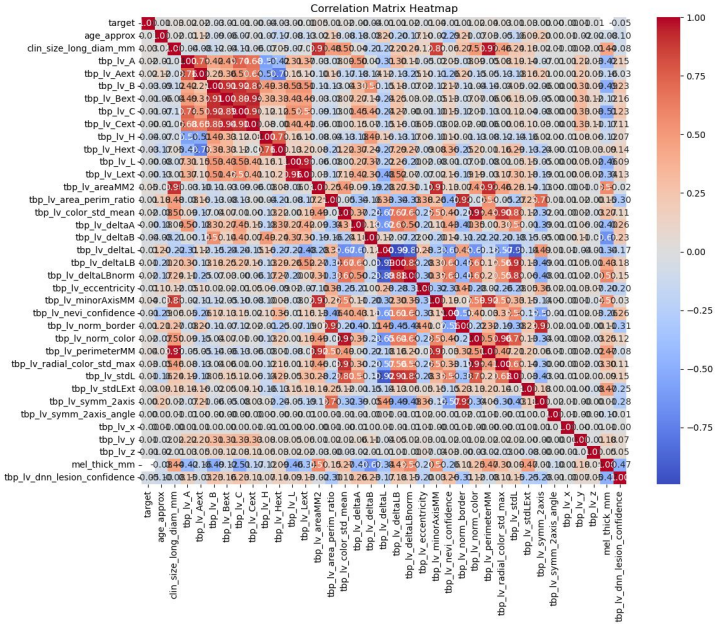


Figure 2. Correlation Matrix Heatmap

Data Reduction

- Less multicollinearity–features can be approximated more credibly
- Overall, remaining represent the correlation of information implicitly utilized by the CNN
  - **Color values** and **deviations** between different categories of color
  - **Area/Perimeter Ratio, Eccentricity, axisAngle,** and the coordinates **(x, y, z)** for size
  - Interactions between the size and color
- Clean further to model what the image processing NN “sees”
  - Remove confidences

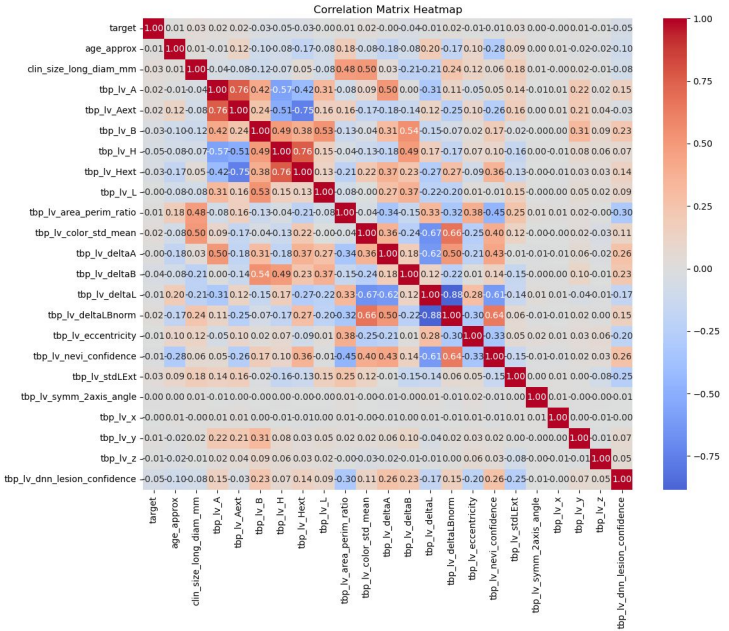


Figure 3. Correlation Matrix Heatmap with Data Reduction

Logistic Regression Fit

- Use sklearn/statsmodels.api fit a logistic regression with numeric covariates
  - Leaves output in odds/probability, $0 < p < 1$, for classifying
- Most significant explanatories: **Diameter, L Color, Area/Perim. Ratio, Color Deviation**
  - Mostly information the neural network is working with and deems important!

🔍 Top 10 statistically significant features (by p-value):

	feature	coefficient	p_value
2	clin_size_long_diam_mm	0.237364	8.005475e-31
19	tbp_lv_y	0.000981	1.877913e-10
16	tbp_lv_stdLExt	0.271393	2.782461e-10
8	tbp_lv_L	0.047199	2.684857e-07
9	tbp_lv_area_perim_ratio	-0.061246	5.380087e-07
10	tbp_lv_color_std_mean	0.328789	2.558963e-06
13	tbp_lv_deltaL	0.221014	1.219136e-05
5	tbp_lv_B	-0.219495	6.547193e-05
14	tbp_lv_deltaLBnorm	0.228704	6.532161e-04
1	age_approx	0.009292	2.239888e-02

Figure 4. Top 10 Statistically Significant Feature

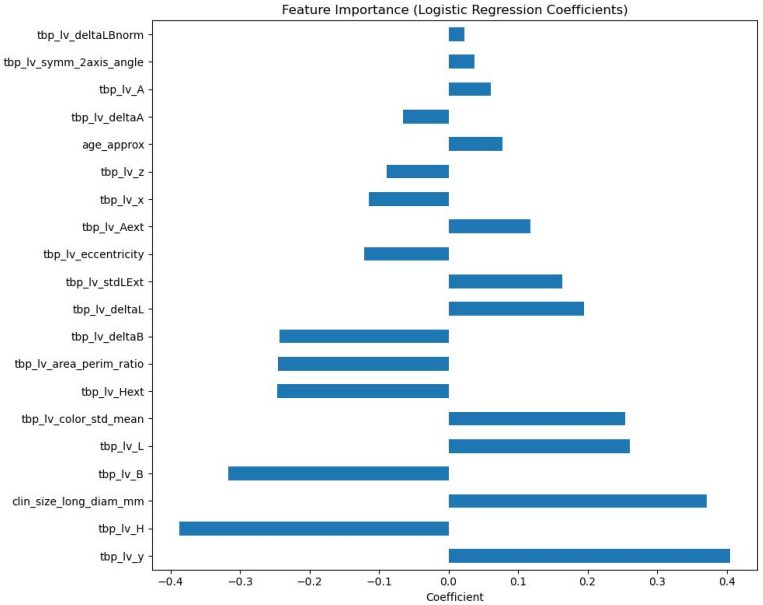


Figure 5. Logistic Regression

ROC Plot/CNN Connection

- **AUC** (Area Under Curve) Value = 0.89
  - 1.00 = Perfect Classification, 0.5 = Random Guess
  - Generally, >0.80 is a “good fitting” model in statistics
- Performs well using features informed by the CNN pipeline
- Reflects patterns observed by the image processing
  - Mathematically consistent, another simpler model (Logistic) would fit nicely

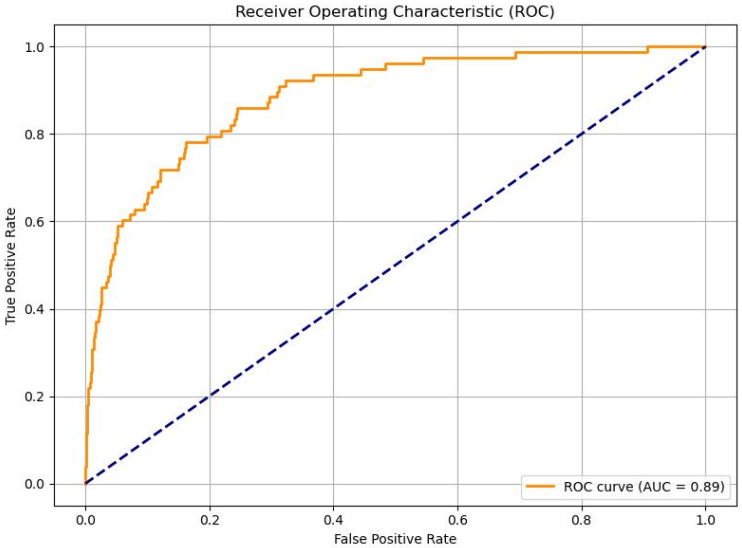


Figure 6. Receiver Operating Characteristic (ROC)

Distribution of Target (train_metadata.csv)

- Huge deficit of malignant lesions
  - 393 compared to 383,411 Benign
- Logarithmic scale needed for human eye to view comparison in lesion type
- Definitely will play a factor in how the CNN will be trained
  - Maybe take a subset of the benign...

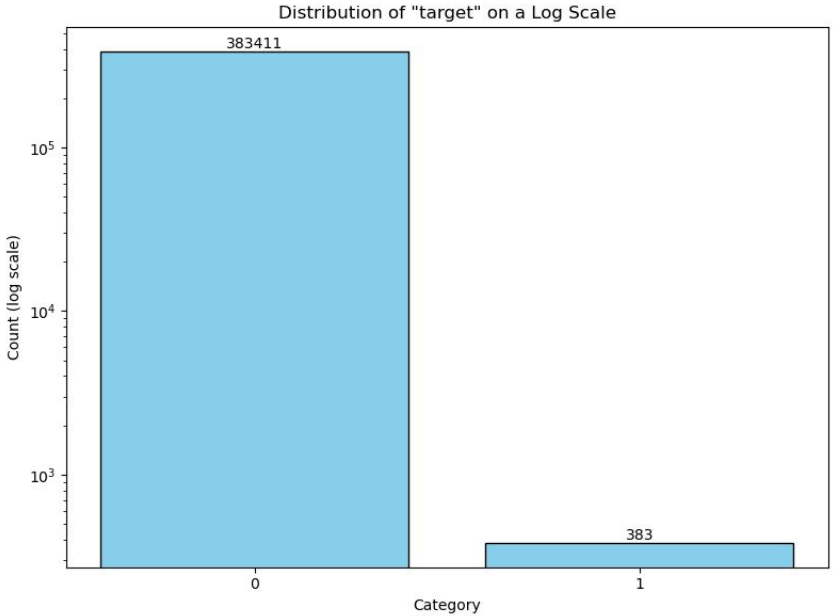


Figure 7. Log-Scaled Bar Chart of Distribution of Target

Distribution of Target by Sex (Log Scale)

Males not only **have more lesion cases** (260,798) than females (122,613), but also **more malignant ones** (274 vs 109).

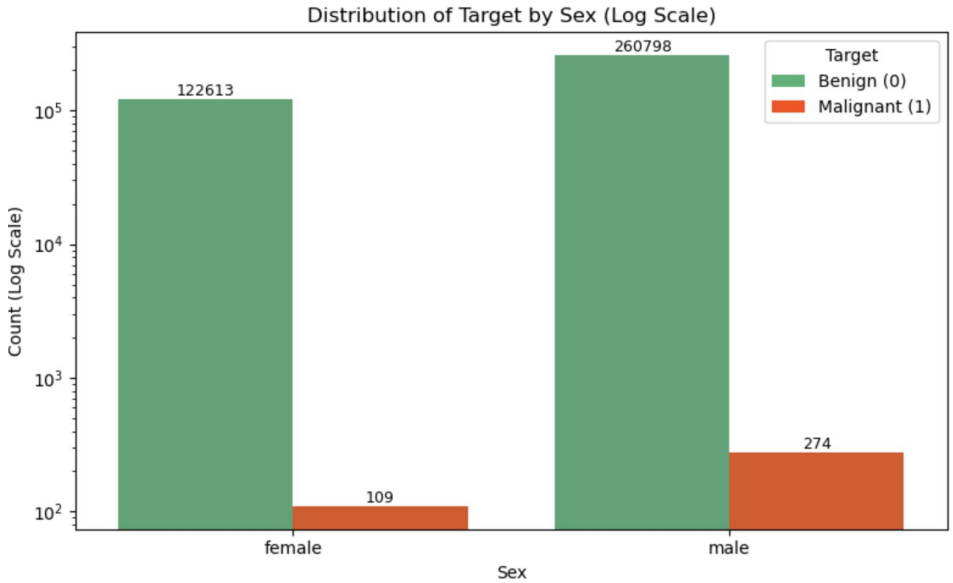


Figure 7. Log-Scaled Bar Chart of Lesion Classification by Sex

Lesion Location Site Proportions

- While most lesions are benign, malignant cases are most concentrated in the **head/neck** (77/11,678 = 0.66%), highlighting their clinical relevance despite low overall frequency.
- While the CNN doesn't currently utilize this information, it's possible it could be merged into the pipeline to classify more accurately.

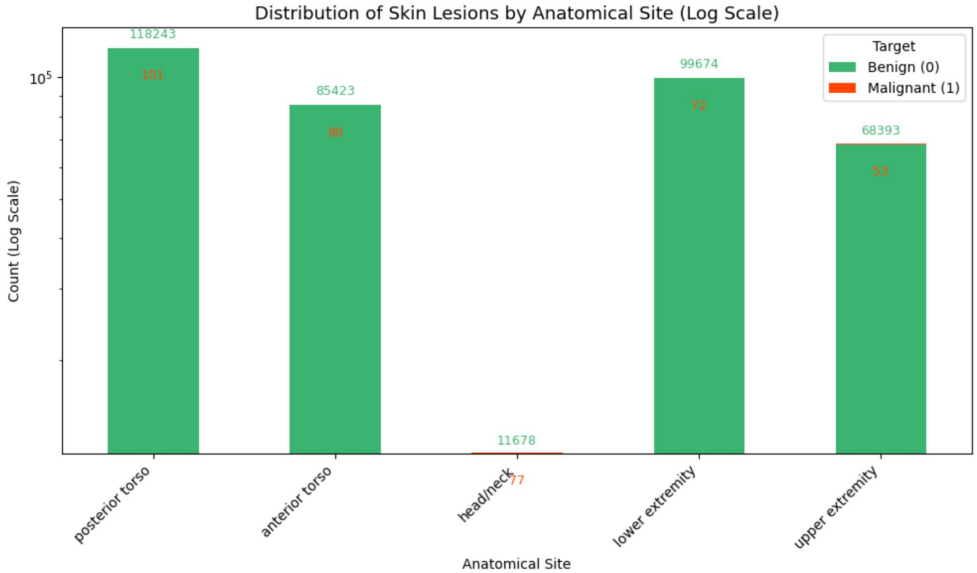


Figure 9. Distribution of Lesions by Anatomical Site with Emphasis on Malignant Clusters (Log Scale)

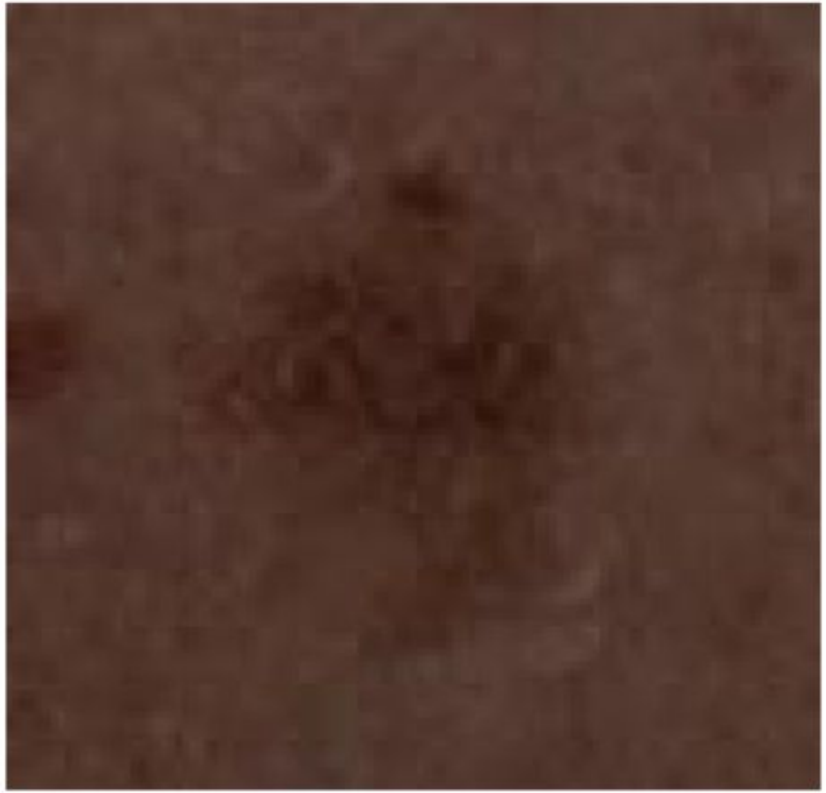
Classification via CNN

- We only sampled portion of the dataset (3%) because the dataset was too large to train model off
- From there, we created a PyTorch dataset to reference the corresponding images based on the prognosis
- To counteract imbalanced data, we used weighted penalties and data augmentation as well as undersampling the benign examples because they far outnumbered the malignant; approx. 1 to 1000
- This was followed by a 30 epoch training cycle to optimize the model, resulting in **Train Loss: 0.0070, Train Acc: 0.9679, Val Loss: 0.0066, Val Acc: 0.9675**

Visualizing Future Maps Using PyTorch

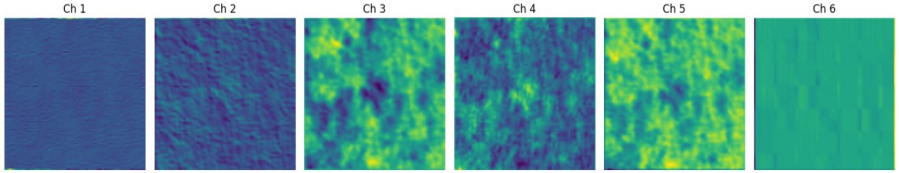
Benign

Malignant

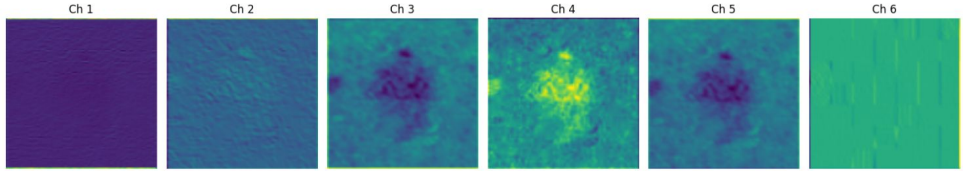


Feature Maps

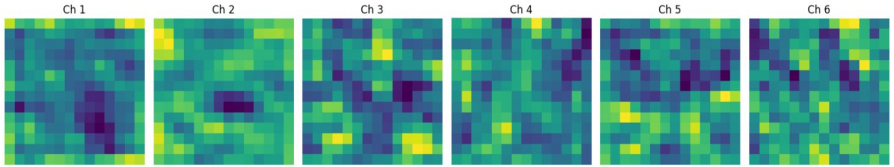
Benign Layer 1



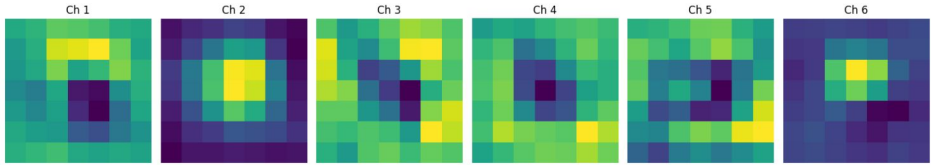
Malignant Layer 1



Benign Layer 14



Malignant Layer 14



Impact & Future Work

Our model demonstrates how **AI can assist in identifying malignant skin lesions**, particularly in **low-resource clinical settings**, using simple, smartphone-like images. Despite using only 3% of the data, our CNN achieved **97% validation accuracy (with weighted loss penalties),** showing a strong potential for real-world deployment.

Looking ahead, we aim to:

- **Scale to the full dataset** to boost robustness and generalizability.
- **Explore multimodal modeling** by integrating **metadata (e.g., age, color deviations)** with image features.
- **Improve interpretability** using **activation maps** and feature visualization to explain model decisions.
- **Ensure fairness across demographics,** reducing healthcare bias in sex and anatomical site predictions.

References

International Skin Imaging Collaboration. SLICE-3D 2024 Challenge Dataset. International Skin Imaging Collaboration https://doi.org/10.34970/2024-slice-3d (2024).

Thank You