DATASCI 281

# SKIN LESION CLASSIFICATION

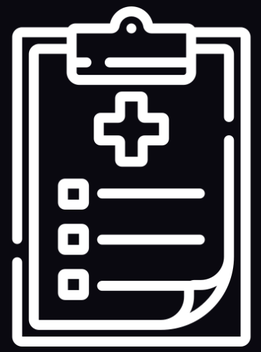
Summer 2024

ELAD OZ, WILLIAM TENG, NITIN GOMATAM, JOANNA YANG, ALEX CHEN

# INTRODUCTION

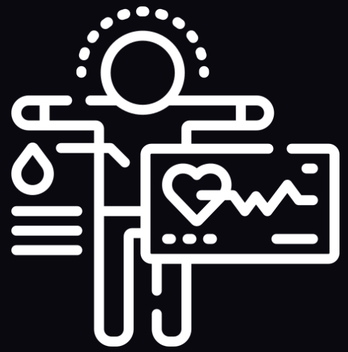
## PURPOSE

A skin lesion is an area of skin that is different from the surrounding skin and is common as as a result of injury or sunburn, but it can also be a potential marker for cancer. Our purpose is to use the MNIST HAM10000 dataset, a dataset comprising of various cancerous and non-cancerous skin lesions, to identify features that can classify skin lesions into these categories.



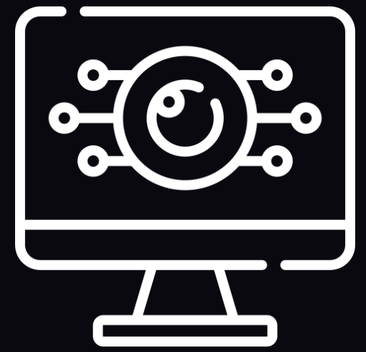
Accurate lesion identification for effective treatment planning

Early and precise diagnosis enhances treatment outcomes



Misdiagnoses can lead to delays in necessary interventions

Support organizations in standardizing diagnostic criteria



Classification methods to support patients & professionals

# CURRENTLY

**CANCER SCREENING RELIES ON VISUAL INSPECTION AND BIOPSY, FOCUSING ON KEY VARIABLES LIKE ASYMMETRY, BORDER IRREGULARITY, COLOR VARIATION, AND SIZE**

# RESEARCH **METHOD**

01

**DATA**

Acquire and review the dataset

02

**FEATURES**

Extract features from images

03

**MODEL**

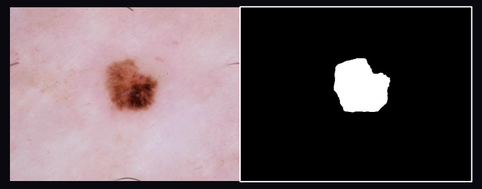
Combine features together and model

# DATA

MNIST HAM10000

# THE **DATASET**

## CANCEROUS

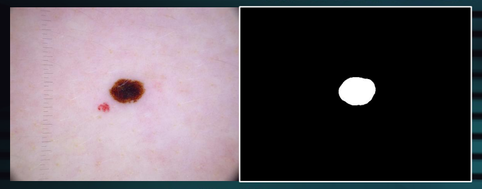


**AKIEC**

Actinic Keratoses &
Intraepithelial Carcinoma
(327 Images)

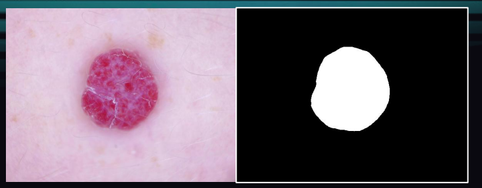
**BCC**

Basal Cell Carcinoma
(514 Images)



**MEL**

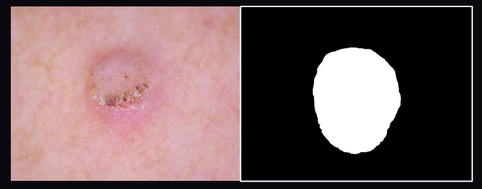
Melanoma
(1113 Images)



**VASC**

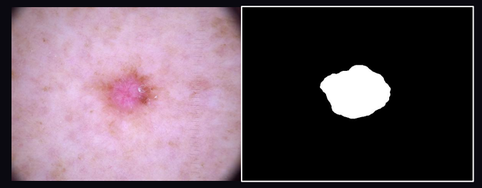
Vascular Lesions
(142 Images)

## NON-CANCEROUS



**BKL**

Benign Keratosis-like Lesions
(1099 Images)



**DF**

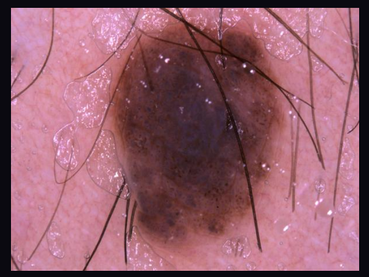
Dermatofibroma
(115 Images)

**NV**

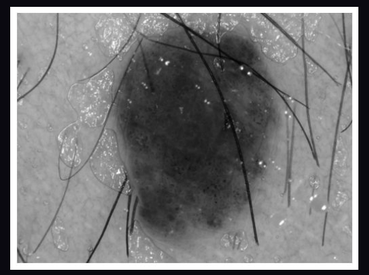
Melanocytic Nevi
(6705 Images)

# PREPROCESSING
# HAIR REMOVAL

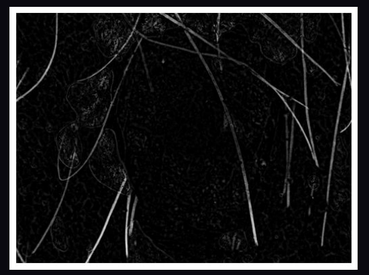
## DULLRAZOR ALGORITHM



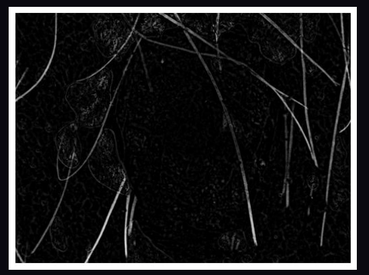
ORIGINAL



GRAYSCALE



BLACKHAT



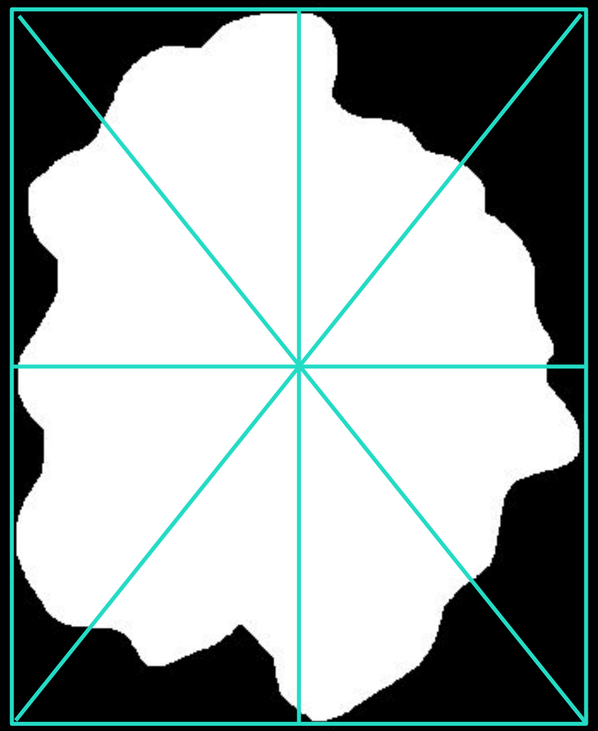
GAUSSIAN BLUR

BINARY THRESHOLD

FINAL IMAGE

# FEATURES

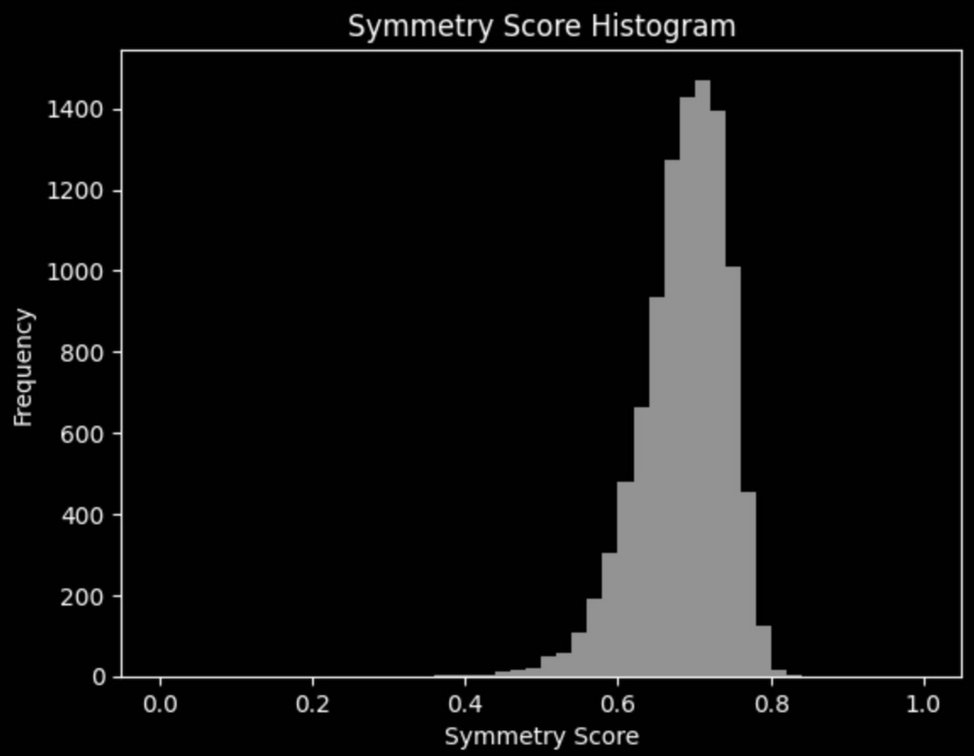
SYMMETRY | ECCENTRICITY | COLOR | TEXTURE | RESNET

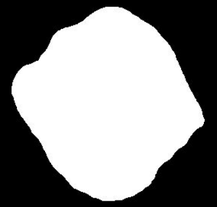


# SYMMETRY

## IMPLEMENTATION STEPS USING CV2:

1. Find contours of lesion and extract largest
2. Compute bounding box around largest contour and extract lesion region
3. Calculate symmetry scores by reflecting lesion and calculating % overlap (0-1)
    - X-axis
    - Y-axis
    - Main diagonal
    - Anti diagonal
4. Average symmetry scores to compute combined symmetry score (0-1)



**MOST SYMMETRICAL**
melanocytic nevi (NV)
symmetry_score = 0.822683

**MOST ASYMMETRICAL**
melanocytic nevi (NV)
symmetry_score = 0.362873

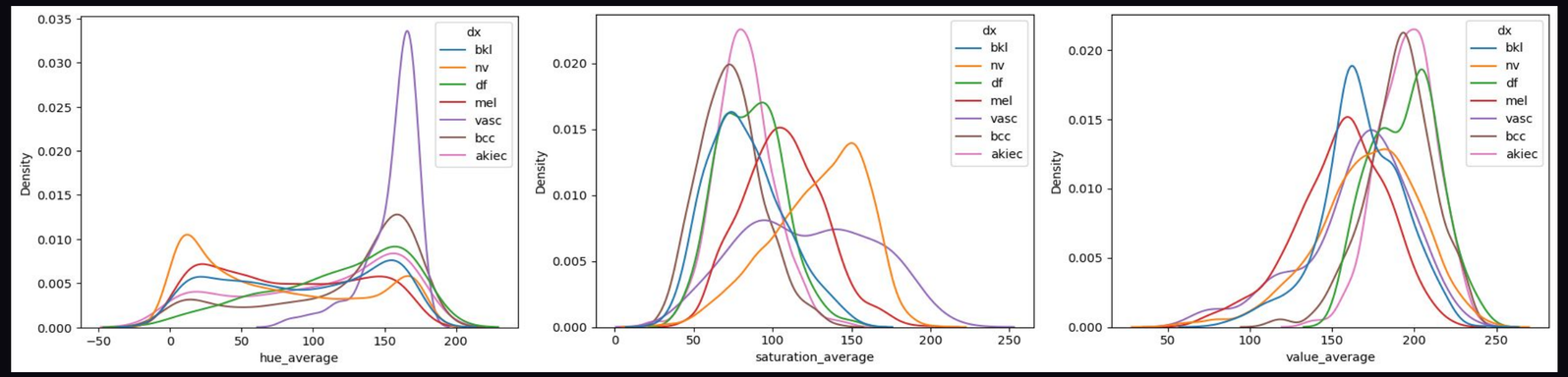
SYMMETRY

# COLOR

# HSV AVERAGE VALUE of Lesions

## IMPLEMENTATION

1. Convert the image from RGB to HSV to capture color variances between different lesions and masked out surrounding skin
2. Originally attempted to use histogram or bin values but model failed in converging

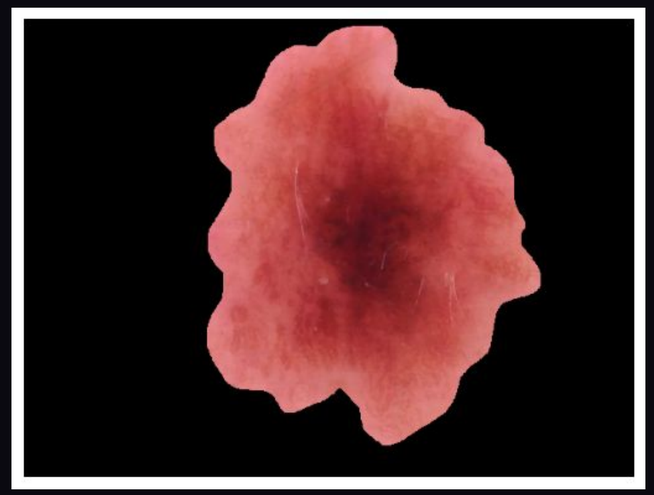


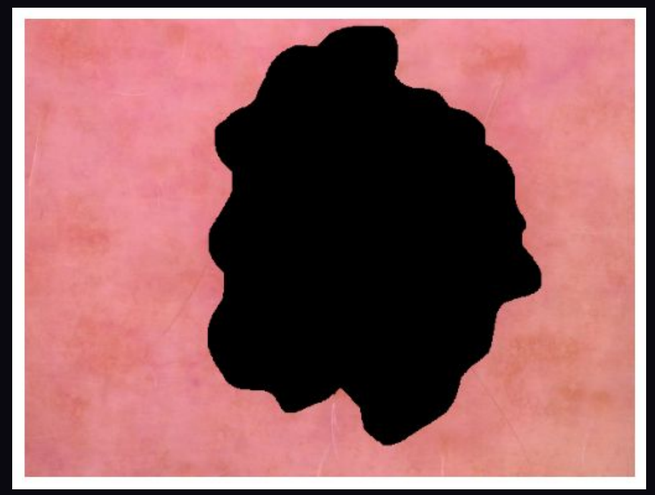
# COLOR

# HSV Color Invariances

**Lesion v. Surrounding Skin**

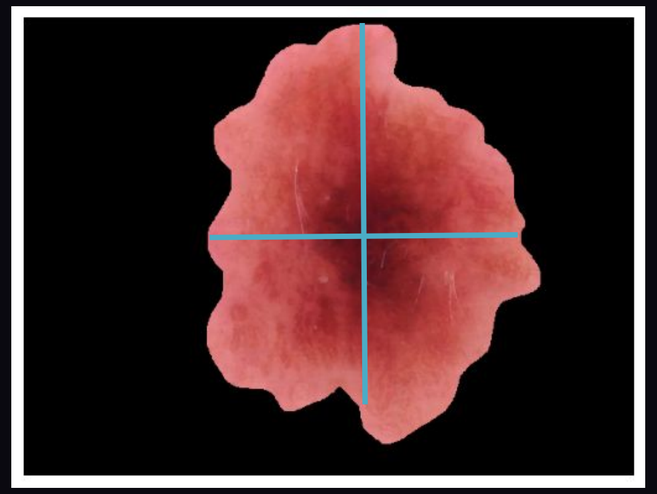
Calculate average and median hue within each lesion

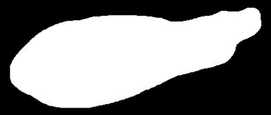




**Variances within Lesion**

Divide lesion into quadrants and compute average and median hue





Eccentricity: 0.300
(BCC - Cancerous)

Eccentricity: 0.946
(BCC - Cancerous)

# ECCENTRICITY

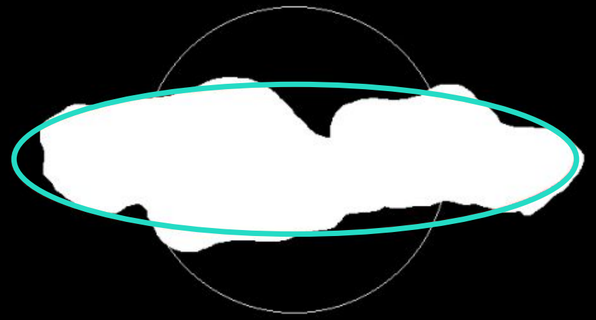
## IMPLEMENTATION STEPS USING SKIMAGE:

1. Find largest lesion in binary image and identify the centre of the lesion
2. Compute the distribution of pixels around the centre
3. Calculate eccentricity scores as a ratio of distance between

How far a skin lesion deviates from being circular (0 most circular; 1 least circular).

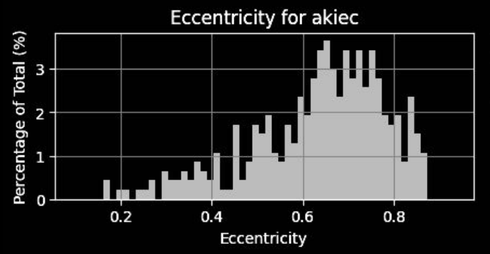
We noticed that each skin lesion type have a common eccentricity that is specific for the class. However upon calculating, BCC ranges across the spectrum, implying it may be less reliable than anticipated as a feature.

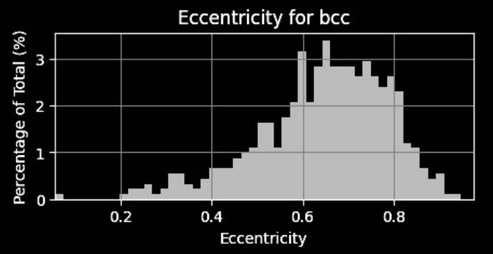
**LEAST ECCENTRIC**
Eccentricity: 0.088
(NV - Non Cancerous)

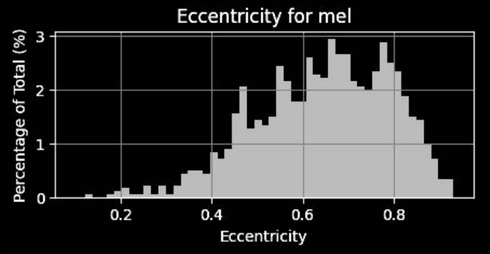


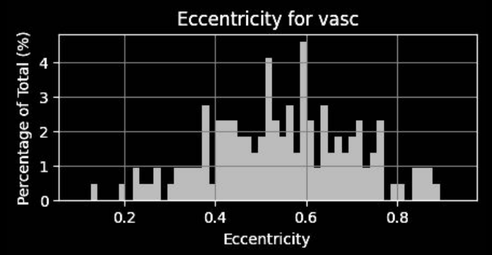
**MOST ECCENTRIC**
Eccentricity: 0.964
(BKL - Non Cancerous)

**CANCEROUS**

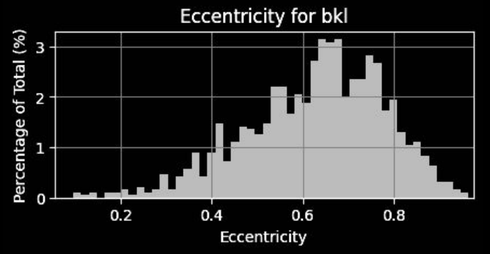


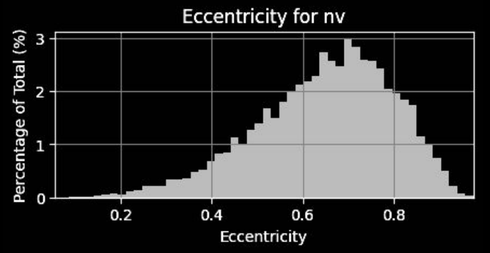



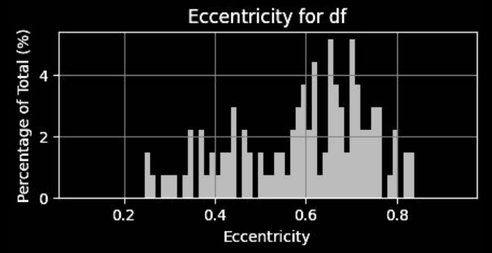



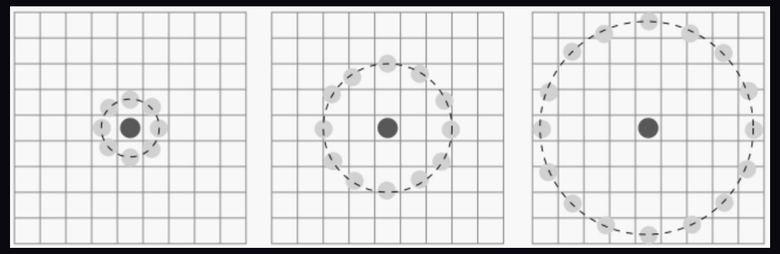




**NON-CANCEROUS**





Linear Binary Patterns

Gabor Filters

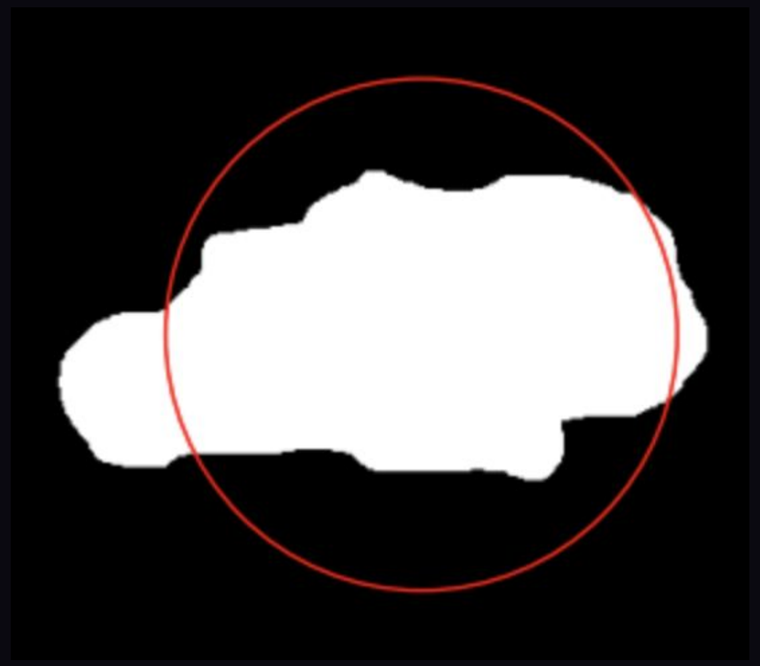
# TEXTURE

## Gabor Filters

- Band pass filters
- Usually used in a collection with multiple sets of hyperparameters, controlling:
  - spatial orientation
  - frequency band & filter profile
- Convolved with image

## Linear Binary Pattern

- Encode region around each pixel with a binary string, each element represents whether a pixel in some location is larger or smaller than region center.

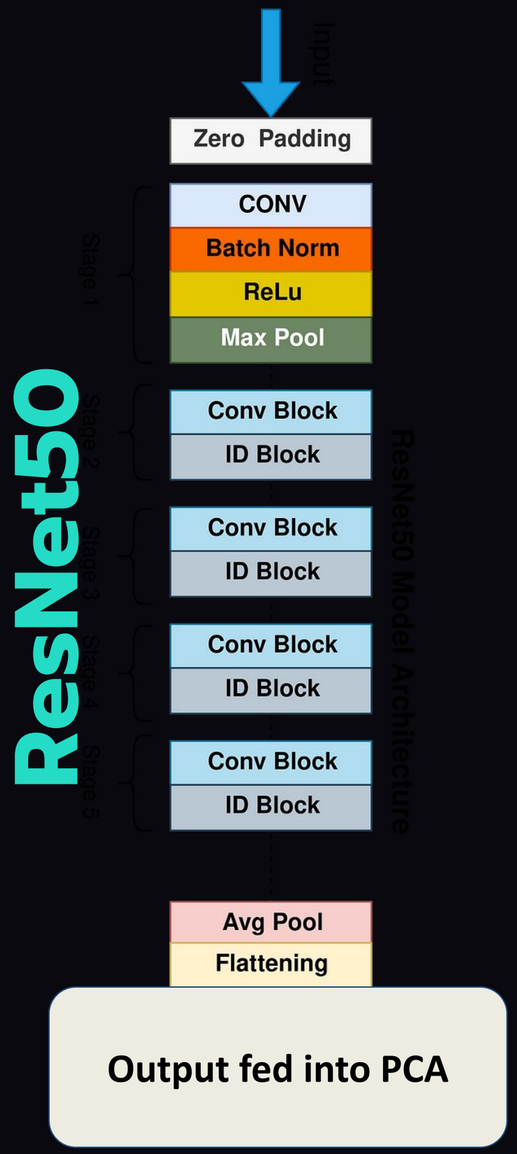


Polar Axes for Aggregation

# TEXTURE

## IMPLEMENTATION STEPS:

1. Identify largest bounded circle & smallest bounding circle given segmentation mask.
2. Compute feature statistics over over concentric rings, aggregate along radial direction.
3. Example statistics calculated:
    - Standard deviation, skew of ring means
    - Slope of ring mean on radius from linear regression
    - Radial bin number for ring with highest mean ring
    - Interquartile range of values over entire circle




# ResNet50

ResNet50 is an advanced image classification model and used in an attempt to retrieve features beyond our own feature extractions. Its architecture includes convolutional layers for feature extraction, identity and convolutional blocks for feature transformation, and fully connected layers for classification.

## IMPLEMENTATION STEPS:

1. Pass images through the pre-trained neural network
2. Obtain output vectors and use PCA to reduce dimensionality
    - Attempted dimensional reduction of 2, 4, and 6 to assess during modeling

Image from [The Annotated ResNet-50]

# MODEL

LOGISTIC REGRESSION | RANDOM FOREST

# MODELLING

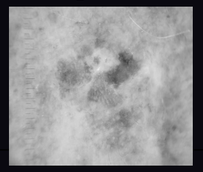
## GETTING THE DATA RIGHT

### DUPLICATE LESION ID REMOVAL

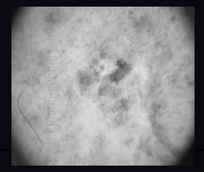
Picked the first image in list for each lesion ID

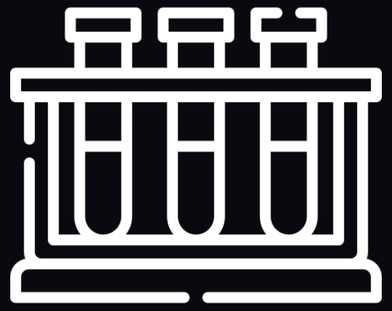
Leaves 7470 samples in total

ISIC_0027419



ISIC_0025030





### TRAIN-TEST SPLIT

70% train, 15% validation, 15% test

### CLASS IMBALANCE

Stratified Splitting: Maintains the original class proportions in both training and test sets.

Class Weights: Adjusts the model to give more importance to minority classes.

SMOTE: Generates synthetic examples for minority classes to balance the dataset

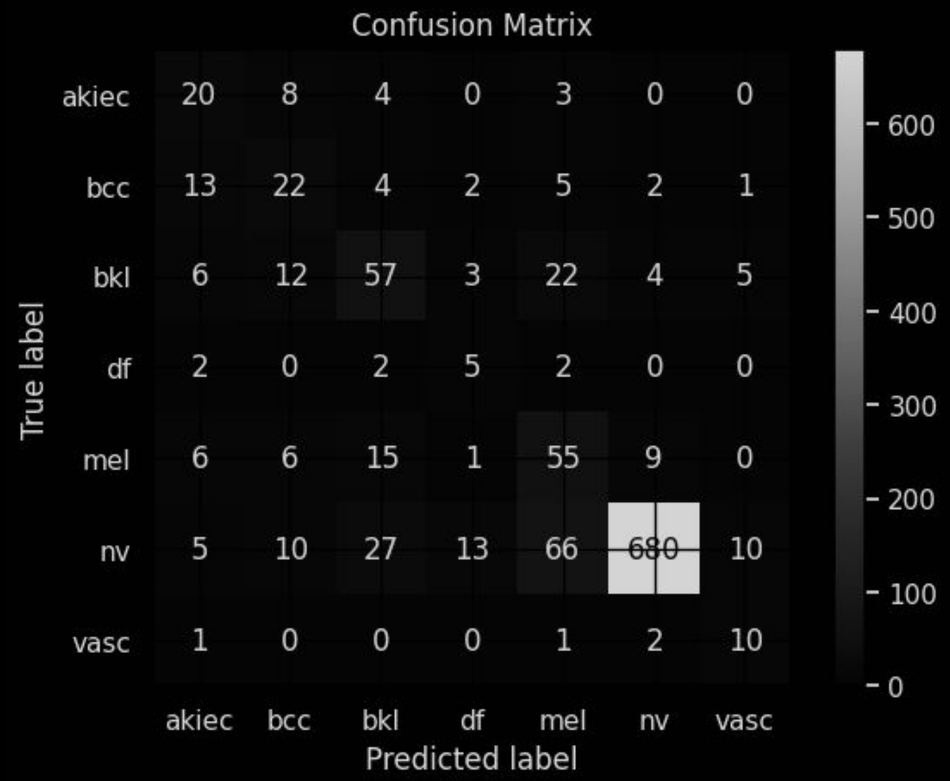
# LOGISTIC REGRESSION

## PRE PROCESSING

- Numerical features standardized
- Categorical features encoded (one-hot)
- Fill missing “age” with average age, and 0 for other features

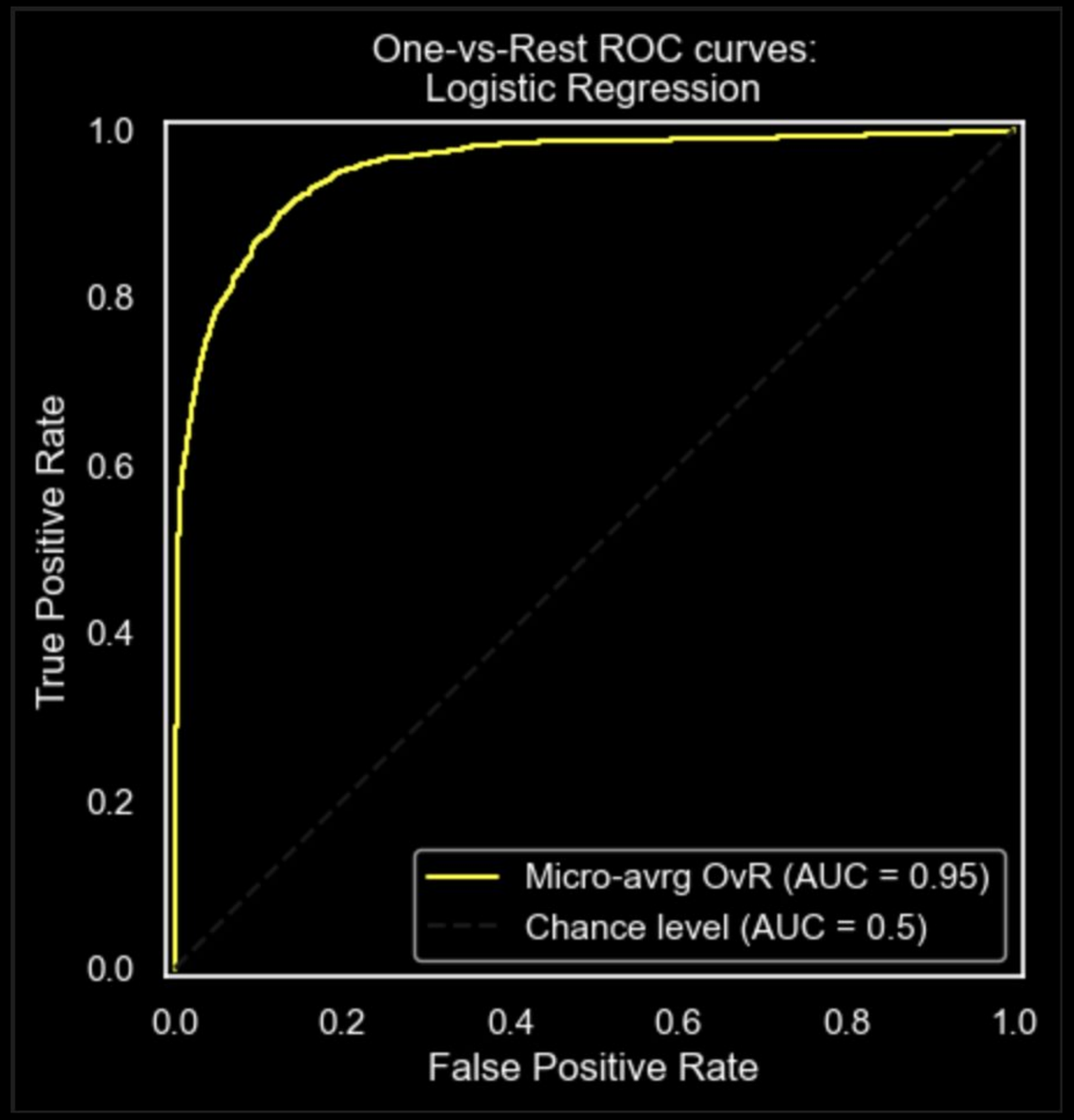
## HYPER PARAMETERS

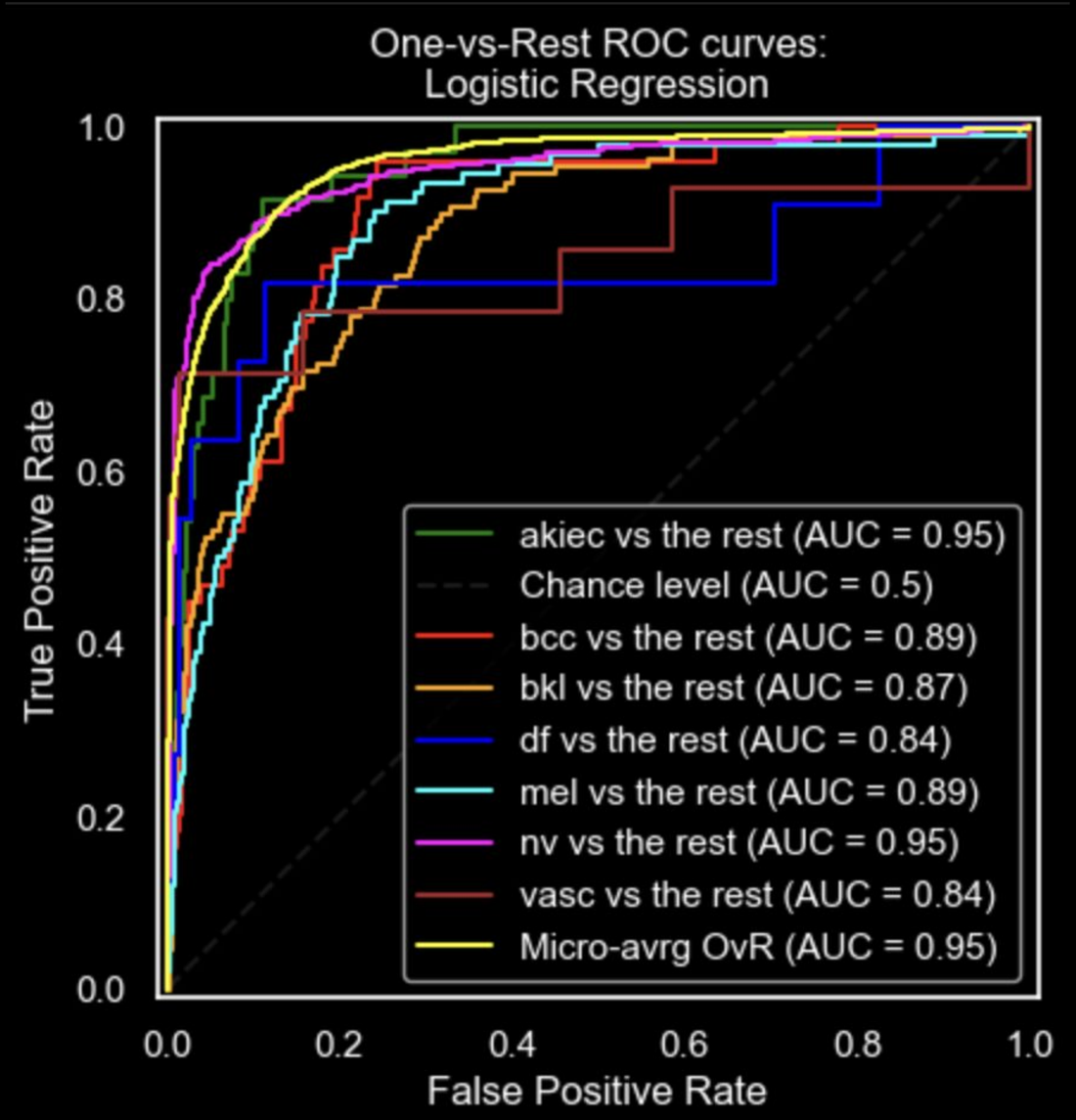
- Max iterations
- Sample weights: balanced classes - proportional to inverse class frequency




| Class | Precision | Recall | F1 |
|---|---|---|---|
| akiec | 0.27 | 0.41 | 0.33 |
| bcc | 0.26 | 0.41 | 0.32 |
| bkl | 0.48 | 0.50 | 0.49 |
| df | 0.19 | 0.45 | 0.26 |
| mel | 0.36 | 0.59 | 0.44 |
| nv | 0.97 | 0.81 | 0.88 |
| vasc | 0.39 | 0.60 | 0.47 |

# LOGISTIC REGRESSION



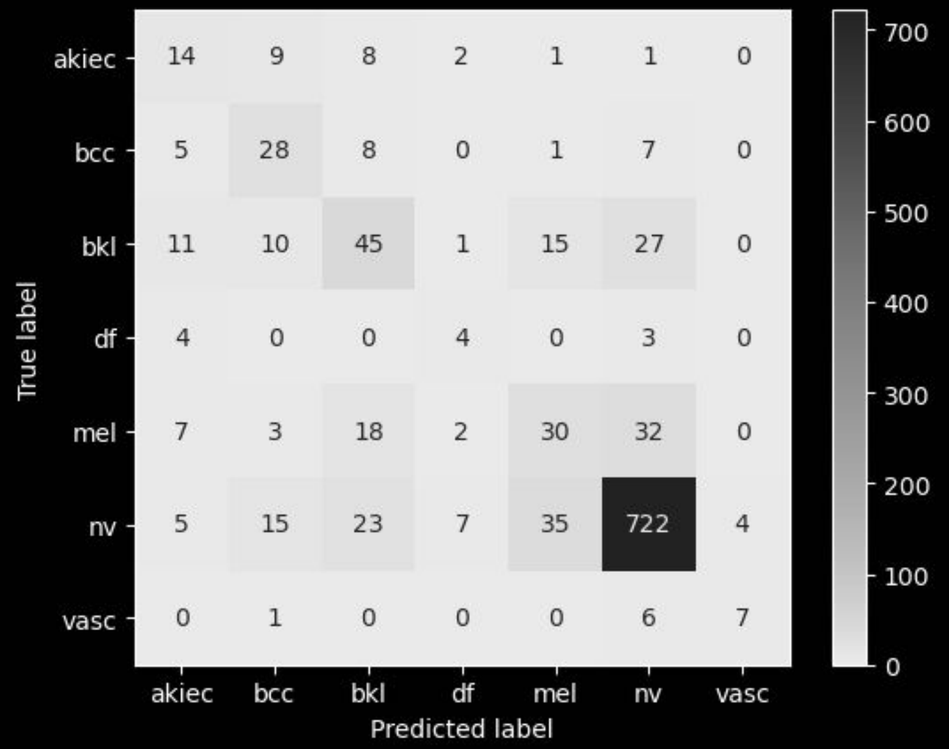
# RANDOM FOREST

## PRE PROCESSING

- Balanced random forest with bootstrap class weighing: draw a bootstrap sample from minority classes
- sample with replacement the same number of samples from majority class
- SMOTE application

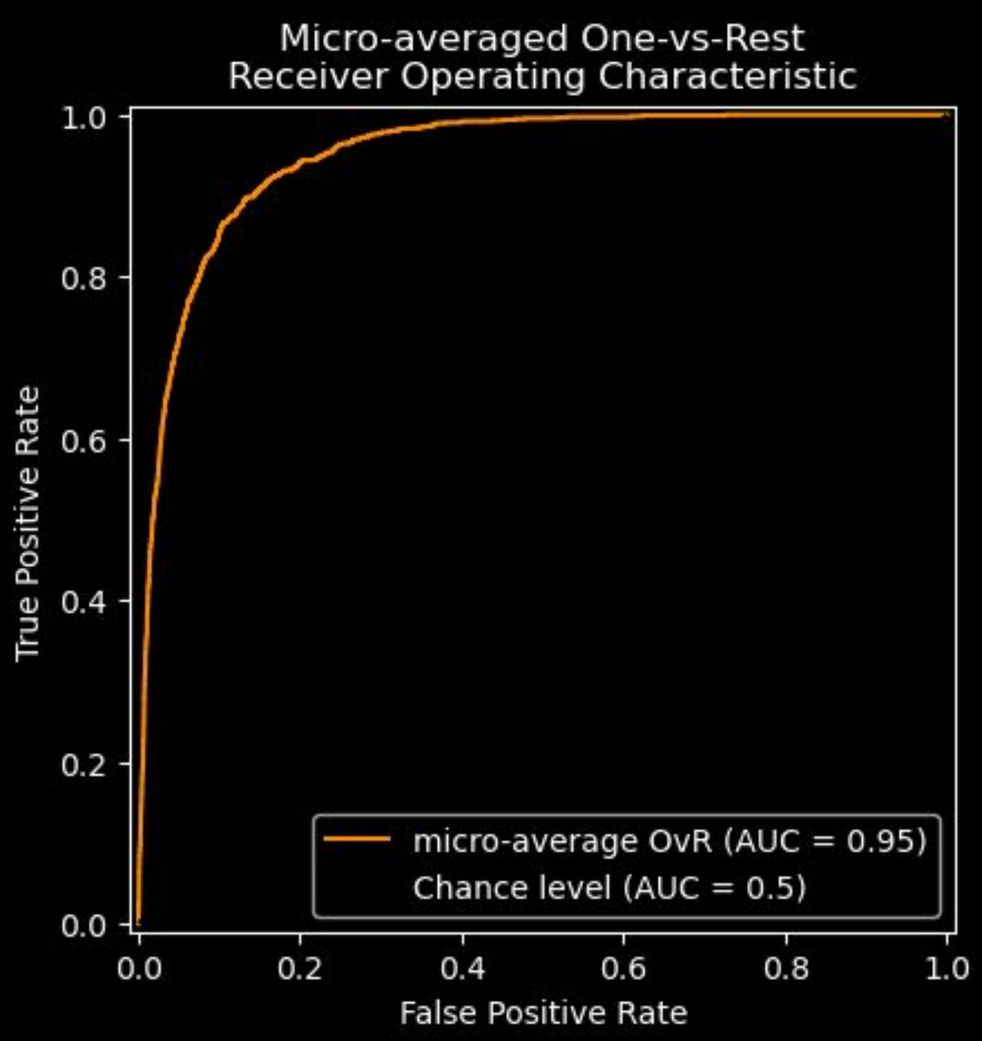
## HYPER PARAMETERS

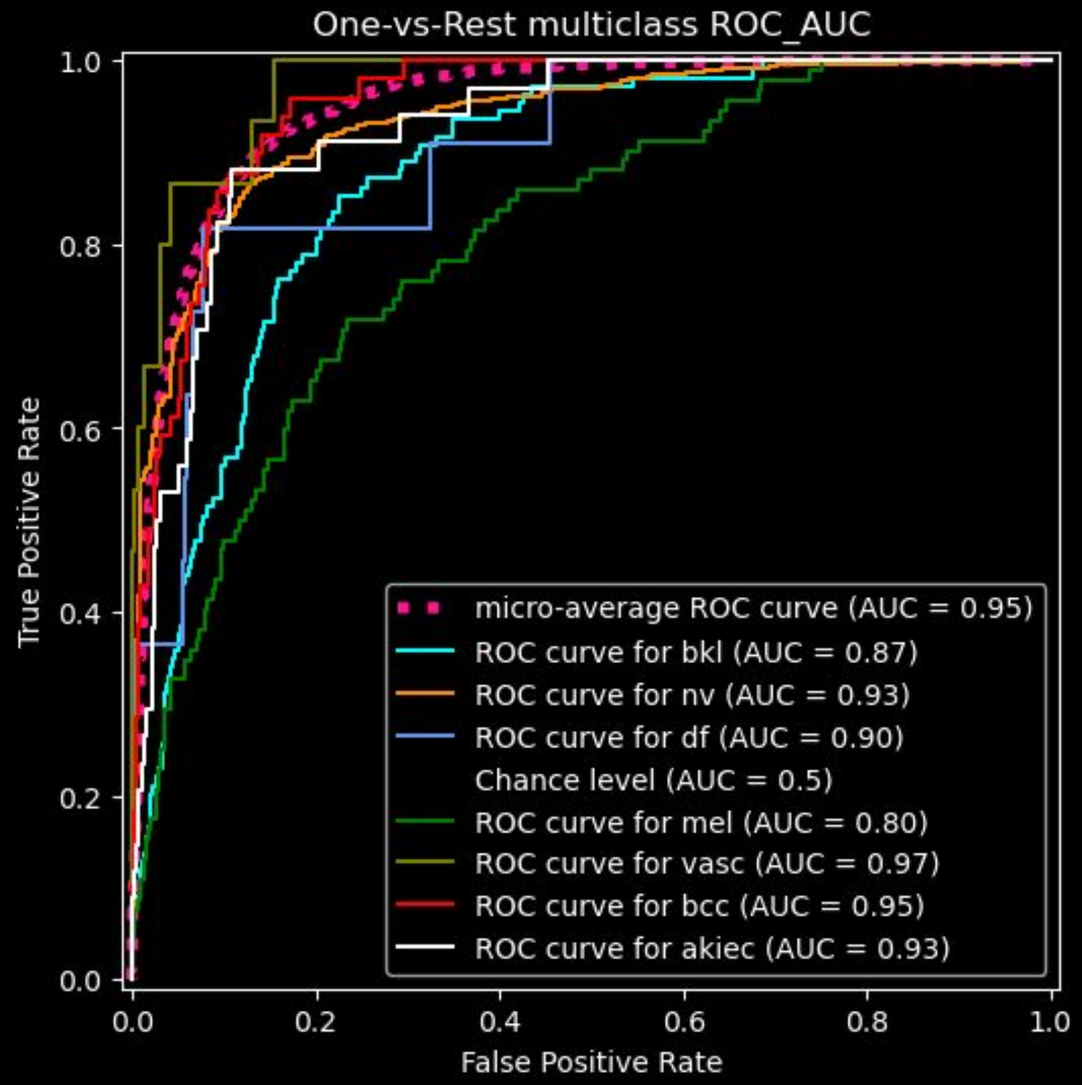
- Grid search over:
  - Number of estimators [100, 200, 300]
  - Max depth [3, 5, 10, 12]
- Best: 300 estimators with max depth of 12




| Class | Precision | Recall | F1 |
|---|---|---|---|
| akiec | 0.30 | 0.40 | 0.34 |
| bcc | 0.42 | 0.57 | 0.49 |
| bkl | 0.44 | 0.41 | 0.43 |
| df | 0.25 | 0.36 | 0.30 |
| mel | 0.36 | 0.33 | 0.34 |
| nv | 0.90 | 0.89 | 0.89 |
| vasc | 0.63 | 0.50 | 0.56 |

# RANDOM FOREST



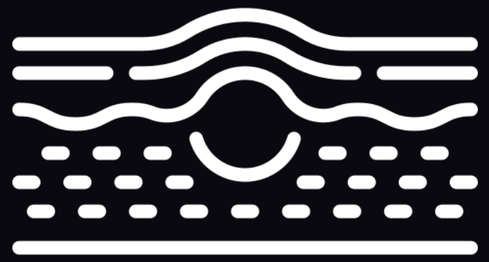
# RESULTS

OVERVIEW

| | VALIDATION ACCURACY [%] | TEST ACCURACY [%] |
|---|---|---|
| LOGISTIC REGRESSION | 72.5 | 75.7 |
| RANDOM FOREST | 77.6 | 76.7 |

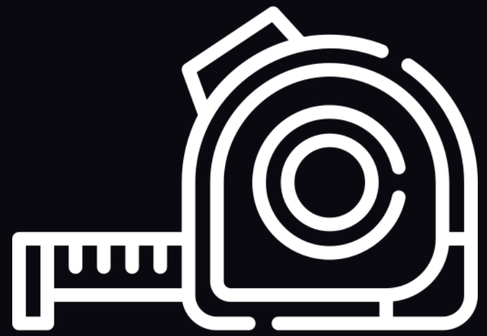
# CURRENT LIMITATIONS

## FUTURE SCOPE



### Lesion Protrusion

Existing data is 2D, therefore unable to reliably determine protrusion



### Lesion Size

Additional size metric to measure relative size of lesions

### Lesion Age

To understand lesions at different stages of development

# CONCLUSION

## Handling Data

In addition to provided metadata, feature extraction was done on common visual indicators of cancer. Addressing class imbalance through class weighting for ensuring that cancerous lesions, which were underrepresented

## Model Performance

The Logistic Regression and Random Forest models, optimized through hyperparameter tuning and class weighting, demonstrated promising results, showcasing the value of combining simpler models with carefully selected features.

## Future Directions

Refining hair removal algorithms, developing advanced feature extraction techniques, finding additional datasets, and exploring neural networks for classification of skin lesions.