SKIN LESION SEGMENTATION AND CLASSIFICATION

ISIC 2019 Dataset · 8 Disease Classes · 25,331 Images

Data Science Practicum I · Regis University

25,331	8	90%+	53:1
Images	Disease Classes	Target Accuracy	Class Imbalance

02

PROBLEM STATEMENT

Why automated skin lesion analysis matters

01

Skin cancer is one of the most common and dangerous cancers globally.

02

Early detection drastically improves patient survival rates.

03

Manual diagnosis is time-consuming and prone to human error.

04

Rare skin diseases are often misdiagnosed due to a lack of examples.

→ *To solve this, we first need to understand our data...*

03

OBJECTIVE

What this project set out to achieve

01	Build a robust multi-class CNN classifier for 8 skin lesion types.
02	Develop a preprocessing pipeline to remove visual artifacts (hair, lighting).
03	Compare techniques to solve extreme data imbalance — Class Weights, SMOTE, CGANs.
04	Evaluate whether lesion segmentation improves classification accuracy.
05	Ensure model decisions are interpretable for clinical trust via Grad-CAM.

04

DATA SOURCES & OVERVIEW

ISIC 2019 — the benchmark dataset for skin lesion research

Source	ISIC 2019 Challenge Dataset
Volume	25,331 dermoscopic images
Classes	8 disease categories
Ground Truth	Confirmed via histopathology (biopsy)
Why Real?	Clinical images from hospitals worldwide

8 DISEASE CLASSES

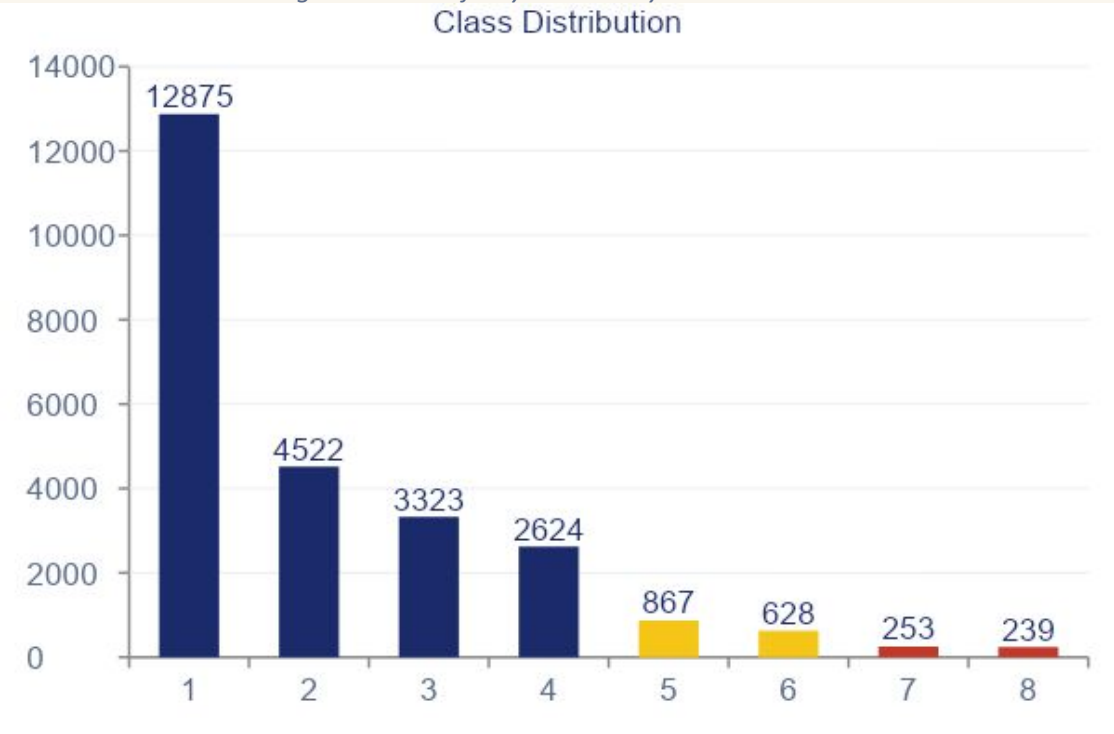
- NV — Melanocytic Nevus
- MEL — Melanoma
- BCC — Basal Cell Carcinoma
- BKL — Benign Keratosis
- AK — Actinic Keratosis
- SCC — Squamous Cell Carcinoma
- VASC — Vascular Lesion
- DF — Dermatofibroma

→ ***The biggest challenge with this data is the extreme class imbalance...***

05

EXTREME CLASS IMBALANCE

The core challenge — 53:1 majority to minority ratio



53:1

Majority to Minority Ratio

12,875

NV — largest class

239

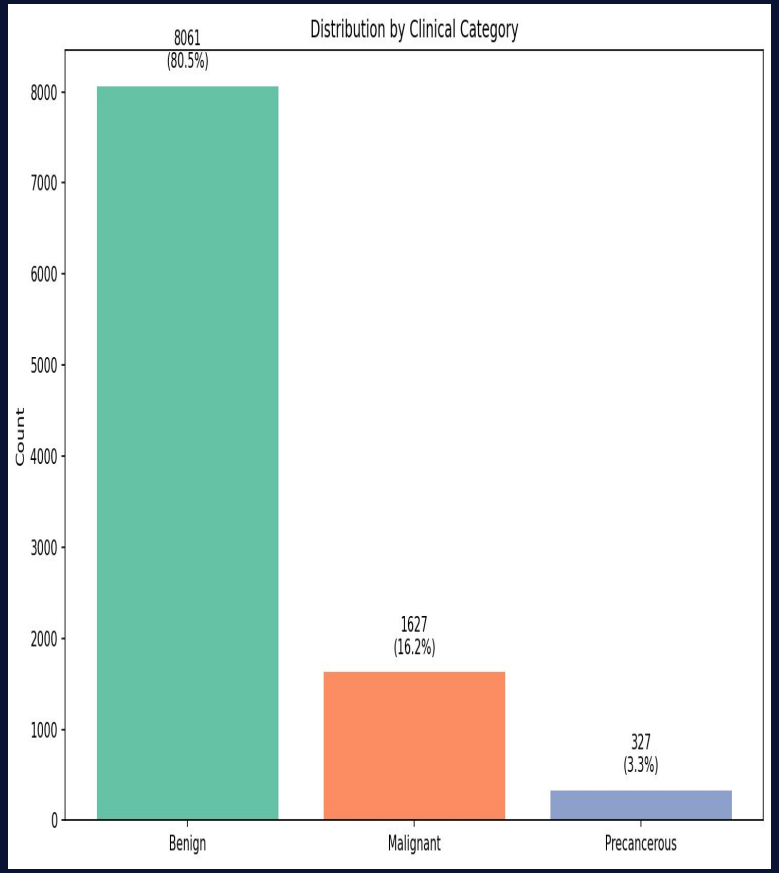
DF — smallest class

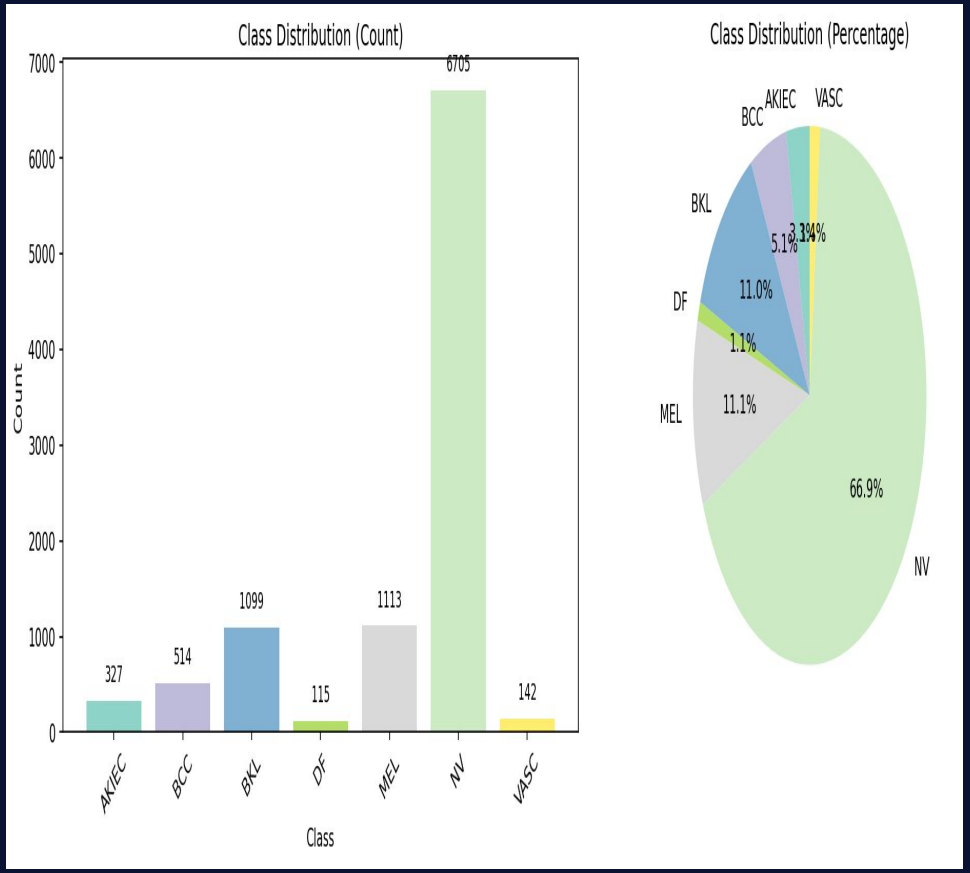
33.4%

Malignant classes combined

→ *Before modeling, we need to clean and standardize the images...*

Class Distribution — Visual





06

IMAGE CHARACTERISTICS

What the raw data looks like before preprocessing

01

Varying Dimensions

Images range from 600×450 to 1024×1024. Median size is 1024×768. Must be standardized before feeding into CNN.

02

Hair Artifacts

88% of sample images contain visible hair structures that can mislead the CNN into learning the wrong features.

03

Brightness Variation

Standard deviation of brightness is 29.13 — very high. Different clinics and cameras create inconsistent lighting.

04

Color Dominance

Red channel dominates across most images due to skin tone and lesion characteristics. Normalization is critical.

→ *These issues are fixed by our preprocessing pipeline...*

07

IMAGE PREPROCESSING PIPELINE

Cleaning images before they reach the model

01

Resize

Bilinear interpolation → 256×256

All images resized to a standard 256×256 pixels. Ensures consistent input dimensions for the CNN regardless of original camera resolution.

02

Hair Removal

Blackhat filter + Telea inpainting

Morphological blackhat filtering detects dark elongated hair strands. cv2.inpaint then replaces those pixels with surrounding skin texture.

03

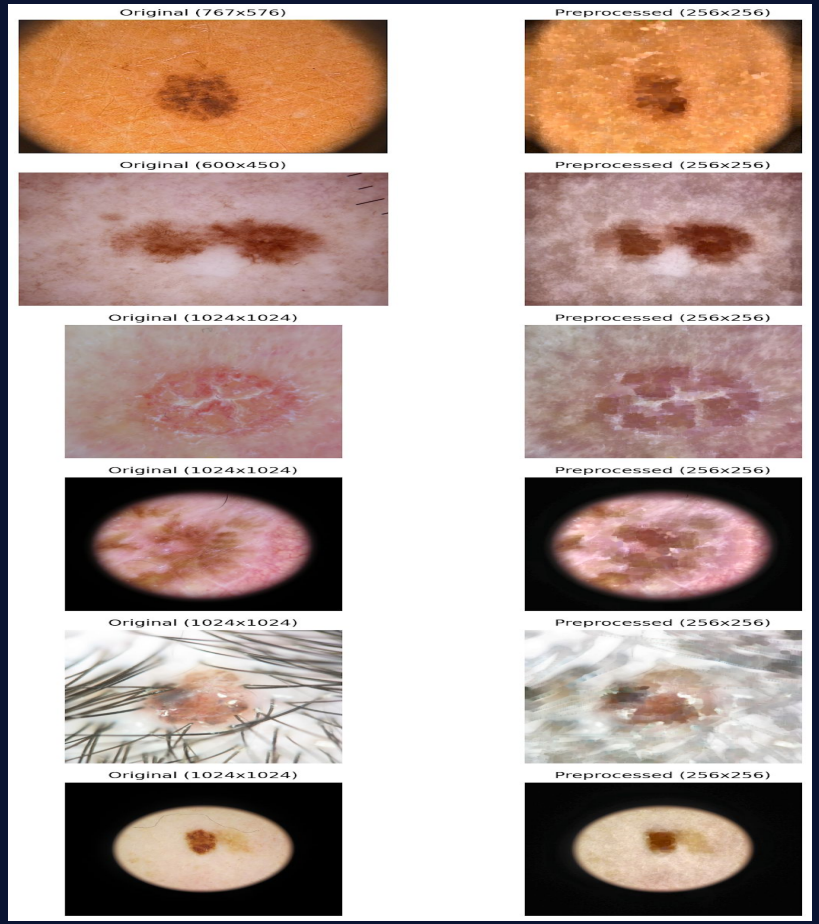
CLAHE Normalization

L-channel in LAB color space

Enhances local contrast without distorting lesion color. Applied only to the Lightness channel — A and B (color) channels stay untouched.

→ ***Now with clean data, let's look at how the model was built...***

Preprocessing Pipeline — Before & After



08

MODEL ARCHITECTURE — EfficientNetB2

Transfer learning with two-phase fine-tuning

TWO-PHASE TRAINING STRATEGY

PHASE 1 — WARM-UP

Learning Rate: 1e-2

Backbone is FROZEN. Only the dense head is trained first. This prevents the randomly-initialized head from destroying the pre-trained ImageNet weights through large gradient updates.

PHASE 2 — FINE-TUNING

Learning Rate: 1e-4

Backbone is UNFROZEN. The entire network trains end-to-end at a very small learning rate. This allows the CNN to adapt ImageNet features to dermoscopic images without forgetting them.

→ But even a great model fails on imbalanced data — here's how we fixed that...

09

SEGMENTATION — BCDU-NET

Isolating the lesion from background skin

01 What is it?

BCDU-Net = Bi-directional ConvLSTM U-Net with Dense blocks. A specialized architecture for medical image segmentation.

02 Why use it?

Background skin confuses the classifier. BCDU-Net creates a binary mask — zeroing out everything except the lesion itself.

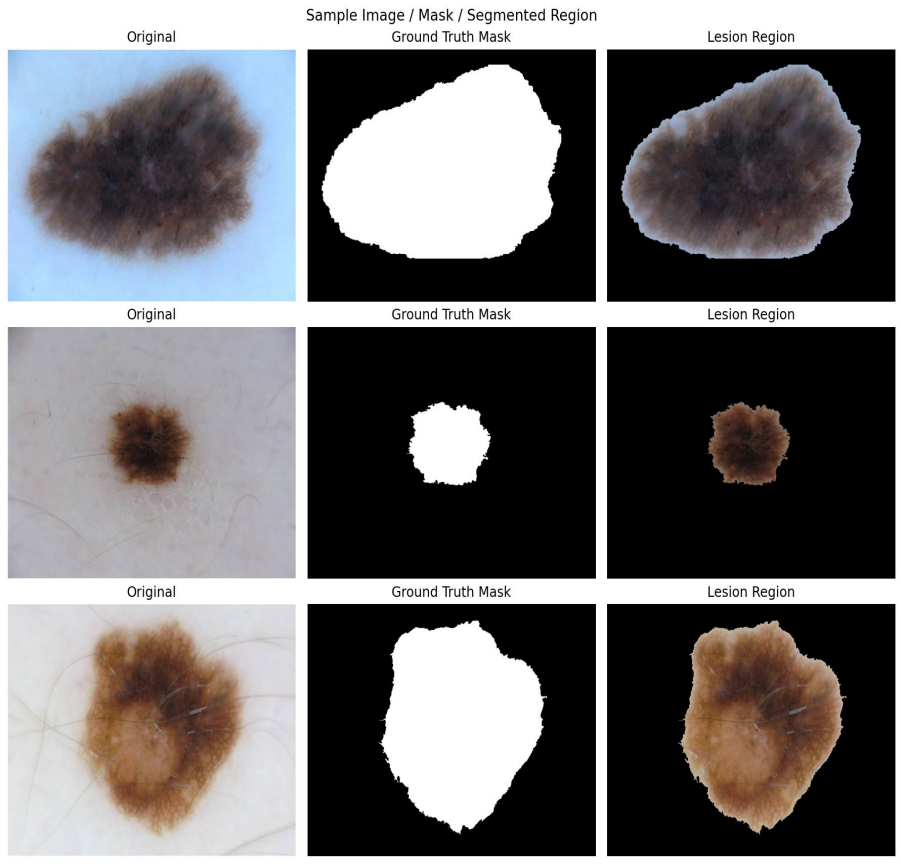
03 How it works

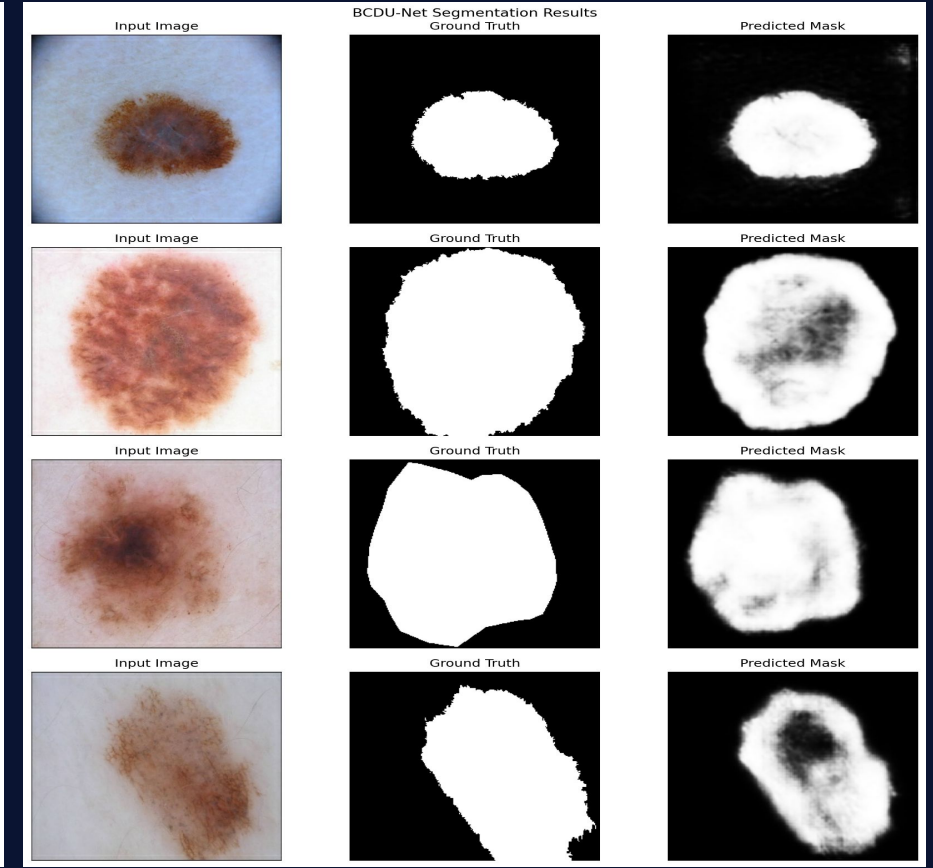
Dense blocks reuse features across layers. The Bi-directional ConvLSTM at the bottleneck captures complex spatial patterns in both directions.

04 Loss Function

Hybrid BCE + Dice Loss. BCE handles pixel-level accuracy. Dice Loss maximizes IoU — critical since the lesion is a small part of the image.

BCDU-Net Segmentation Results





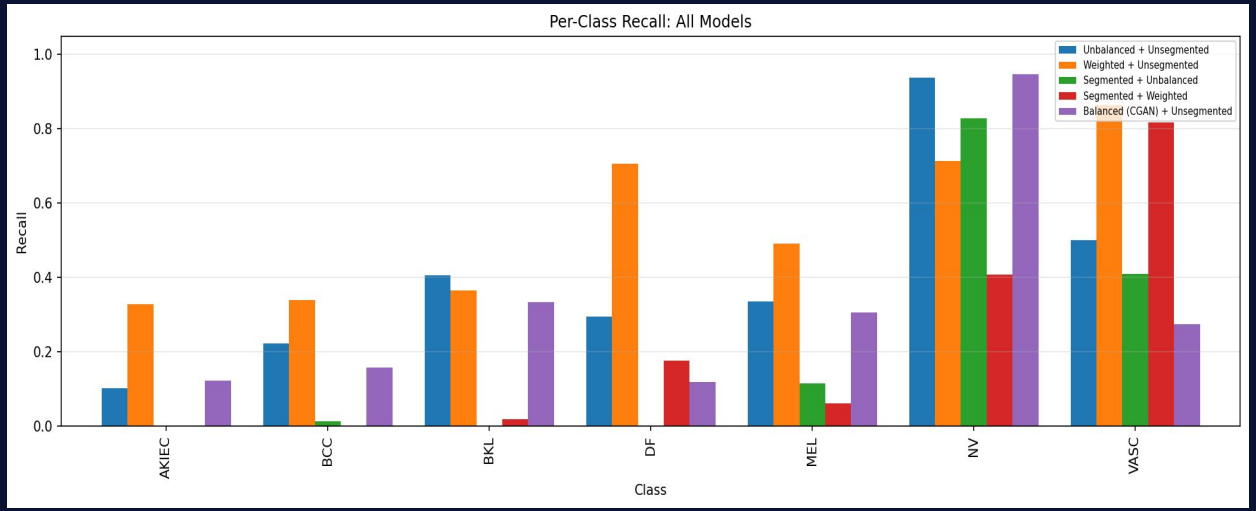
TACKLING CLASS IMBALANCE — 4 STRATEGIES

Systematic comparison of balancing approaches

Strategy	Accuracy	Macro F1	DF Recall	Verdict
Baseline (Unbalanced)	73%	0.57	0.33	High accuracy — but misses 67% of rare tumors
Class Weights	62%	0.49	0.69	Easy to add — but trades precision, F1 actually drops
SMOTE (Feature Space)	—	↑	↑	Better balance — but backbone frozen, no new pixel learning
CGAN (Pixel Space)	Best	0.79	Best	Generates real new images — best end-to-end retraining ✓

Per-Class Recall — Balancing Strategy Comparison

Baseline misses rare cancers near 0% recall on AKIEC, only ~20% on BCC and MEL. A model with 73% accuracy is still dangerous. Class weighting fixes minority classes recall jumps 3–5x on rare classes like DF and AKIEC just by reweighting the loss. No strategy wins everything every model still dominates on NV (the majority class). Balancing is a tradeoff, not a cure.



RESULTS & EVALUATION

Macro F1 — not Accuracy — is the true measure of success

73%

Baseline Accuracy (misleading)

0.79

Best Macro F1 — CGAN model

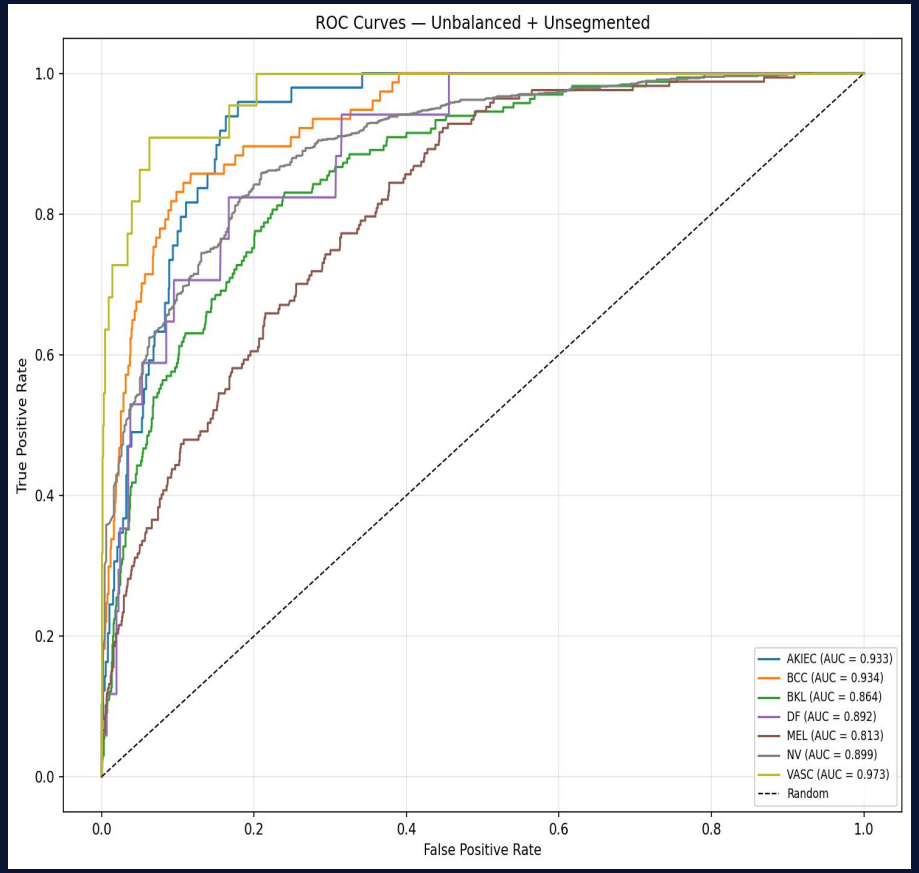
33%

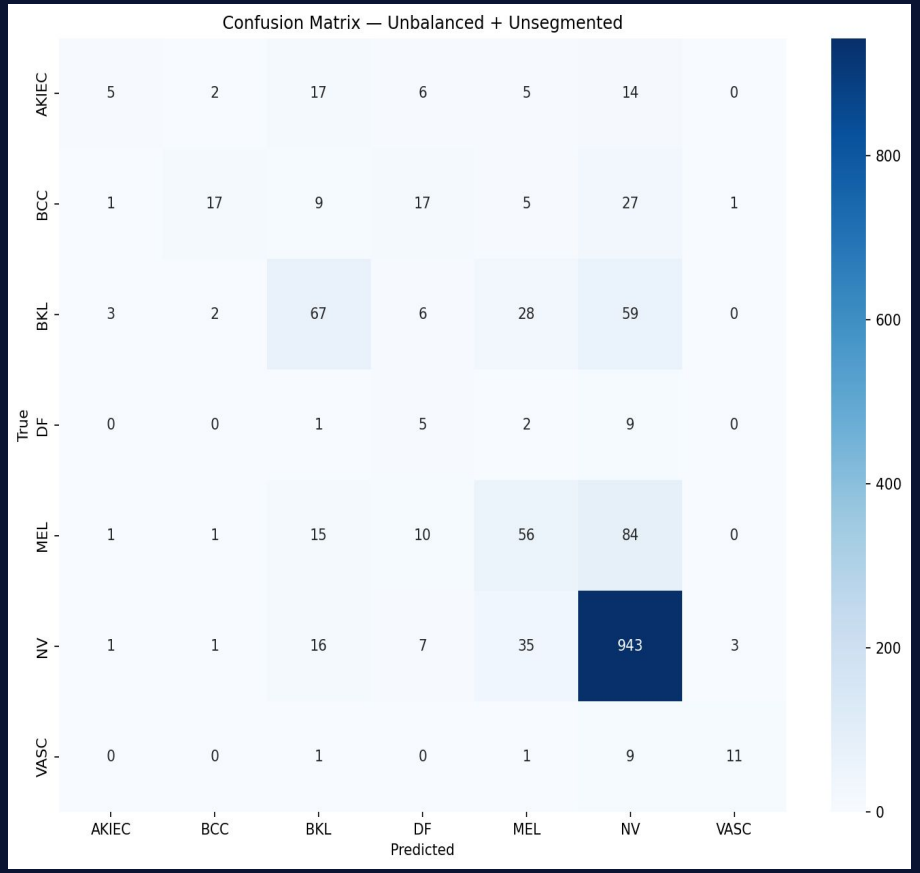
DF Recall — Baseline (terrible)

↑ High

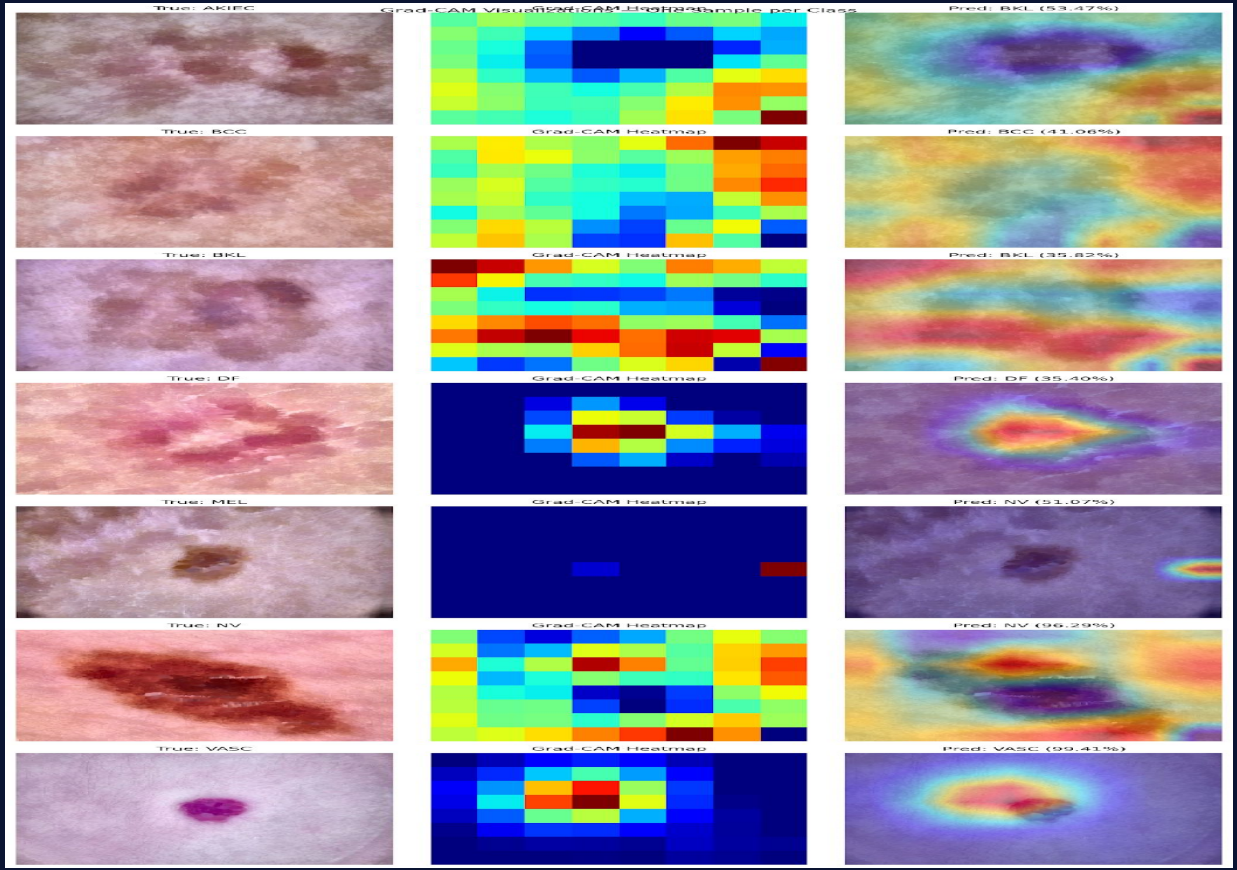
DF Recall — CGAN (fixed!)

Model Evaluation — ROC Curves & Confusion Matrix

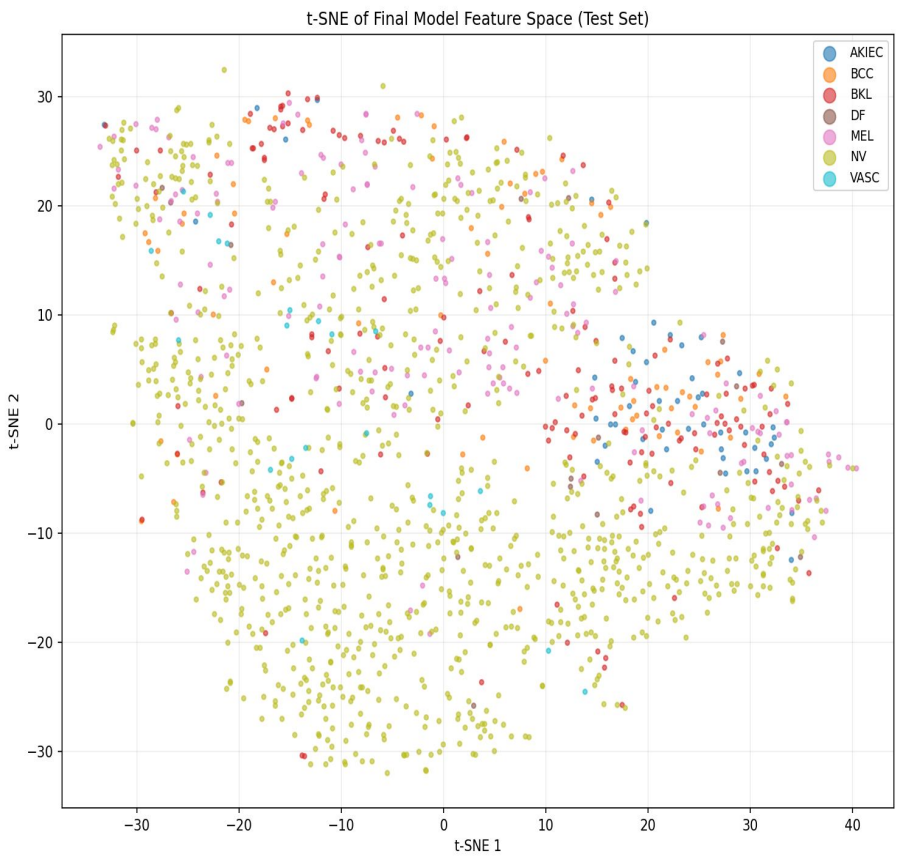


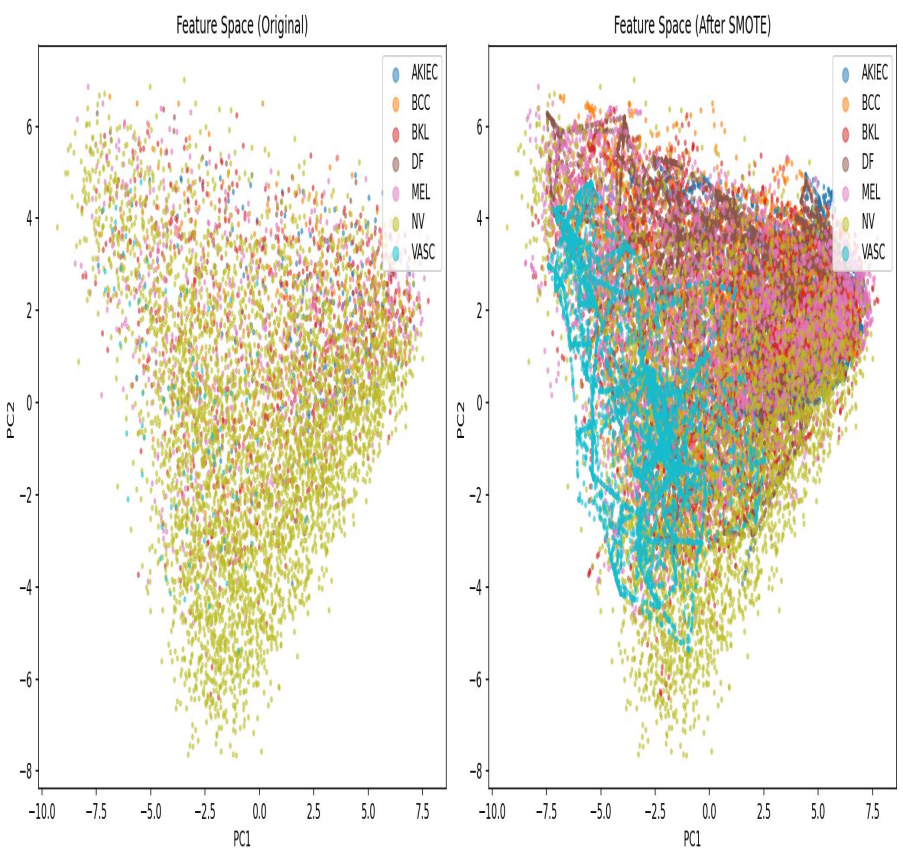


Grad-CAM — What the Model Looks At



Feature Space — t-SNE & SMOTE Analysis





KEY INSIGHTS & LEARNINGS

What this project teaches about building medical AI

01 Accuracy is a Liar in Healthcare

73% accuracy sounds great — until you find out 67% of rare tumors are missed. Always use Macro F1 + Per-Class Recall.

02 Don't SMOTE Pixels

SMOTE on raw images creates ghosting artifacts. Apply it in the CNN's 1408-D feature space instead — much cleaner results.

03 Class Weights Backfire

Weighting the loss function trades majority precision for minority recall. F1 can actually drop — from 0.57 to 0.49 in our case.

04 Segmentation = Forced Attention

BCDU-Net masking forces the CNN to only look at the lesion. Acts as powerful domain-specific regularization.

05 Always Warm Up Before Fine-Tuning

Unfreezing immediately destroys ImageNet weights. Freeze first, train the head, then unfreeze at a tiny learning rate.

06 CGANs Win on Rare Classes

CGANs generate real pixel-level images — DF went from 239 to 9,012 samples, enabling full CNN retraining on rare classes.

THANK YOU

Skin Lesion Segmentation & Classification — Data Science Practicum I

EfficientNetB2 · BCDU-Net Segmentation · SMOTE · CGAN Balancing

Grad-CAM Interpretability · Macro F1 Evaluation · 53:1 Imbalance Solved

github.com/arjun7973/Data-Science-Practicum-1