XAI for Skin-Cancer Classification

via B-cosification

TOMMASO FEDERICI 2214368
LUCIA FORNETTI 2214370

Computer Vision 2025-2026

Table of Contents

RESEARCH CONDUCTED

OUR PROJECT

Our starting points

1

B-cos Networks: Alignment is All We Need for Interpretability

Moritz Böhle Mario Fritz Bernt Schiele

2

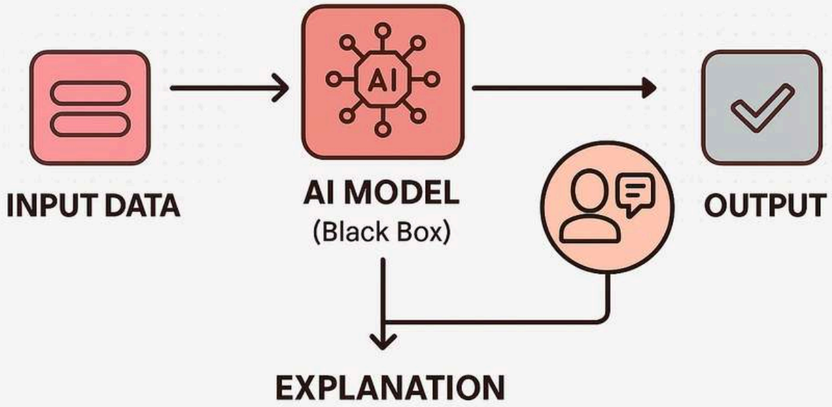
B-cosification: Transforming Deep Neural Networks to be Inherently Interpretable

Shreyash Arya[*,1,2], **Sukrut Rao**[*,1,2], **Moritz Böhle**[*,†,1,3], **Bernt Schiele**[1]

eXplanble AI

WHAT IS XAI?

Processes and methods for understanding and trusting the results generated by machine learning algorithms.



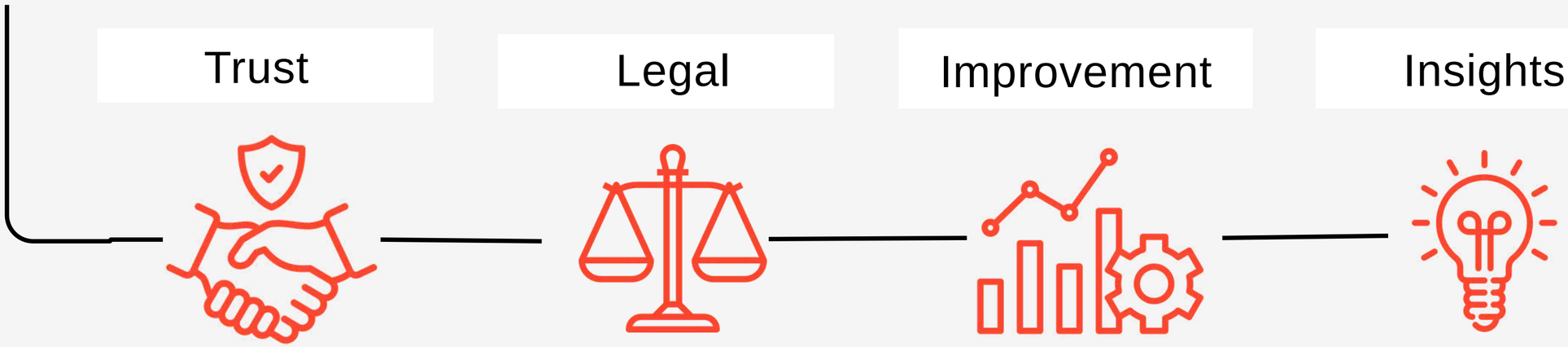
WHY WE NEED XAI?

Trust

Legal

Improvement

Insights



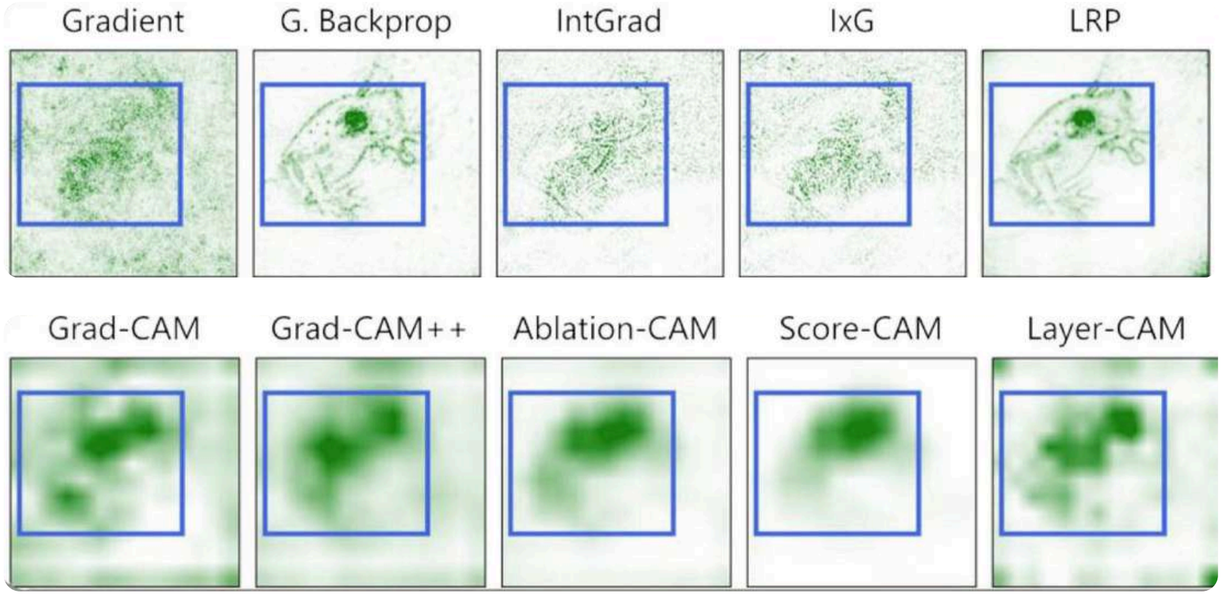
eXplanble AI

THE CRISIS OF POST-HOC INTERPRATATION

- Post-hoc methods explain a black box

10 different algorithms
=
10 different outputs

Which one to trust?



from: https://neurips.cc/media/neurips-2024/Slides/95051.pdf

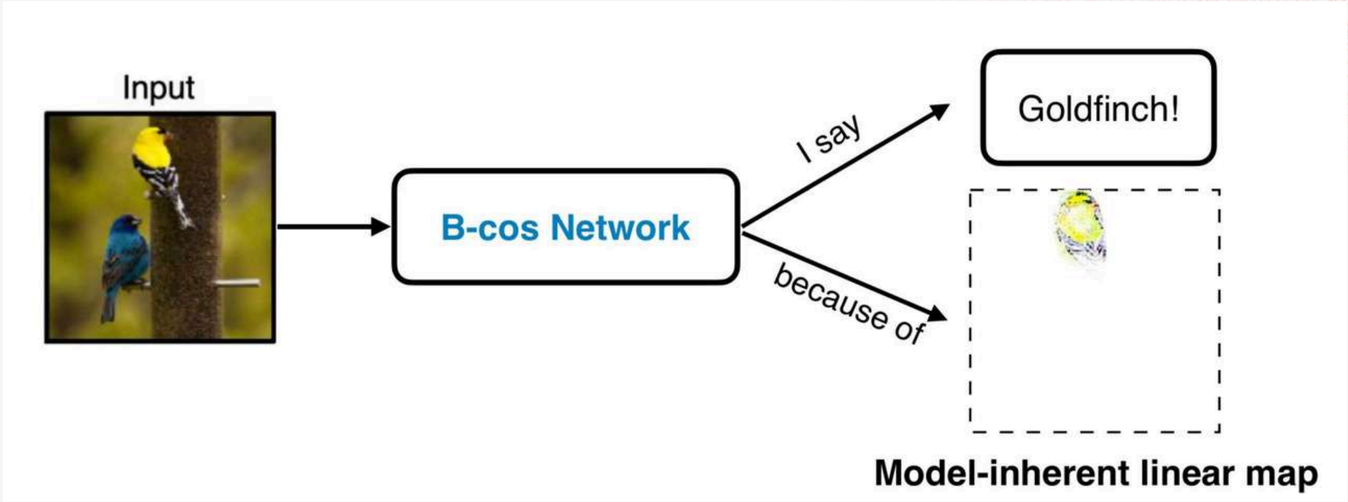
eXplanble AI

INHERENT INTERPRETABILITY: B-COS NEURAL NETWORK

- We want a faithful and interpretable explanation

BLACK-BOX ⇨ **WHITE-BOX**

- The model is intrinsically interpretable



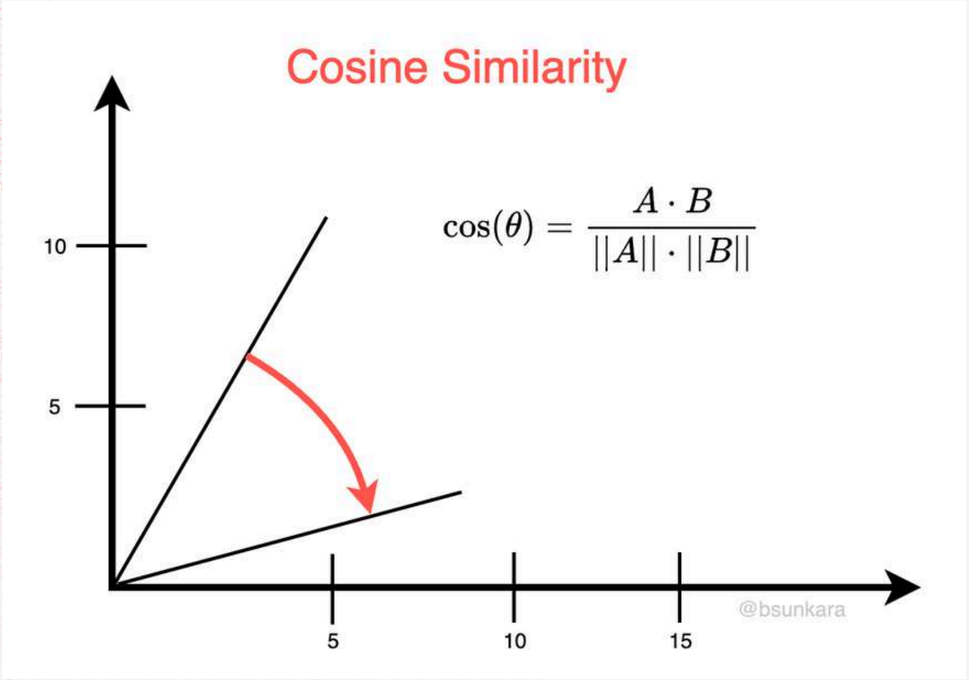
from: https://neurips.cc/media/neurips-2024/Slides/95051.pdf

B-cos Neural Network

LINEAR TRANSFORMATION

$$f(x, w) = w^T x = ||w|| \, ||x|| cos(x, w)$$

Alignemnt pressure: Hyperparameter B forces the model to prioritize directional alignment between weights and inputs.



B-COS TRANSFORMATION

$$f(x, w) = w^T x = ||\hat{w}|| \, ||x|| |cos(x, w)|^{B} \times sign(cos(x, w))$$

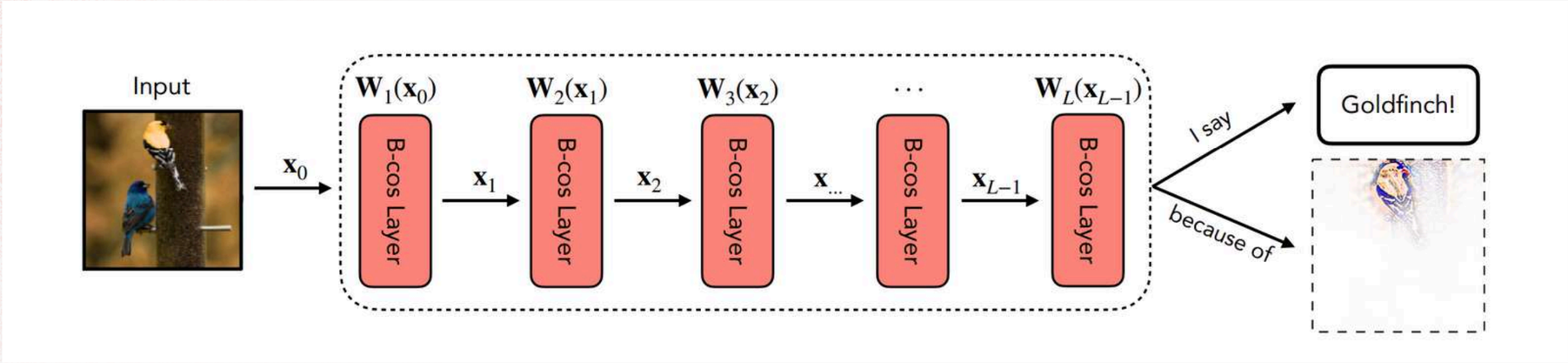
B-cos Neural Network

WEIGHTS AS PROTOTYPES

Weights act as visual prototypes rather than abstract directions due to the alignment pressure.

MAXOUT FOR EXPRESSIVITY

To gain performance, B-cos NN can use MaxOut as non-linearity (ReLU is a specific case).



from: https://neurips.cc/media/neurips-2024/Slides/95051.pdf

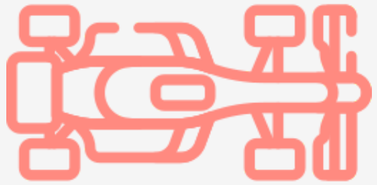
B-cosification

B-cos TRAINING FROM SCRATCH	**B-cosified** CONVERSION & FINE-TUNING
The network is designed and trained with B-cos layers from the start.	Converts pre-trained models by replacing linear layers with B-cos layers.
It offers maximum fidelity and interpretability.	It allows existing powerful models to become interpretable with minimal accuracy loss.
MaxOut as non-linearity.	Maintains ReLU to be more aligned with the original model.
Training from scratch can be costly.	Significantly fewer training steps than full retraining.

Recap & Our Case of Study

WHITE-BOX

Inherent Interpretability



B-COSIFICATION

Fewer training steps than native B-cos NN

XAI FOR MEDICAL TASKS

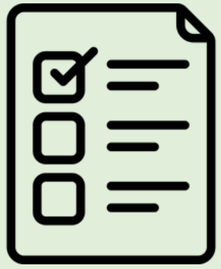
Our goal: apply this framework to the medical field

XAI for Skin-Cancer Classification

OUR PROJECT

Task & Motivation

THE CLINICAL NEED FOR XAI



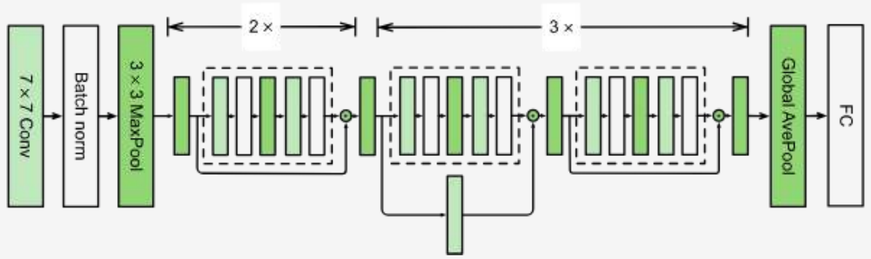
- Implementation of an **intrinsically interpretable** model for the **classification of skin lesions** from dermoscopic images.

- **Supporting** dermatologists in the early identification of malignancies to improve patient outcomes.
- Replacing "black-box" outputs with transparent logic that clinicians can **easily interpret**.

Models, Tools & Dataset

MODEL BACKBONE

- ResNet50 (pretrained on ImageNet)



TOOLS & FRAMEWORKS

- XAI Library: Official B-cosification repository (GitHub).
- Environment: Kaggle for training.

kaggle

DATASET: HAM10000

- Composition: 10,015 dermatoscopic images of pigmented lesions, divided in 7 classes
- Technical Challenge: Significant class imbalance (e.g., melanoma vs. nevi)

Training Strategy

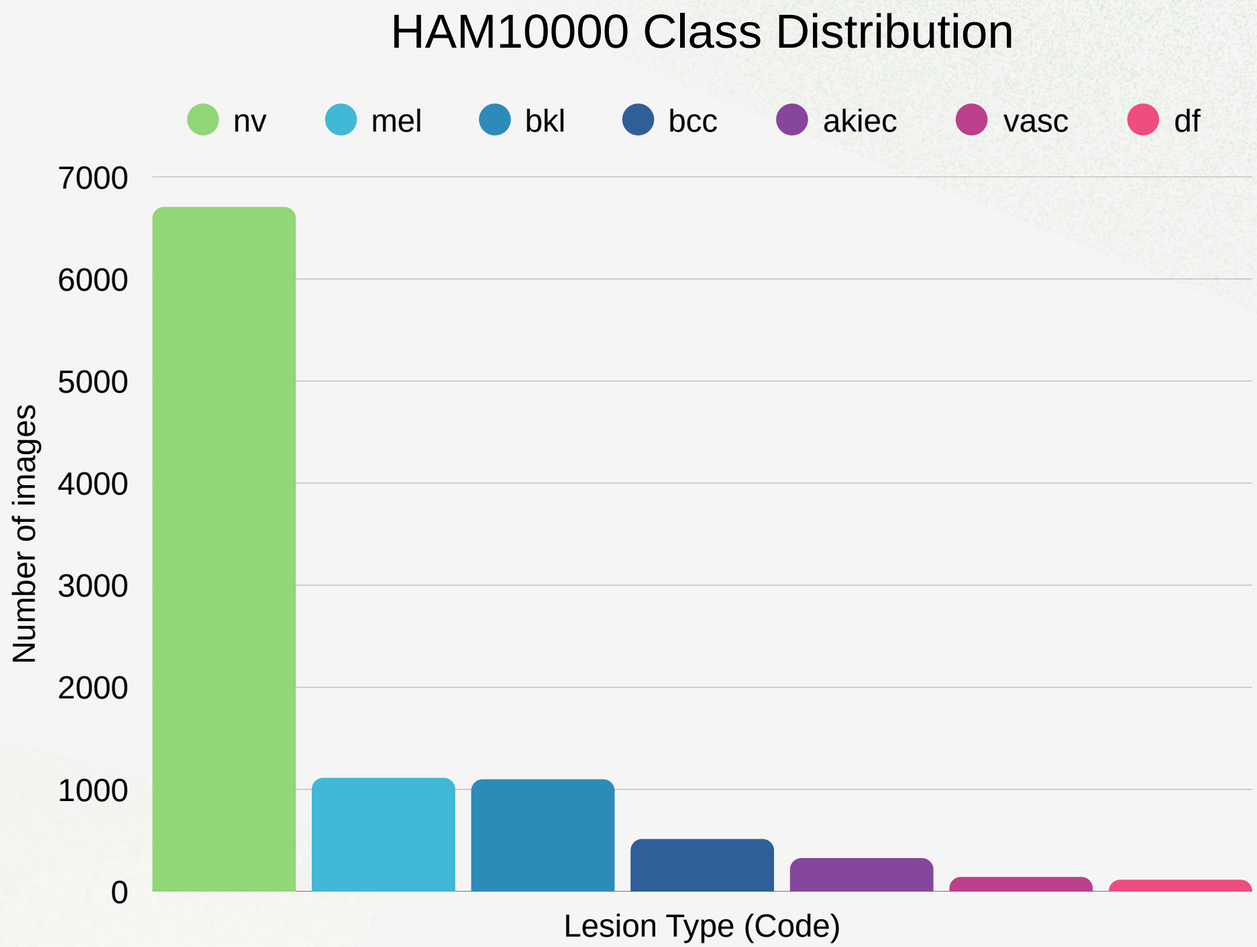
MENAGE DATA IMBALANCE

WEIGHTEDRANDOMSAMPLER

to mitigate the significant **imbalance** in the dataset

DATA AUGMENTATION

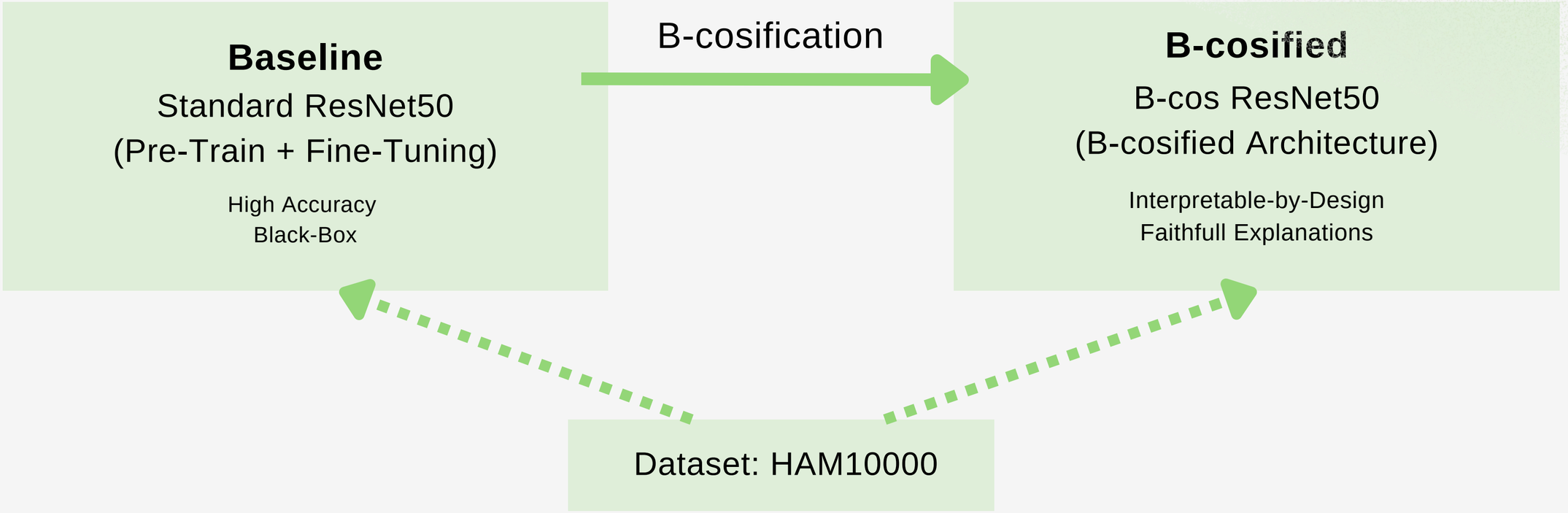
to manage imbalance and improve **generalization**



Training Strategy

TRAINING APPROACH

- Baseline training & B-cosification



Training Strategy

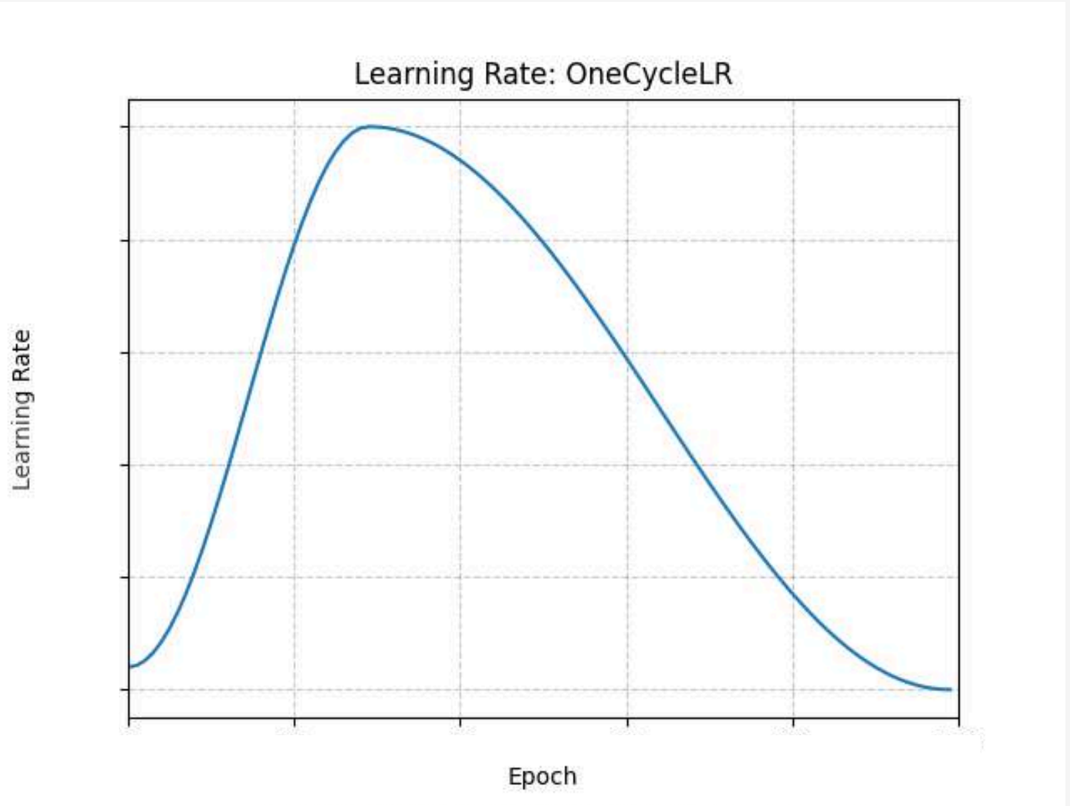
OPTIMIZATION

ONECYCLELR

To overcome local minima and converge towards the optimum more quickly

FULL FINE-TUNING

Adopted to bridge the significant domain shift from natural to dermoscopic images and to realign the model's weights after b-cosification



Evaluation

SELECTED METRICS

- Metrics chosen based on the official ISIC 2018 challenge - Task 3

INDIVIDUAL PERFORMANCE

- Accuracy
- Sensitivity (Recall)
- Specificity
- F1-score (Macro/Weighted)
- MAP

PROBABILISTIC ANALYSIS

- AUC (Area Under ROC)
- AUC80 (specific for melanoma diagnosis, integrated between the 80-100% sensitivity)

AGGREGATE

- Mean AUC across all classes
- AUC Malignant vs Benign

PREDICTIVE POWER

- PPV (Precision)
- NPV

Evaluation

COMPARISON BETWEEN THE TWO MODELS

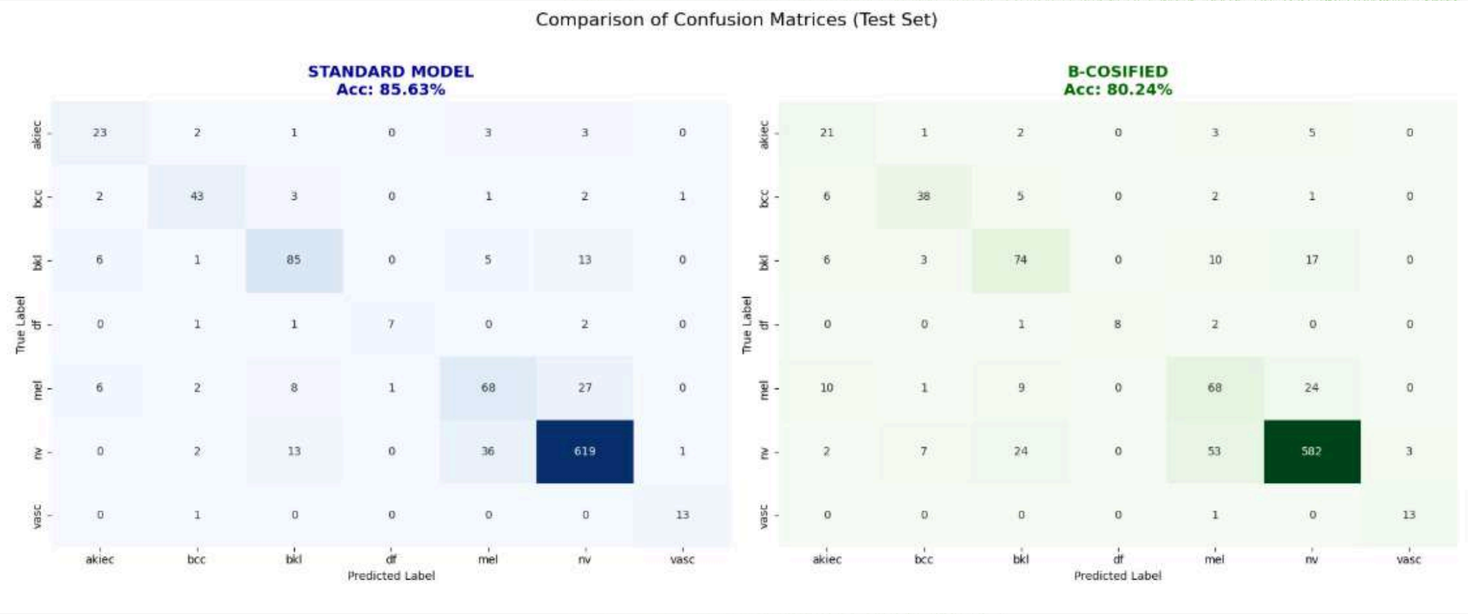
- The **trade-off** between Performance & Interpretability.
- B-cosified model shows a systematic **decrease** in performance across all metrics.
- But it is still **robust**.

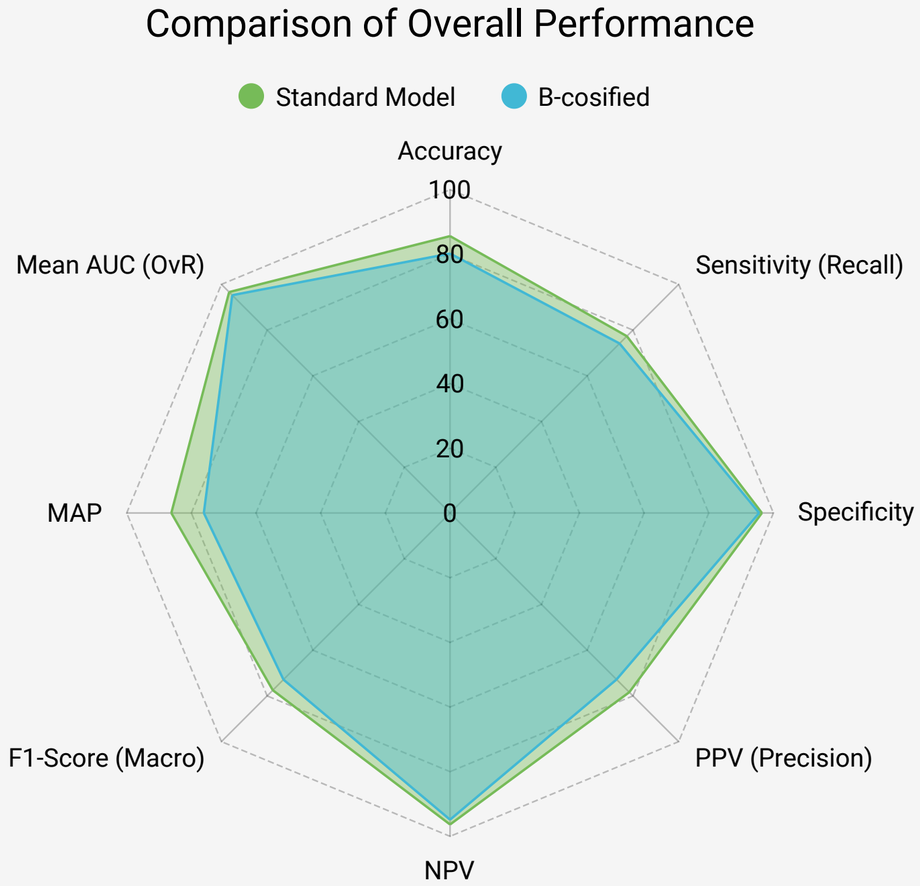
Metric Category	Metric	Standard Model	B-cos Model
Overall	**Accuracy**	85.63%	80.24%
	Macro F1-Score	77.52%	72.78%
	MAP (Mean Avg Precision)	86.20%	76.15%
Clinical	**Sensitivity (Recall)**	77.33%	74.14%
	Specificity	96.45%	95.60%
	PPV (Precision)	78.39%	72.82%
	NPV (Neg. Pred. Value)	96.34%	94.85%
Probabilistic	**Mean AUC (OvR)**	96.62%	95.26%
	Binary AUC (Mal vs Ben)	0.9427	0.9215
	pAUC (Sensitivity ≥ 80%)	0.8926	0.8331

Evaluation

COMPARISON BETWEEN THE TWO MODELS

- Confusion matrix and radar visualization

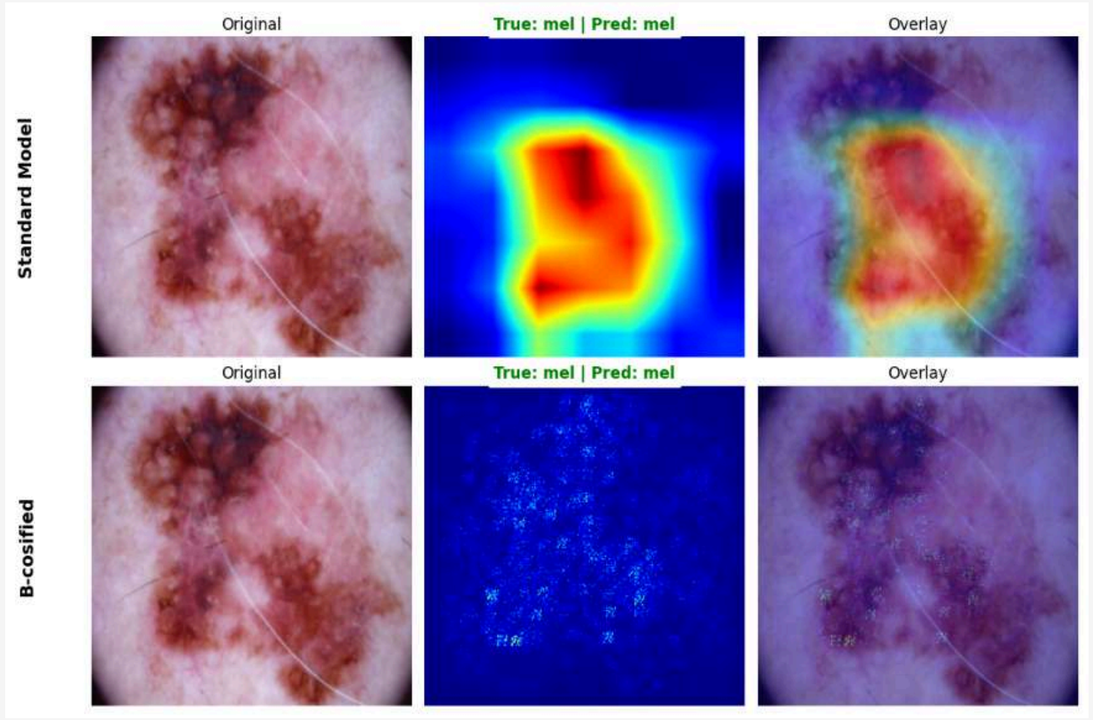


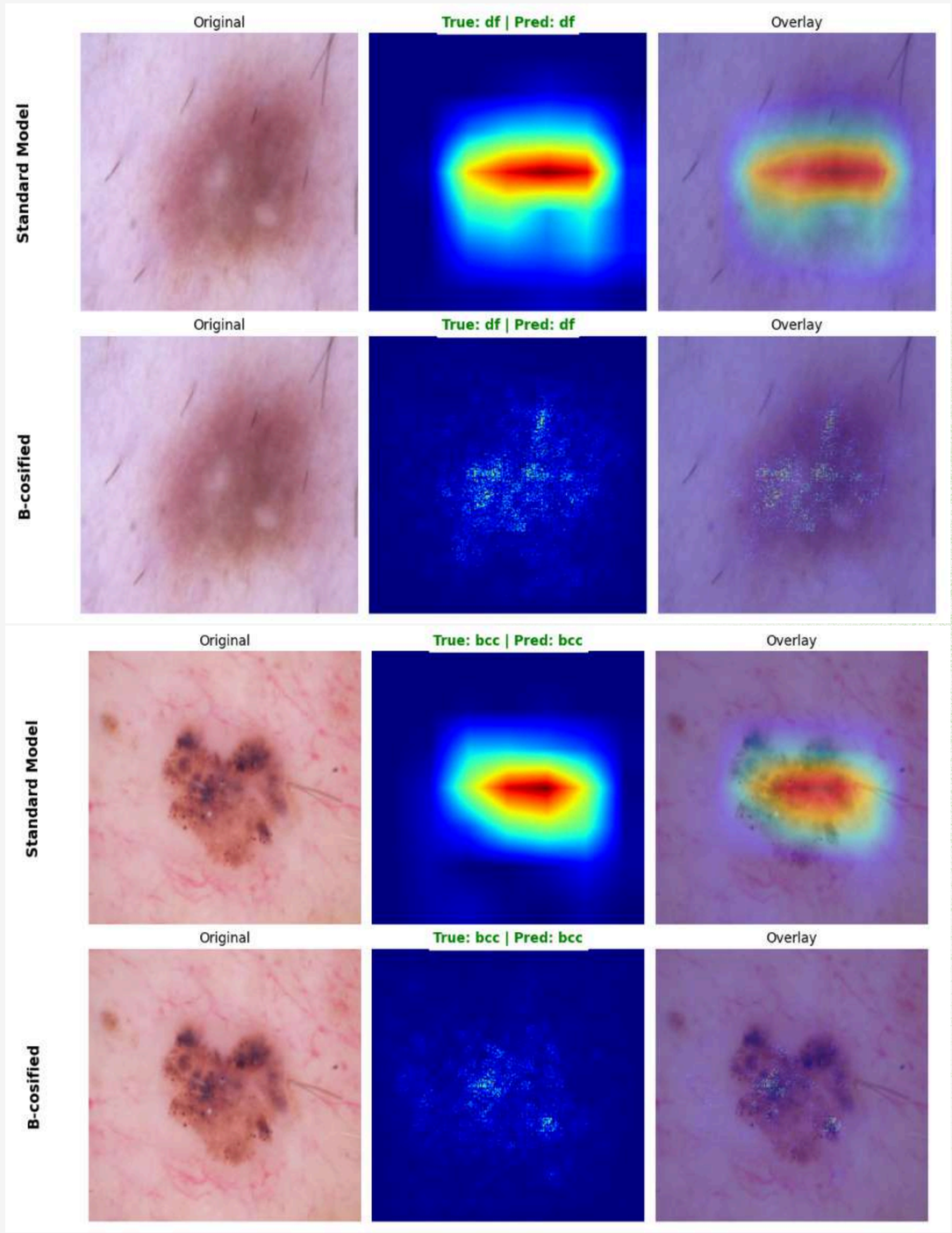


Evaluation

COMPARISON OF THE EXPLANATIONS

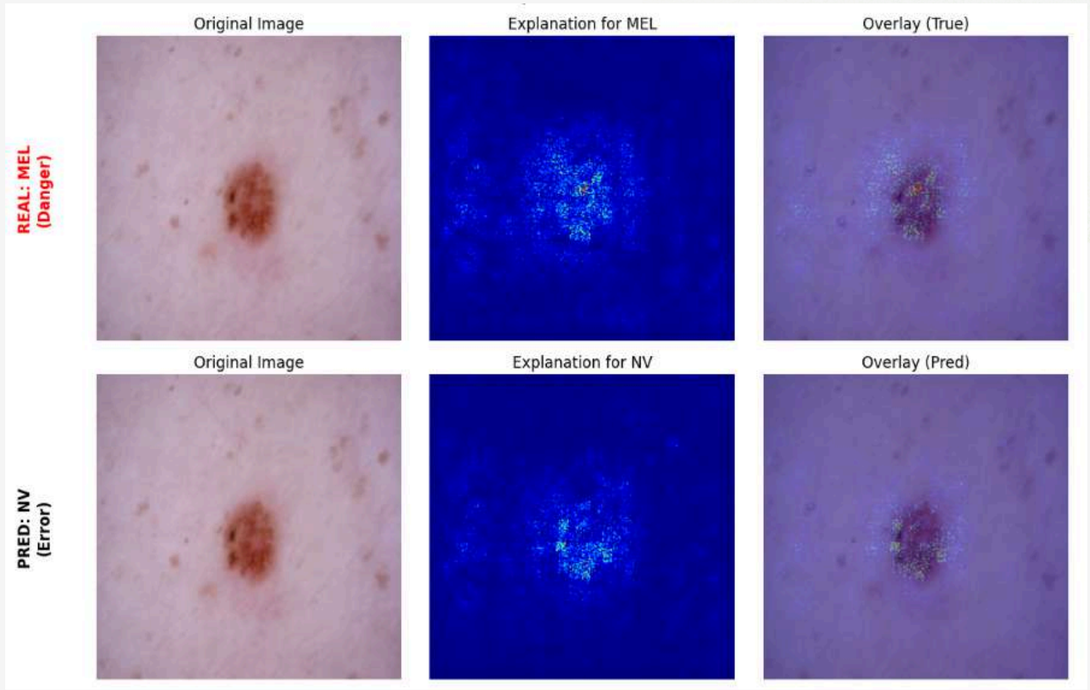
- WHERE (GradCAM) vs WHAT (B-cos)

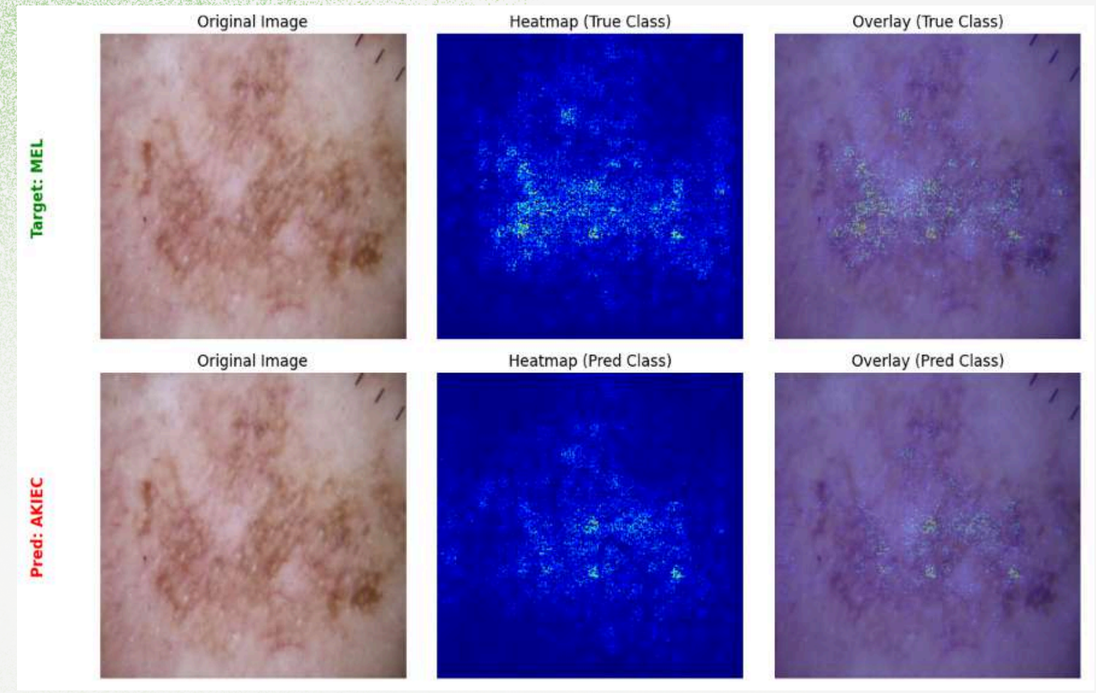


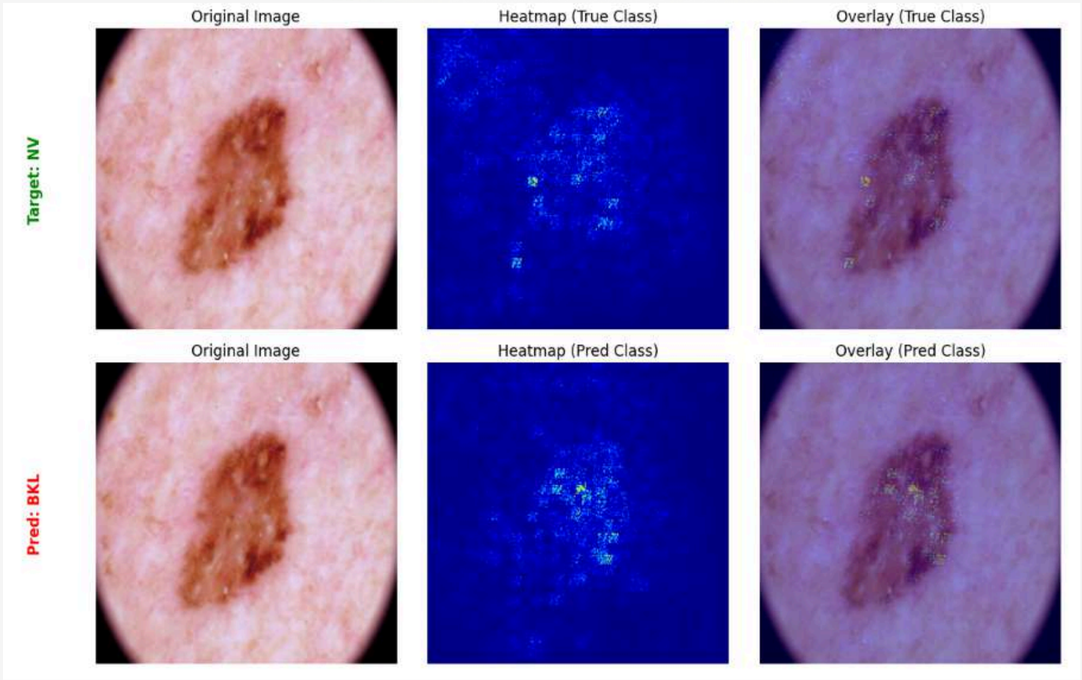


Evaluation

WHEN B-COS FAILS







Thank You For Your Attention

QUESTIONS?