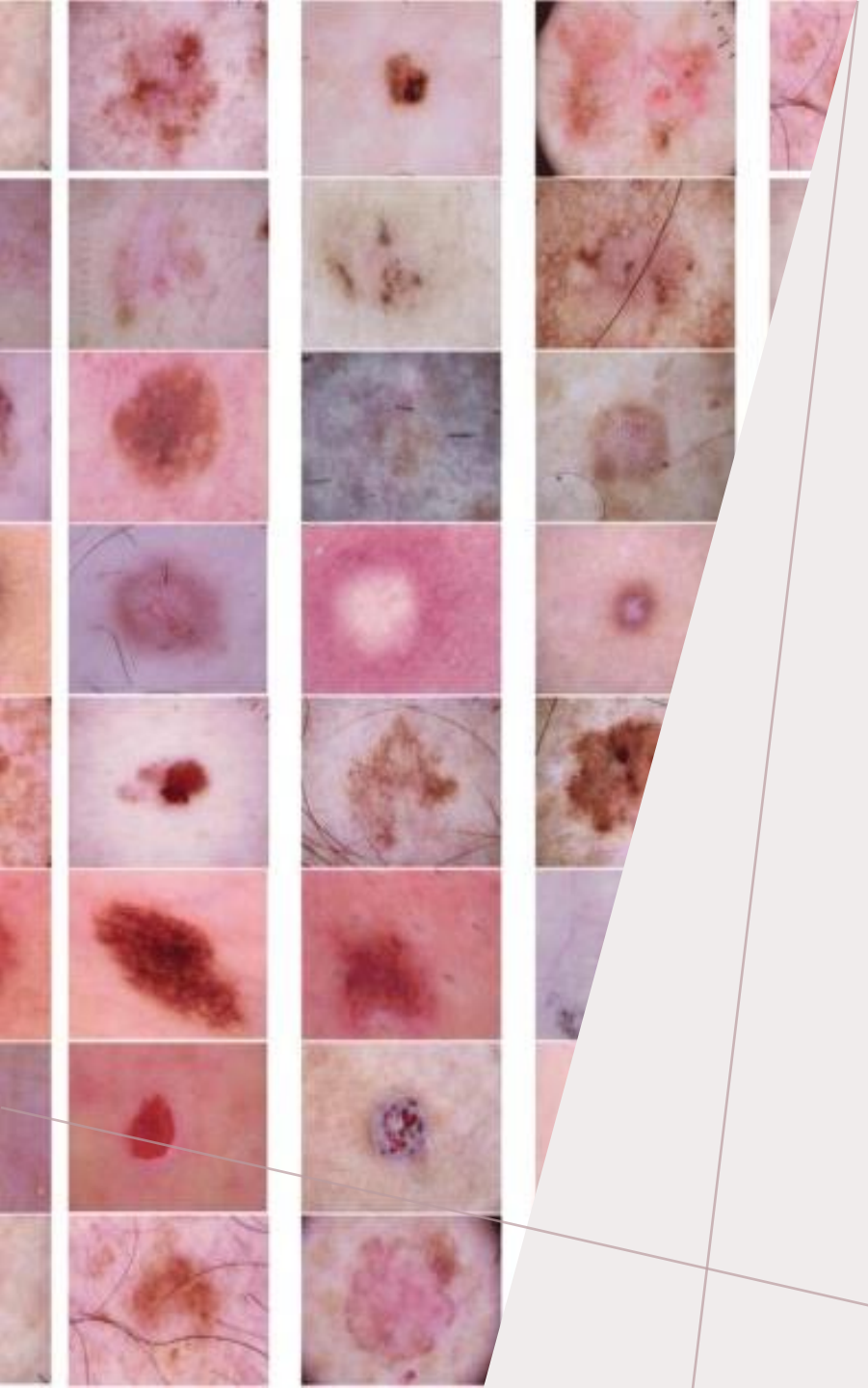


CAD PROJECT

BINARY AND MULTI-CLASS CLASSIFICATION OF SKIN LESION USING MACHINE LEARNING

Prepared By:

Mahdi Islam

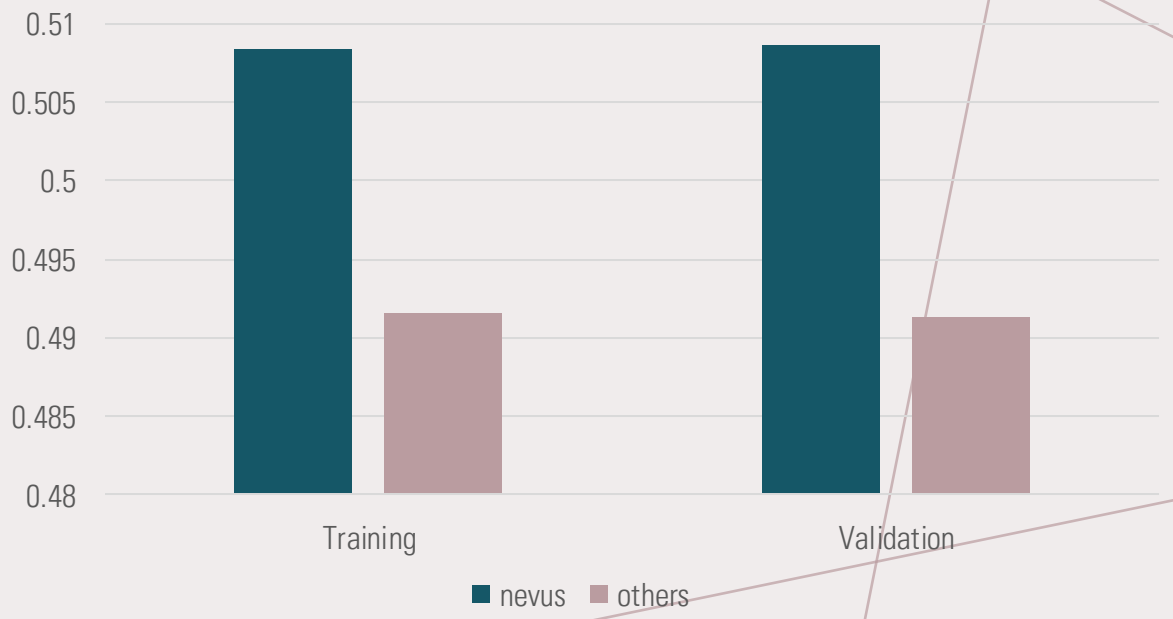
Musarrat Tabassum

Skin Lesion Dataset

Lesion	Images	
	Train	Validation
Nevus	7725	1931
Others	7470	1865
Total	15195	3796

Balanced Dataset

Binary Dataset

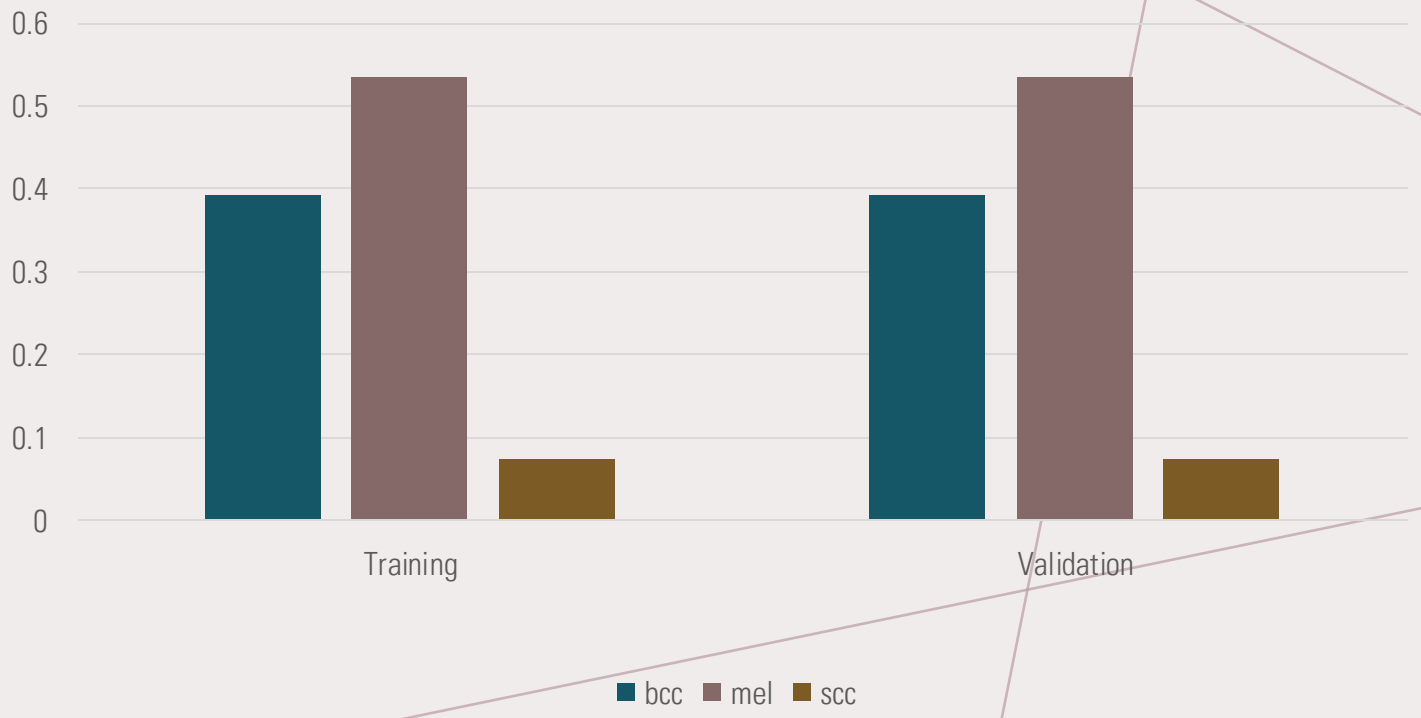


Skin Lesion Dataset

Multi-class Dataset

Lesion	Images	
	Train	Validation
BCC	1993	498
Melanoma	2713	678
SCC	376	94
Total	5082	1270

Imbalanced Dataset



Flowchart

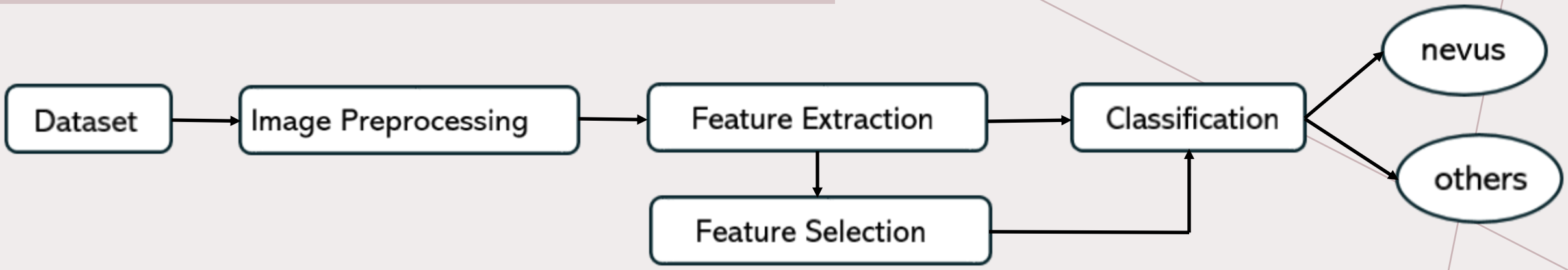


Fig: Binary Classification

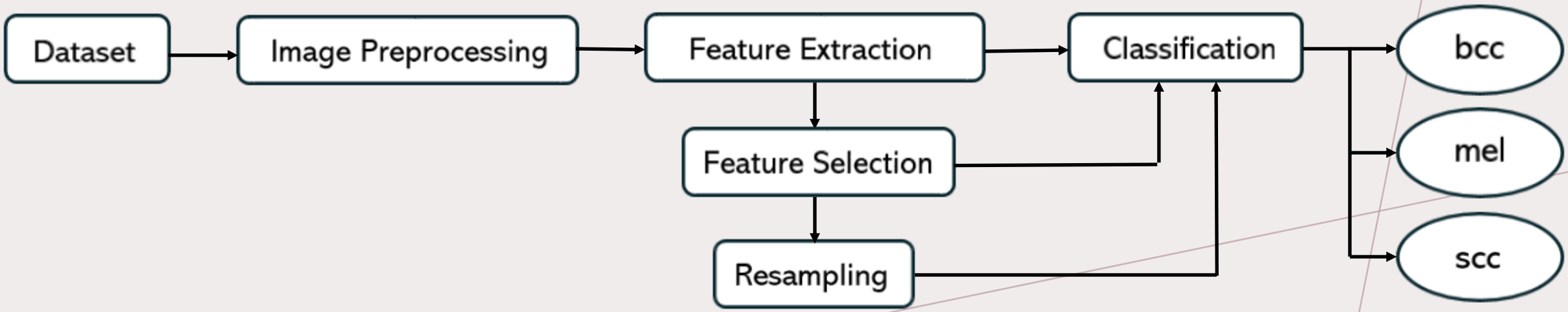
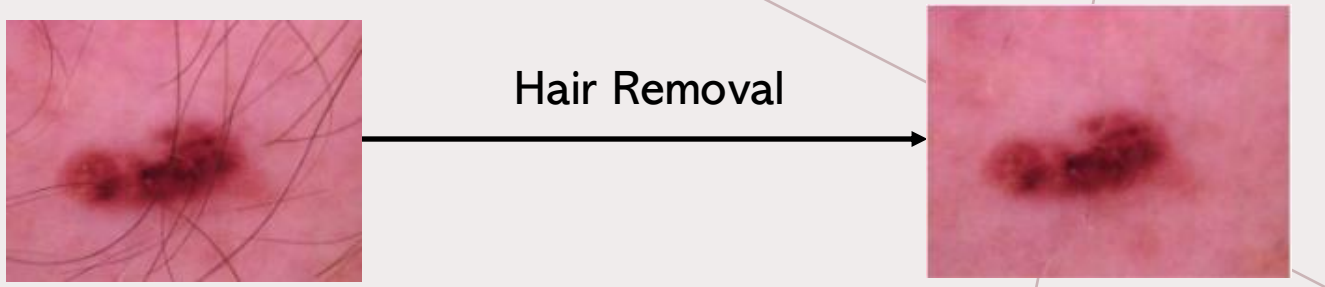


Fig: Multi-class Classification

Image Preprocessing

- **Hair Removal**
  - Blackhat Filter using multiple oriented structural element + Threshold



- **Color Normalization**
  - Achieve consistent color balance across images to minimize lighting and color variation effects.

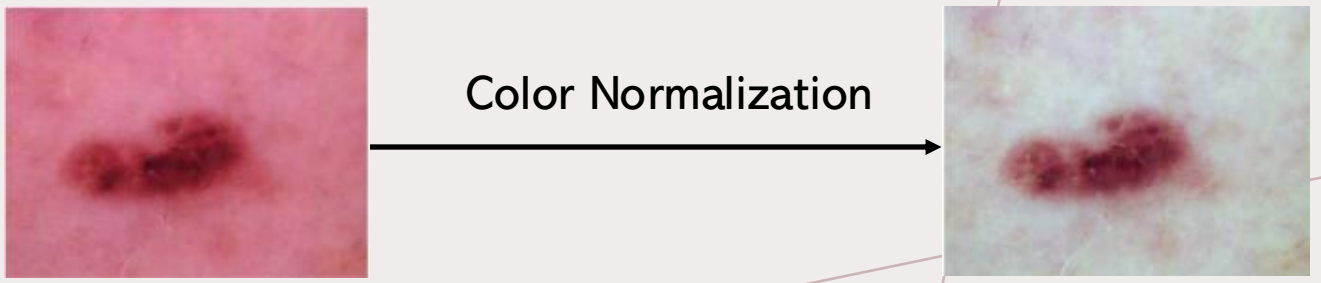
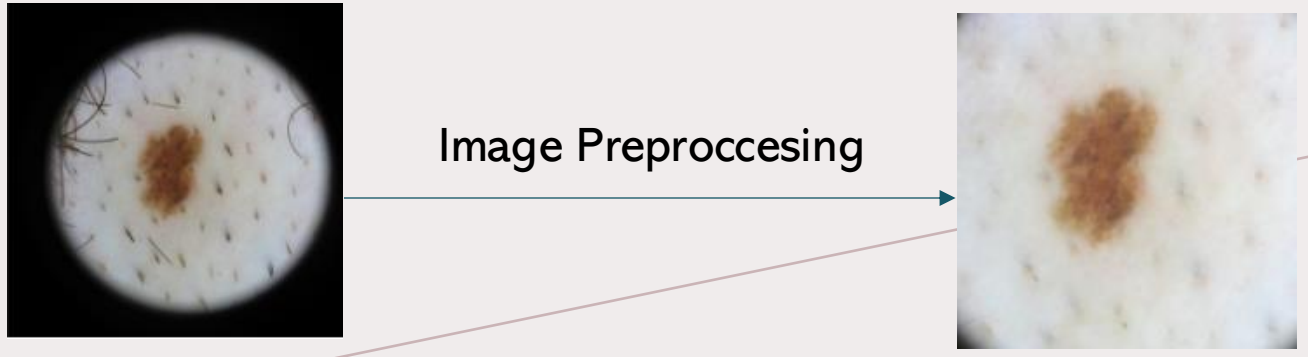


Image Preprocessing

- **Region of Interest (ROI)**
  - search for the first and the last position in the diagonal of the image where the mean value of the pixels is higher than the threshold.



Feature Extraction

- **Color Features**
  - Color Stats
  - Color Histogram
  - Color Moments
  - Entropy
- **Texture Features**
  - Local Binary Pattern (LBP)
  - Gray Level Co-occurrence Matrix (GLCM)
  - Haralick
- **Gradient Features**
  - Wavelets
  - HOG

Feature Extraction

Color Features

- Color stats -(min, max, kurtosis, median, 25th percentile, 75th percentile, normalized std) for each RGB, HSV, and LAB channel.
- Color moments -(mean, skewness, variance) from each RGB, HSV and LAB channel of the image.
- Color histogram features for each RGB, HSV, and LAB color spaces.
- Entropy for each channel.
- Gray scale entropy.

Feature Extraction

Textue Features

- Local Binary Pattern (LBP) from the blue channel of the image (P=16, R=2).
- Gray Level Co-occurrence Matrix (GLCM) from the blue channel of the image ('correlation', 'homogeneity', 'contrast', 'energy', 'dissimilarity).
- Haralick features from the blue channel of the image.

Gradient Features

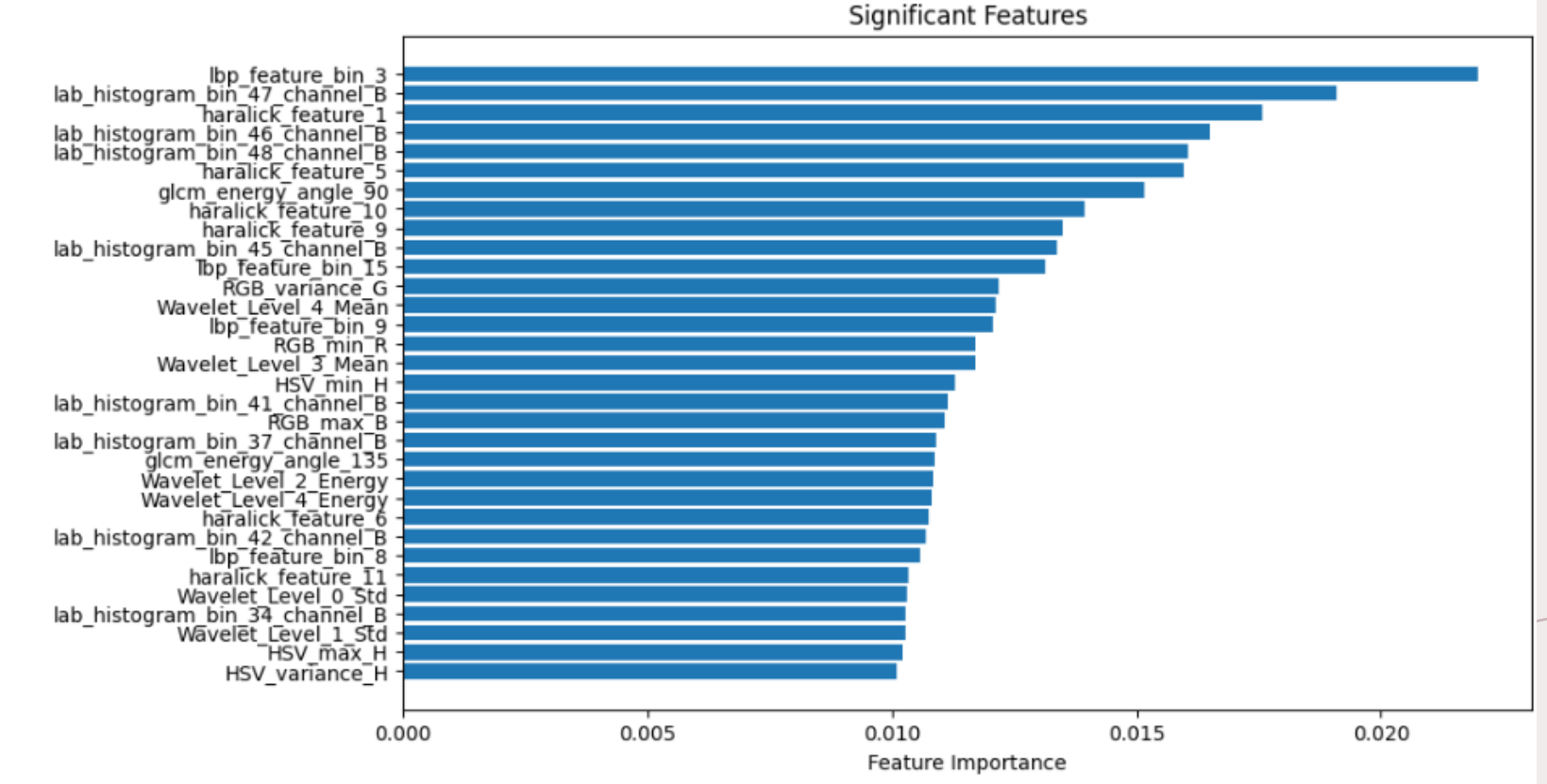
- Wavelet features from the grayscale image.
- Extracting Histogram of Gradients (HOG) and then applying PCA to reduce its dimension.

Feature Extraction

Types of Features		Number of Features
Color Features	Color Stats	63
	Color Moments	27
	Color Histogram	576
	Entropy	9
	Gray scale entropy	1
Texture Features	Local Binary Pattern	18
	GLCM	20
	Haralick	13
Gradient Features	Wavelets	15
	HOG+PCA	100
Total		842

Feature Engineering

Feature Importance extracted by training a random forest with all the features.



Machine Learning Classifier

MAIA

Single Classifiers Implemented

Random Forest | LightGBM | XGBoost | Support Vector Machine (SVM)

Ensemble of the classifiers

Stacking

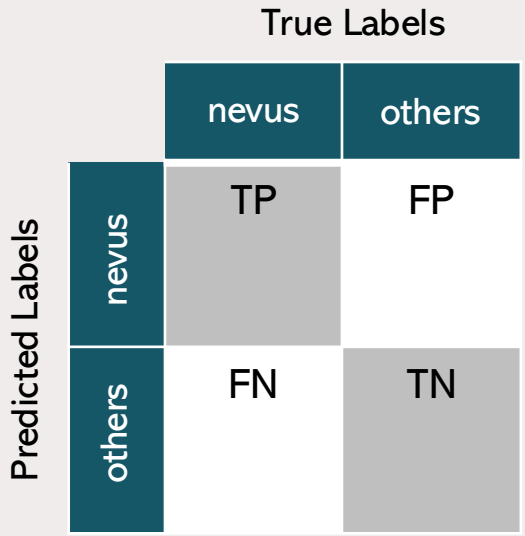
- Random Forest
- LightGBM
- XGBoost
- SVM

Majority Voting

Evaluation Metric

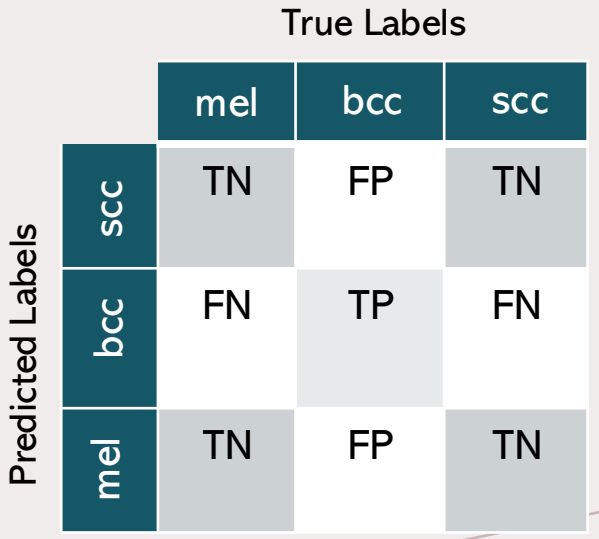
Binary Classification

- Accuracy $= \frac{TP+TN}{TP+FP+TN+FN}$

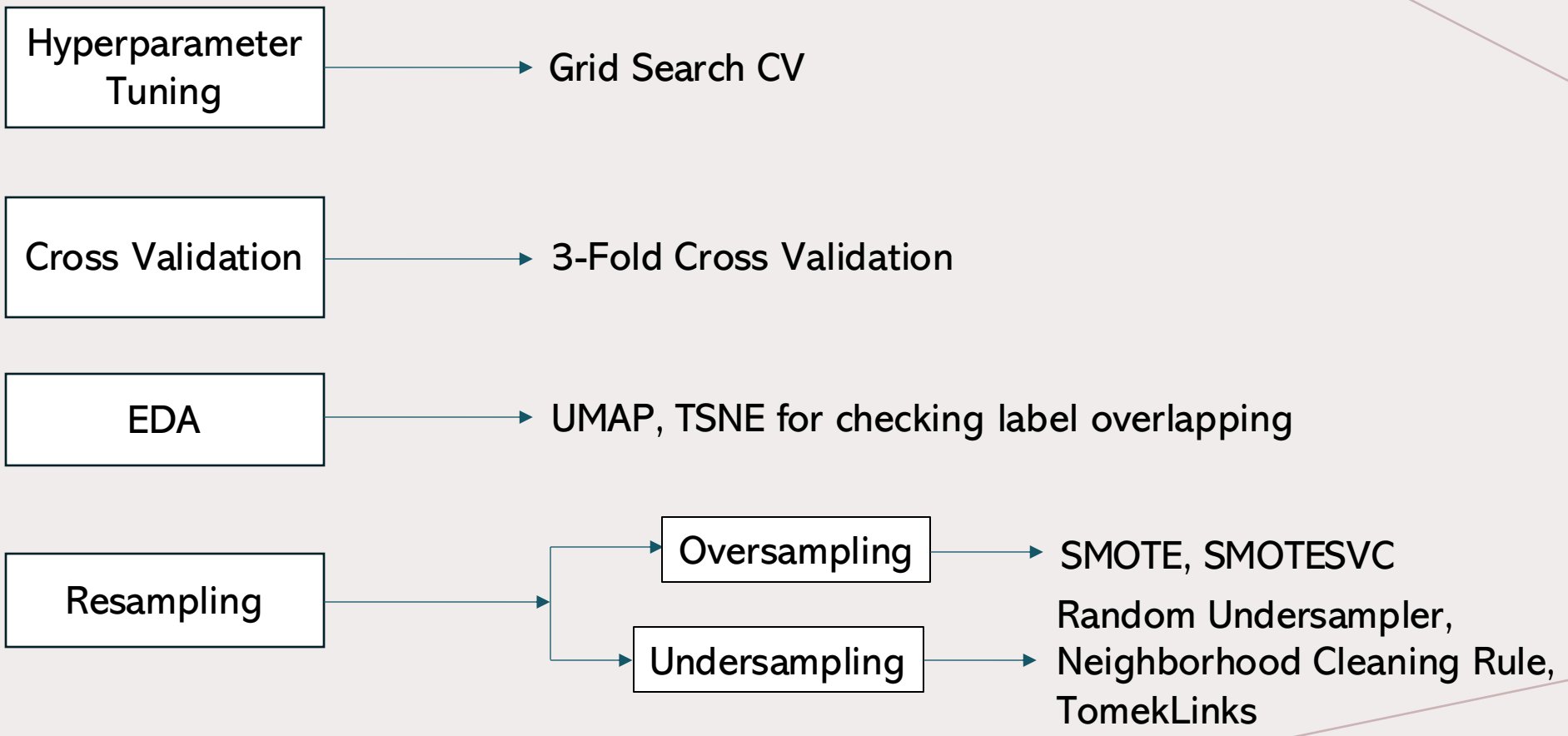


Multi-class Classification

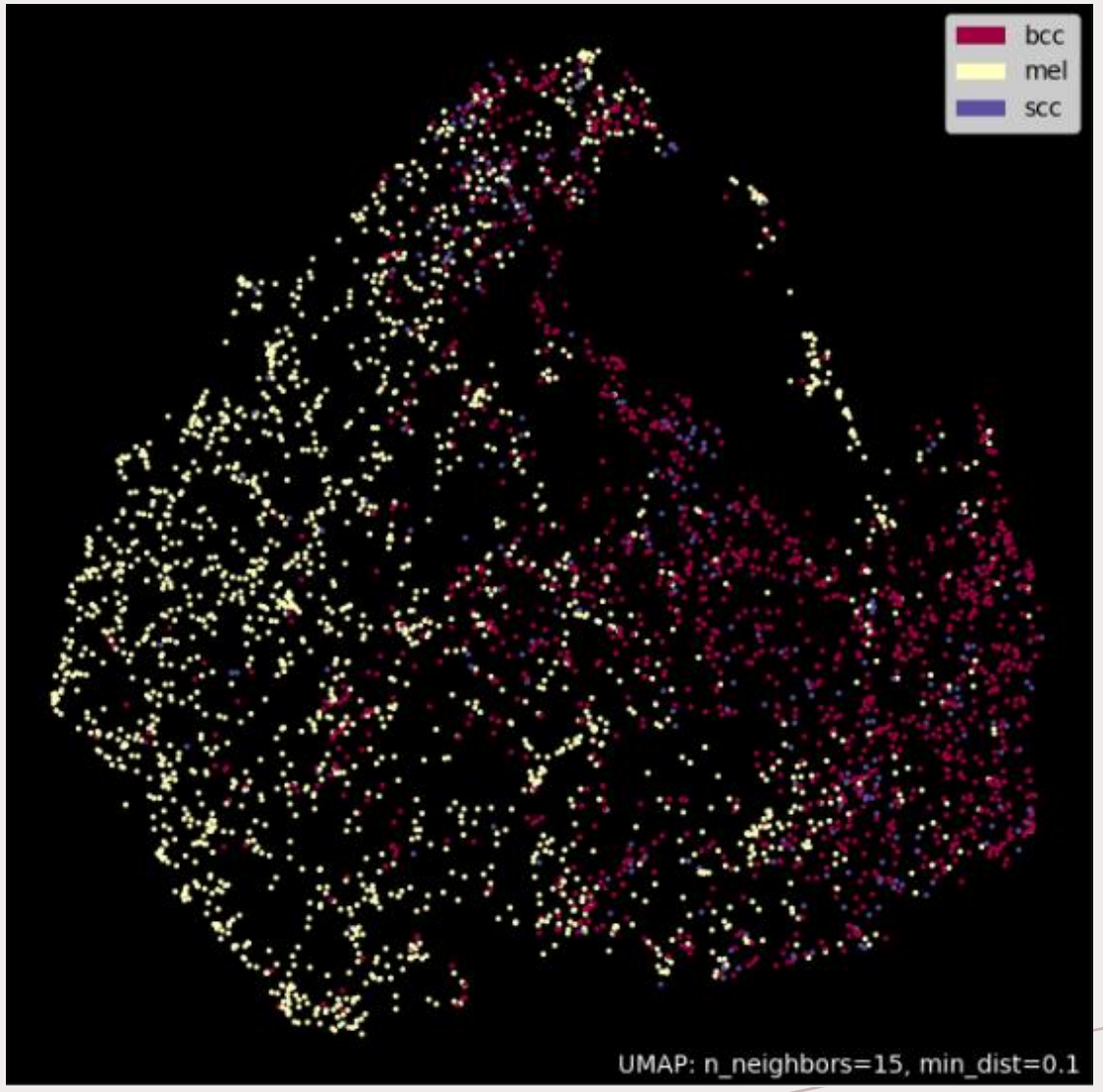
- Kappa Score $= \frac{P_o - P_e}{1 - P_e}$
- Balanced Accuracy $= \frac{1}{3} * \frac{TP}{TP+FN}$



Parameters and Project Design



Data Imbalance Issue



scc and bcc labels are overlapped!!!

Result Analysis: Binary Challenge

	Accuracy		F1 Score	
Model	With Feature Selection	Without Feature Selection	With Feature Selection	Without Feature Selection
Random Forest	.7984	.8184	.795	.815
LightGBM	.80	.8263	.80	825
XGBoost	.8160	.8244	.8165	.825
SVM	.8236	.8276	.825	.83
Stacking	.8329	***.8363***	.8355	***.84***
Voting	.8278	0.8355	.825	.835

Result Analysis: Multiclass Challenge

	Kappa Score		Balanced Accuracy	
Model	With Feature Selection	Without Feature Selection	With Feature Selection	Without Feature Selection
Random Forest	.6558	.6598	.6191	.6121
LightGBM	.6848	.707	.6479	.6593
XGBoost	.7006	.7079	***. 6914***	.6622
SVM	.704	.6873	.6902	.6404
Stacking	***.708***	.7065	. 6829	.6842
Voting	.697	.7128	.6727	.6670

Conclusion

Key Findings

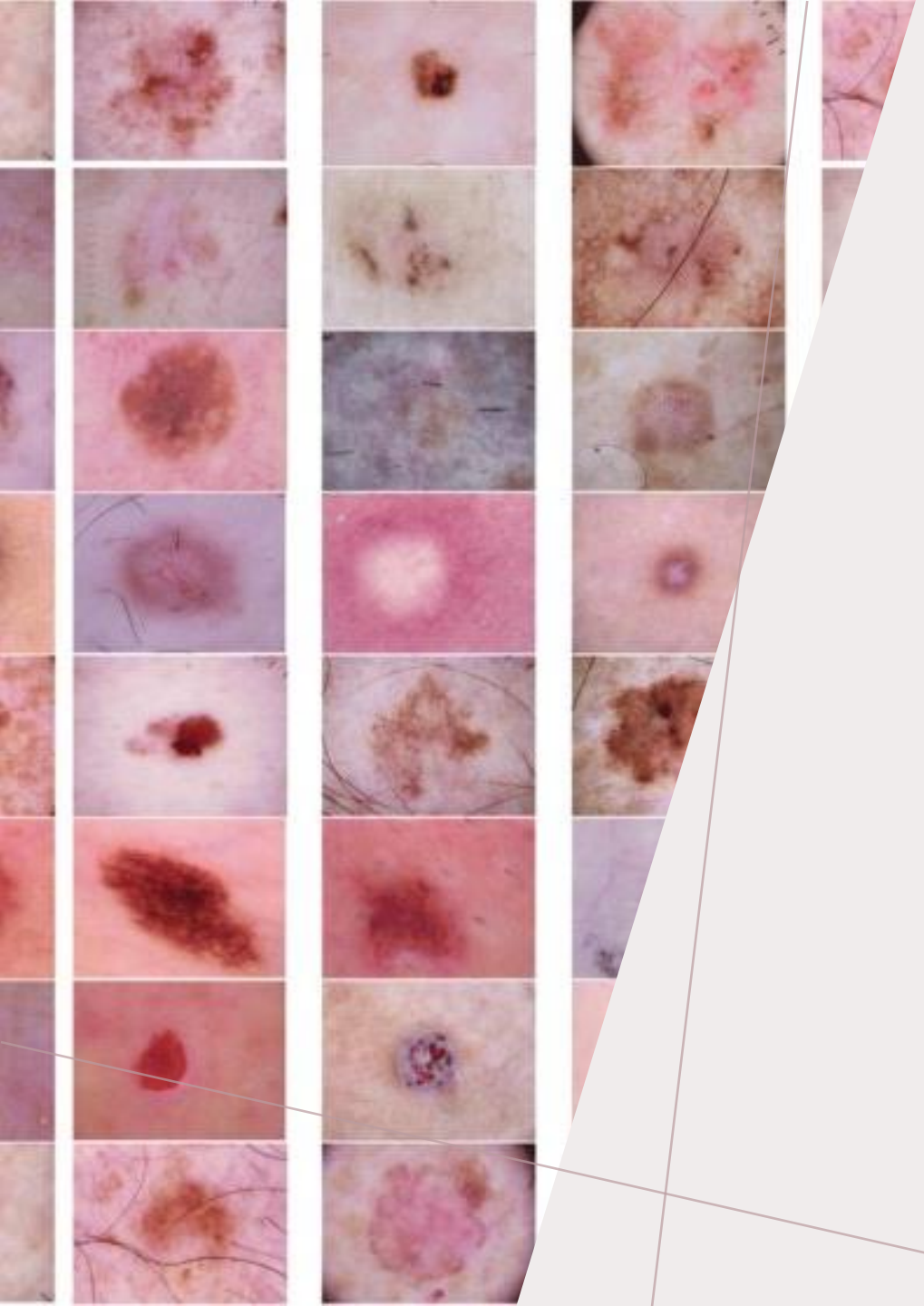
- Color and texture features are significant for this project.
- Imbalance problem could not be solved due to severe overlap between labels (scc and bcc).
- Feature selection did not drop the performance.

Future Scope

- Multi-resolution framework.
- Feature extraction after lesion segmentation using deep learning.

Reference

- newaz-aa/Sampling-algorithms-experimental-analysis: The performance of different sampling techniques are provided in this repository. The work is associated with the paper titled, "A Comprehensive Evaluation of Sampling Techniques for Addressing Class Imbalance Across Diverse Datasets".
- A. Javaid, M. Sadiq and F. Akram, "Skin Cancer Classification Using Image Processing and Machine Learning," 2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST), Islamabad, Pakistan, 2021, pp. 439-444, doi: 10.1109/IBCAST51254.2021.9393198. keywords: {Image segmentation;Thresholding (Imaging);Melanoma;Feature extraction;Skin;Random forests;Principal component analysis;Skin lesion segmentation;contrast stretching;features extraction;features reduction;features normalization;features scaling;wrapper method;SMOTE sampling;skin cancer classification;random forest classifier},
- Oversamplers — smote_variants 0.5.1 documentation
- UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction — umap 0.5 documentation
- imbalanced-learn documentation — Version 0.12.4



Erasmus Mundus

MAIA

THANK YOU

Universitat de Girona