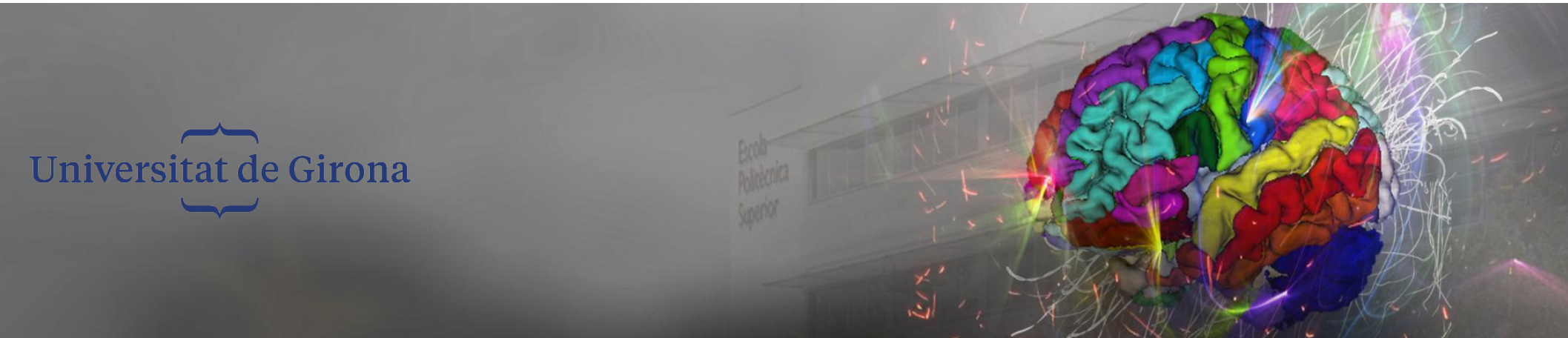
MAIA

CAD Project: Skin Lesion Classification

Kazi Saeed Alam, Gonzalo Esteban Mosquera Rojas



Objectives

Skin Lesion Classification : Binary and Three-Class Problem

Questions to be solved. What is the best pipeline for classification?

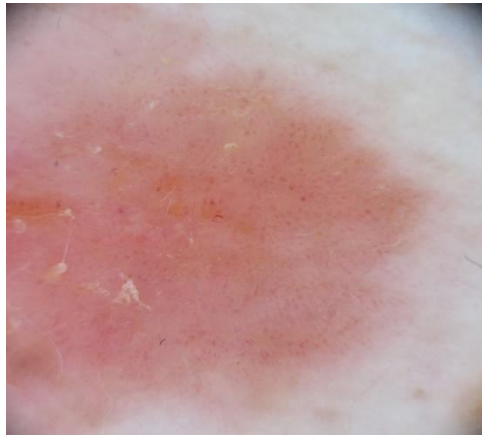
- With or without pre-processing?
- Hair Removal and/or vignette removal?
- Data Augmentation or Over Sampling/SMOTE? (for imbalanced class Problem)
- Feature Extraction Techniques?
- Feature Selection?
- Classifier Selection?
- Ensemble Classifiers?

Overview

- Binary problem: nevus vs others
- Three class problem: mel (melanoma), bcc (basal cell carcinoma), scc (squamous cell carcinoma)
- Two main ways of tackling the problems: classification with raw images and with preprocessed images.
- 11 different classifiers trained on 7 different feature set.
- Class Imbalance treatment: Data augmentation and oversampling.

melanoma

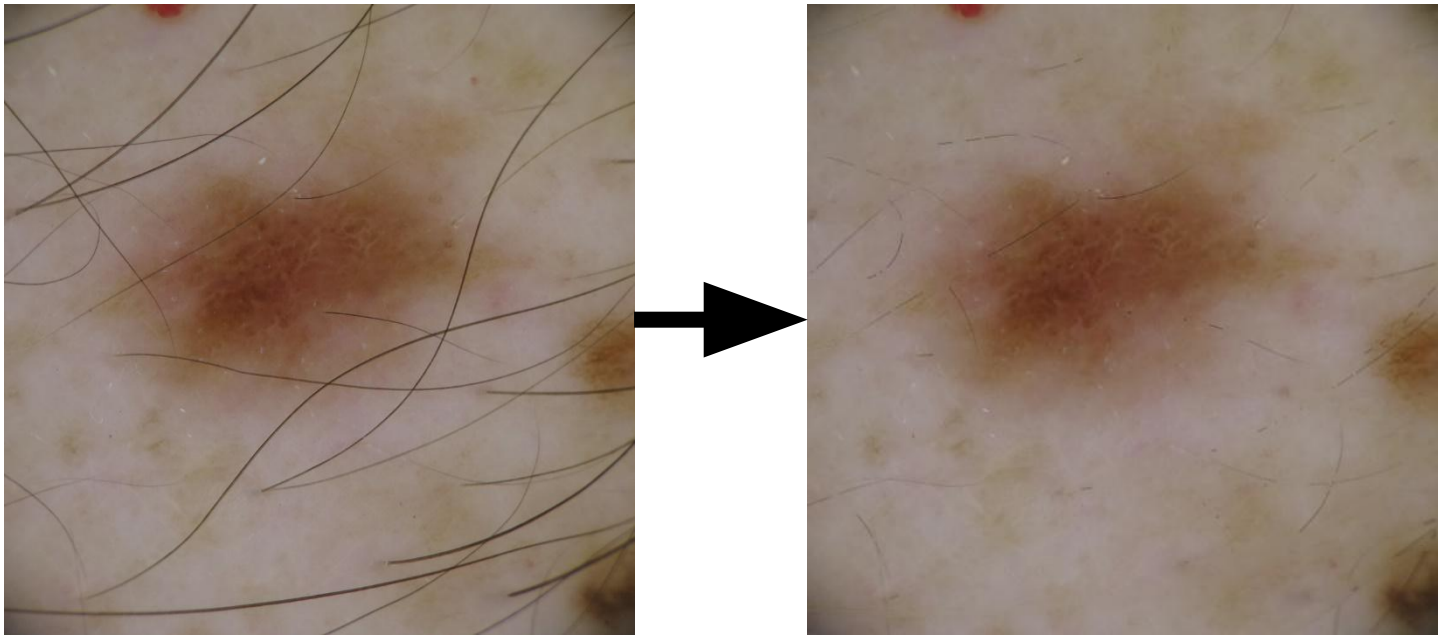
bcc



scc

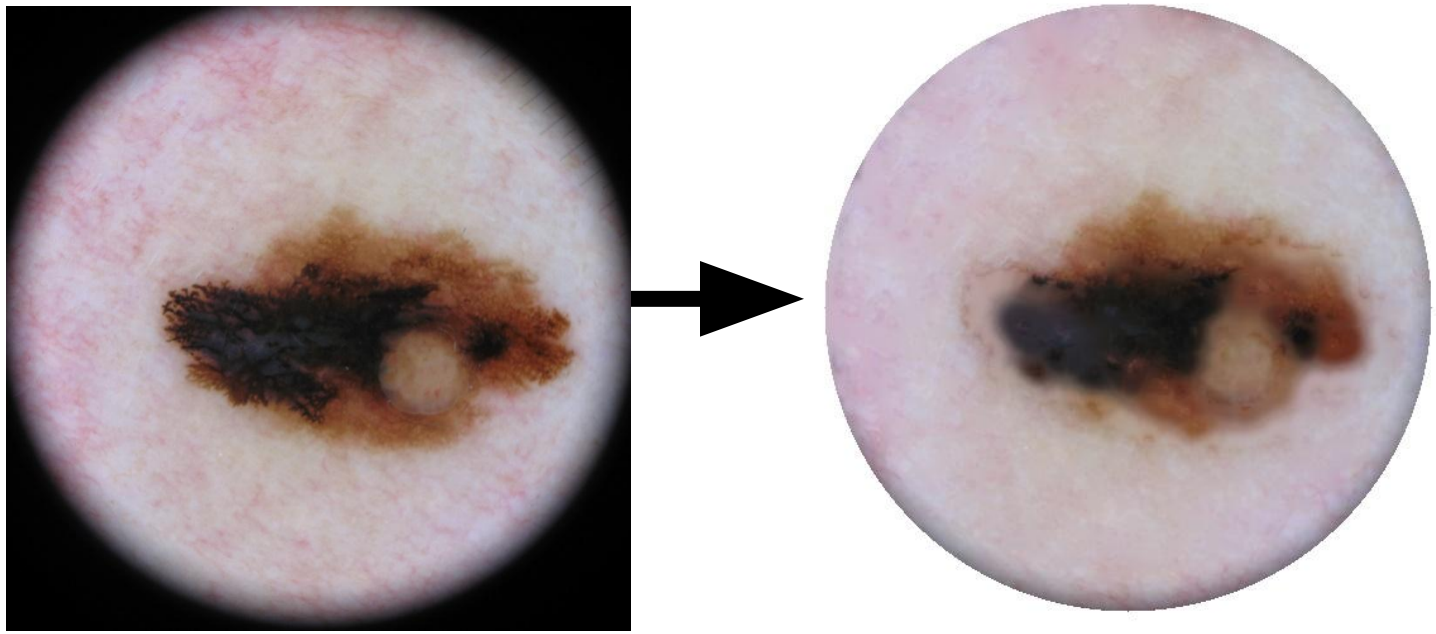
Preprocessing (Hair Removal)

- Sum of rotating Top-Hats Transforms
- Morphological Operation
- Image Inpainting



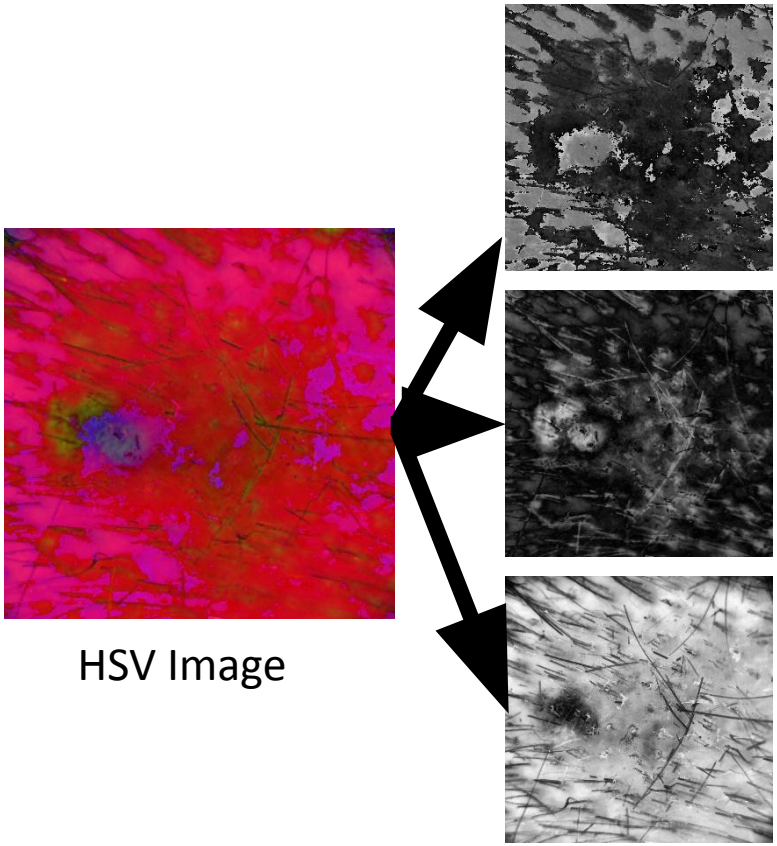
Preprocessing (Vignette Removal)

- Creating Circular Mask
- Counting the number of black pixels around the mask
- Reduce the radius according to a threshold



Preprocessing (Histogram Equalization)

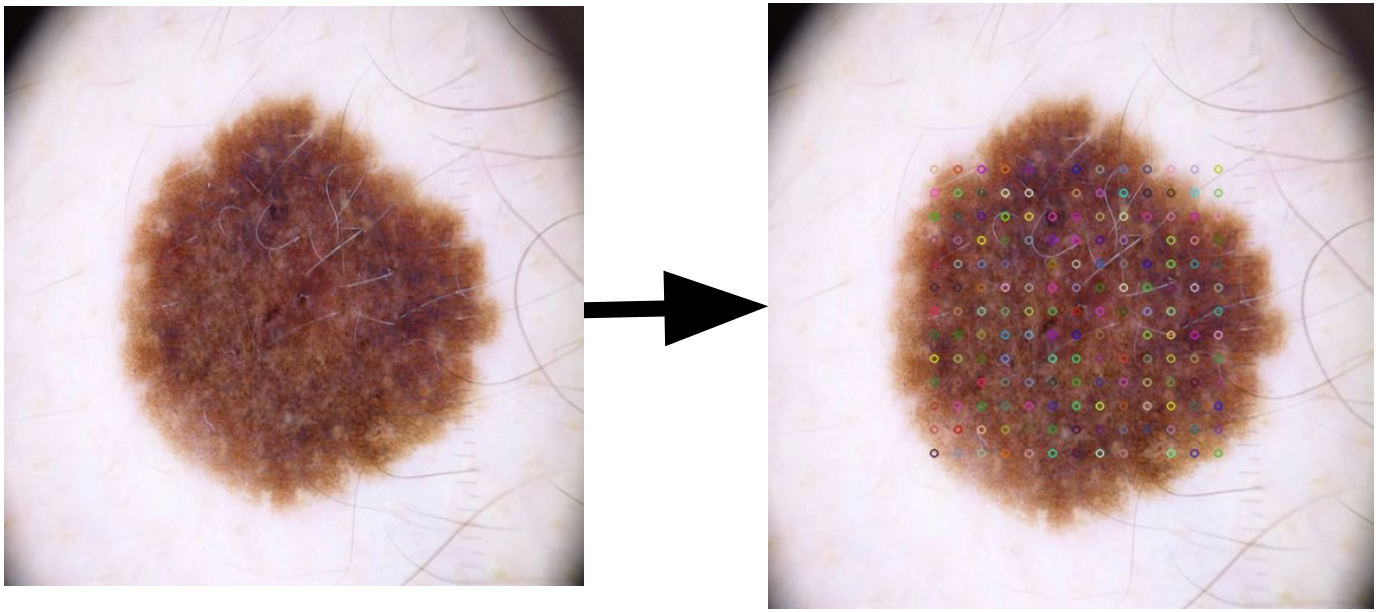
- Contrast limited adaptive histogram equalization.
- Histogram equalization on smaller tiles of images(8*8).



HSV Image

Feature Extraction: SIFT + BoW

- 200 randomly chosen images are used for vocabulary creation
- The number of clusters for BoW was kept 200
- Dense sift descriptors were used



Feature Extraction: Color Histograms

- Concatenated histogram of R, G and B channels
- Histogram bins: 64
- Number of features: 192

- For each channel, the following statistics were extracted: min, max, mean, std, skew. Channels analyzed (RGB and HSV)
- Number of features: 30
- In some experiments for the three class problem, histogram equalization was performed in the image.

Feature Extraction: GLCM Texture features

- GLCM Matrix along 3 distances (1,2,3 pixels) and 4 angles $(0, \frac{\pi}{4}, \frac{\pi}{2}, 3\frac{\pi}{2})$
- Mean and std of GLCM Matrix properties (contrast, dissimilarity, homogeneity, energy, correlation) along the chosen configuration.
- Number of features: 10

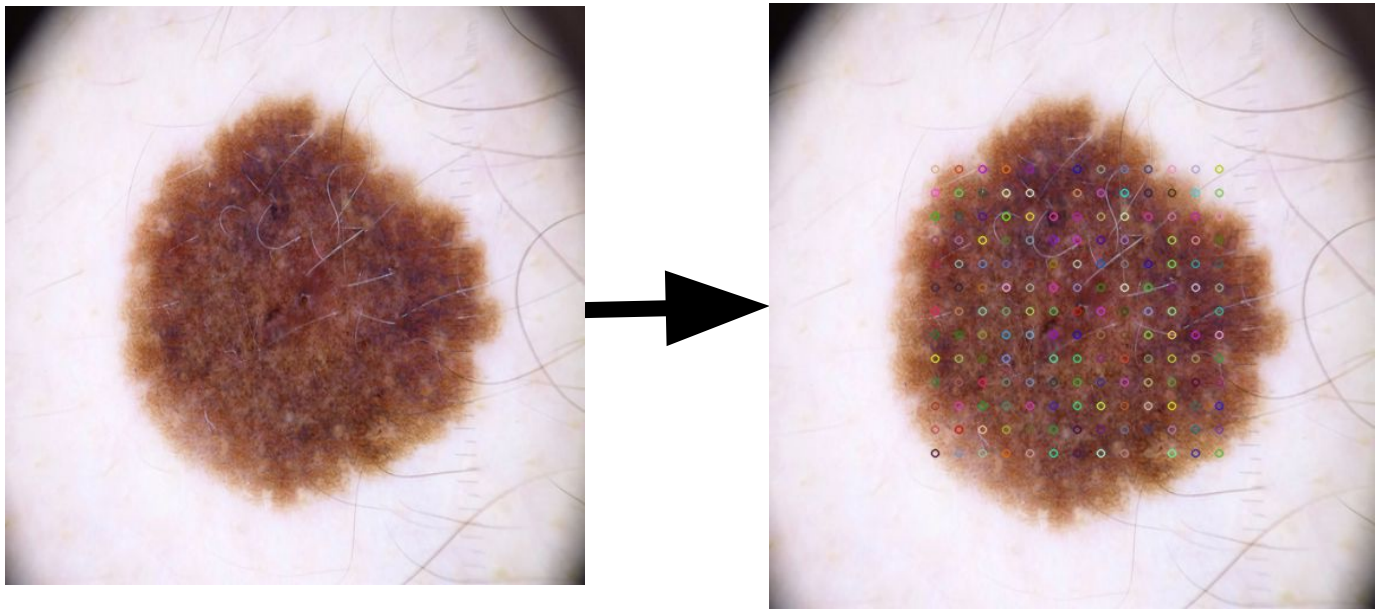
Feature extraction: LBP Texture features

- LBP concatenated histograms for 3 cases (P,R): (8,1), (16,2), (24,3).
- Number of features: 54

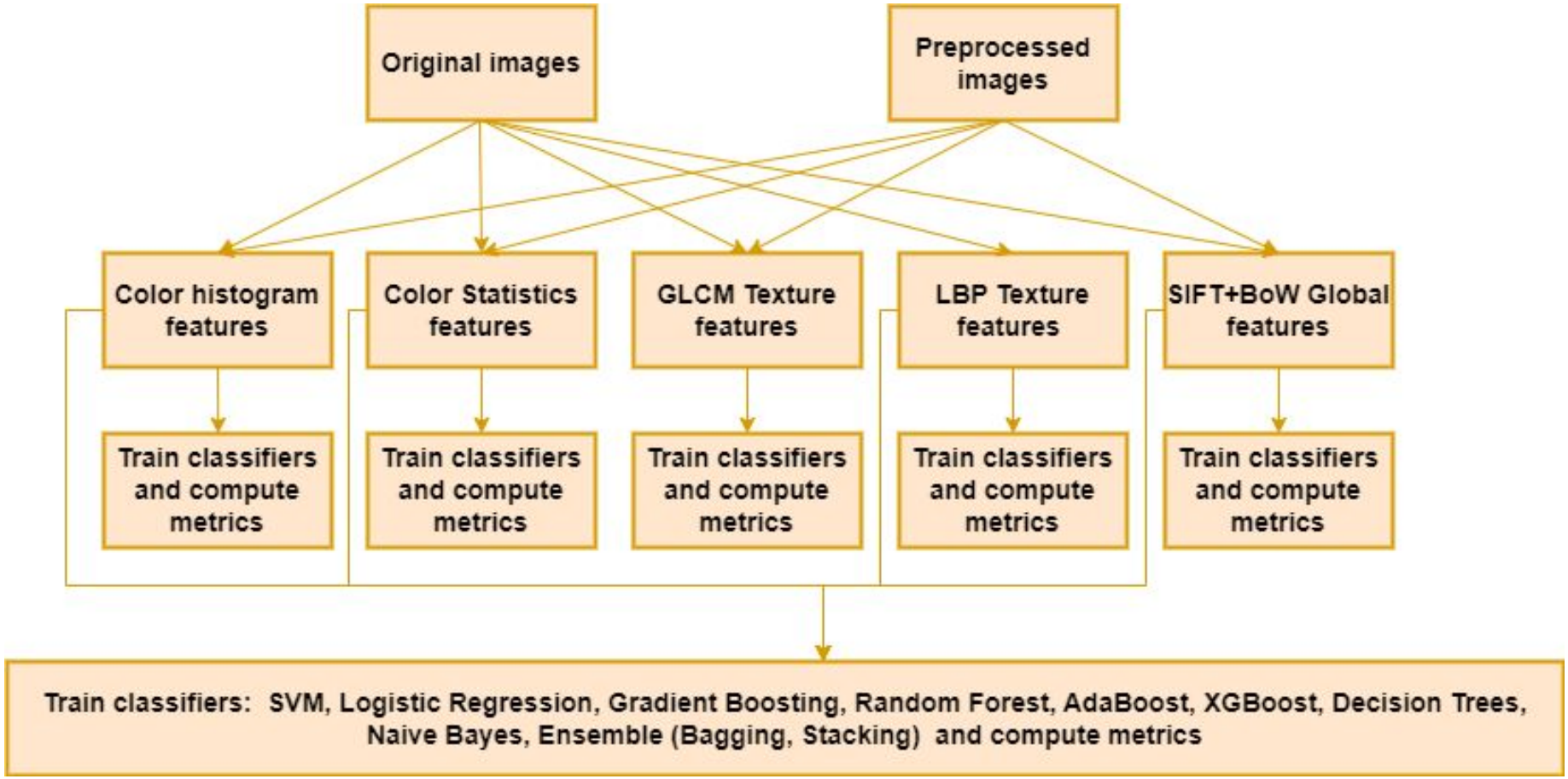
- Statistics of LBP histograms for the 3 chosen cases: mean, std, kurtosis, skew, entropy:
- Number of features 15

Feature Extraction: SIFT + BoW

- 200 randomly chosen images are used for vocabulary creation
- The number of clusters for BoW was kept 200 (Features)
- Dense sift descriptors were used
- Total number of features: 500



Two class Problem: Scheme Diagram



Two class problem: Results (I)

Case – 1 : Without Preprocessing

Features	SVM	LR	Gboost	RF	Adaboost	Decision Trees	XGBoost	Naive Bayes	KNN	Ensemble	
										Bagging	Stacking
CH	0,750	0,674	0,768	**0,801**	0,730	0,713	0,765	0,628	0,750	0,803	**0,804**
CS	0,782	0,739	0,768	**0,793**	0,734	0,716	0,763	0,629	0,753	0,793	**0,795**
GLCM	0,696	0,649	0,696	**0,717**	0,649	0,641	0,688	0,574	0,682	0,725	**0,728**
LBPH	0,734	0,726	0,713	**0,745**	0,699	0,666	0,715	0,653	0,722	0,745	**0,761**
LBPS	0,672	0,667	0,685	**0,712**	0,661	0,660	0,700	0,588	0,682	0,713	**0,715**
SIFT + BoW	0,757	0,743	0,749	**0,762**	0,738	0,676	0,745	0,726	0,739	0,756	**0,768**
All	**0,822**	0,804	0,807	**0,820**	0,773	0,728	0,805	0,723	0,788	0,820	**0,830**

Two class problem: Results (I)

Case – 2 : With Preprocessing

Features	SVM	LR	Gboost	RF	Adaboost	Decision Trees	XGBoost	Naive Bayes	KNN	Ensemble	
										Bagging	Stacking
CH	0,767	0,671	0,778	**0,801**	0,738	0,695	0,770	0,684	0,736	0,795	**0,800**
CS	0,784	0,714	0,774	**0,793**	0,741	0,700	0,773	0,630	0,763	0,786	**0,795**
GLCM	0,725	0,691	0,717	**0,727**	0,706	0,641	0,721	0,634	0,698	0,734	**0,736**
LBPH	**0,748**	0,724	0,737	0,745	0,714	0,665	0,737	0,657	0,722	0,750	**0,754**
LBPS	0,699	0,694	0,696	**0,714**	0,673	0,643	0,692	0,570	0,696	**0,722**	0,715
SIFT + BoW	**0,760**	0,739	0,743	0,759	0,733	0,664	0,741	0,714	0,731	0,750	**0,768**
All	**0,817**	0,793	0,805	0,815	0,778	0,711	0,806	0,743	0,779	0,817	**0,830**

Two class problem: Results (III)

- Taking only the original images, an ensemble of classifiers with majority voting was constructed under the following three scenarios:
- Ensemble of the 3 best standalone classifiers for each feature set, and then ensemble of the resulting ensemble classifier from each feature set.
- Ensemble of the best standalone classifier for each feature set
- Ensemble of the best overall classifier for each feature set (ensemble of stacking classifiers)

Ensemble Model	Balanced accuracy
Model 1	0,781
Model 2	0,802
Model 3	0,824

Two class problem: results (IV)

- Another experiment was carried out, where the best standalone model for the best feature set (dataset without preprocessing, SVM, all features), was retrained on all possible combinations of features to determine the optimum number of features using SelectKBest function. It was obtained that with 415 features, SVM obtains an accuracy score of 0.8245 as opposed to 0.822 with 500 features.

Two class problem: results (V)

- Overall, results are very similar with and without preprocessing. However, models trained on dataset **without preprocessing** perform slightly better results and it is computationally less expensive, so it was decided to use this approach for the final pipeline selection.

Two class problem: results (VI)

- **Random Forest** classifier proved to be the best model along all experiments as a standalone classifier. However, for all cases **Stacking classifier** gets the highest performance, so it was decided to keep this model for the final pipeline selection.

Two class problem: results (VII)

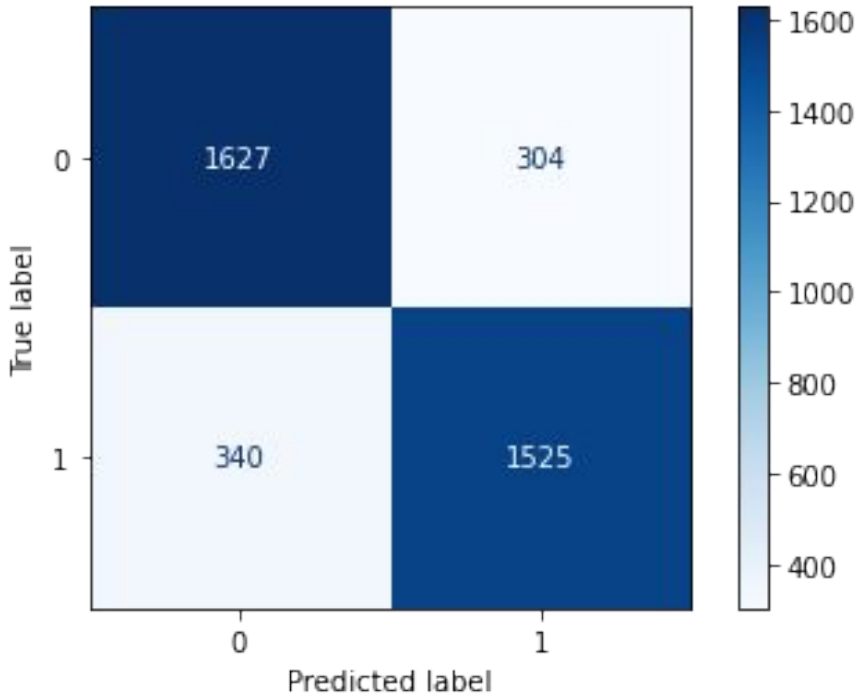
- Finally, when evaluating all experiments, using a **Stacking classifier** with **all features** yields the highest balanced accuracy, therefore, this is used as the final pipeline for generating predictions on test set.

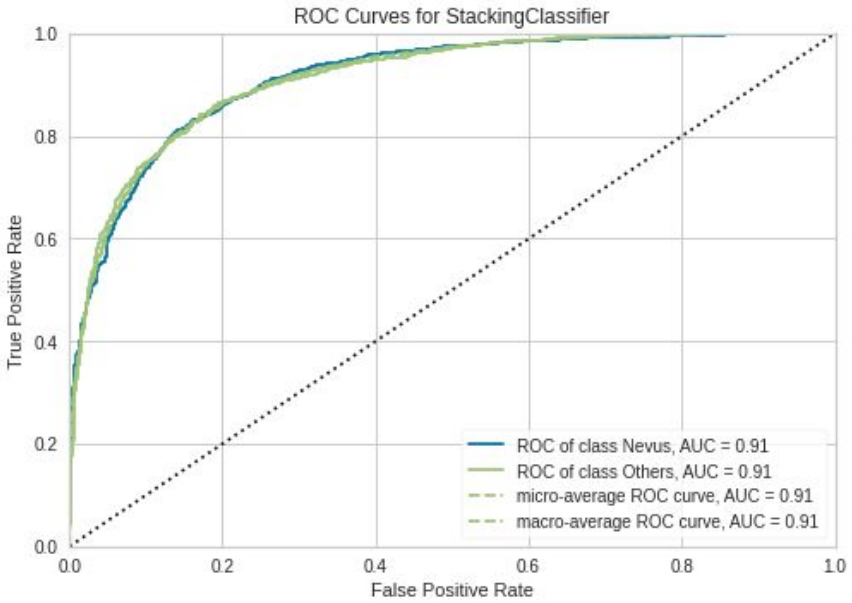
Two class problem: Best model

- Stacking of classifiers trained on data without preprocessing and using all features.
- Stacking of classifiers: using the output of each individual classifier as input of a final classifier
- Classifiers used for the stacking: LR, KNN, Decision Tree, SVM, Bayes, Gboost, RF, Adaboost
- Final Classifier: LR

Two class problem: Best model

Class	Precision	Recall	F1-Score
0	0.83	0.84	0.83
1	0.83	0.82	0.83

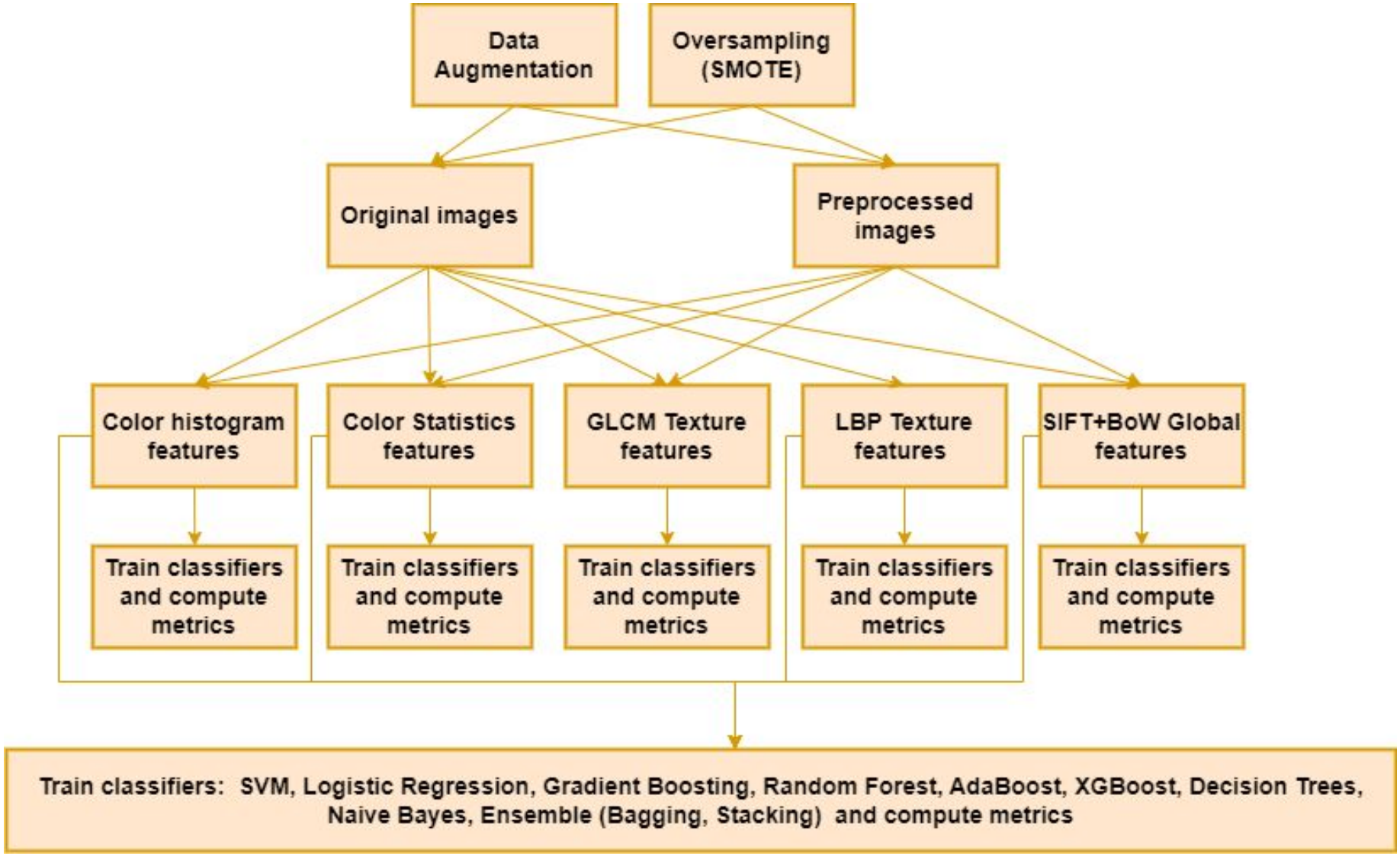




Two class problem: conclusions

- The overall results for the validation set are good. The misclassification errors for both classes are very similar, which is theoretically correct as the classes are balanced.
- Using an ensemble of stacking classifiers trained on each feature set did not perform better than a single stacking classifier trained on all features.

Three class Problem: Scheme Diagram



Three class Problem: Results(I)

- Case – 1 : Without Preprocessing + Data Augmentation (Balanced Accuracy)

Feature Extractor	SVM	LR	GBoost	RF	Adaboost	Decision Trees	XGBoost	Naive Bayes	KNN	Ensemble	
										Bagging	Stacking
CHF	0.667	0.54	0.686	**0.744**	0.565	0.621	0.651	0.428	0.674	0.746	**0.756**
CSF	0.648	0.534	0.627	**0.696**	0.581	0.588	0.622	0.429	0.651	0.688	**0.71**
GLCM	0.507	0.464	0.51	**0.562**	0.476	0.514	0.508	0.454	0.518	0.539	**0.564**
LBPH	0.464	0.457	0.467	**0.515**	0.402	0.475	0.446	0.419	0.473	0.51	**0.522**
LBPS	0.409	0.372	0.408	**0.451**	0.401	0.433	0.432	0.34	0.431	0.441	**0.462**
SIFT + BoW	0.587	0.56	0.57	0.586	0.527	0.473	0.561	0.517	**0.596**	0.566	**0.614**

Three class Problem: Results(I)

- Case – 2 : Without Preprocessing + Smote (Balanced Accuracy)

Feature Extractor	SVM	LR	GBoost	RF	Adaboost	Decision Trees	XGBoost	Naive Bayes	KNN	Ensemble	
										Bagging	Stacking
CHF	0.69	0.543	0.682	**0.71**	0.592	0.591	0.635	0.457	0.698	0.731	**0.733**
CSF	0.634	0.541	0.643	**0.673**	0.574	0.558	0.635	0.458	0.631	0.693	0.645
GLCM	0.546	0.506	0.558	**0.615**	0.506	0.527	0.548	0.41	0.578	0.625	**0.613**
LBPH	**0.598**	0.575	0.567	0.592	0.523	0.498	0.571	0.405	0.557	**0.611**	0.588
LBPS	0.521	0.466	0.496	0.507	0.478	0.439	0.496	0.362	0.509	**0.525**	0.501
SIFT + BoW	**0.589**	0.518	0.482	0.517	0.458	0.475	0.494	0.499	**0.589**	0.521	0.576

Three class Problem: Results(I)

- Case – 3 : With Preprocessing + Data Augmentation (Balanced Accuracy)

Feature Extractor	SVM	LR	GBoost	RF	Adaboost	Decision Trees	XGBoost	Naive Bayes	KNN	Ensemble	
										Bagging	Stacking
CHF	0.619	0.501	0.65	0.682	0.595	0.572	0.644	0.459	0.645	**0.702**	**0.706**
CSF	0.602	0.52	0.584	**0.652**	0.536	0.543	0.576	0.437	0.632	0.636	**0.669**
GLCM	0.55	0.503	0.568	**0.59**	0.492	0.477	0.55	0.446	0.552	**0.611**	0.595
LBPH	0.572	0.579	0.548	0.589	0.514	0.495	0.535	0.463	0.556	**0.597**	**0.599**
LBPS	0.497	0.472	0.527	**0.554**	0.459	0.46	0.523	0.349	0.492	0.546	**0.562**
SIFT + BoW	0.577	0.541	0.562	**0.591**	0.508	0.442	0.554	0.484	0.575	0.588	**0.599**

Three class Problem: Results(III)

- Case – 4 : With Preprocessing + Smote (Balanced Accuracy)

Feature Extractor	SVM	LR	GBoost	RF	Adaboost	Decision Trees	XGBoost	Naive Bayes	KNN	Ensemble	
										Bagging	Stacking
CHF	0.647	0.518	0.637	0.658	0.57	0.545	0.623	0.458	0.656	**0.666**	**0.661**
CSF	0.622	0.511	0.611	**0.644**	0.54	0.549	0.616	0.434	**0.629**	0.654	0.623
GLCM	**0.575**	0.496	0.534	0.58	0.504	0.505	0.531	0.43	0.541	**0.582**	0.557
LBPH	**0.581**	**0.591**	0.576	0.555	0.537	0.482	0.538	0.451	0.539	0.549	0.558
LBPS	0.504	0.489	0.523	**0.528**	0.48	0.441	0.483	0.353	0.507	0.538	**0.524**
SIFT + BoW	**0.555**	0.52	0.482	0.507	0.477	0.427	0.481	0.47	0.552	0.521	**0.558**

Three class Problem: Results(IV)

- **Without any Processing** yields better results
- **Data Augmentation** provides better results so far than SMOTE Oversampling
- **Color Histogram (BGR)** is the best Feature Extractor so far
- Case – 5 (Hair Removal Only with Data Augmentation, CHF)

Feature Extractor	SVM	LR	GBoost	RF	Adaboost	Decision Trees	XGBoost	Naive Bayes	KNN	Ensemble	
										Bagging	Stacking
CHF	0.635	0.495	0.644	**0.716**	0.562	0.579	0.641	0.409	0.642	0.701	**0.723**

- Standalone hair removal is better than combined with vignette removal
- Still without any preprocessing is better performer, so decided to stick with it

Three class Problem: Results(IV)

- Case – 6 : Features Combined (Balanced Accuracy)
- Combined the feature extractors (CHF, CSF, GLCM, SIFT, LBP)
- Different combinations were tried but the best one was using **CHF, CSF, GLCM**

Feature Extractor	SVM	LR	GBoost	RF	Adaboost	Decision Trees	XGBoost	Naive Bayes	KNN	Ensemble	
										Bagging	Stacking
[CHF, CSF, GLCM]	0.601	0.583	576	**0.647**	0.525	0.601	0.568	0.443	0.643	0.624	**0.687**

- **SelectKBest Features** was used, the highest accuracy received was **0.731** combining **286** Features
- Improved the overall result but still lacks behind standalone CHF

Three class Problem: Results(IV)

- Case – 7 : Ensemble Classifiers (Balanced Accuracy)
- Majority and Average voting was used
- **3 best models** from each case-I per feature extractor was used
- Equal votes are untied choosing the lower class

Ensemble Model	**Balanced Accuracy**
CHF (RF, Gboost, SVM)	0.714
CSF (SVM, Gboost, RF)	0.656
GLCM (RF, KNN, Gboost)	0.553
Ensemble of Ensemble	0.681
Average Voting	**0.743**

- Average Voting gives improved performance with CHF (RF, Gboost, SVM), still not the best

Three class Problem: Results(V)

- Case – 8 : Color Histogram Analysis (Balanced Accuracy)
- As color histogram yields the best result, so other **color spaces** were also explored with adaptive histogram equalization

Feature Extractor	SVM	LR	GBoost	RF	Adaboost	Decision Trees	XGBoost	Naive Bayes	KNN	Ensemble	
										Bagging	Stacking
CHF_BGR	0.656	0.552	0.668	**0.738**	0.614	0.617	0.670	0.450	0.689	0.739	**0.741**
CHF_HSV	0.679	0.561	0.685	**0.751**	0.587	0.594	0.670	0.457	0.674	0.740	**0.758**
CHF_Luv	0.654	0.546	0.687	**0.729**	0.575	0.606	0.675	0.387	0.676	0.744	**0.748**
CHF_XYZ	0.641	0.563	0.677	**0.713**	0.578	0.585	0.668	0.451	0.652	**0.723**	0.634

- HSV color space overcomes the best result of BGR color space

Three class Problem: Results(V)

- Case – 8 : Color Histogram Analysis (Cohen Kappa Score)
- As the problem is imbalanced class problem, **cohen kappa score** was also measured for the best case scenario

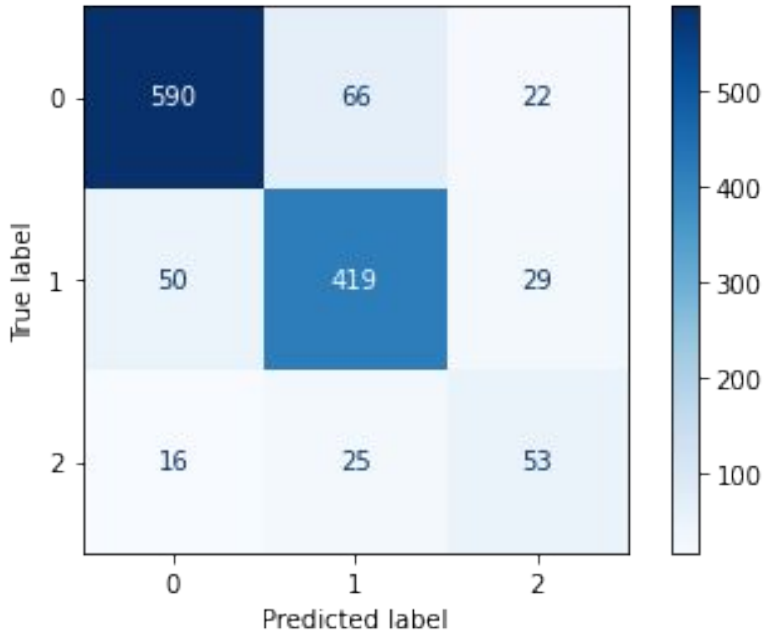
Feature Extractor	SVM	LR	GBoost	RF	Adaboost	Decision Trees	XGBoost	Naive Bayes	KNN	Ensemble	
										Bagging	Stacking
CHF_BGR	0.508	0.396	0.533	0.658	0.423	0.439	0.523	0.282	0.534	0.640	**0.674**
CHF_HSV	0.545	0.374	0.572	0.679	0.434	0.401	0.556	0.283	0.537	0.655	**0.708**
CHF_Luv	0.521	0.349	0.564	0.658	0.422	0.439	0.549	0.063	0.514	0.661	**0.693**
CHF_XYZ	0.486	0.394	0.529	0.617	0.414	0.395	0.502	0.287	0.499	0.612	**0.634**

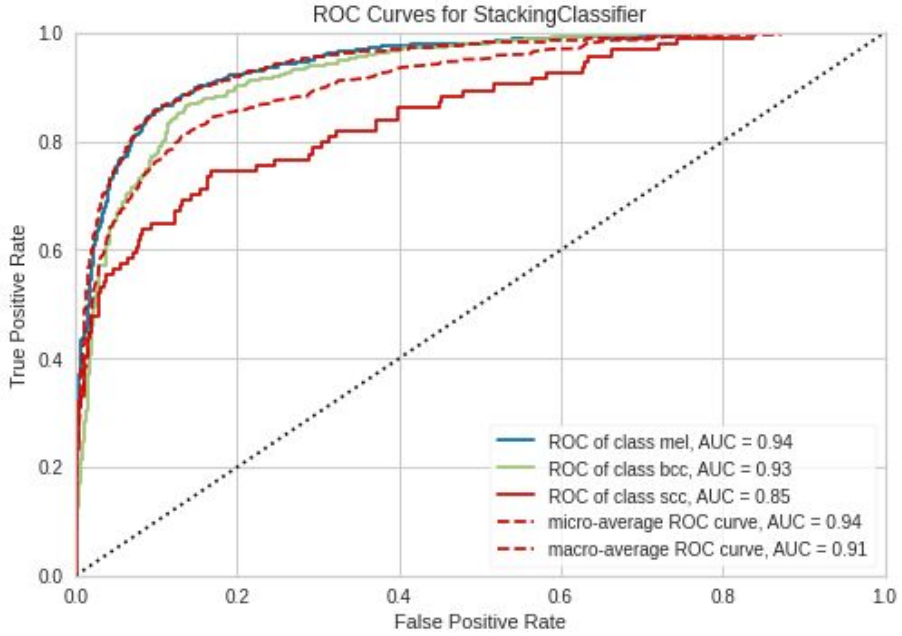
- HSV Color space Histogram is the best!

Three class Problem: Best Model

- Classification Report on Validation Set (Best Model)

Class	Precision	Recall	F1-Score
0	0.9	0.87	0.88
1	0.82	0.84	0.83
2	0.51	0.56	0.54





Three class Problem: Conclusions

- Winner of Three Class Classification Problem Configuration:
  - Without preprocessing (Hair/Vignetter Removal)
  - Contrast Enhancement by **Histogram Equalization**
  - Using **Data Augmentation**
  - Feature Extraction: **Color Histogram**
  - Color Space: **HSV**
  - Classifier: **Stacking (Ensemble) / Random Forest (StandAlone)**

- The overall balanced accuracy is decent but due to very low number of samples from scc class, it hampers the overall performance.