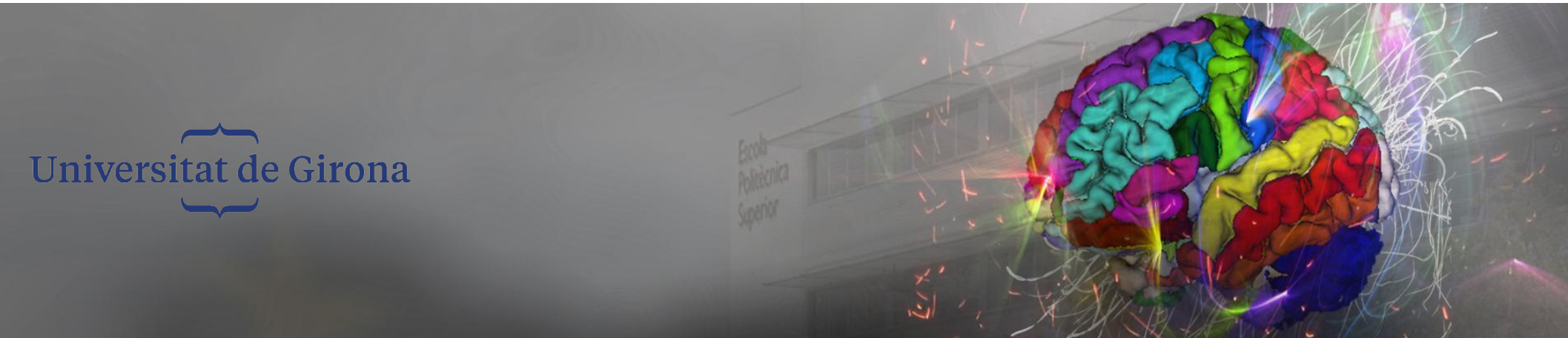
MAIA

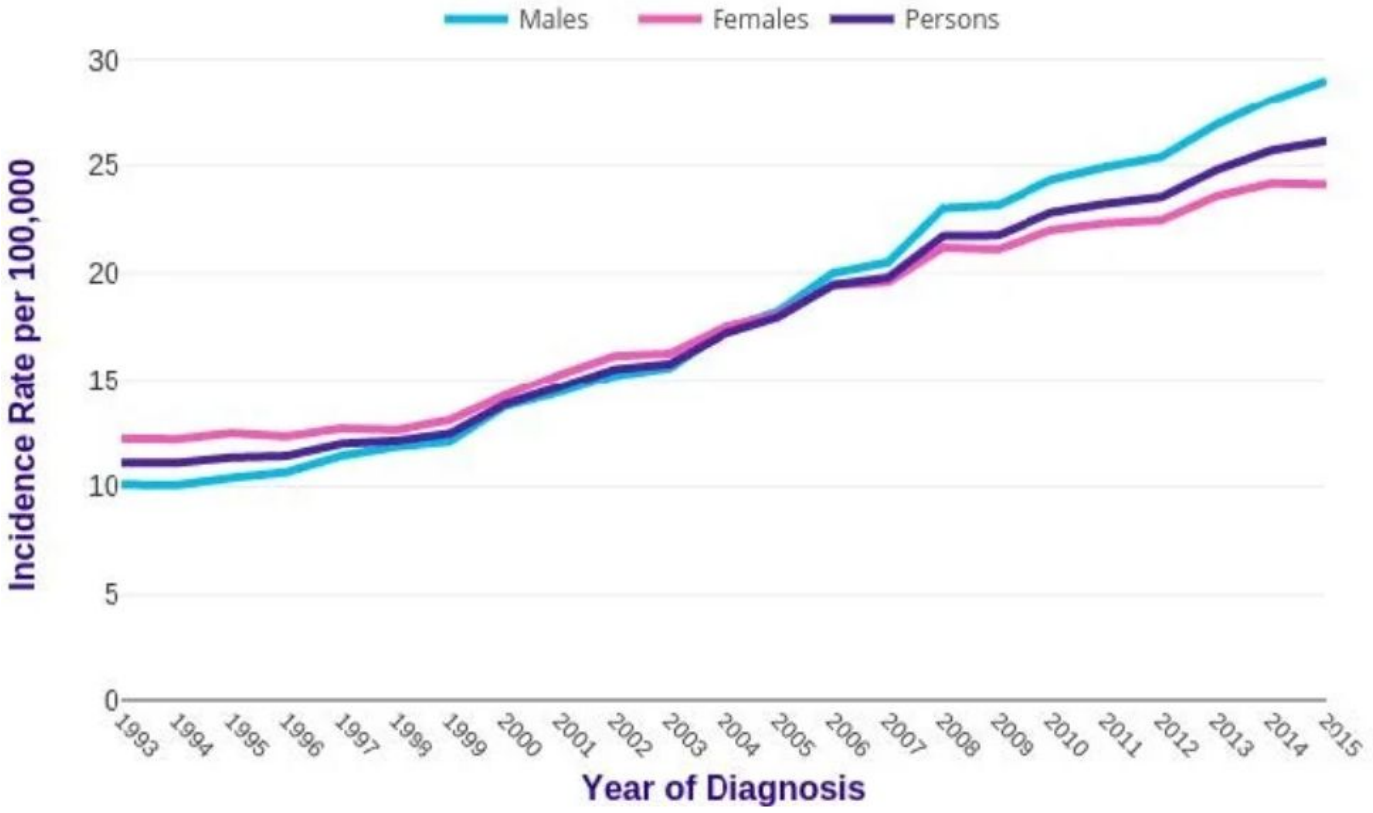
CAD Project 1: Skin Lesions Classification

Anastasiia Rozhyna
Aroj Hada



Introduction

- Skin cancer is one of the most common cancers in the world
- Melanoma is the deadliest form of skin cancer
- Annual cases of melanoma have increased by nearly 50% to over 287,000



Challenges

- **Challenge 1**

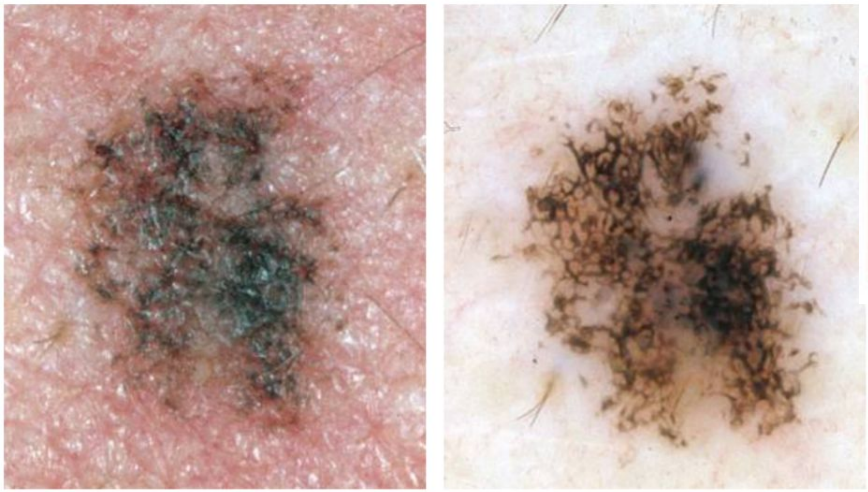
  The binary problem of classifying Nevus images vs all the others. We will give you 6000 images, 3000 being nevus, 3000 being a combination of the others to train the system. The test set will consist on 1015 images.

- **Challenge 2**

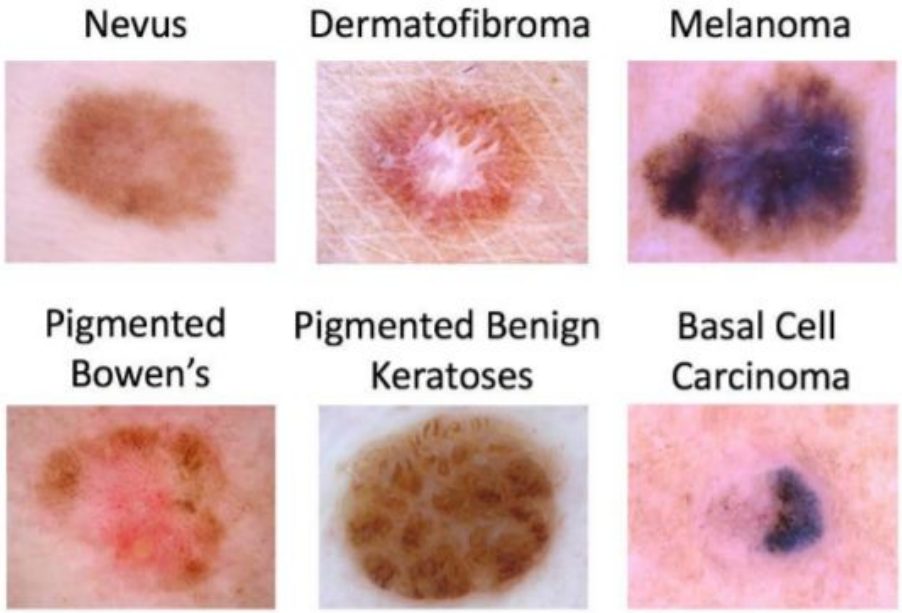
  A three-class problem consisting on the classification of melanoma vs benign keratosis vs basal cell carcinoma. The training set will consist on 1000 images for the first two classes and 500 for the last one (imbalanced problem). The test set consist on 226 images

Datasets

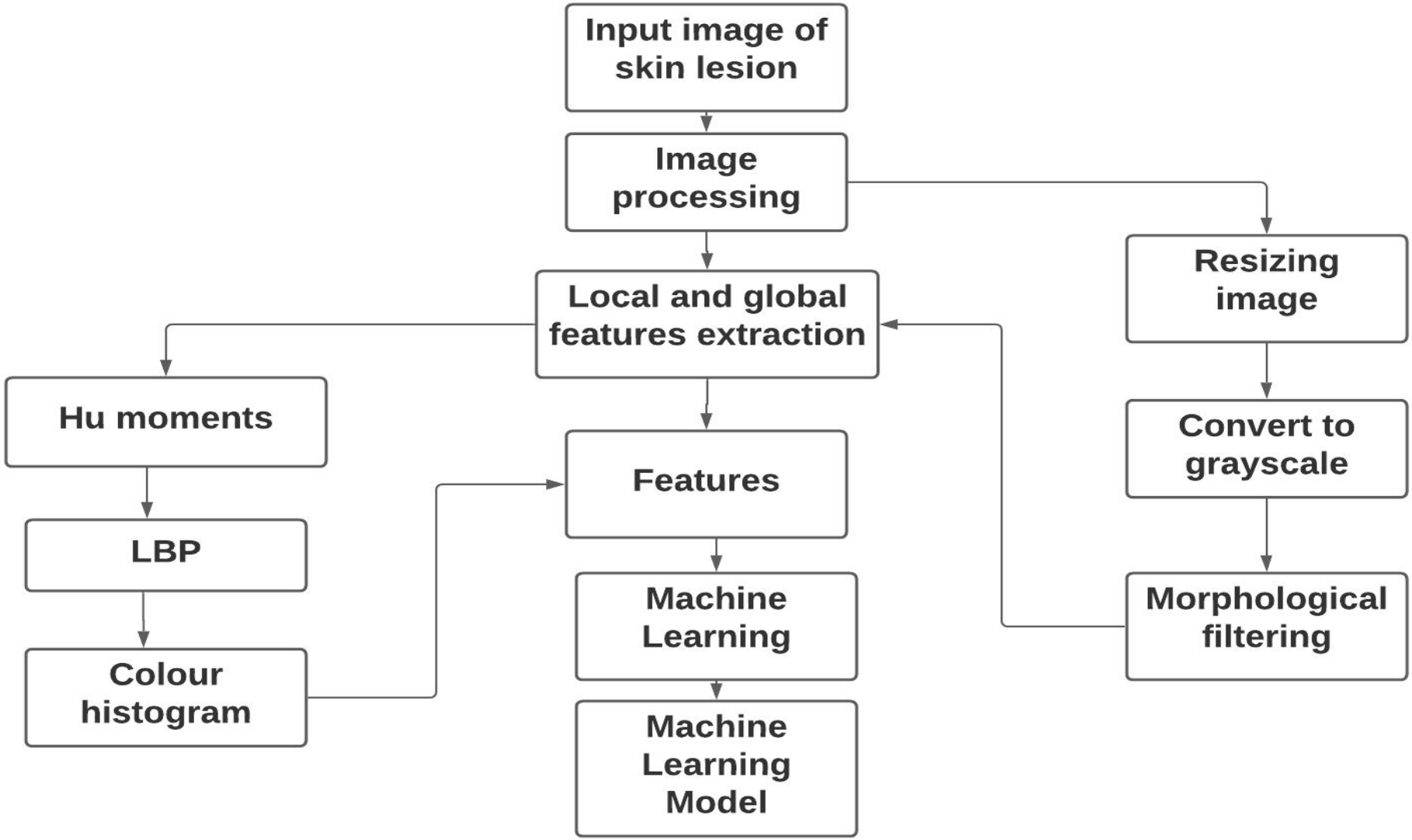
- The input data are dermoscopic images in JPEG format
- The lesion images come from the HAM10000 Dataset



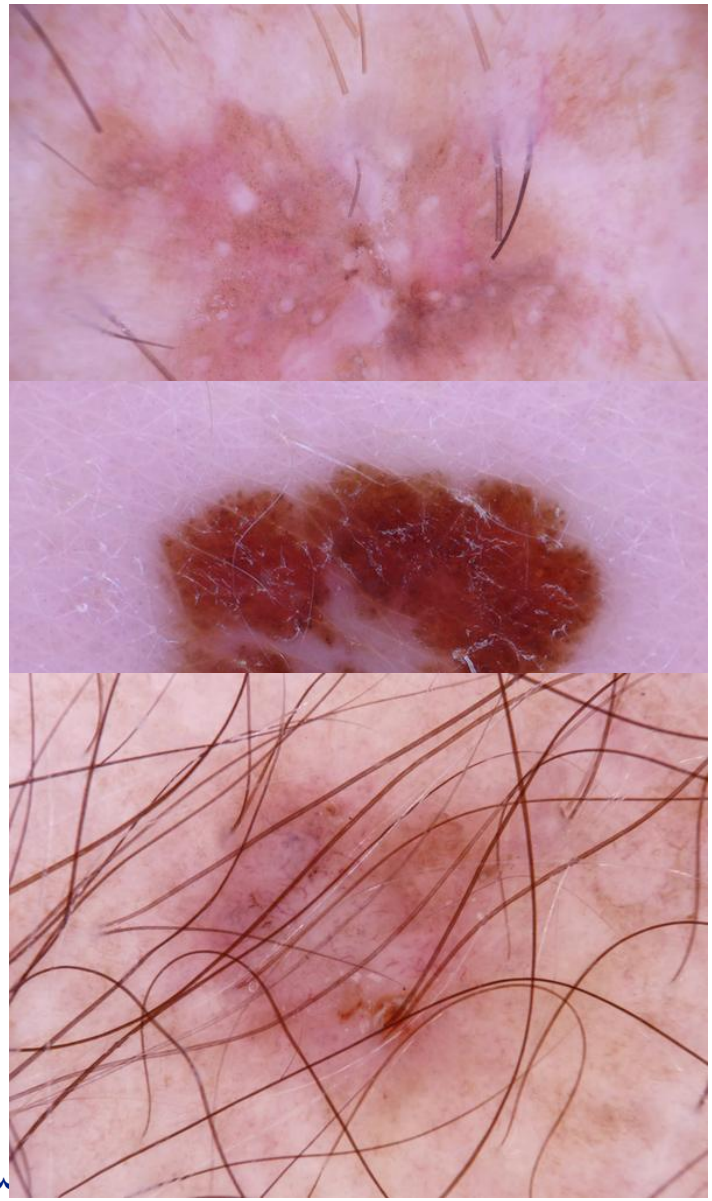
Clinical Image v. Dermoscopy image

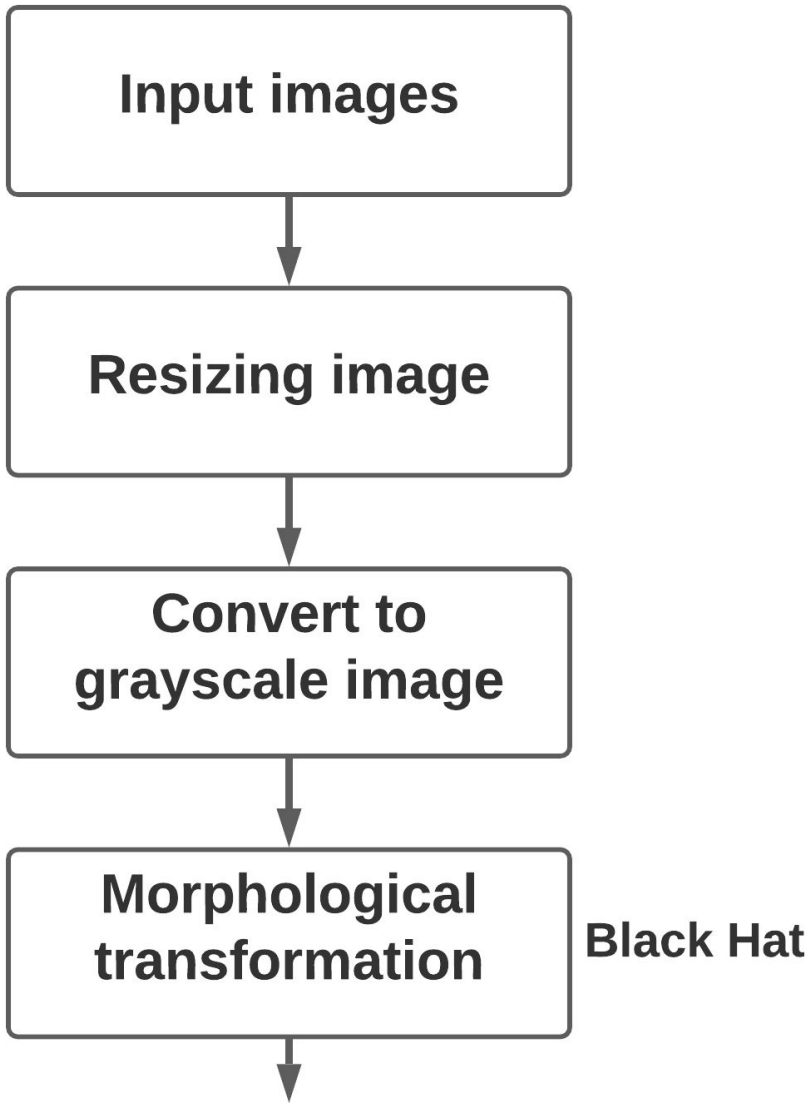


General Pipeline



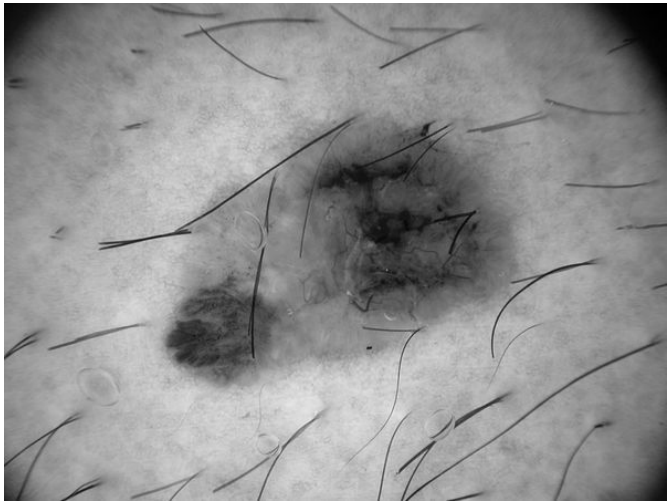
Preprocessing



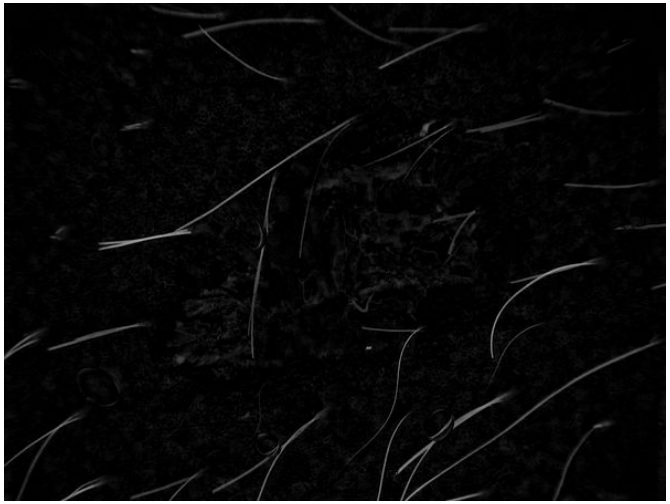


Hair removal

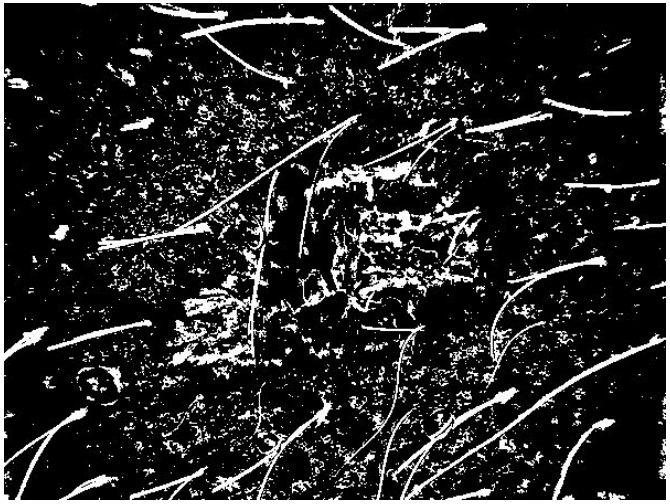
Original Image



GreyScale Image



Blackhat Transformation



Intensify hair contours
with thresholding

Inpainting original image
depending on last image

Feature Extraction

- Hu moments
- LBP
- Colour histogram
- HoG
- GLCM
- Haralick
- SIFT

Feature Engineering

- Data Normalization: Min-Max scaling
- PCA (Principal Component Analysis)

Problem	Accuracy with PCA on best model	Accuracy without PCA on best model
2-class problem	0.868	0.784
3-class problem	0.752	0.69.3

ML Algorithms tried

Algorithms	Parameters
• Random Forest	n_estimators : [500, 1000, 2000]: max_depth : [3, 5, 7]
• Gradient Boosted Trees	n_estimators in [500, 1000, 2000]: max_depth in [3, 5, 7]: learning_rate in [0.01, 0.1]:
• SVM	kernel in ['linear', 'rbf']: C in [1.0, 10.0, 100.0, 1000.0]
• Logistic regression	penalty in ['l1', 'l2']: C in [1.0, 10.0, 100.0, 1000.0] :
• KNN	n_neighbors in [5, 10, 15]: weights in ['uniform', 'distance']:

Final Results and Discussions

Metrics achieved different model on validation set

Model	Results	
	2-class Classification	3-class Classification
Random Forest	Accuracy=80.66 Kappa= 0.613 F-1 score= 0.804	Accuracy: 0.694 Kappa: 0.633 F1-score: 0.692
Gradient Boosted Trees	**Accuracy=86.8** **Kappa= 0.735** **F-1 score= 0.867**	**Accuracy: 0.752** **Kappa: 0.727** **F1-score: 0.752**
SVM	Accuracy=80.7 Kappa= 0.613 F-1 score= 0.806	Accuracy: 0.744 Kappa: 0.705 F1-score: 0.744
Logistic regression	Accuracy=80.1 Kappa= 0.602 F-1 score= 0.801	Accuracy: 0.652 Kappa: 0.571 F1-score: 0.650
KNN	Accuracy=80 Kappa= 0.6 F-1 score= 0.8	Accuracy: 0.592 Kappa: 0.407 F1-score: 0.594

Conclusions

- The total number of models for each run is 49.
- Best model was Gradient Boosting for both the challenges
- Feature combination with the highest metrics: Color Histogram + LBP + Hu moments
-

Metric	2- class classification	3- class classification
Accuracy	0.864	0.752
Kappa	0.728	0.727
F1-score	0.864	0.752
Recall	0.864	0.752
Precision	0.866	0.755

Possible improvement

- Joining the local and global features
- Data augmentation
- Symmetry based features

References

1. Hair Removal : https://github.com/sunnyshah2894/DigitalHairRemoval
2. An Overview of Melanoma Detection in Dermoscopy Images Using Image Processing and Machine Learning
   Nabin K. Mishra, M. Emre Celebi
3. Computer aided Melanoma skin cancer detection using Image Processing Shivangi Jaina, Vandana Jagtap, Nitin Pisea