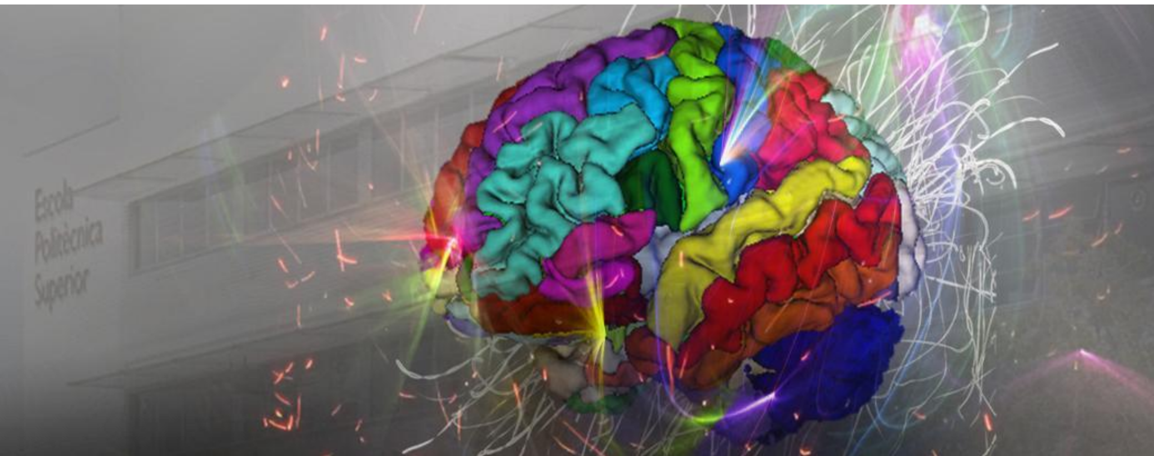
CAD: Skin Lesion Classification

Manasi Kattel
Vladyslav Zalevskyi

Universitat de Girona

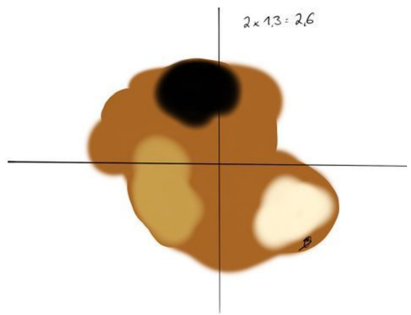


Content

1. Literature review: ABCD rule
2. Preprocessing
   a. Hair removal
   b. Segmentation
3. Feature Extraction
   a. Color
      i. Color preprocessing
   b. Texture
   c. Shape
4. BoW
5. Challenge 1
   a. Feature selection/dimensionality reduction
   b. Results and experiments
6. Challenge 2
   a. Feature selection/dimensionality reduction
   b. Results and experiments
7. Conclusions

Universitat de Girona

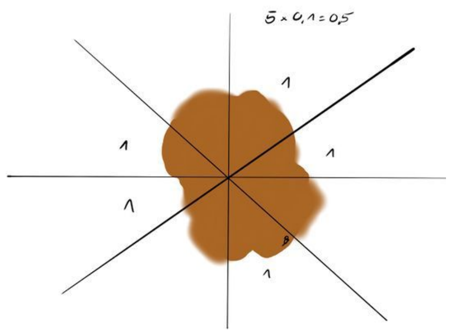
Literature Review: ABCD Rule



A

Asymmetry

- Perform segmentation
- Extract features from the lesion mask
- Highest weight



B

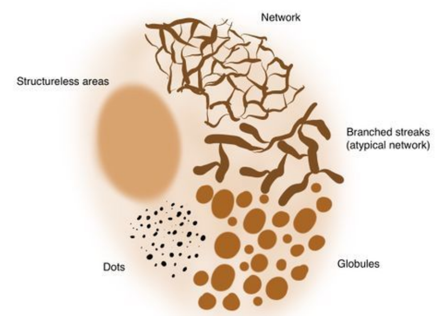
Border

- Perform segmentation
- Analyze textures and color at the lesion border
- Lowest weight

C

Color

- Transform to different color spaces
- Extract features per channel
- Try BoW



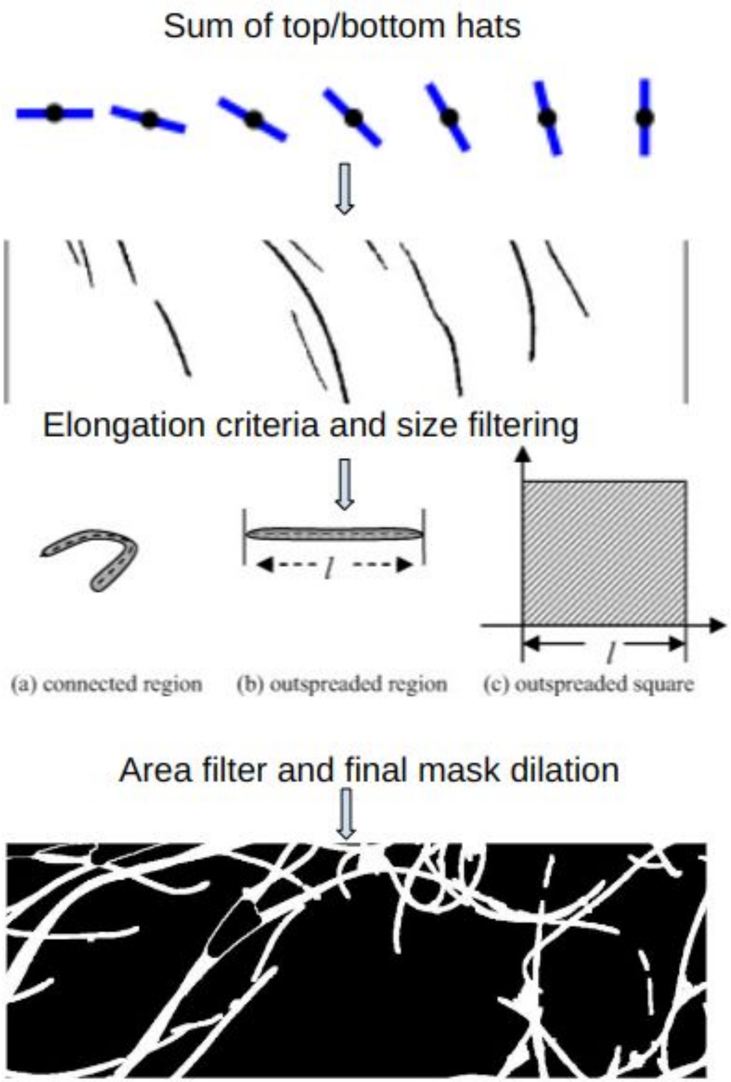
D

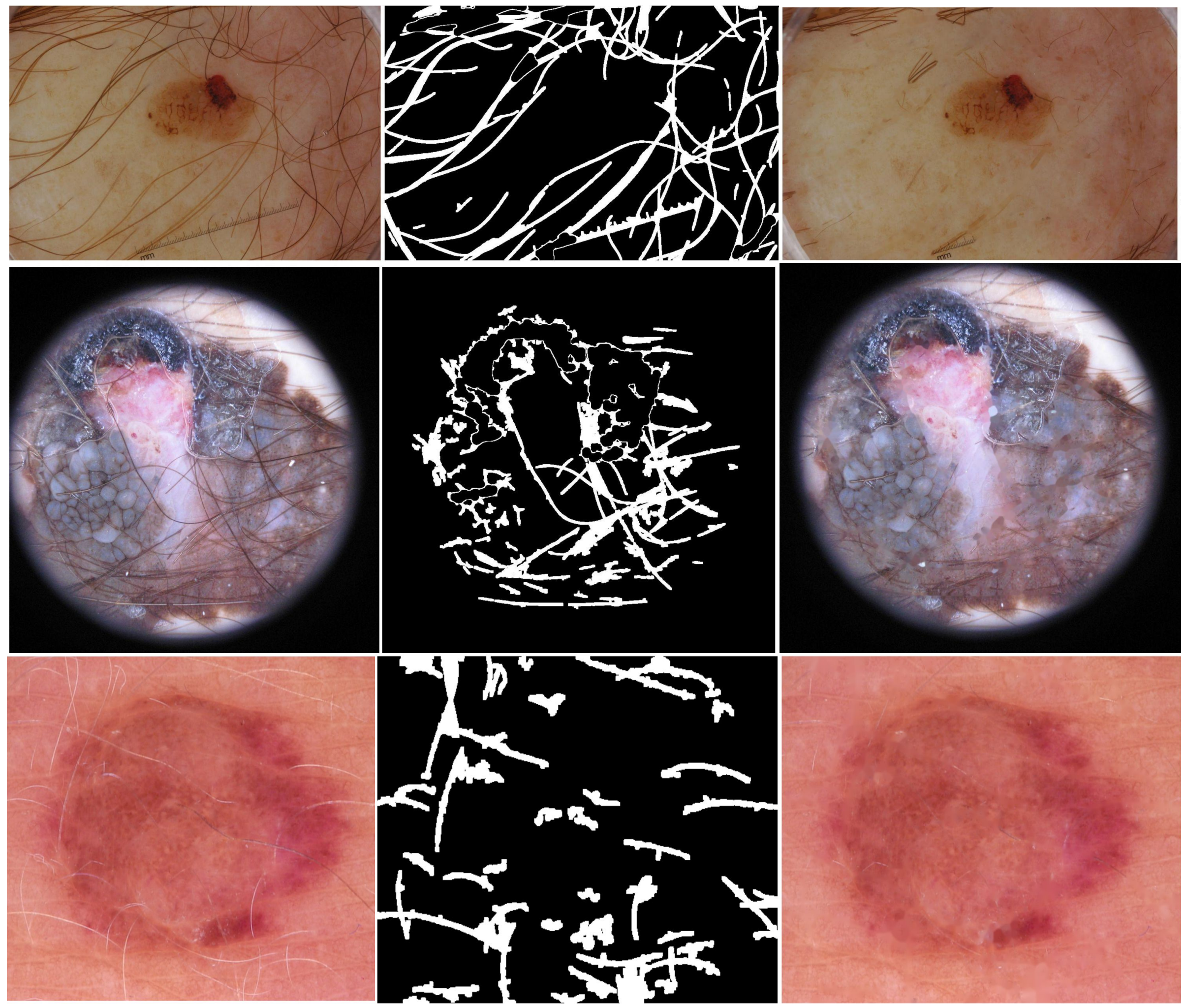
Dermoscopic structures

- Perform segmentation
- Analyze textures and color at the lesion border
- Try BoW

Universitat de Girona

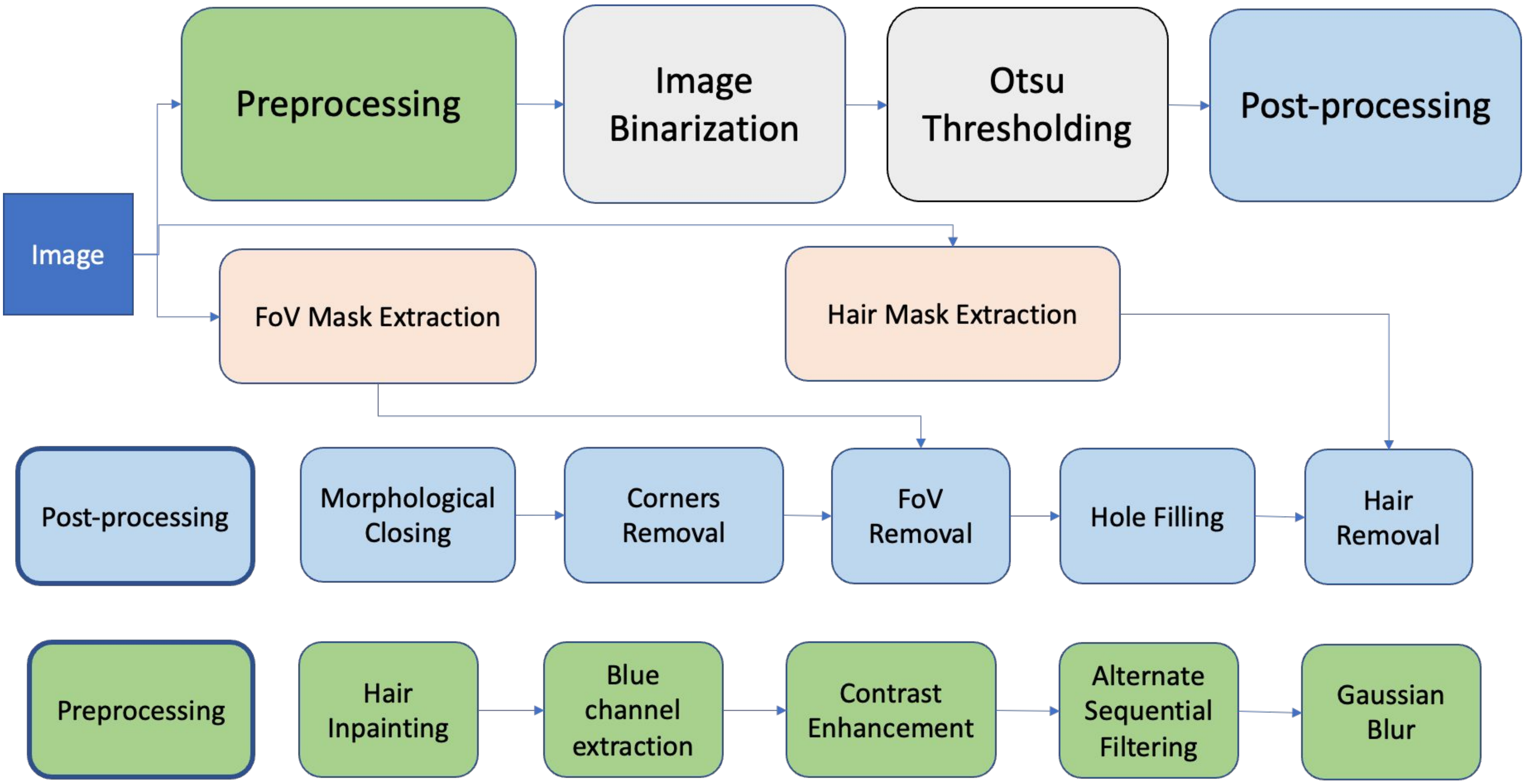
Preprocessing: Hair Removal





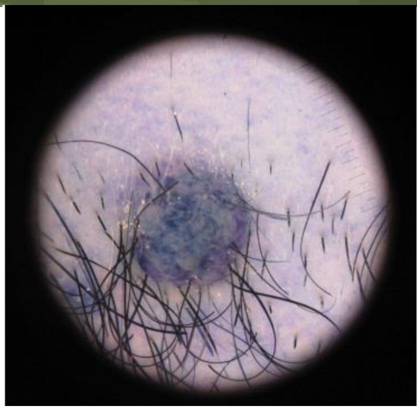
Universitat de Girona

Preprocessing: Segmentation Pipeline

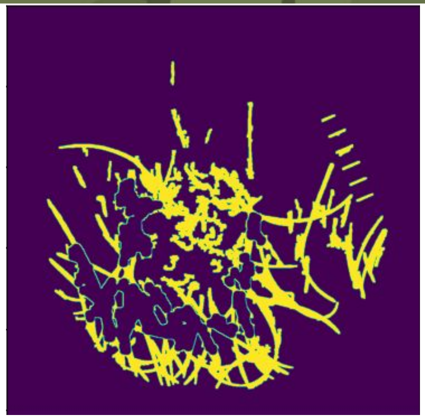


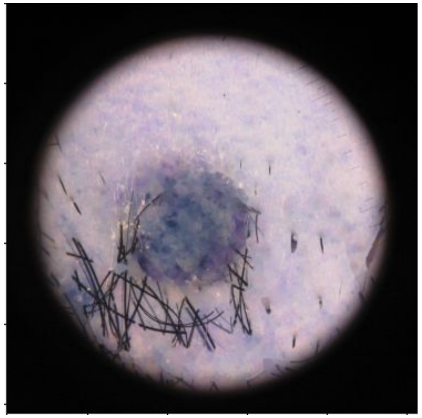
Universitat de Girona

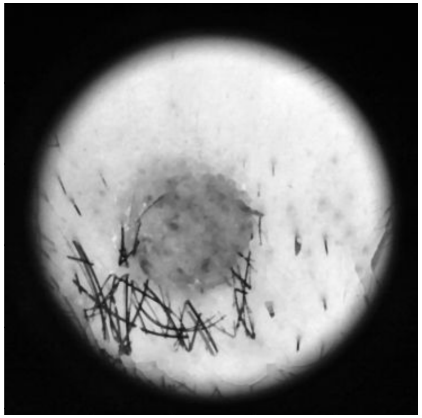
Preprocessing: Segmentation Pipeline


Original Image

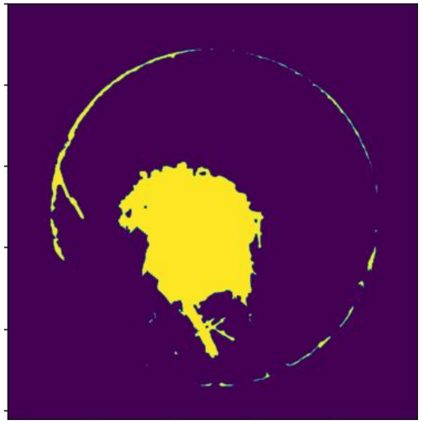

FoV Mask

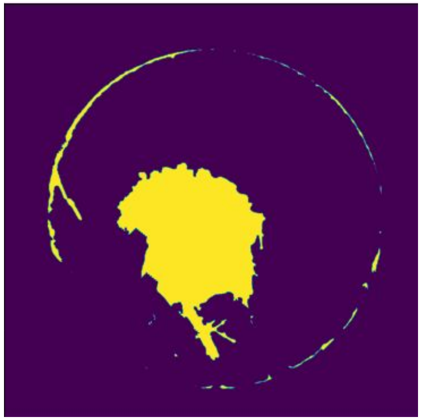

Hair Mask


Inpainted


Enhanced and smoothed


Otsu Thresholding


Remove FoV

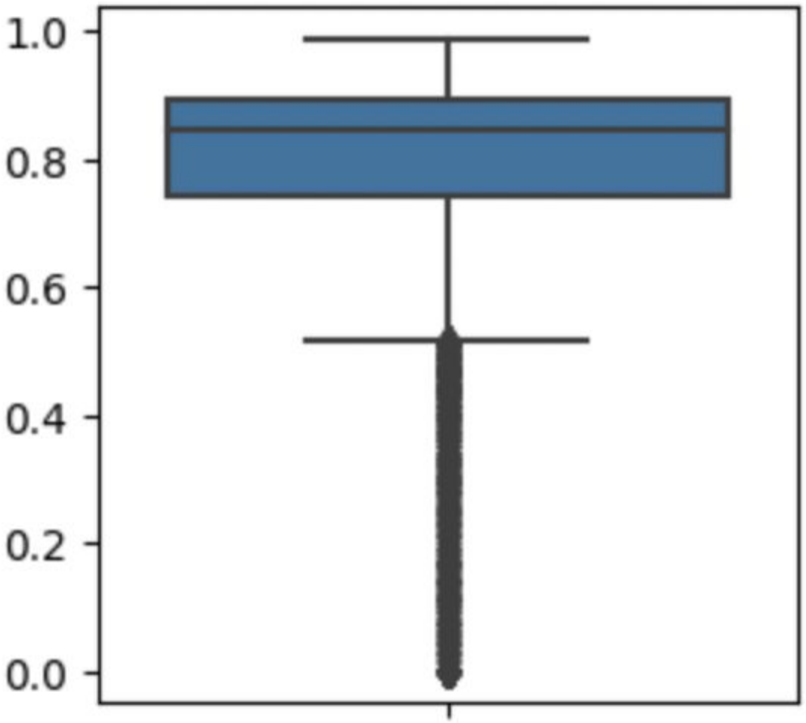

Fill Holes

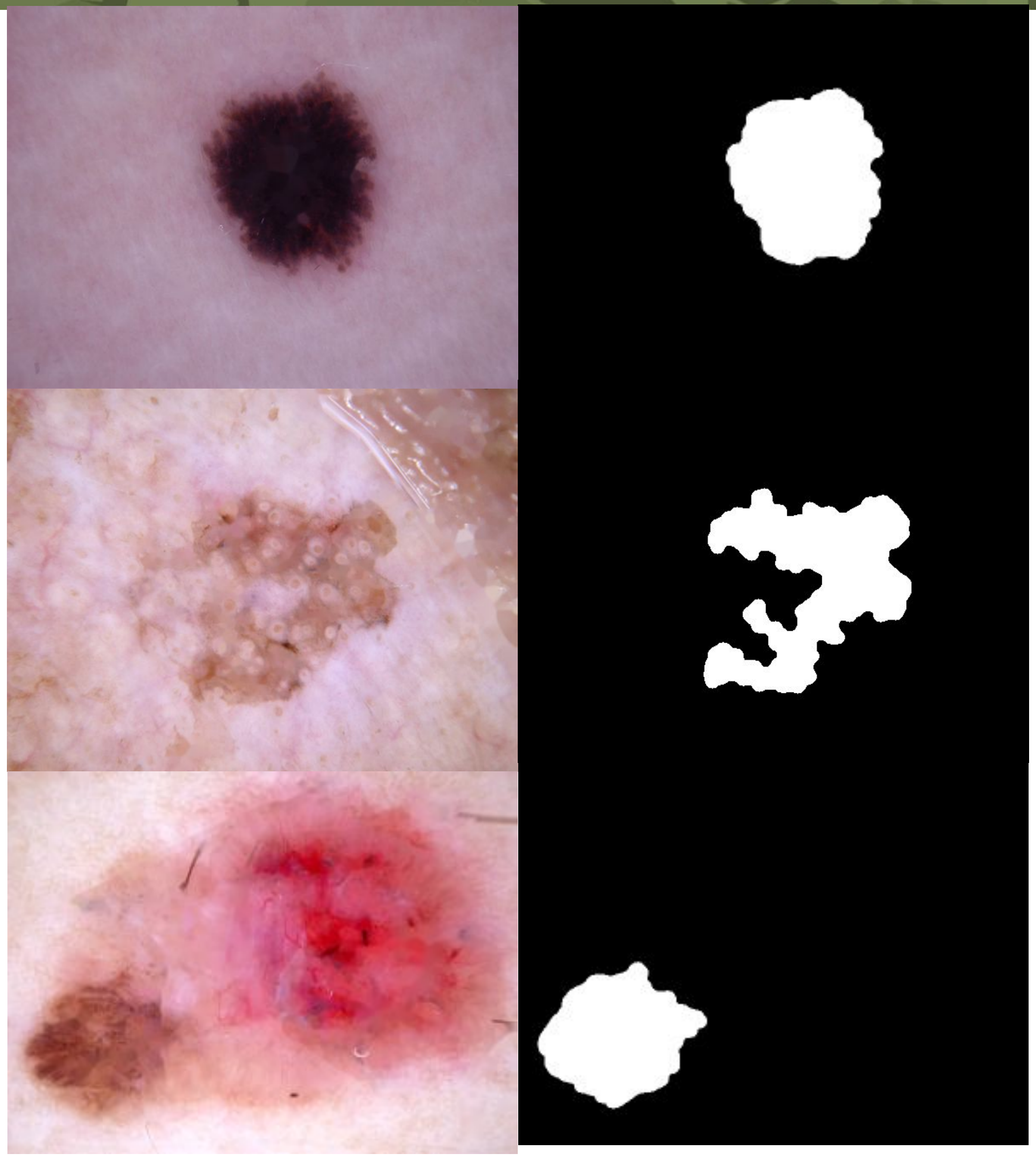

Final Segmentation Mask

Universitat de Girona

Preprocessing: Segmentation

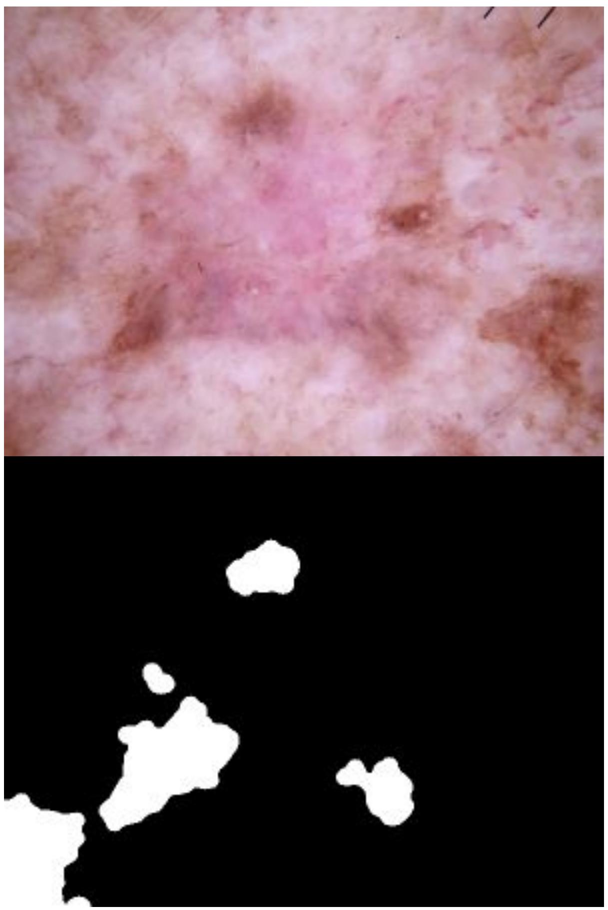
Dice scores of the developed segmentation algorithm reported on the HAM10000 dataset

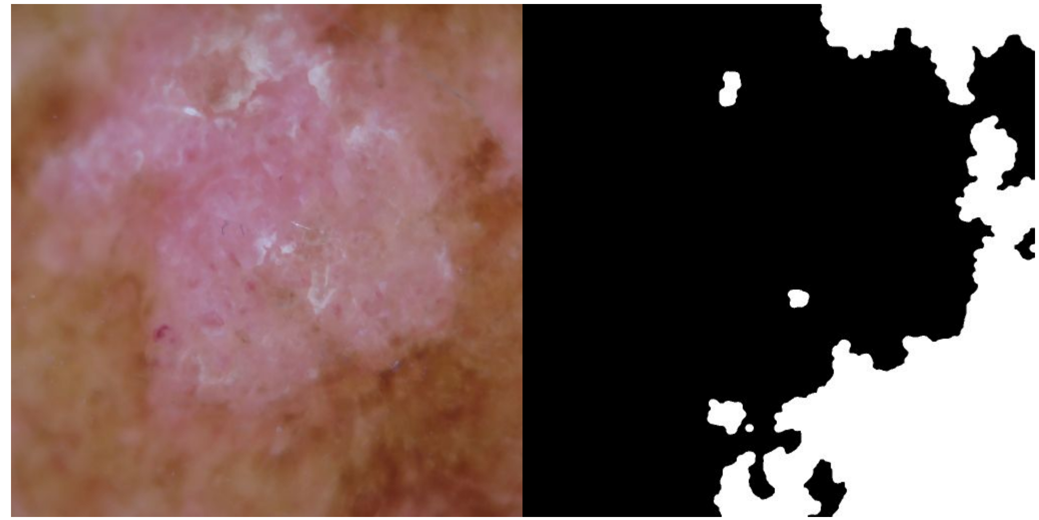


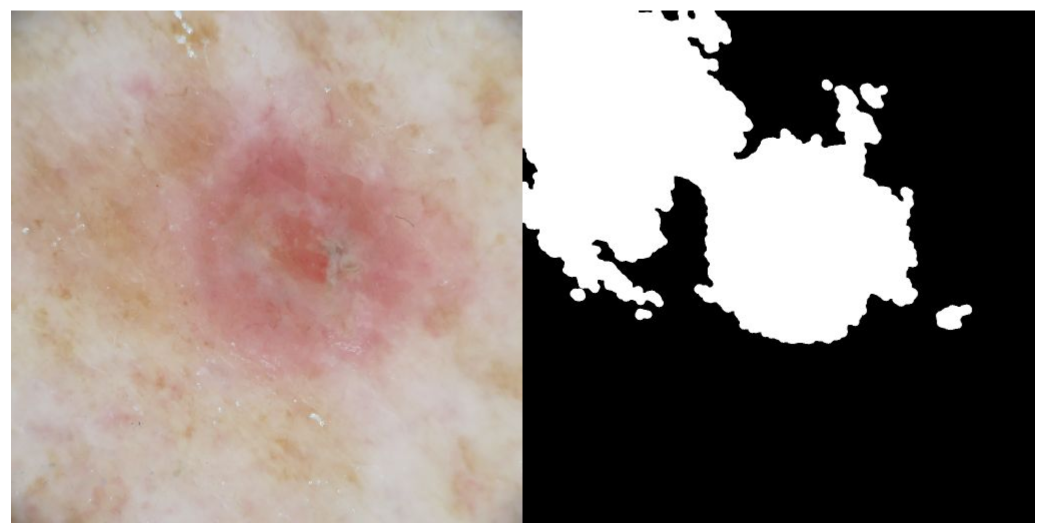


Segmentation

Segmentation algorithm fails for the three class problem

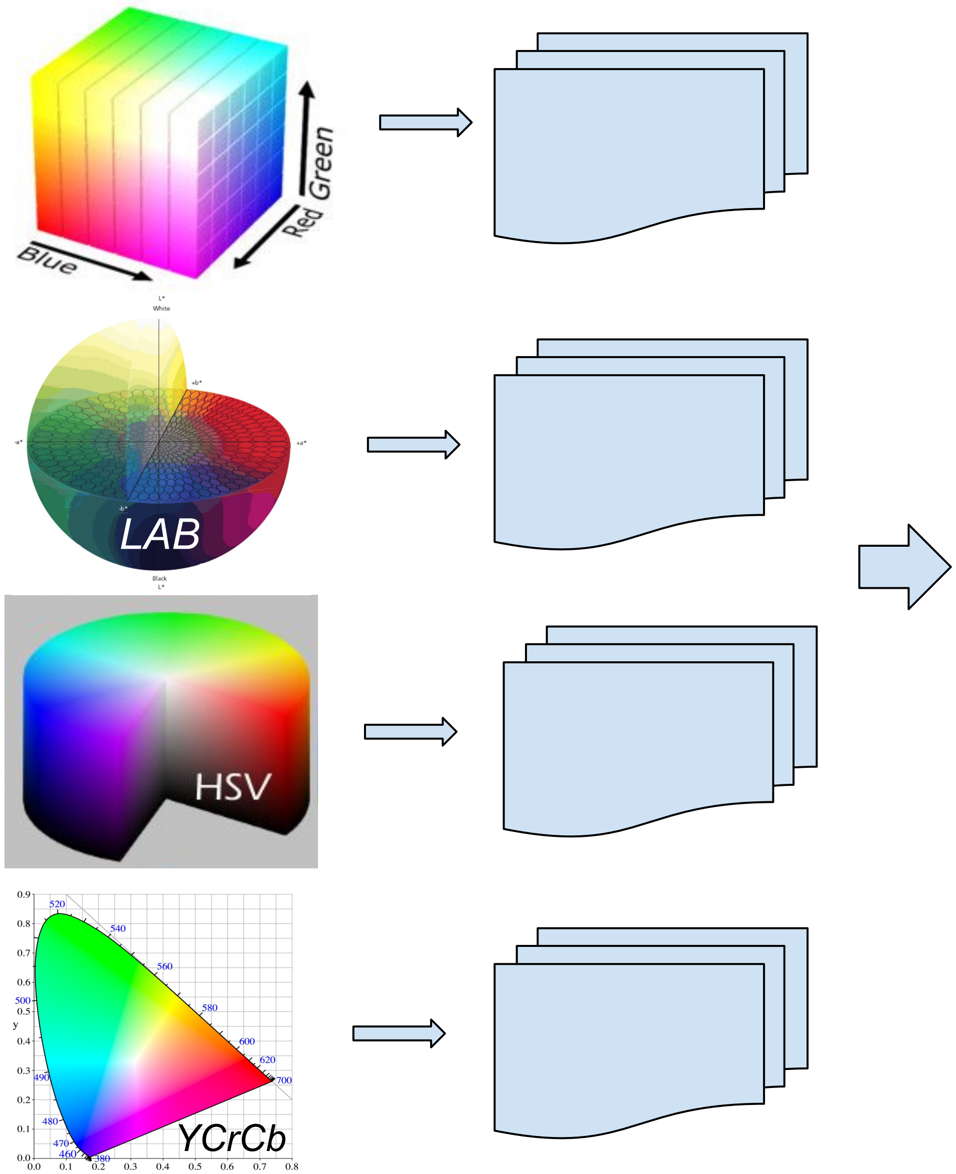






Universitat de Girona

Features Extraction: Color



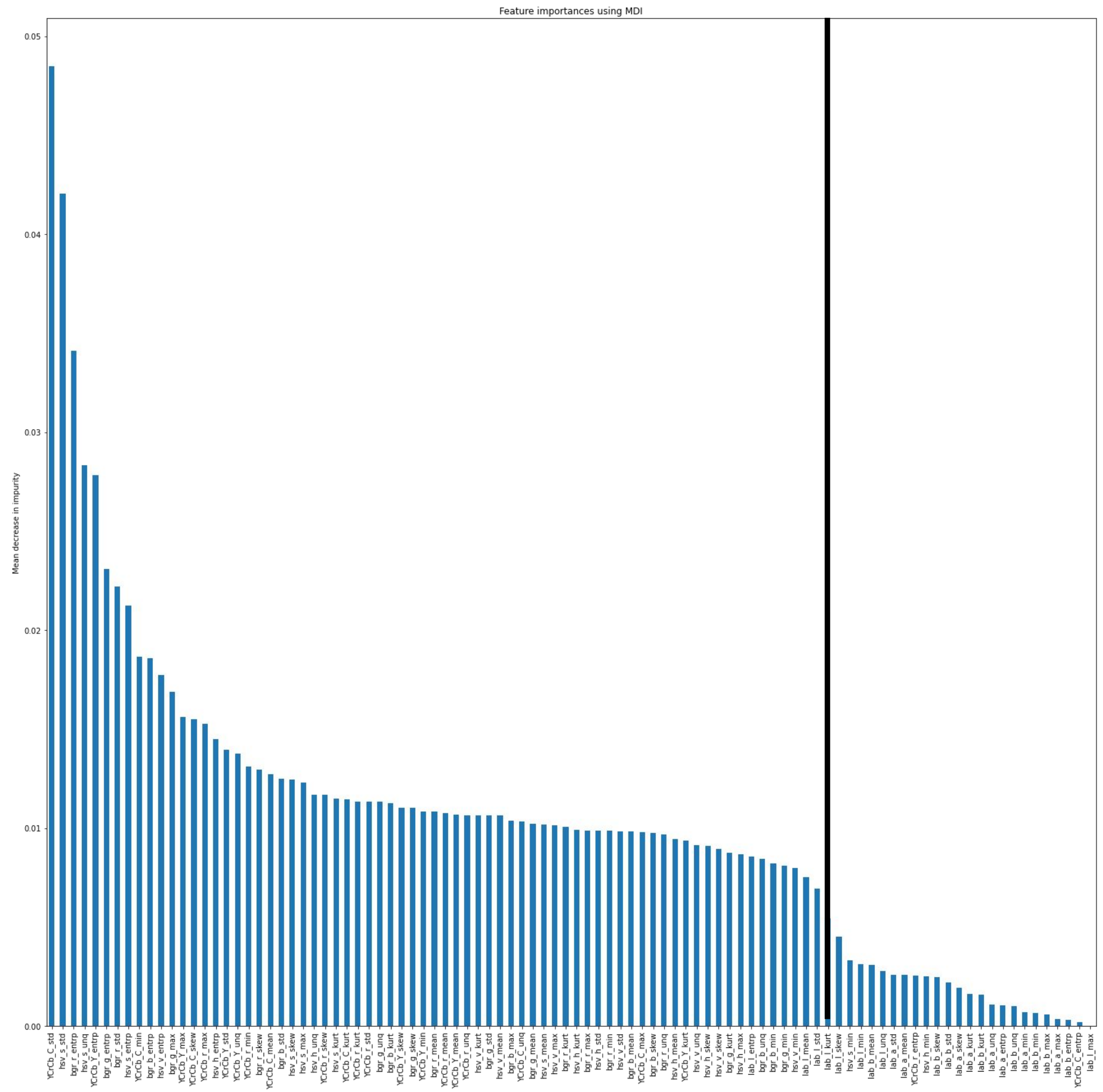
Channel-wise color features

1. Mean
2. Variance
3. Skewness
4. Kurtosis
5. Max
6. Min
7. Entropy
8. Number of unique values

Universitat de Girona

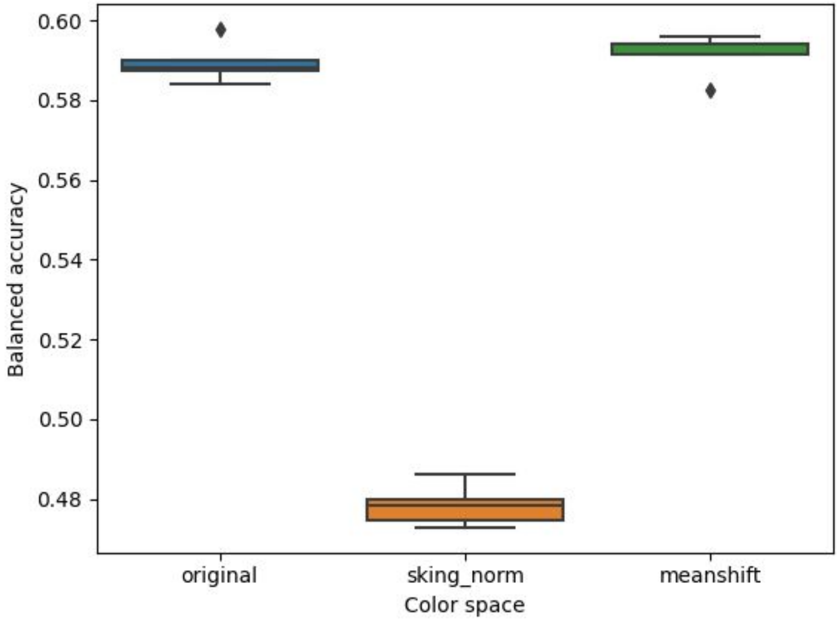
Features Extraction: Color

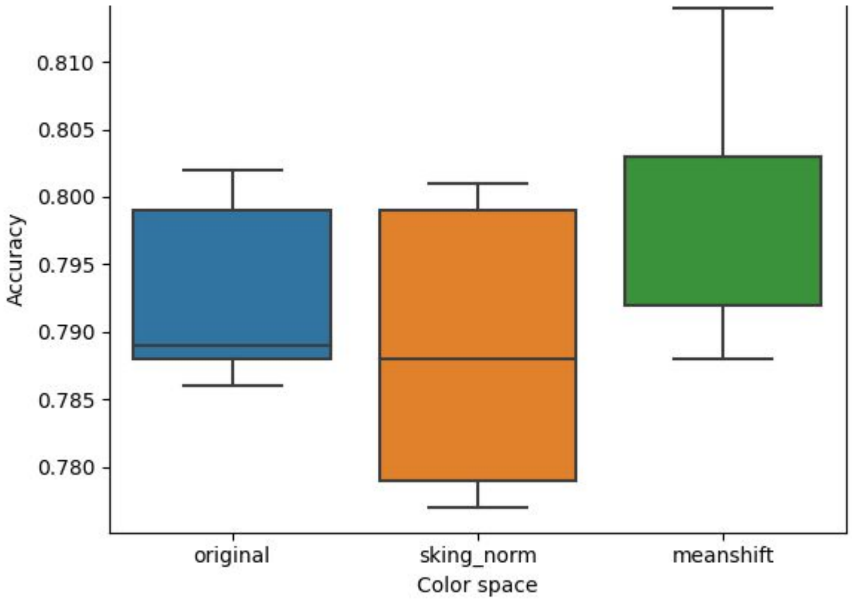
- Lab* colorspace features were the weakest (removal of these features led to the improvement of the weighted f1 from 0.7881 to 0.7974) and decreased number of features from 96 to 72



Universitat de Girona

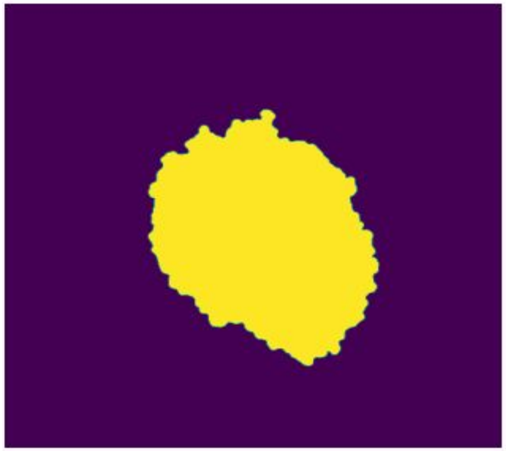
Preprocessing: Color Normalization



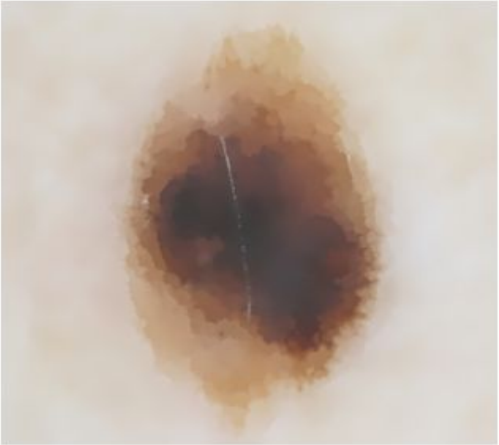


Original Image

Skin Color Normalization



Segmentation



Meanshift

Features Extraction: Texture

GLCM with:

- Distances [2, 5, 7, 10, 15]
- Angles [0, 45, 90, 135]

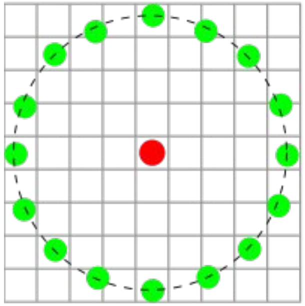
GLCM examples

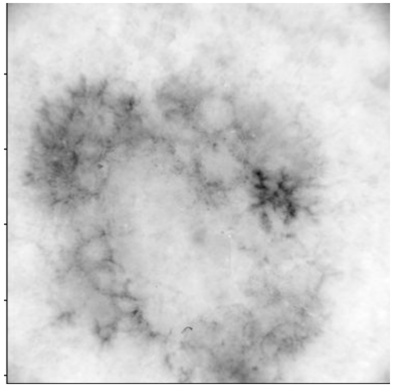
GLCM features

1. Contrast
2. Dissimilarity
3. Homogeneity
4. Energy
5. Correlation
6. Angular Second Moment (ASM)

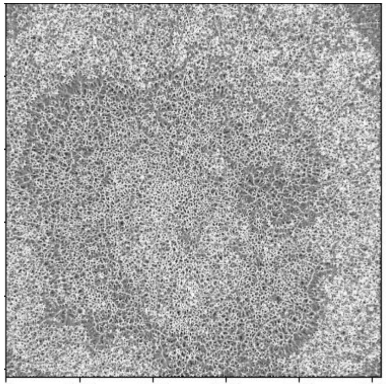
LBP histograms

9 different radius and
number of points combinations

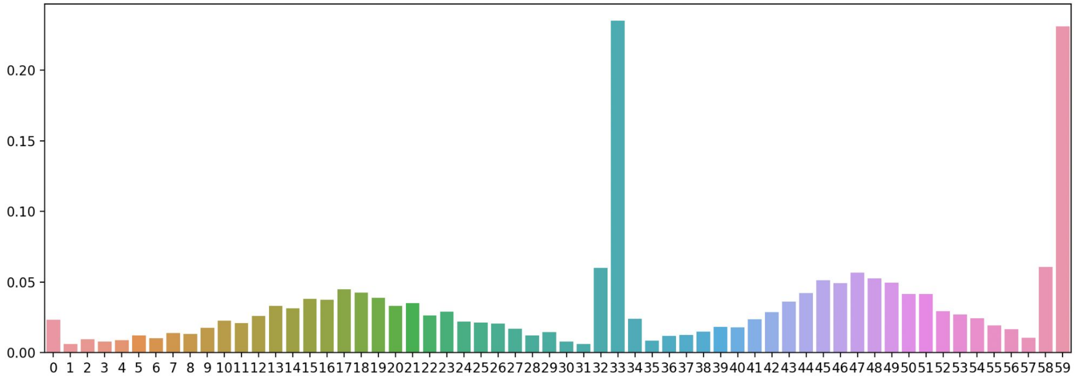




Gray scale image



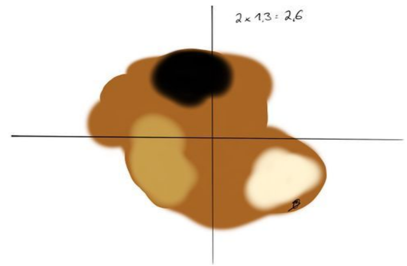
LBP image

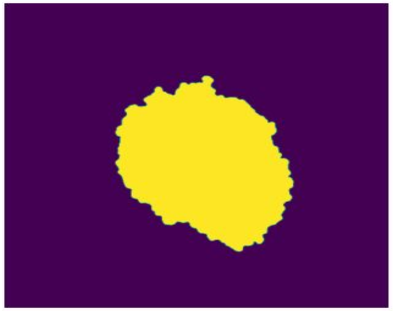


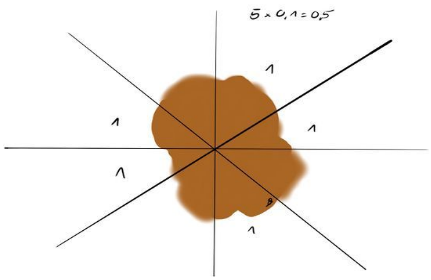
Feature Extraction: Shape

Asymmetry and border features

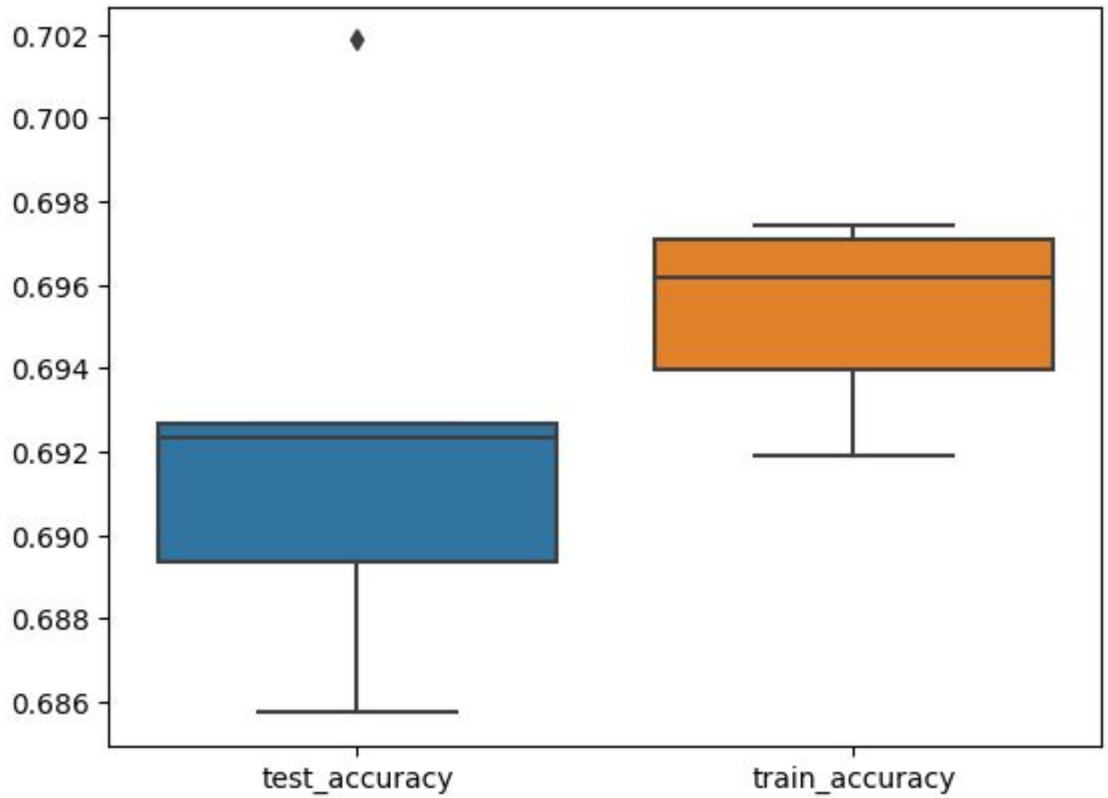
1. Number of lesions in the mask
2. Mean and std of their areas
3. Area
4. Perimeter
5. Circularity
6. Eccentricity
7. Aspect ration
8. Compactness index
9. 7 hu moments







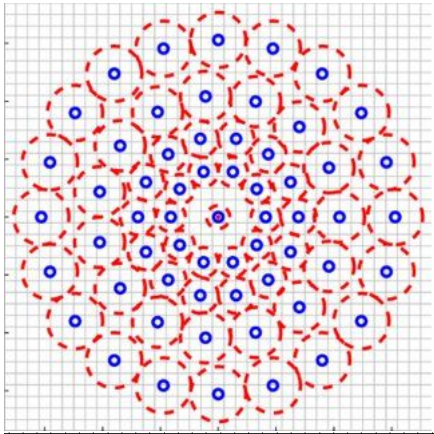
Five-fold CV on full train set of challenge 1 results on only shape features



Universitat de Girona

Feature Extraction: BoW

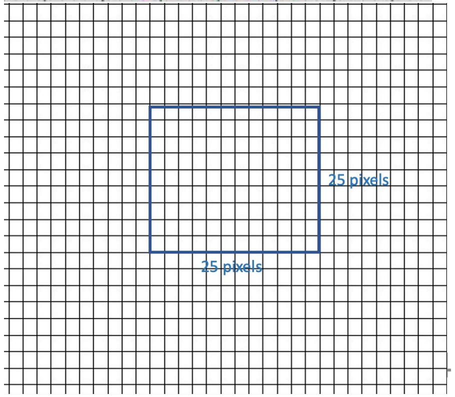
Sample Dense Keypoints → Calculate Descriptors → Cluster Descriptors with k=|vocabulary| → Histogram of clustered descriptors → Calculate TF-IDF



Descriptors experimented with:

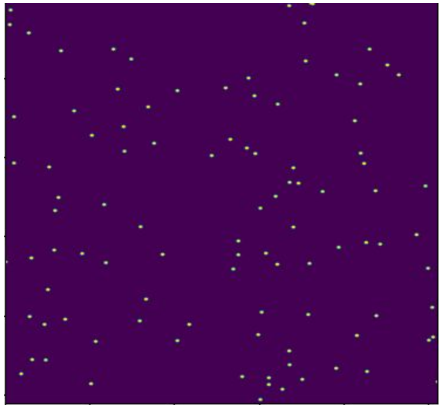
1. Brisk: constructs the feature descriptor of the local image through the gray scale relationship of random point pairs in the neighborhood

2. Color, GLCM, LBP: Calculate the features within patch size of 25 centred at the keypoint

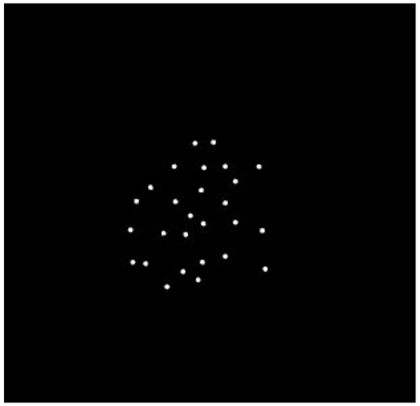


Feature Extraction: BoW

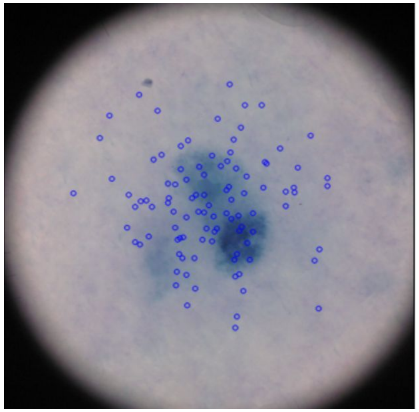
Keypoint Sampling Strategy	Accuracy for Texture Descriptors (challenge 2)	Accuracy for Color Descriptors (challenge 2)	Comments
1. Random within segmentation mask	0.5818	0.6323	Segmentation not good enough for challenge 2
2. Random within centered radius as mask (radius 100)	0.5717	0.6606	Better for color features
3. Gaussian sampled at the centre of the image	0.5959	0.62424	Better for texture features



1



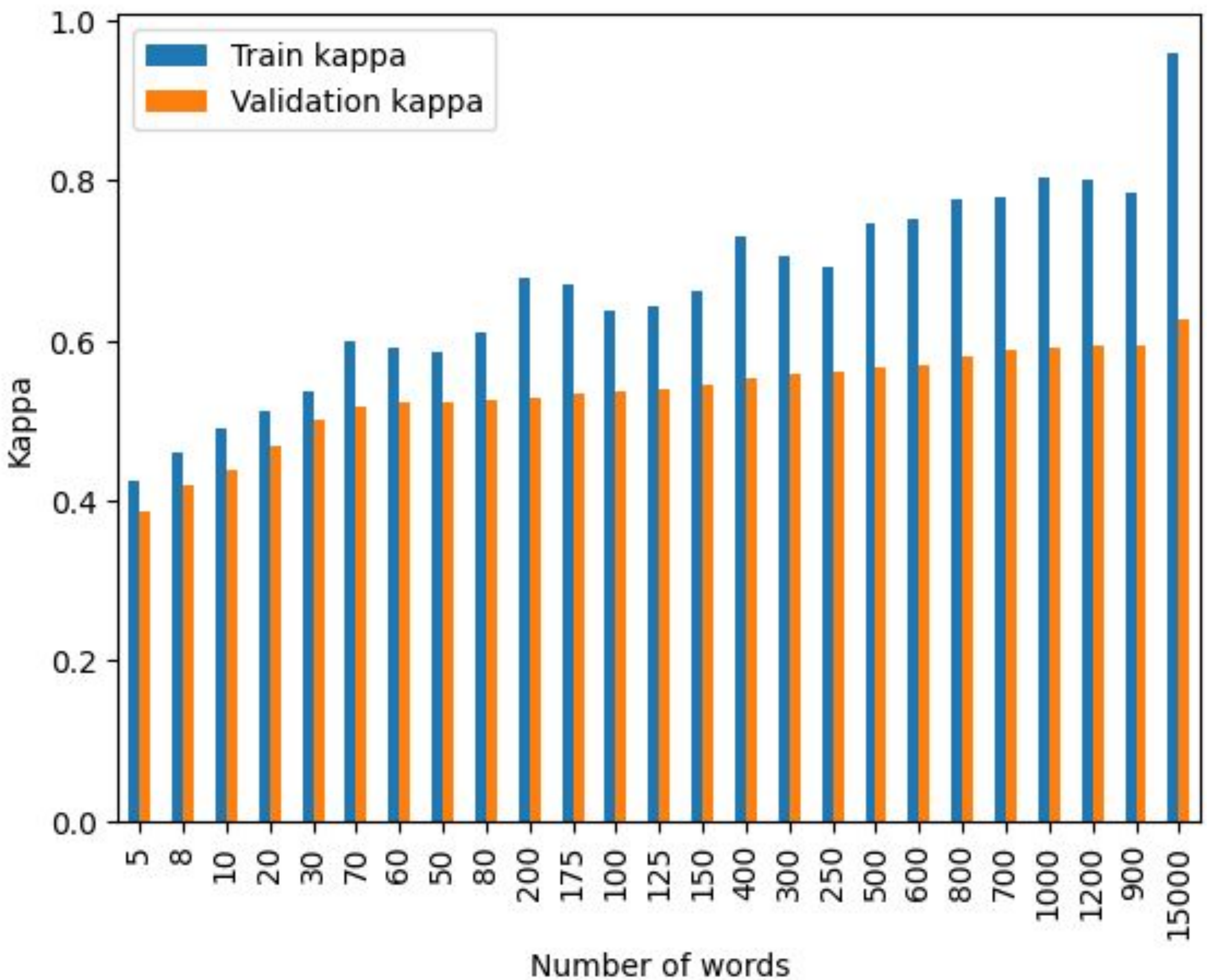
2



3

Feature Extraction: BoW

Vocabulary size experiment: 100 words are enough



Universitat de Girona

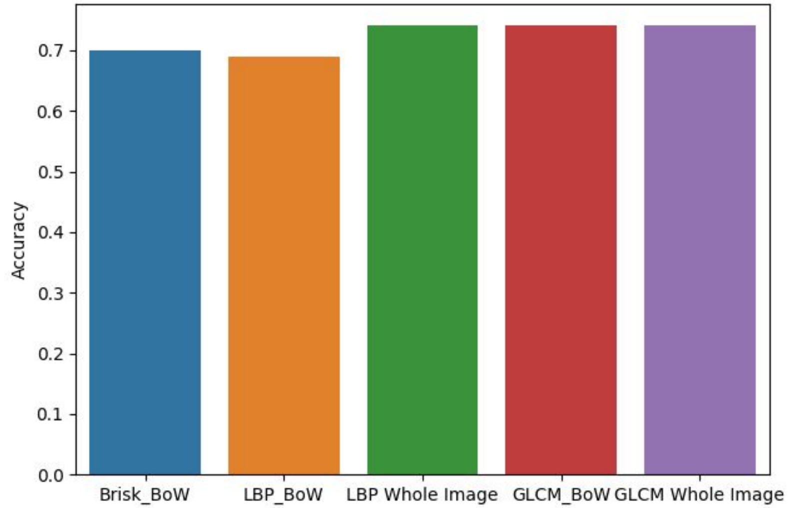
Feature Extraction: BoW

Binary Problem

- BoW not better than whole image features

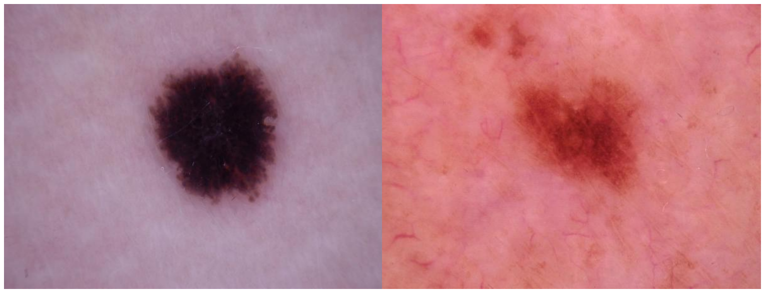
3 Class Problem

- BoW Improved the validation accuracy by ~0.5

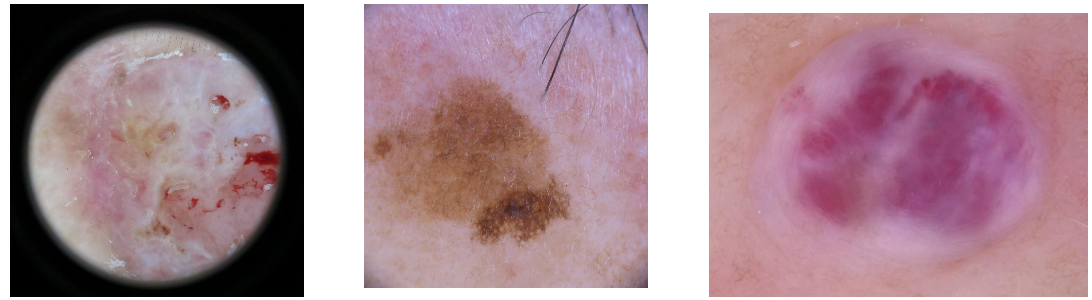


Universitat de Girona

Challenge 1: Overview and Features



Nevus Images



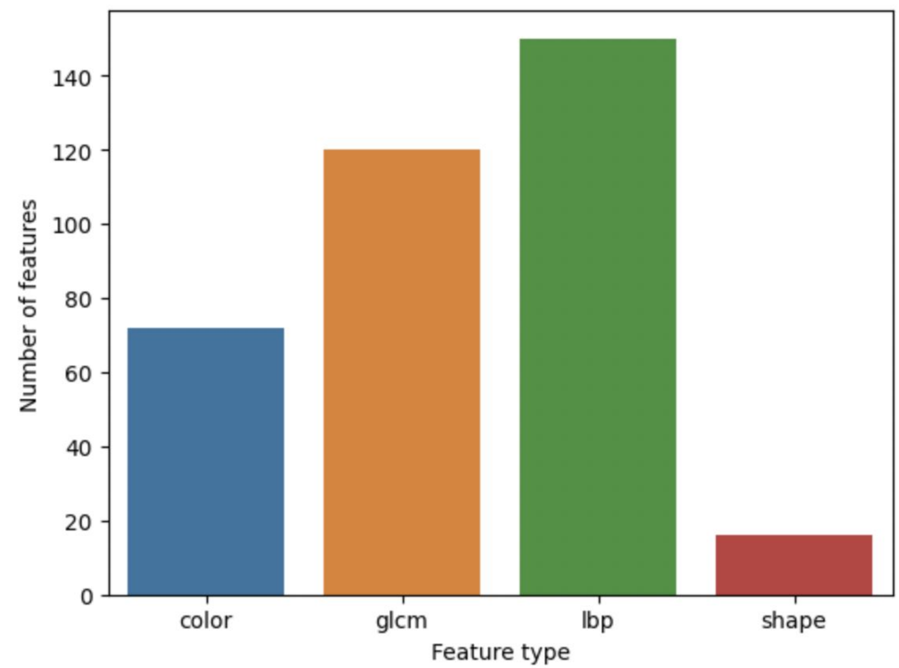
Others Images

Overview: Binary classification problem; balanced huge dataset

Total: 358 features of color, texture and shape

No BoW features didn't bring significant improvement

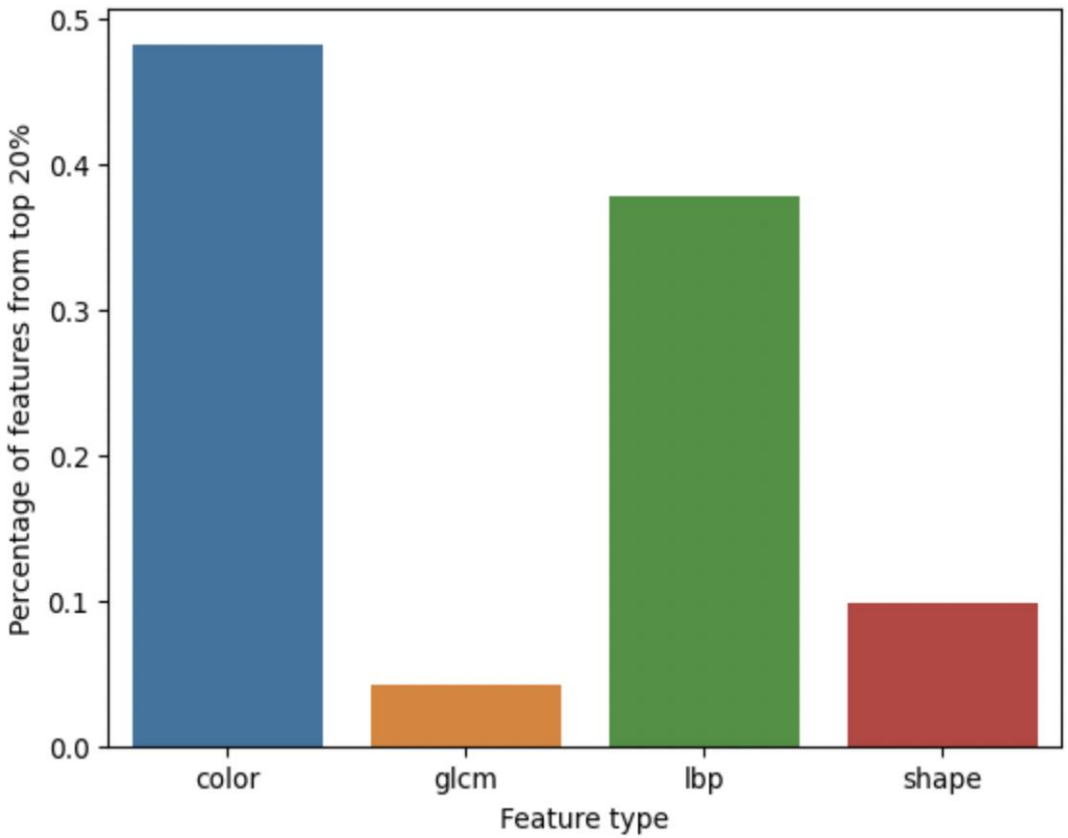
Explored: reducing feature size to tackle the curse of dimensionality

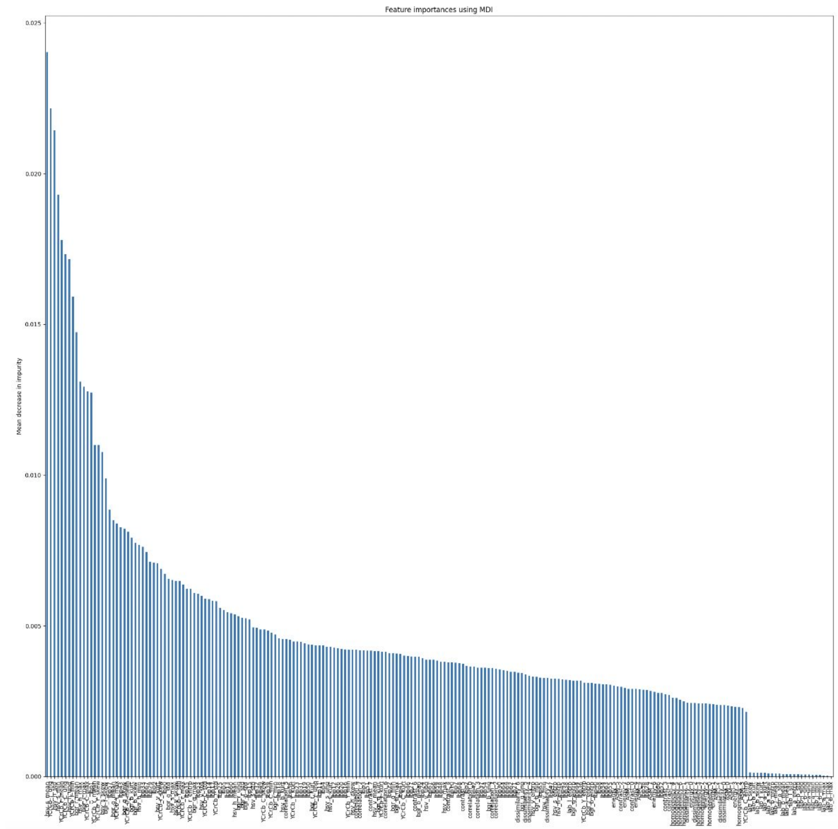


Universitat de Girona

Challenge 1: RF Feature Selection

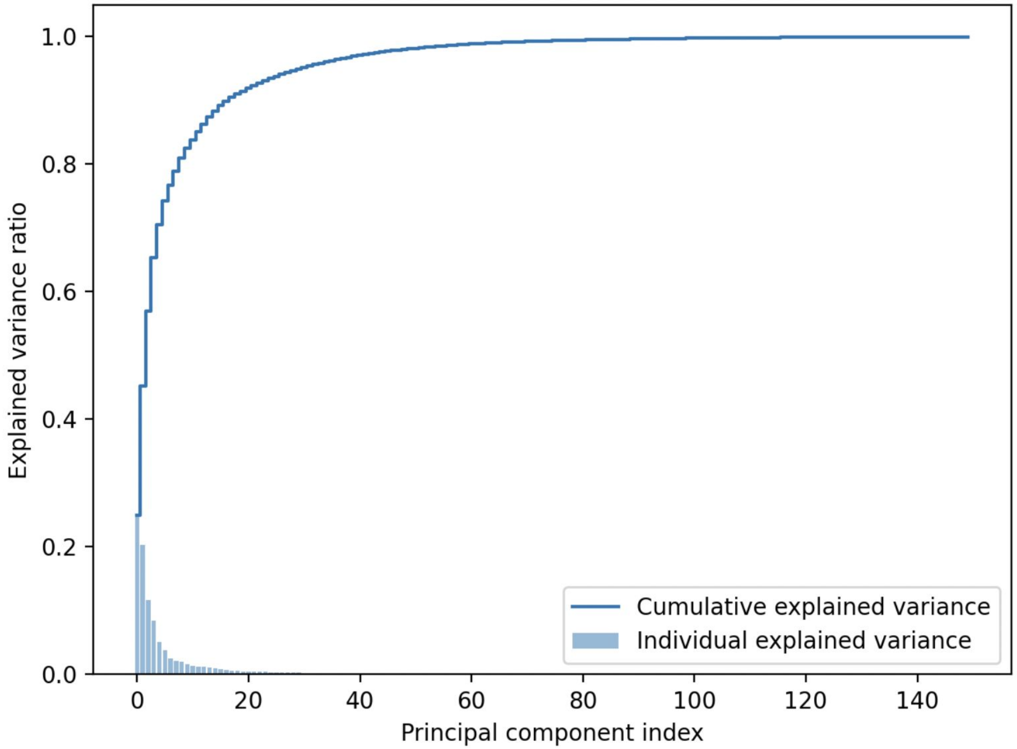
- Lab color space features were removed.
- Selecting k-best features didn't improve the validation accuracy

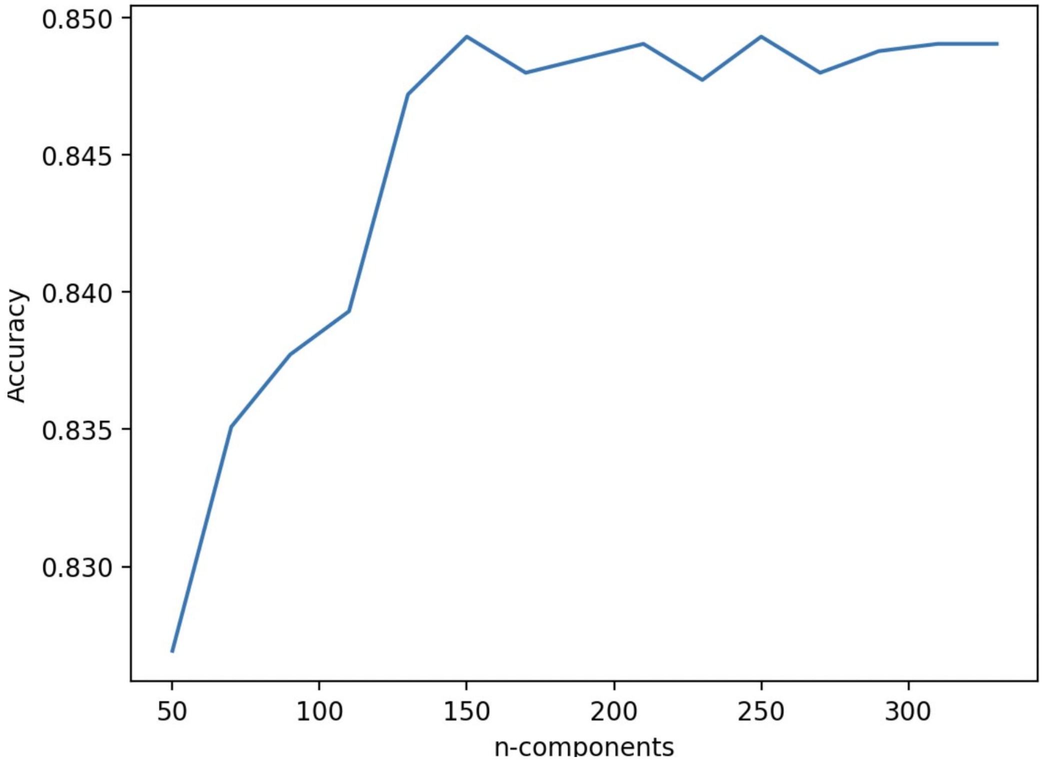




Challenge 1: PCA Dimensionality Reduction

150 principal components chosen as final set of features from 358 features



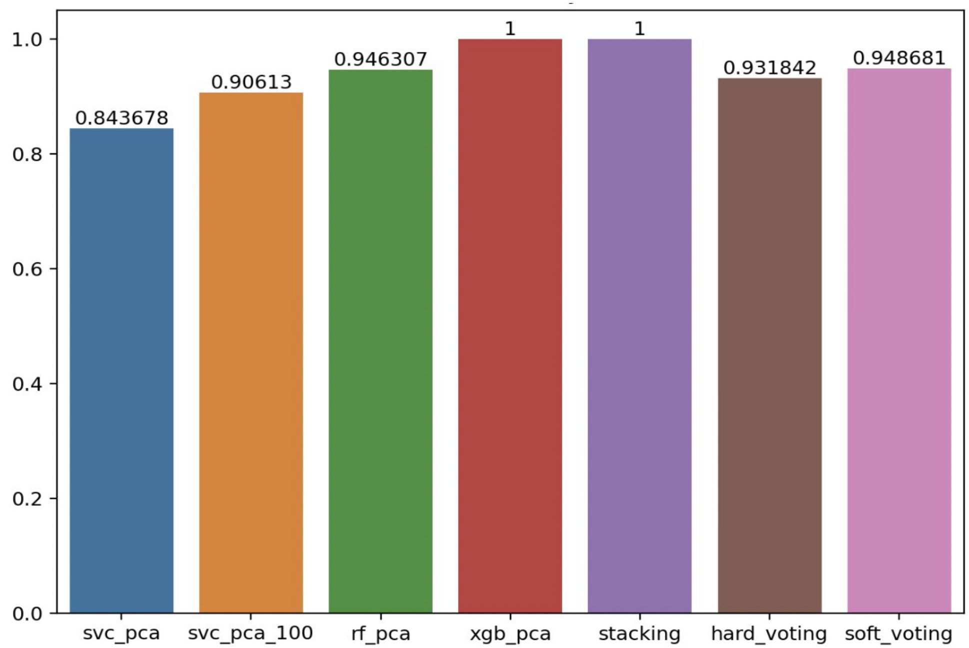


Universitat de Girona

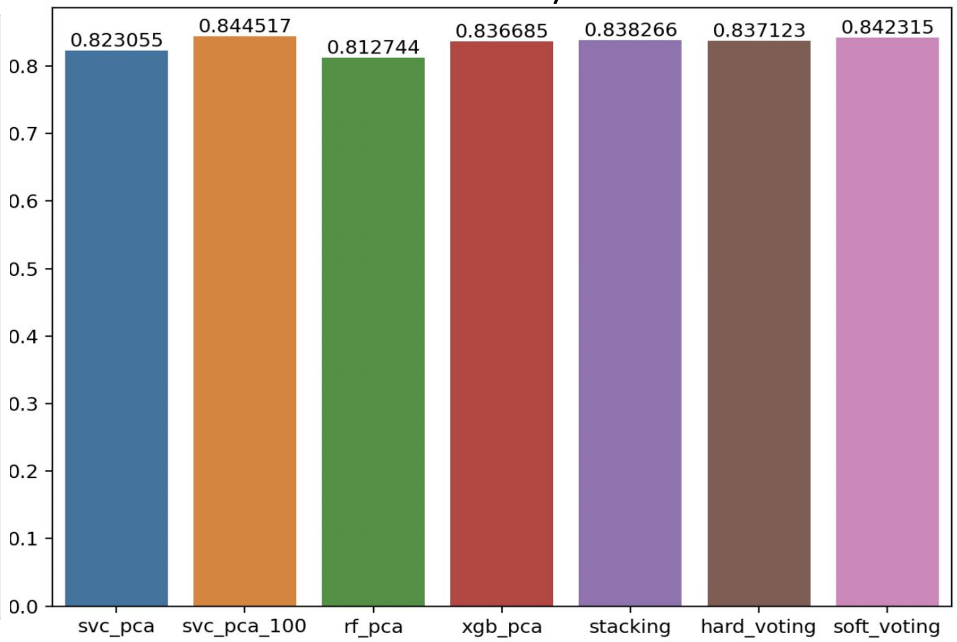
Challenge 1

- All extracted features reduced with PCA (150 components) were used
- Soft-voted Ensemble of tuned SVM and XGBoost classifiers was used as the best trade-off between training (less overfitting) and validation(generalization) accuracy.

Training Accuracy



Validation Accuracy



Universitat de Girona

Challenge 2: Overview and Features

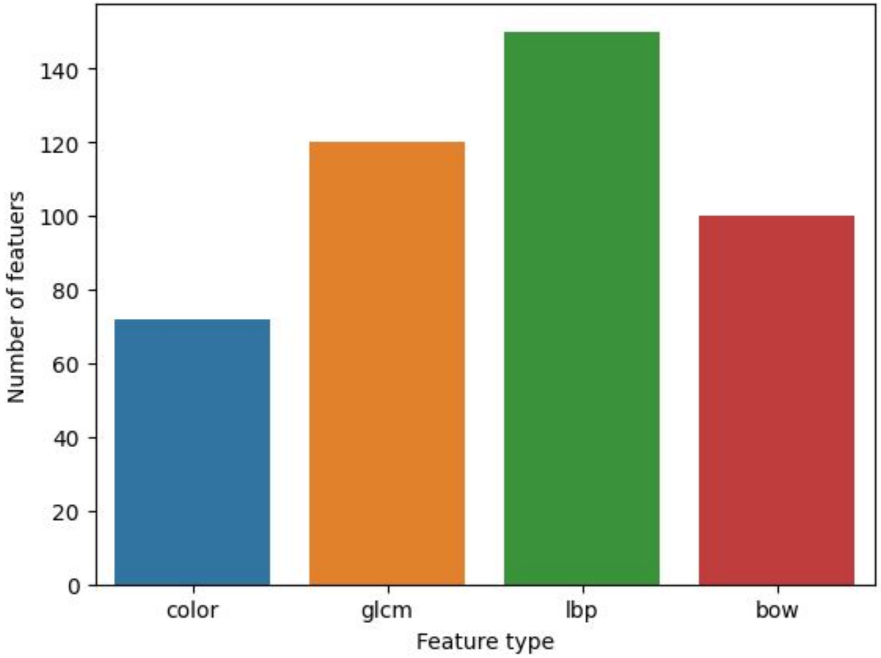
Basal Cell	Squamous Cell	Melanoma

Challenges: multi class, less data, highly imbalance data set.

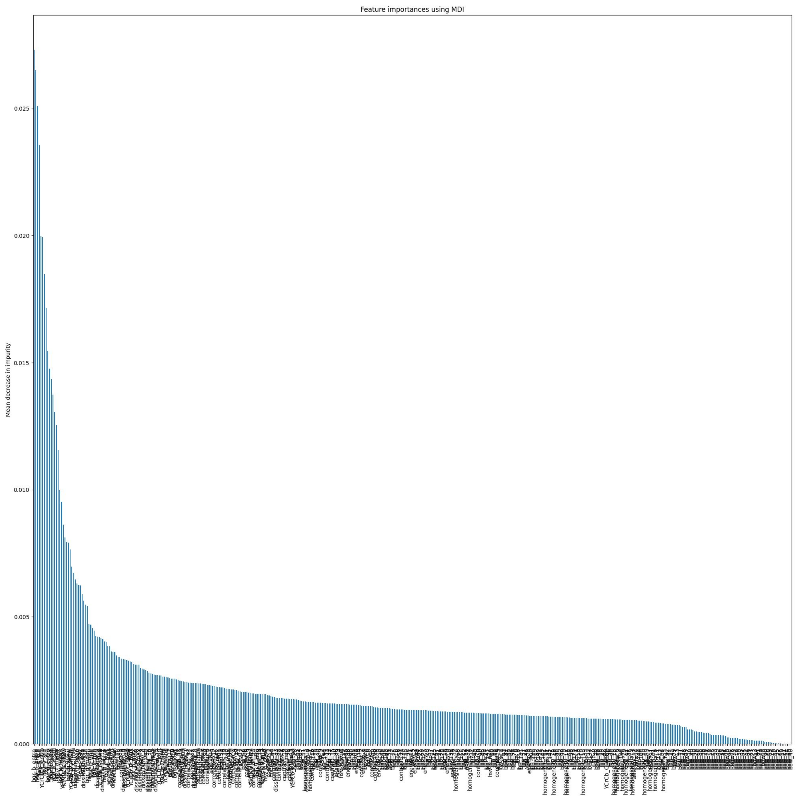
Total: 442 features of color (both global and BoW tf-idf) and texture

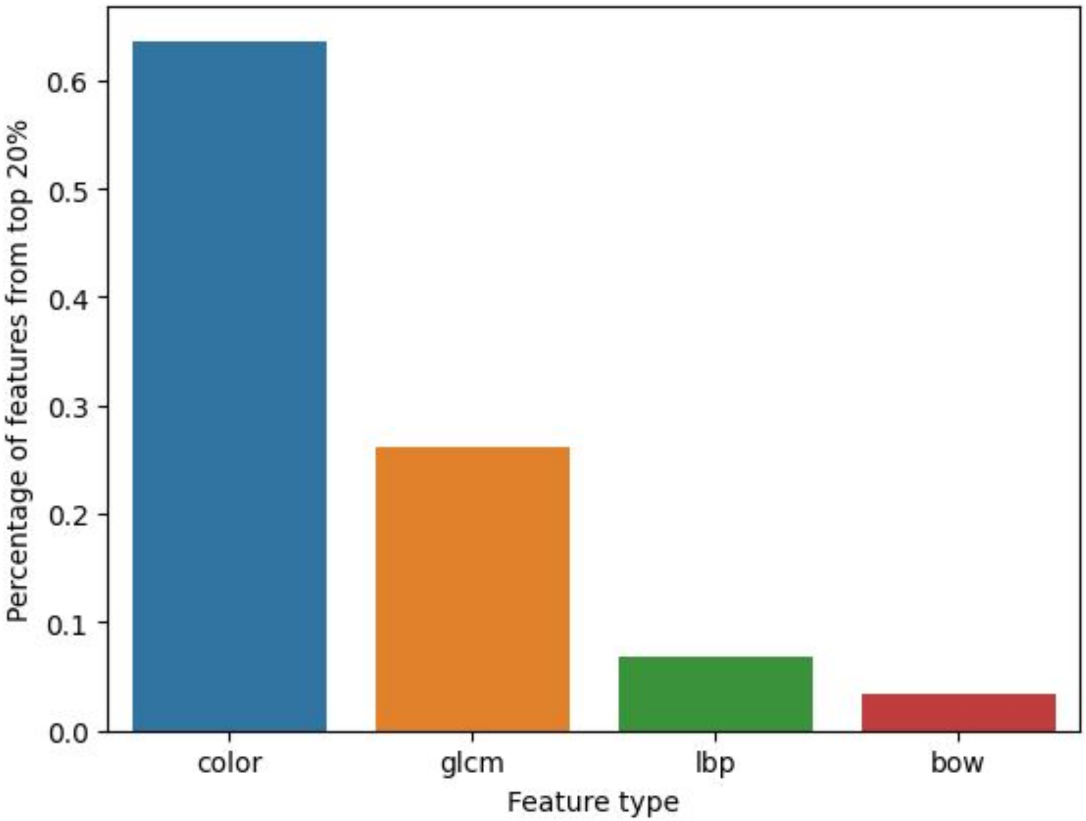
No shape features since the segmentation results were poor

Explored: reducing feature size to tackle the curse of dimensionality and techniques to sive the imbalance



Challenge 2: RF Feature Selection





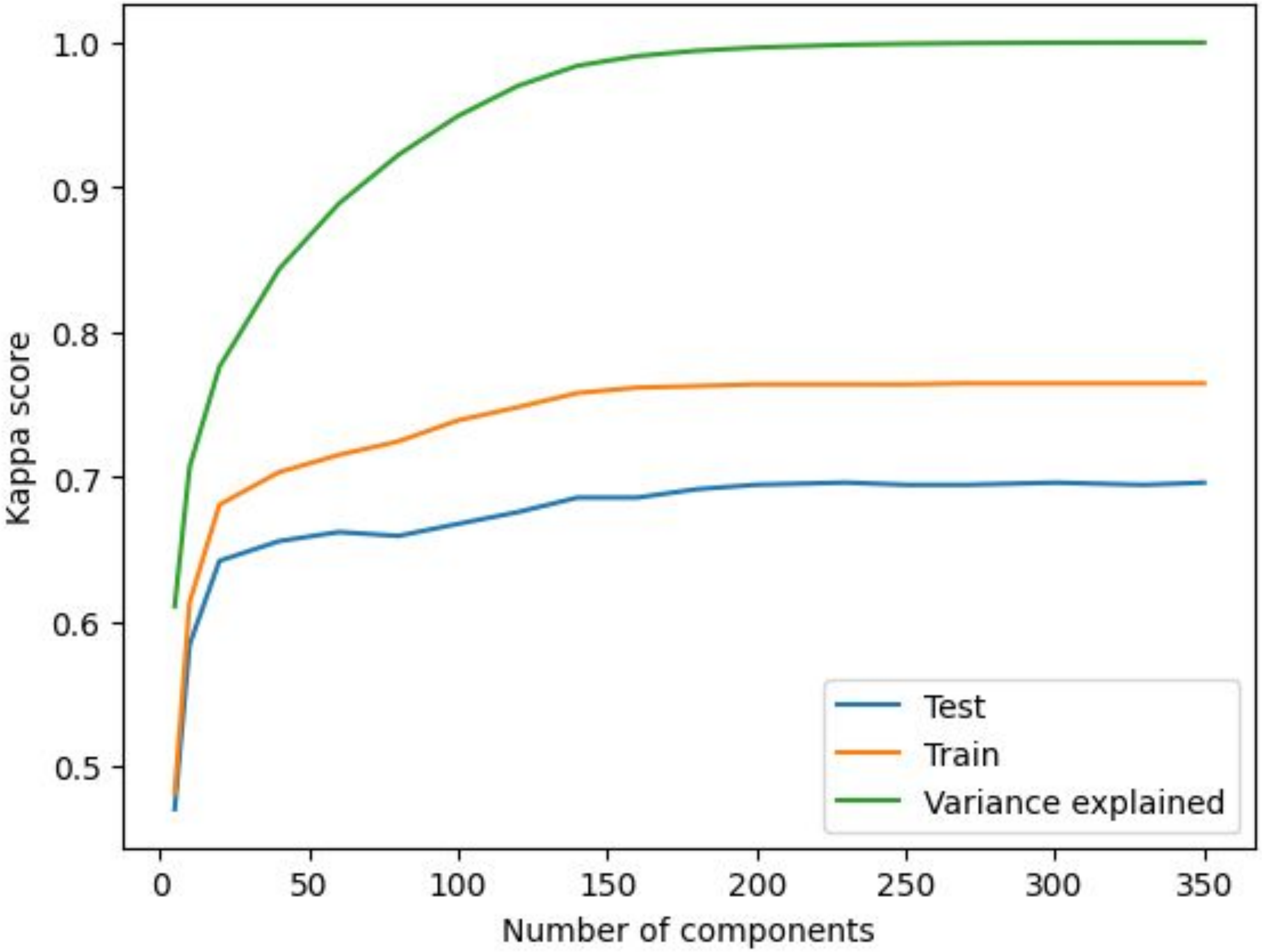
Analysis of feature importances of top 20% RandomForest features shows that global color and glcm texture features were the most prominent ones with BoW and LBP features still having an important contribution.

A further investigation of a features set composed of these 88 top 20% features was done (referred to as rf_fs) .

Challenge 2: PCA Dimensionality reduction

Validation set kappa score for the different number of components in PCA decomposition of all 442 features.

A further investigation of a features set composed of these 150 PCA features was done (referred to as pca) .



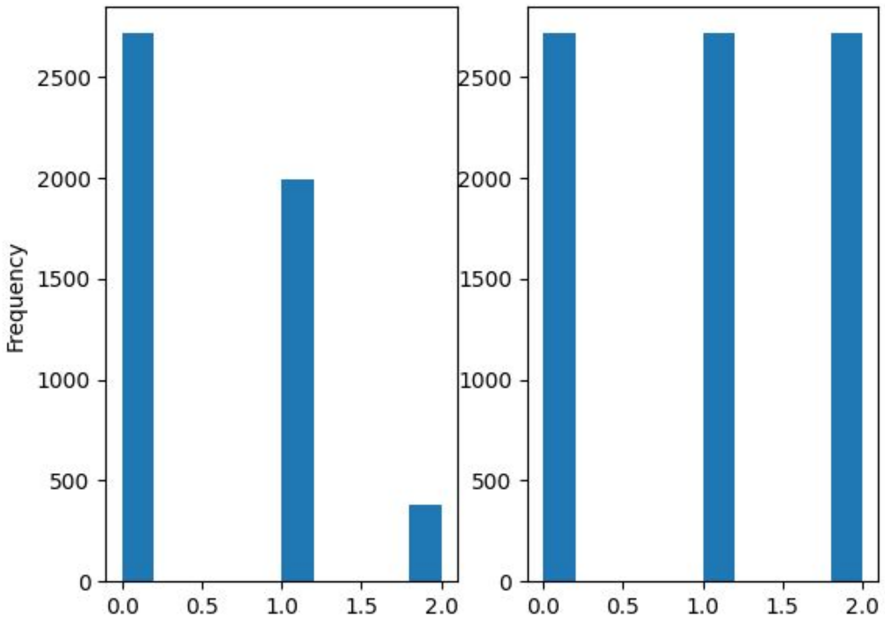
Universitat de Girona

Challenge 2: Imbalance Problem

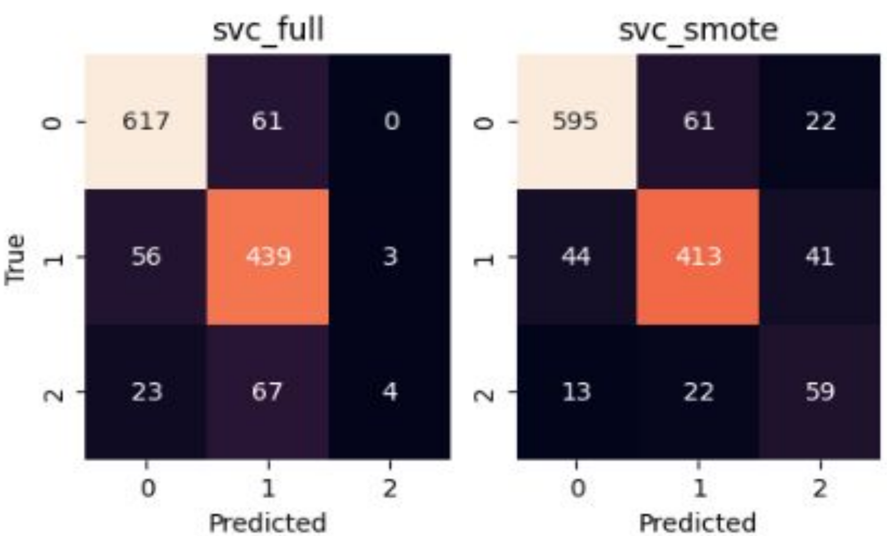
Data-wise we have tried:

- oversampling
- undersampling
- Synthetic Minority Oversampling Technique (the only one to show any improvement)

Model-wise: use of <u>balanced class weights</u> in all of the classifiers we were testing

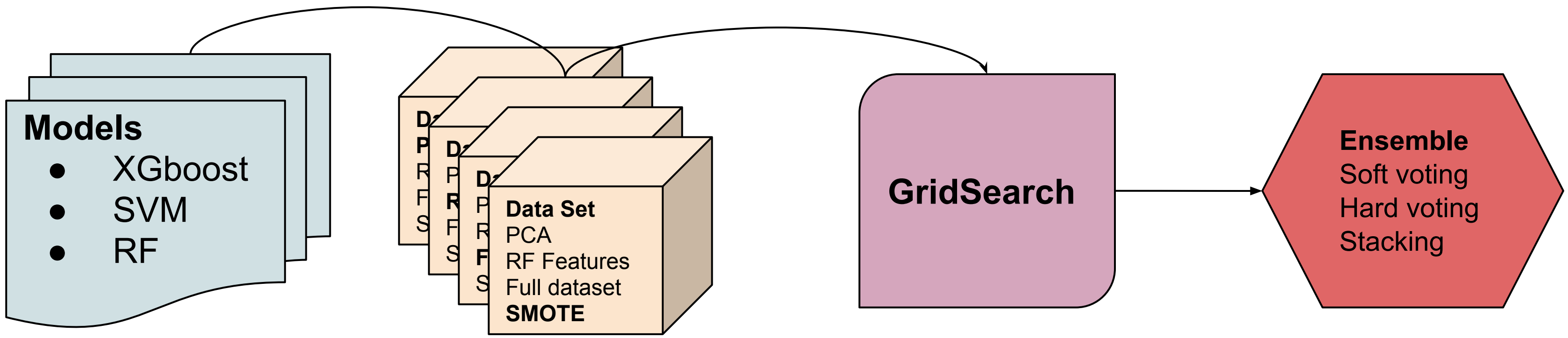


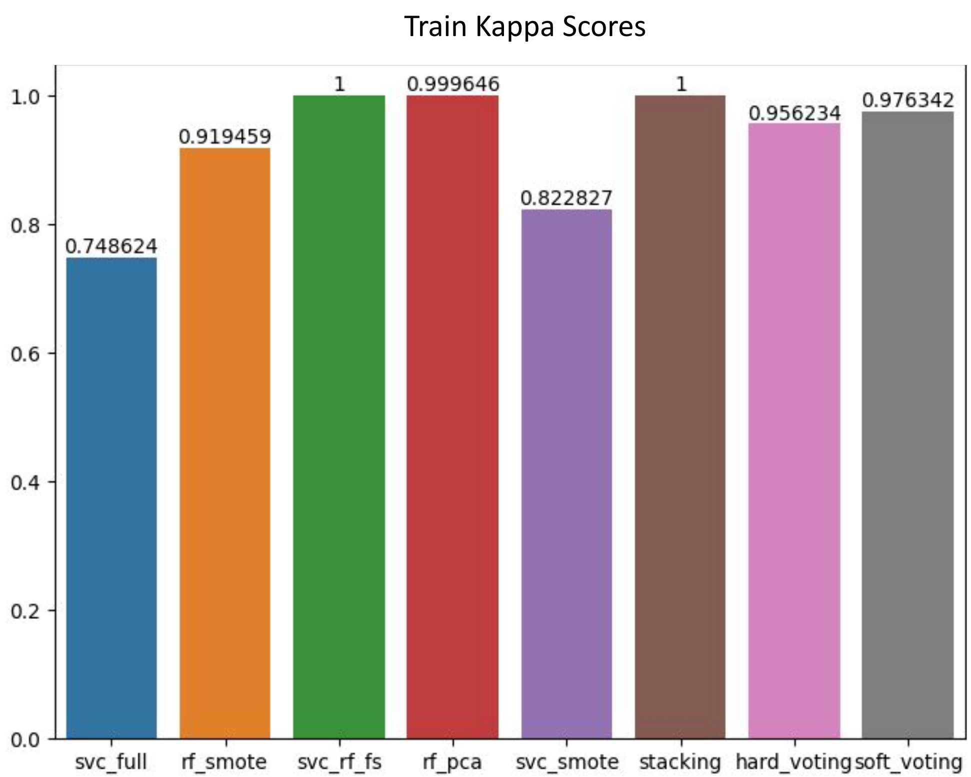
Train set class distributions before and after SMOTE

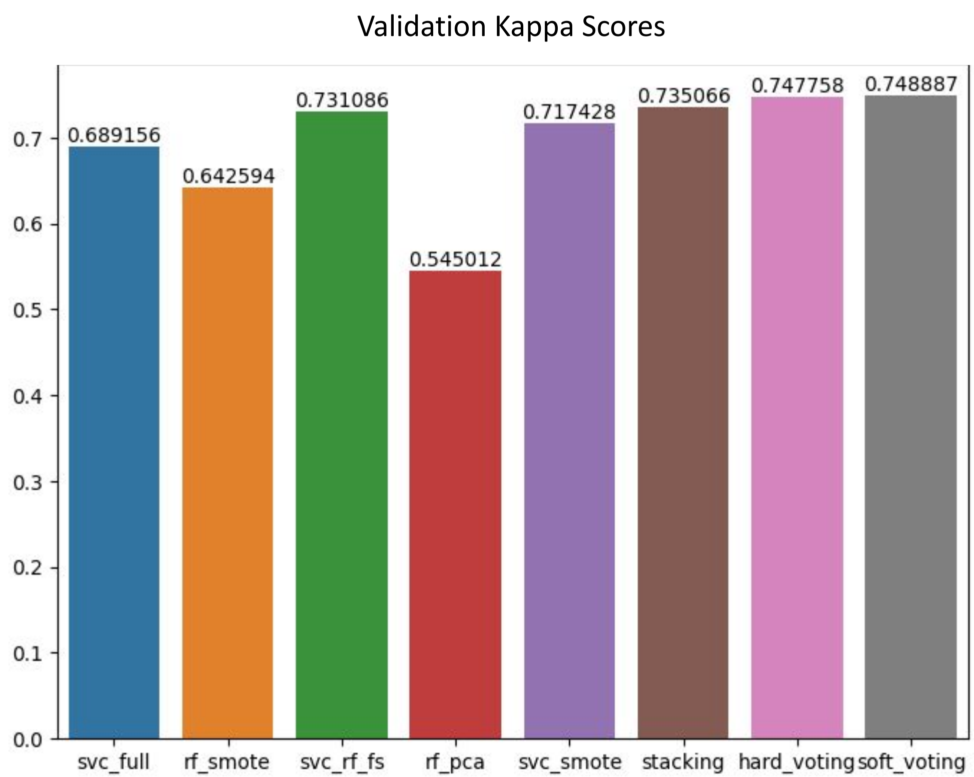


Validation set Confusion Matrices

Challenge 2: Ensembling





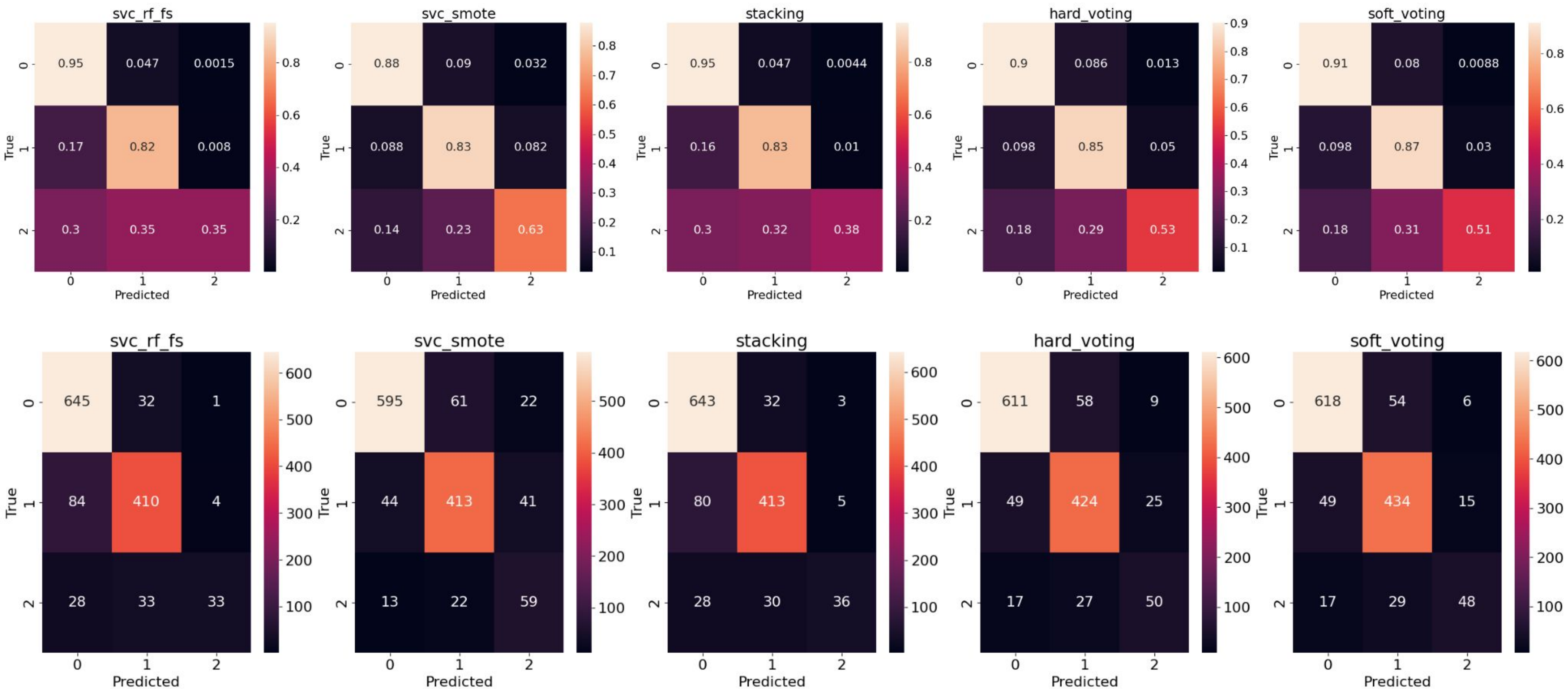


Universitat de Girona

Challenge 2: Final model

Taking a closer look at the confusion matrices of the top 5 model/dataset combinations from the previous slide we can notice that **SVM on SMOTE data achieved the best results**. It has the smallest overfitting while maintaining the best proportion between 3 classes.

Therefore, we have selected it as our final model for the challenge 2 with the validation **kappa of 71.**

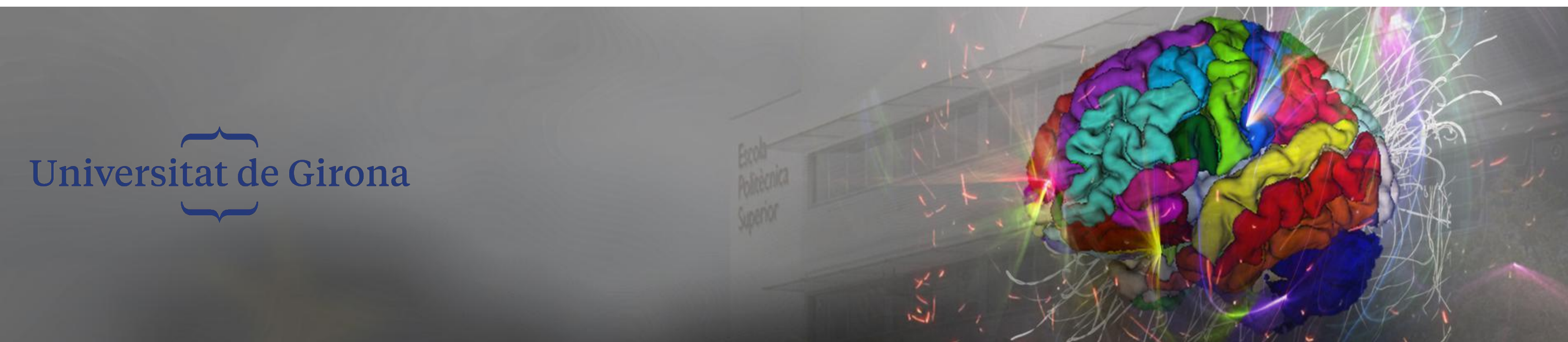


Conclusions

- Color features are the most discriminative for both problems
- Segmentation of lesions can lead to better results but is quite challenging, especially for malignant lesions
- BoW was able to improve performance on 3 class problem, due to increased importance of the small variations in color information between lesion types
  - However for 2 class problems global color features were more effective
- Adding additional features (like texture, shape or BoW) improved the results however also led to increased overfitting
- Data imbalance for 3 class problem was better solved with balance weights SVM on SMOTE data, however the minority class still was considerably underdetected
  - needs more distinctive features

CAD: Skin Lesion Classification Going Deeper

Manasi Kattel
Vladyslav Zalevskyi

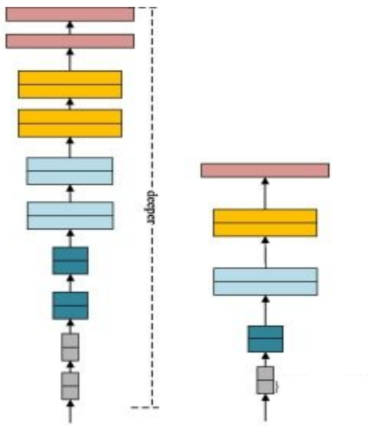


Content

1. Literature review
   a. Current SoTA pipelines
2. Models explored
3. Image preprocessing and data augmentation pipelines
4. Challenge 1
   a. Results and experiments
5. Challenge 2
   a. Experiments: loss functions
   b. Results
6. Ensembling
7. “Pretext learning”
8. Conclusions

Universitat de Girona

Literature Review

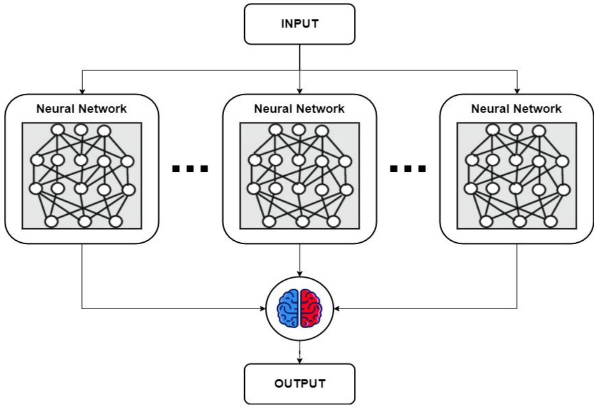


Size matters

ISIC 2019 winning olution [1]:

ensemble of Multi-Res EfficientNets + SEN154 2

SIIM-ISIC Melanoma Classification winning solution [2]: ensembles of EfficientNet B3-B7, se_resnext101, resnest101



Ensembling matters

Instead of following monstrous ensembles and models we focused on:

- Single model architectures of different styles (convolutional and transformer)
- Tuning the models and the data
- Focus on losses, augmentations and ensembling
- Pretext learning

Literature Review

BACC of Different DCNN Models on The ISIC 2018 Skin Lesion Classification Challenge Test Set.

model	BACC	model	BACC	model	BACC
VGG-11	0.769	**DenseNet-169**	0.836	**RegNetX-3.2G**	0.842
VGG-13	0.771	**DenseNet-201**	0.829	**RegNetX-4.0G**	0.834
VGG-16	0.745	**DenseNet-161**	0.837	**RegNetX-8.0G**	0.831
VGG-19	0.750	**EfficientNet-b0**	0.838	**RegNetX-16G**	0.835
ResNet-18	0.812	**EfficientNet-b1**	0.842	**RegNetX-32G**	0.832
ResNet-34	0.825	**EfficientNet-b2**	0.853	**RegNetY-400M**	0.839
ResNet-50	0.834	**EfficientNet-b3**	0.845	**RegNetY-800M**	0.846
ResNet-101	0.838	**EfficientNet-b4**	0.842	**RegNetY-1.6G**	0.850
ResNet-152	0.835	**EfficientNet-b5**	0.843	**RegNetY-3.2G**	**0.858**
SENet-50	0.832	**EfficientNet-b6**	0.848	**RegNetY-4.0G**	0.848
SENet-101	0.845	**EfficientNet-b7**	0.847	**RegNetY-8.0G**	0.846
SENet-152	0.835	**RegNetX-400M**	0.823	**RegNetY-16G**	0.849
SENet-154	0.838	**RegNetX-800M**	0.828	**RegNetY-32G**	0.851
DenseNet-121	0.832	**RegNetX-1.6G**	0.833		

dataset	**7-PT**	**ISIC 2017**	**ISIC 2019**
Best model	RegNet Y-800M	RegNet Y-1.6G	RegNetY -8.0G
Balanced accuracy	0.652	0.743	0.59

For the transformers we chose Swin architecture

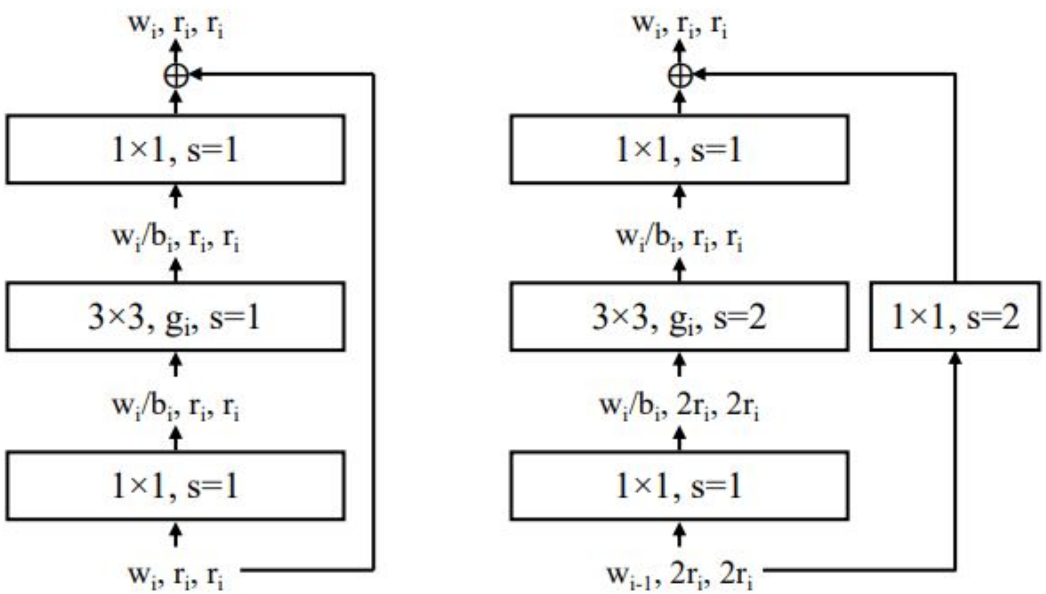
- still one of the best performing single-model architectures on ImageNet
- not very extensive research into transformers and skin lesion cad (not like for convnets)
- easily available with PyTorch

[3]

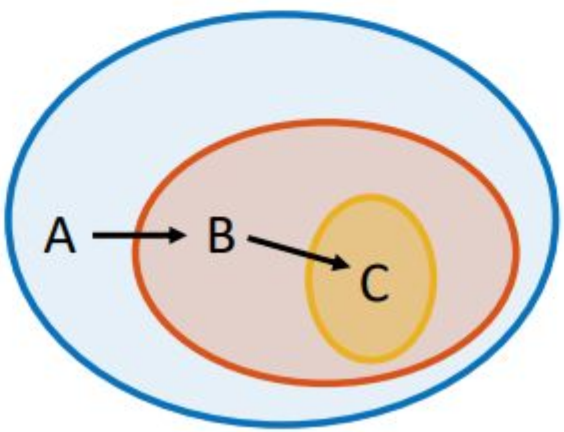
RegNetY

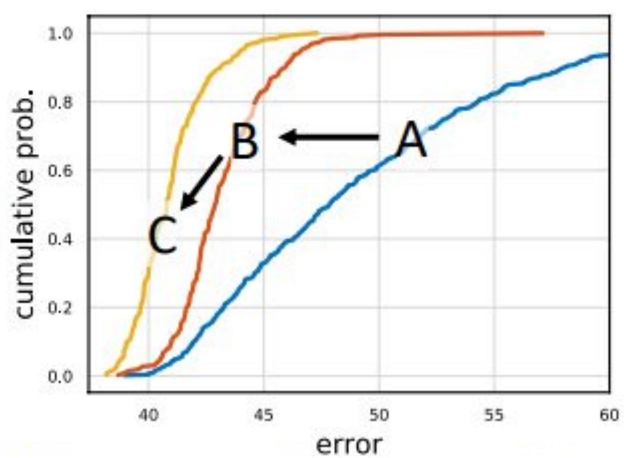
RegNet is a network design space made up of

- Model architectures
- **Different parameters that define a space of possible model architectures**
- Parameters can be the width, depth, groups, etc. of the network.



(a) X block, s=1 **(b)** X block, s=2



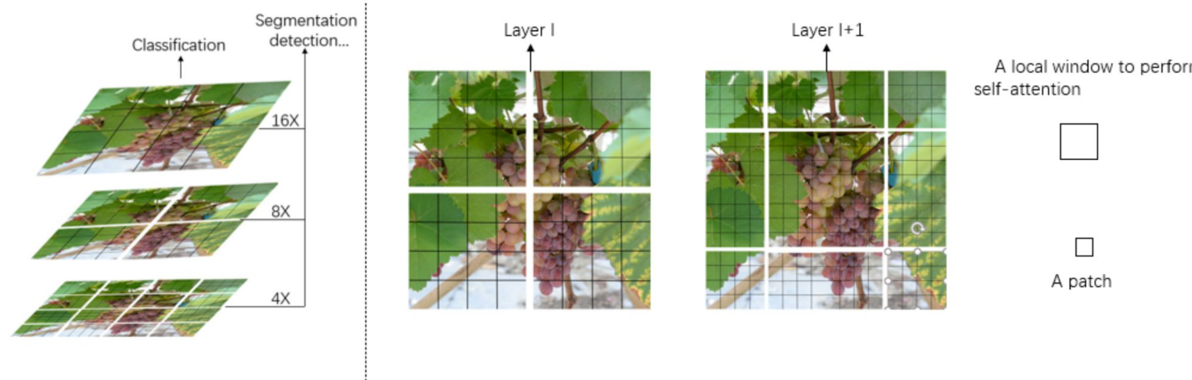


[4]

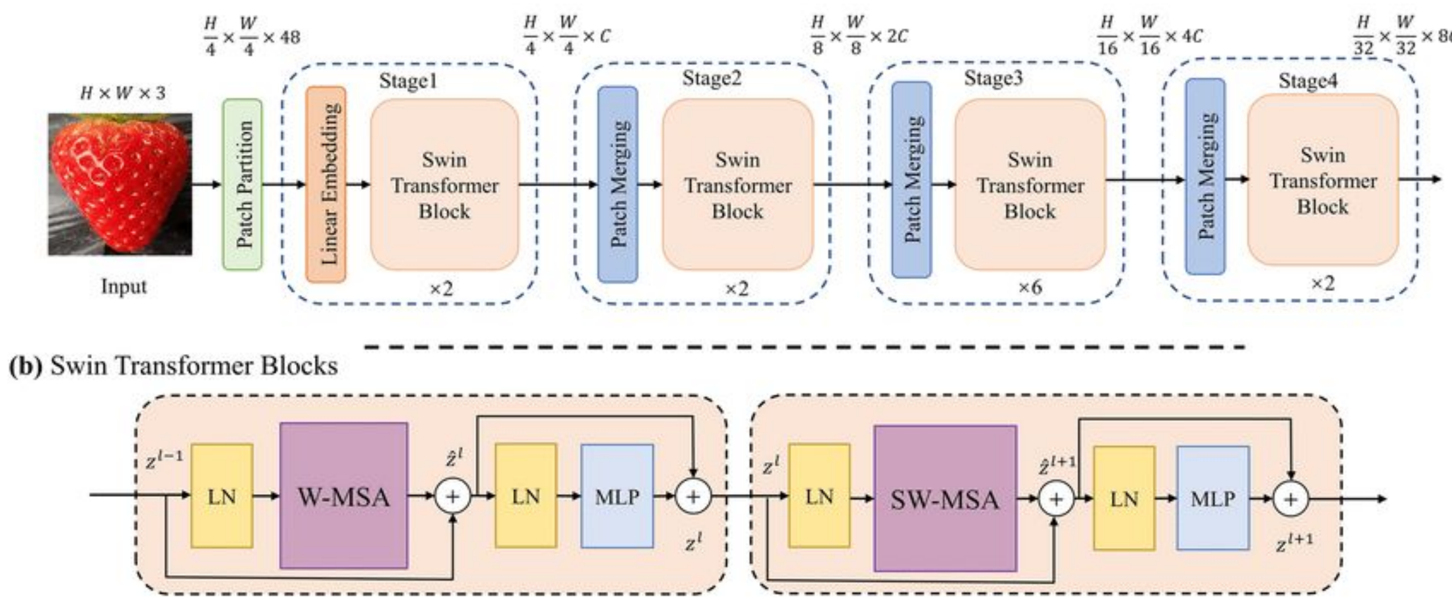
Swin Transformer

State-of-the-art performance in vision tasks; two key concepts

1. **hierarchical feature maps:** allows fine-grained prediction
2. **shifted window attention:** improves complexity



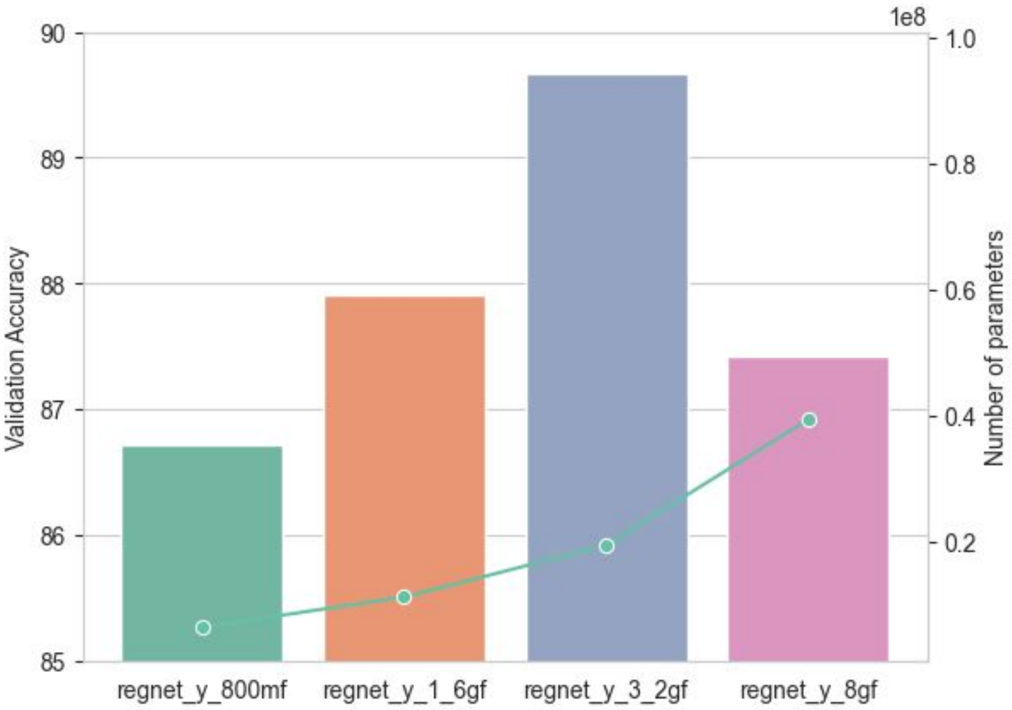
(a) (b)

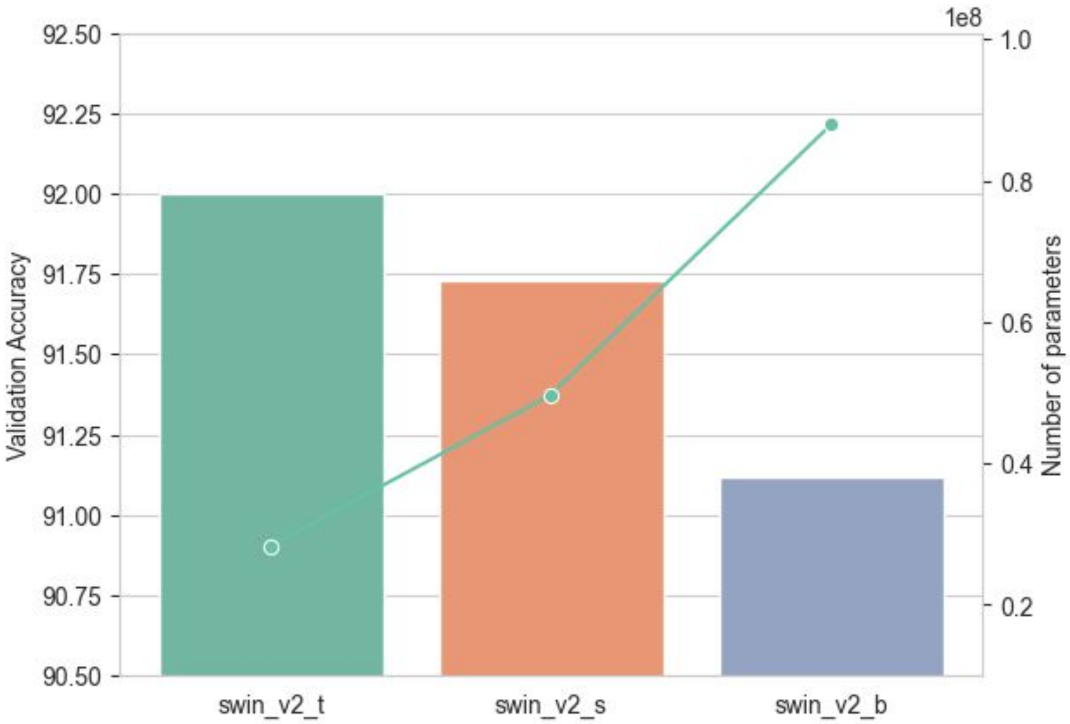


[5, 6]

Universitat de Girona

Model sizes experiment





Size greater than regnet_y_3_2gf, started overfitting, and smaller were underfitting!

Size greater than swin_tiny started overfitting!

Universitat de Girona

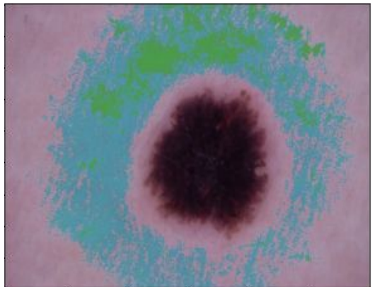
Augmentation

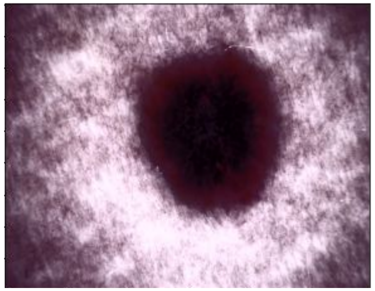
Modified randaugment [3]: 21 transformations(13 colour and 8 shape)

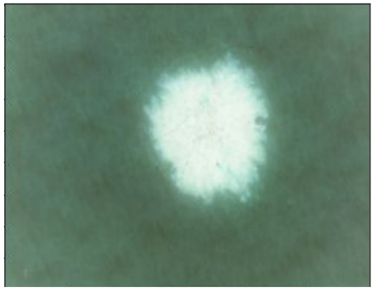
- Randomly select one transformation from {color} transformations, and then randomly select one transformation from {shape} transformations

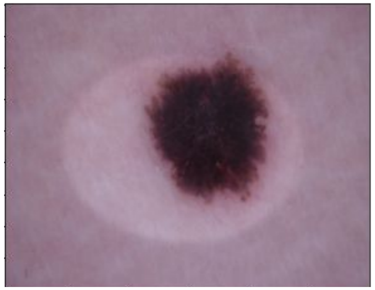
Color transformations

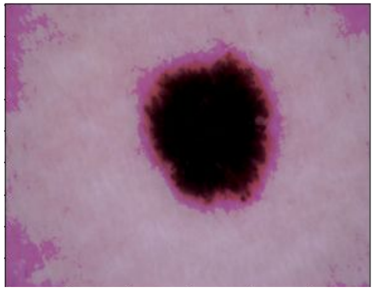

Auto-contrast


Polarize

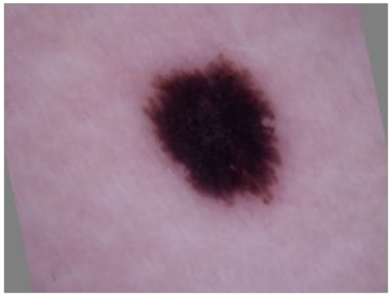

Equalize YUV

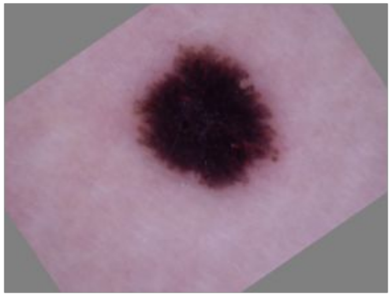

Invert

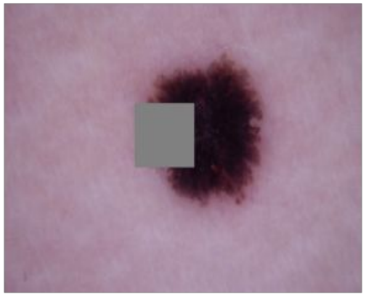

Mixup

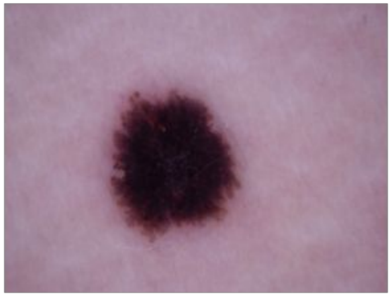

Solarize-add

Shape transformations


Shear

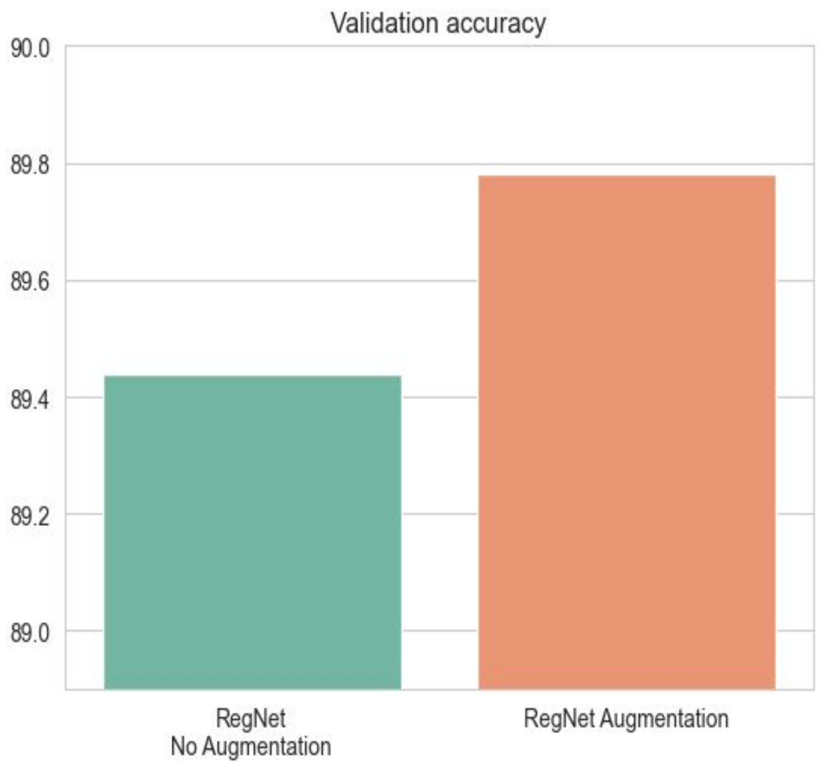

Rotate

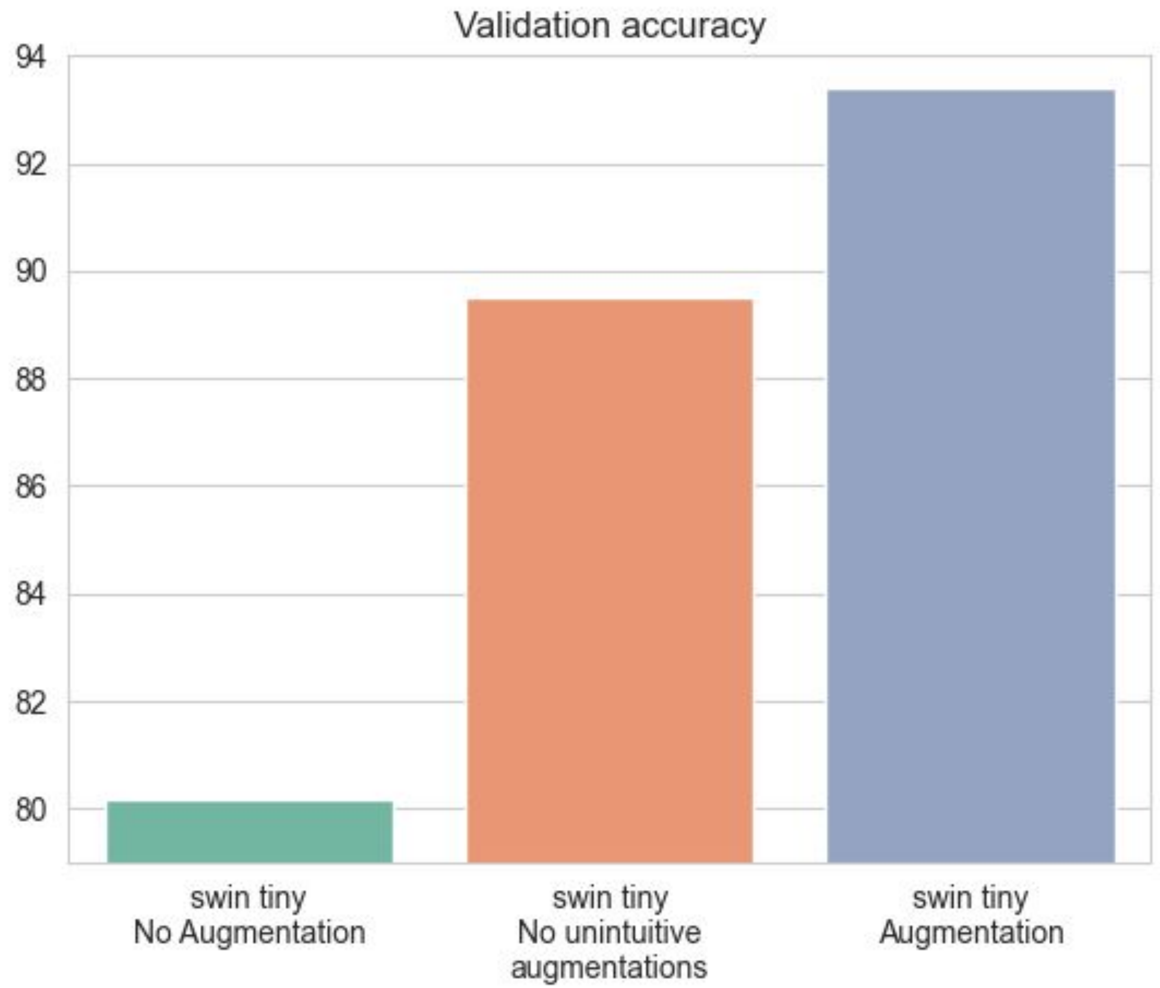

Cutout


Flip

Challenge 1: Augmentation

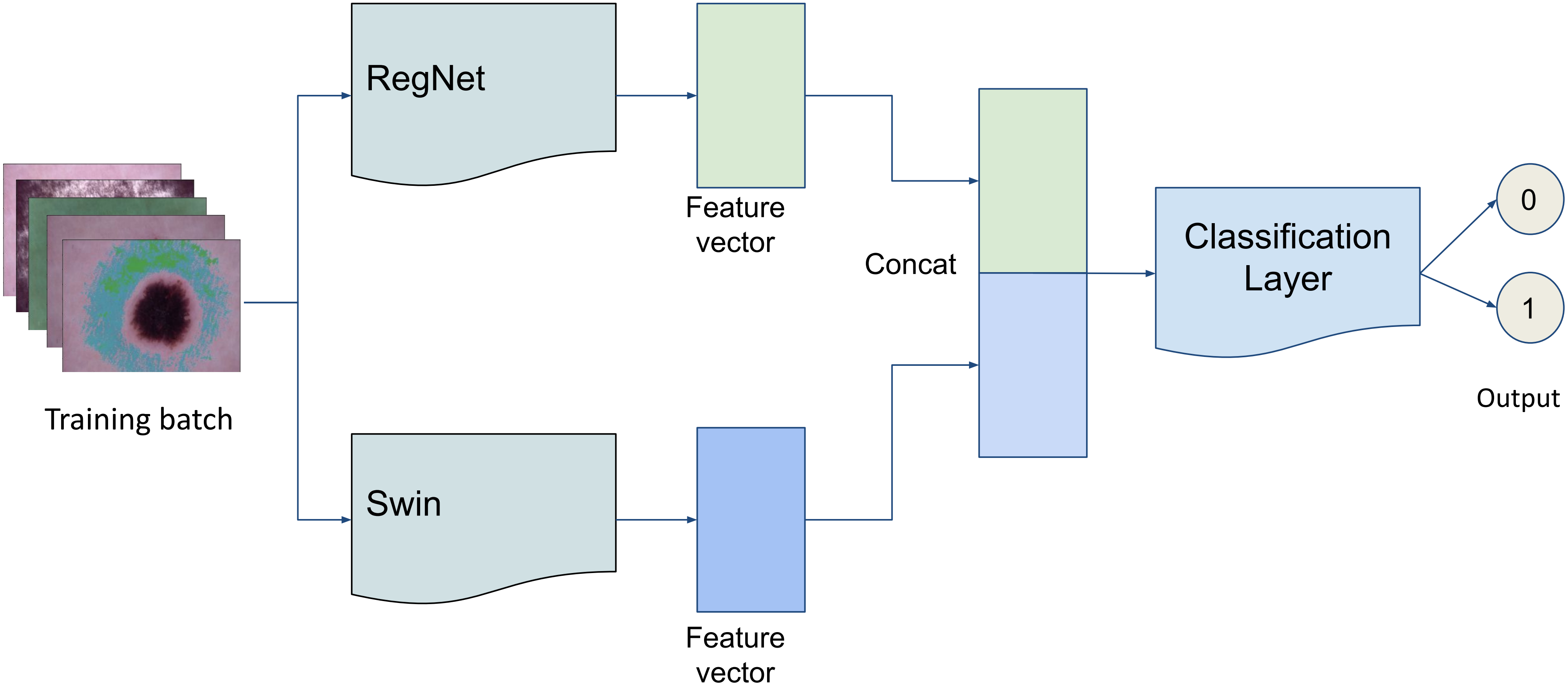
- Experiments on challenge 1: binary problem





Universitat de Girona

Challenge 1: Ensembling



Challenge 1: Ensembling

Freeze	Freeze the pretrained network and only train the linear layer
No Freeze	Do not freeze any layer on the ensemble model

Loss functions

Challenge 1: Cross-entropy loss.

Challenge 2: Losses that tackle class imbalance.

1. Focal loss

$$\mathrm{FL}(p_{\mathrm{t}}) = -(1 - p_{\mathrm{t}})^{\gamma} \log(p_{\mathrm{t}}).$$

- where $-\log(p_t)$ is the cross entropy loss
- $(1 - p_t)^{\gamma}$ is the modulating factor to down-weight easy examples and thus focus training on hard negative.
- The focusing tunable parameter γ smoothly adjusts the rate at which easy examples are down weighted.

Universitat de Girona

Loss functions

2. MWNL Loss [1]:

- Overcomes the class imbalance issue in sample number and classification difficulty
- Improves the accuracy of melanoma classification by adjusting the weight of the loss

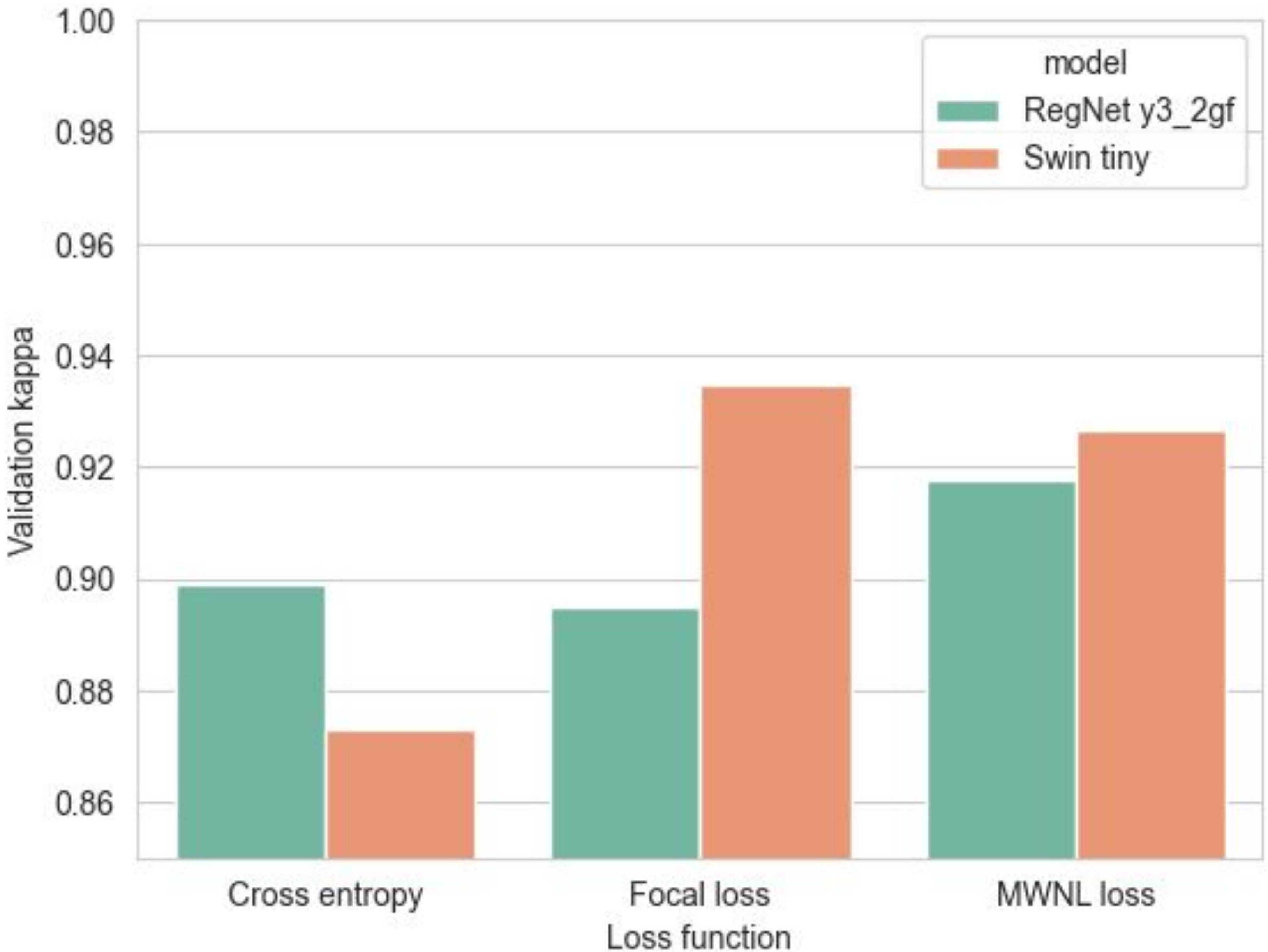
$$\mathrm{MWNL}(z,y) = -C_y\left(\frac{1}{N_y}\right)^{\alpha}\sum_{i=1}^{C} Loss_i.$$

where

$$Loss_i = \begin{cases} (1-p_i^t)^r \log(p_i^t) & p_i^t > T \\ G^* & p_i^t \leq T \end{cases}$$

Universitat de Girona

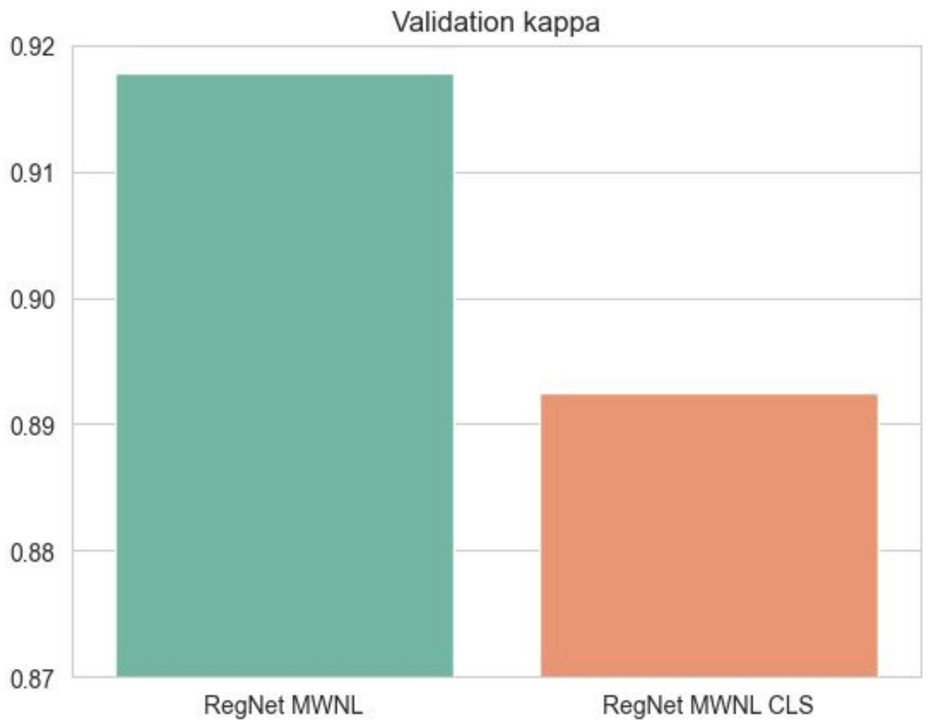
Challenge 2: Loss functions

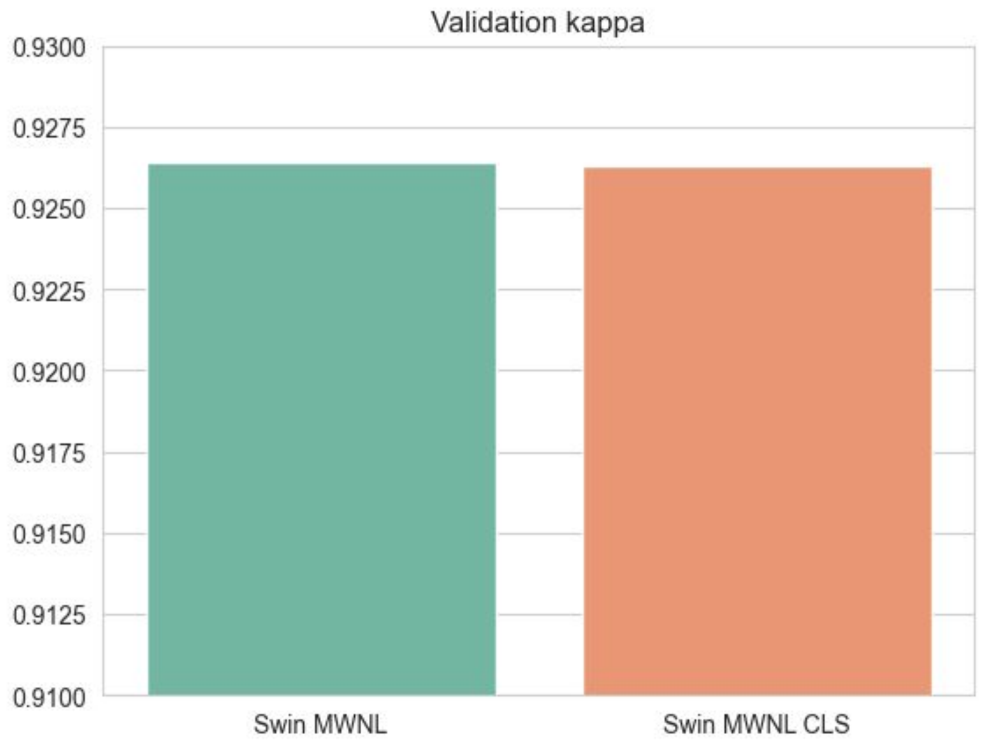


Universitat de Girona

Challenge 2: Cumulative Learning strategy

- First train the network on the originally imbalanced data.
- Then change the training gradually to a re-balancing mode.

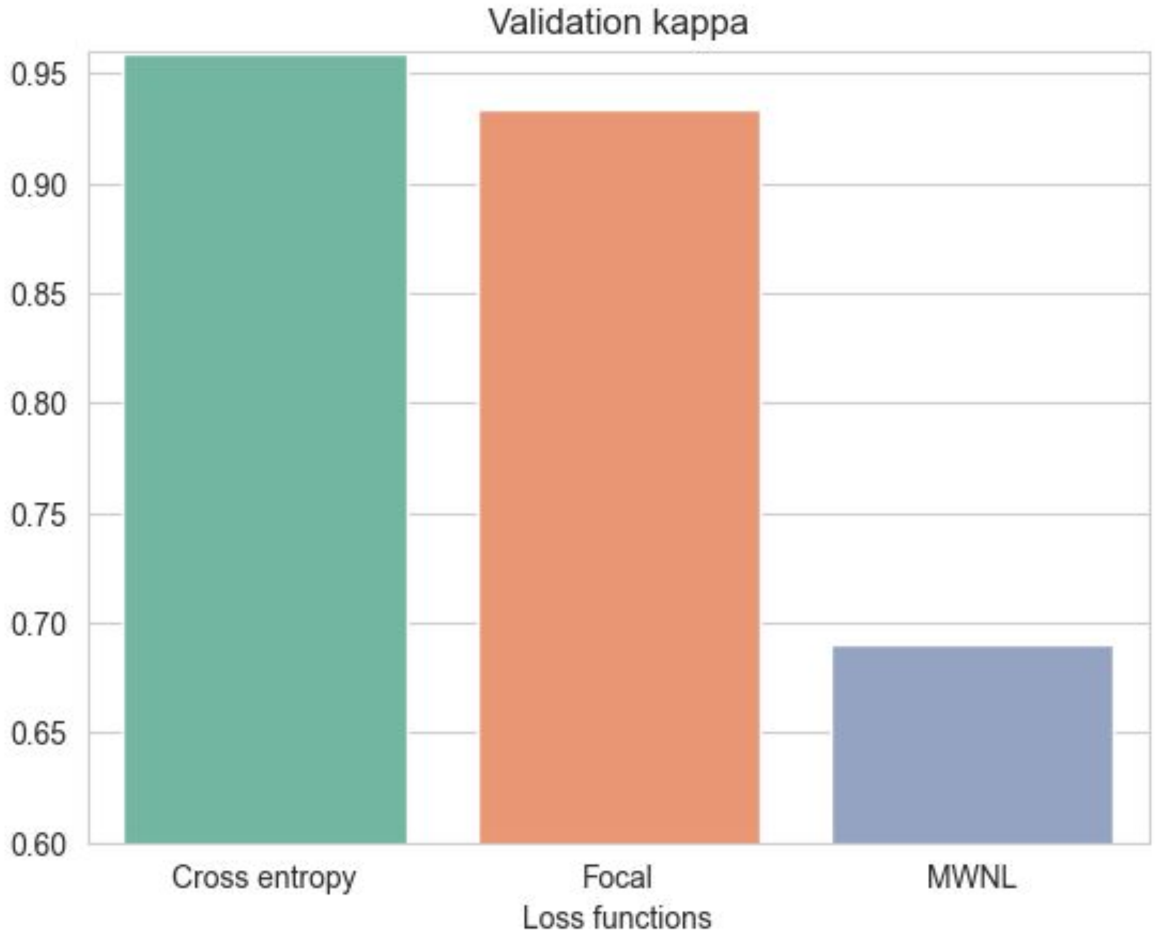




Universitat de Girona

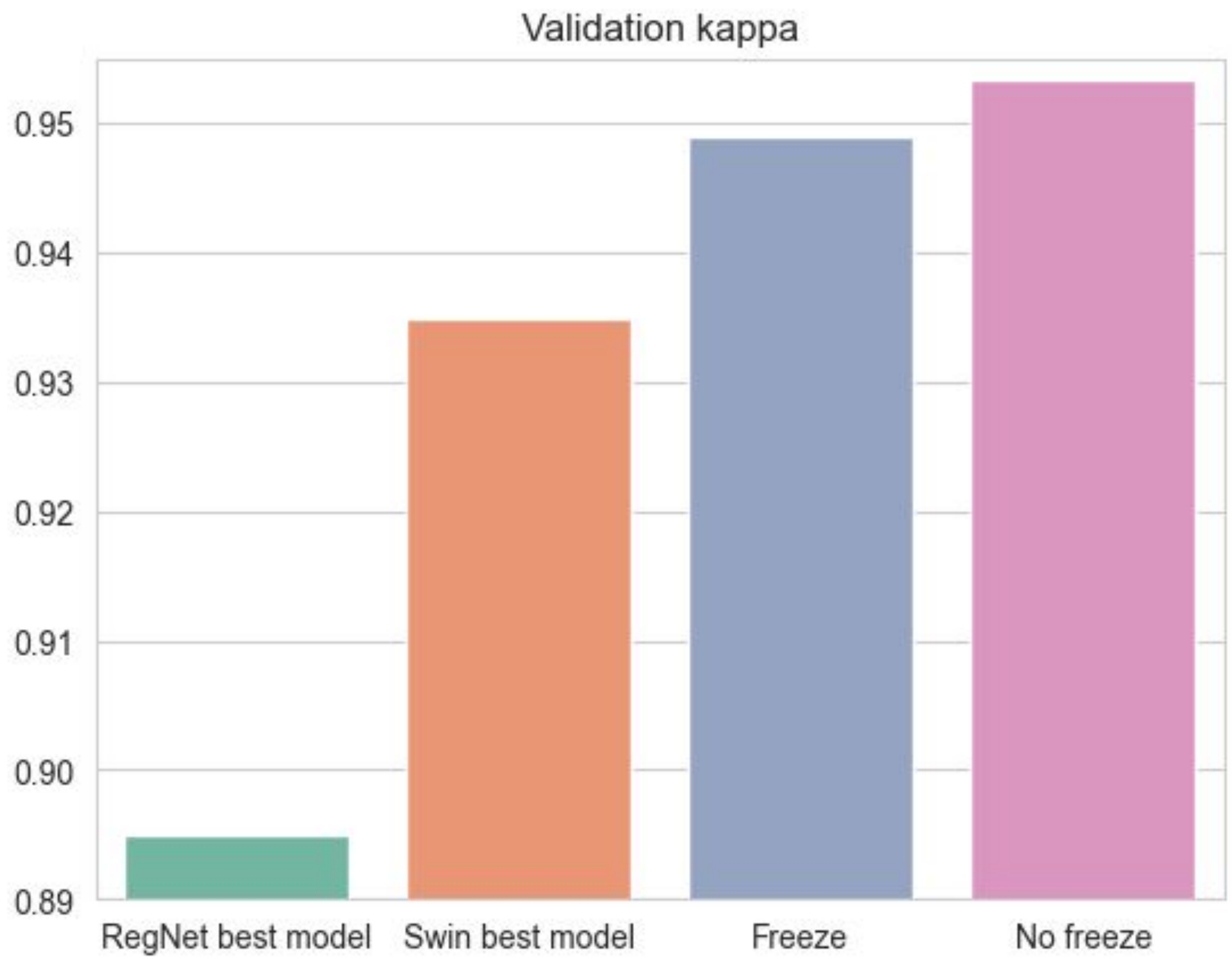
Balanced Sampling

- Weighted sampling of images to get balanced number of images in each batch (swin-tiny)



Challenge 2: Ensembling

Freeze	Freeze the pretrained network and only train the linear layer
No Freeze	Do not freeze any layer on the ensemble model



Universitat de Girona

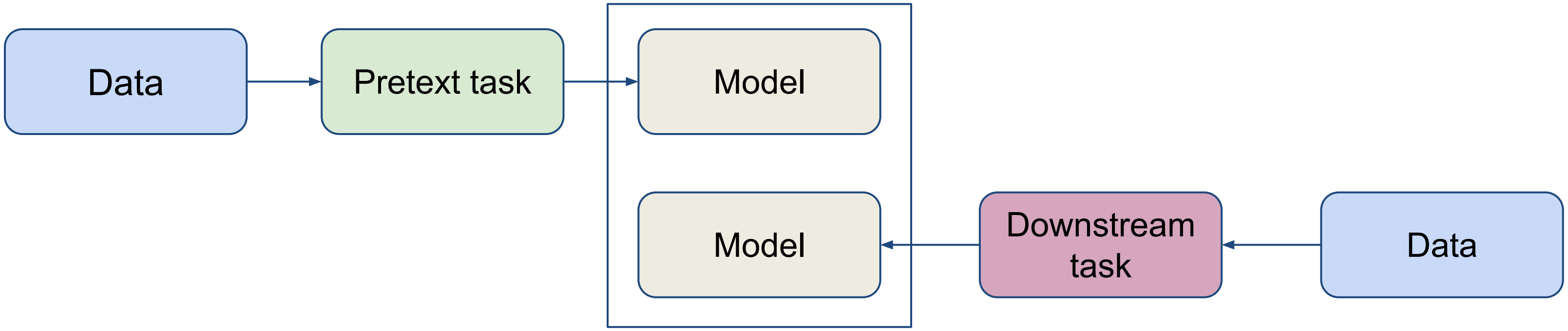
“Pretext learning”

Involves training a model for a task other than what it will actually be trained and used for. This Pretext Training is done prior to actual training of the model.

Needed to be performed with our tested models.

Pretext task to learn:

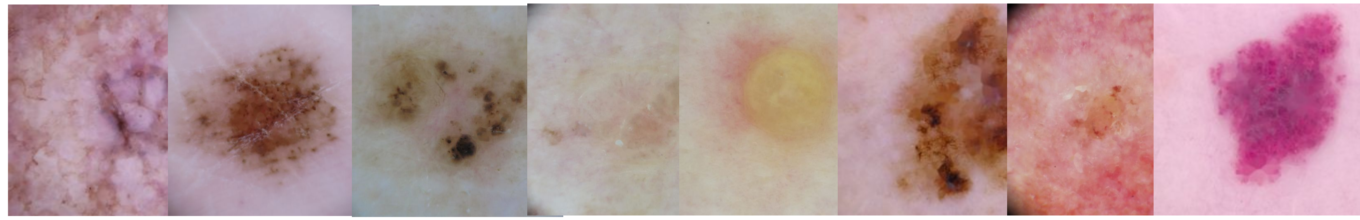
- lesion size
- lesion colors
- abcd scores
- other relevant patient medical data



Shared architecture/weights

“Pretext learning”

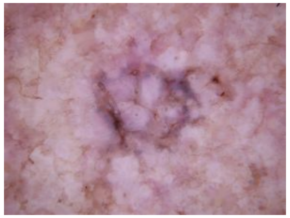
Pretext task
Eight class classification



ack nevus bcc bkl def mel scc vac

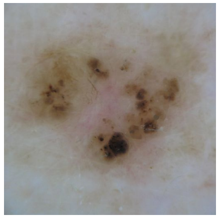
Challenge 1
Binary classification

nevus

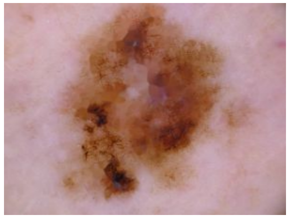


ack/bcc/bkl/def/mel/scc/vac

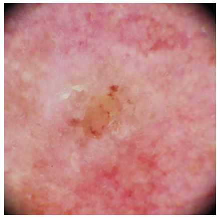
Challenge 2
Three-class classification



bcc



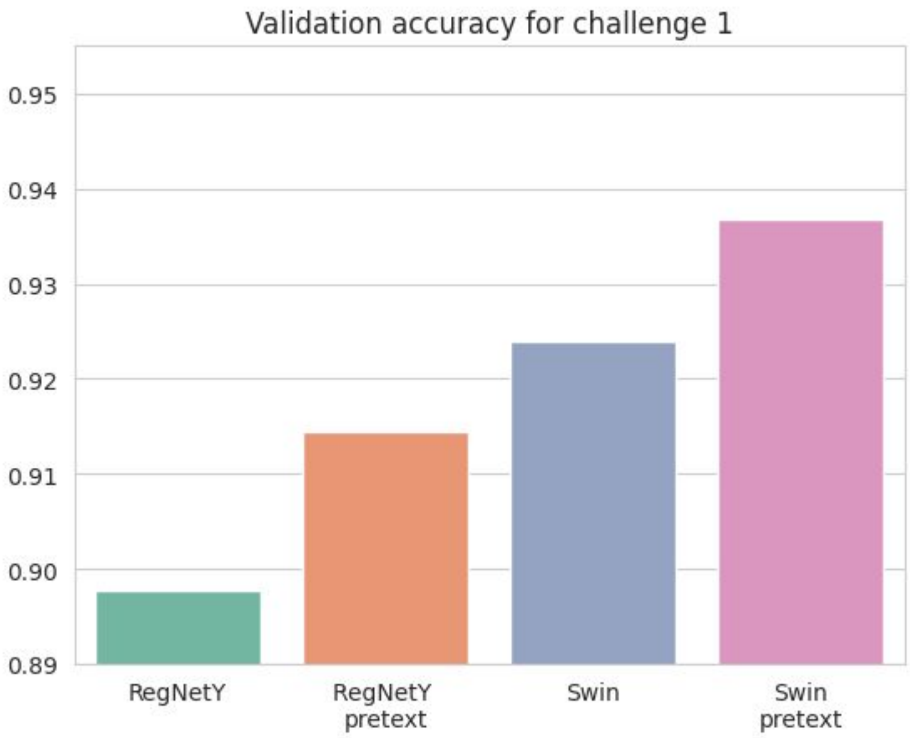
mel

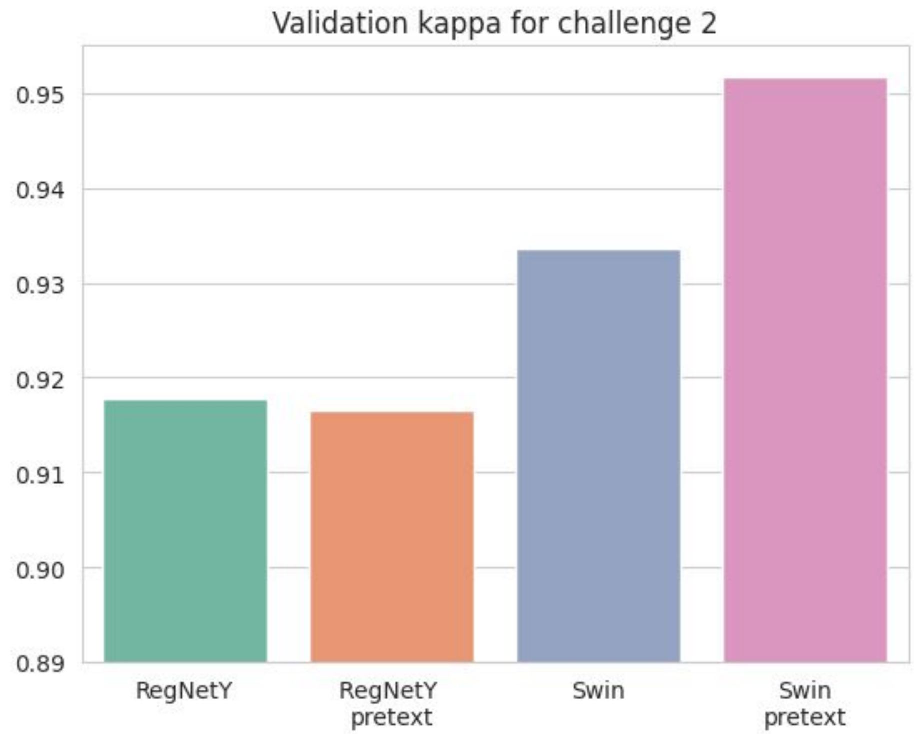


scc

Universitat de Girona

“Pretext learning” results





Both Swin and RegNetY improved performance with the pretext task for challenge 1.

Only Swin was able to maintain information learned during pretext training at challenge 2 training due it it’s bigger size and memory.

RegNetY - 0.818

Swin - 0.835

Final models

Challenge 1

Ensemble (learnable feature fusion)

- RegNetY-3.2GF (with pretext initialization)
- Swin-v2-Tiny (with pretext initialization)

RandAugment

Cross entropy loss

Validation accuracy: 0.936

Challenge 2

Ensemble (learnable feature fusion)

- RegNetY-3.2GF (without pretext initialization challenge 1 transfer learning)
- Swin-v2-Tiny (with pretext initialization and challenge 1 transfer learning)

RandAugment

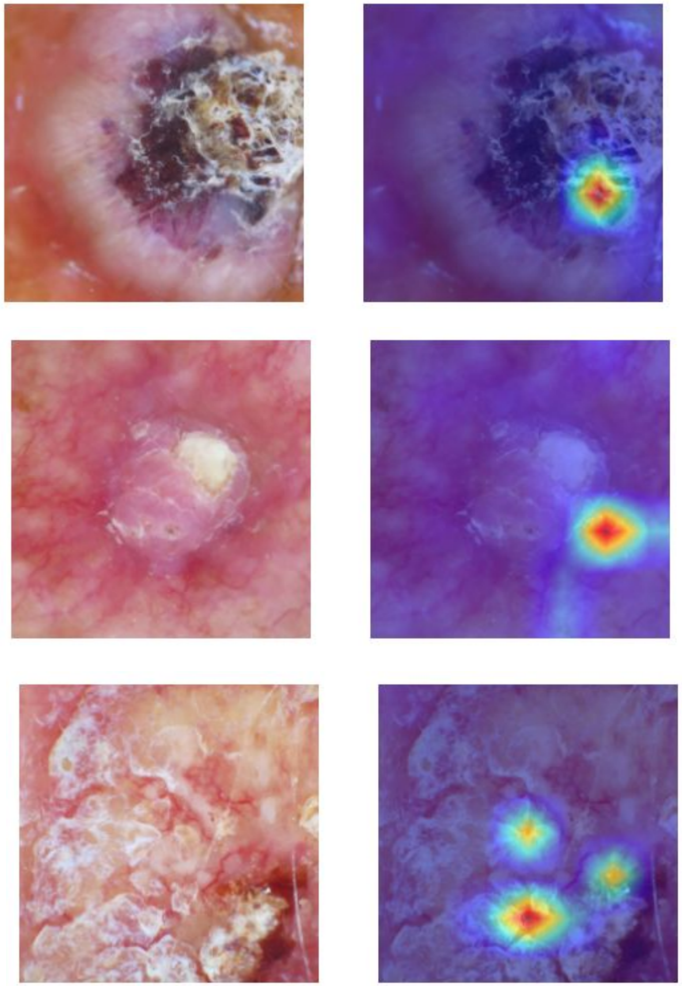
MWNL loss

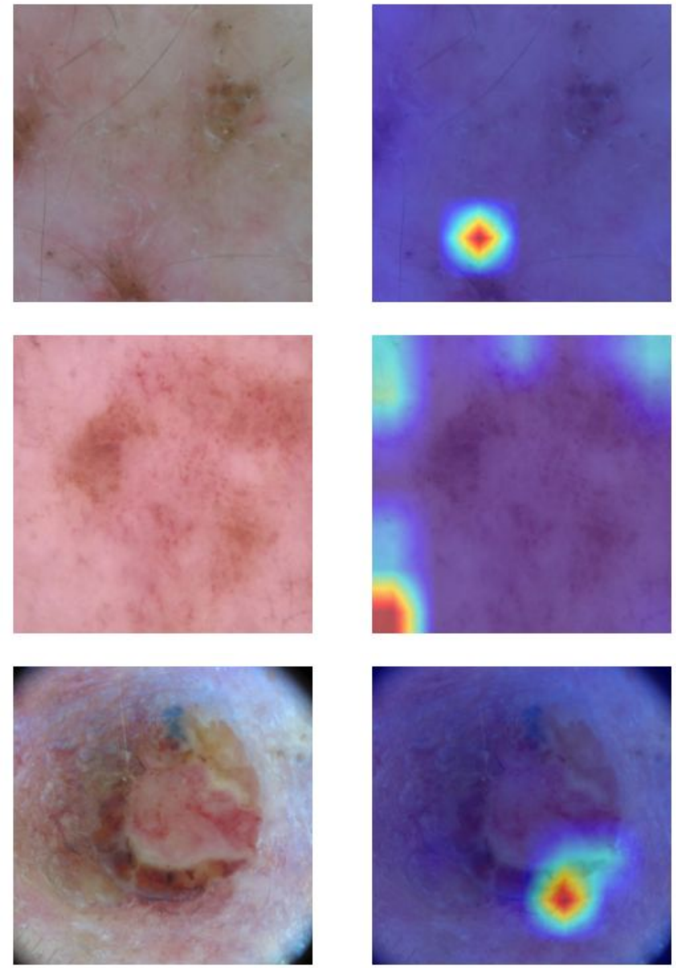
Validation kappa: 0.9533

Universitat de Girona

Grad-CAM

Grad-CAM of Correctly vs. incorrectly classified skin lesions





Conclusions

- Strong augmentations push models to learn a more robust set of features.
- Ensembling is a powerful tool that allowed us to combain and benefit from 2 different feature embeddings of convolutional and transformer models.
- Balanced sampling did help training the models and so did using sample-weight sensitive losses like focal or mwnl did.
- Bigger model sized are more prone to overfitting so the size needs to be fine-tuned depending on the problem and dataset.
- Pretext learning has great potential to improve the results, however the more training or fine tuning we perform over the model the more the initial weights change; only swin was able to benefit from it after challenge 1 and 2 fine tuning.

References

[1]https://www.kaggle.com/c/siim-isic-melanoma-classification/discussion/175412

[2] https://challenge.isic-archive.com/landing/2019/

[3] Yao, Peng & Shen, Shuwei & Xu, Mengjuan & Liu, Peng & Zhang, Fan & Xing, Jinyu & Shao, Pengfei & Kaffenberger, Benjamin & Xu, Ronald. (2021). Single Model Deep Learning on Imbalanced Small Datasets for Skin Lesion Classification.

[4] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He and P. Dollár, "Designing Network Design Spaces," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 10425-10433, doi: 10.1109/CVPR42600.2020.01044.

[5]Wang J, Zhang Z, Luo L, Zhu W, Chen J, Wang W. SwinGD: A Robust Grape Bunch Detection Model Based on Swin Transformer in Complex Vineyard Environment. Horticulturae. 2021; 7(11):492. https://doi.org/10.3390/horticulturae7110492

[6] Zheng, Hao & Wang, Guohui & Li, Xuchen. (2022). Swin-MLP: a strawberry appearance quality identification method by Swin Transformer and multi-layer perceptron. Journal of Food Measurement and Characterization. 16. 1-12. 10.1007/s11694-022-01396-0.