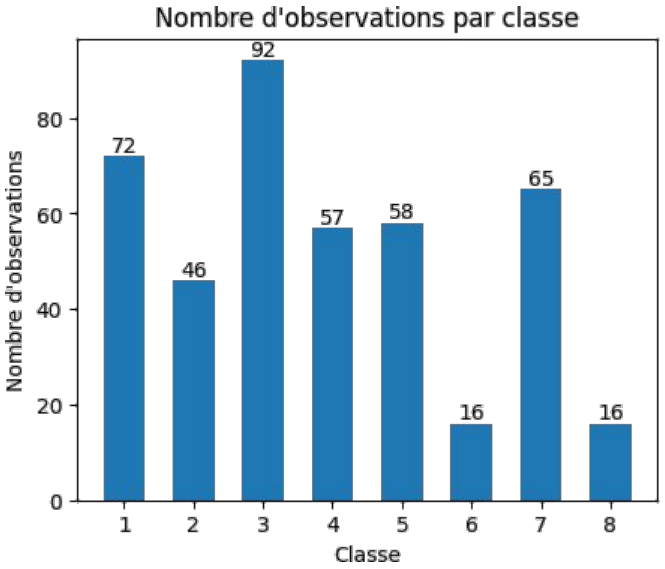
BGDIA703
Restitution data challenge

12 juin 2023

Stratégie

Jeu d'entraînement réduit : 422 images



- ↘ Pas de Deep Learning (SotA)
  - → Extraction features + ML classique
- ↖ Utilisation de méthodes robustes de sélection d'hyperparamètres
  - → Cross validation avec **16 folds**
  - → Leave One Out

Modèle

- Classifieur **SVM** avec kernel non linéaire
- Combinaison de **7 features**

Méthodologie

1. Sélection d'un classifieur
2. Optimisation de features avec le classifieur retenu
3. Optimisation de l'ensemble des features avec Leave One Out

Sélection d'un classifieur

Premières features

PFTAS : Parameter Free Threshold Adjacency Statistics
Mahotas

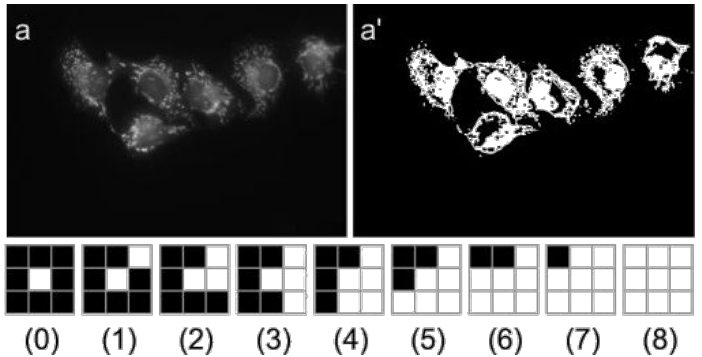
- Seuillage **Otsu** + **TAS**
- Crée **162 features** (appliqué sur images couleur)

Haralick
Mahotas

- Informations de texture
- Meilleurs résultats avec distance =2, moyenne et point-to-point (i.e. max() - min()), et calcul 14ème feature
- Crée **28 features**

Statistiques de couleur

- Calcule 4 moments par canal : μ, σ, γ_1, κ ⇒ **12 features**



Hu Moments
OpenCV

- Calcule 7 moments ⇒ **7 features**

Nicholas A Hamilton et al. *Fast automated cell phenotype image classification* BMC Bioinformatics, March 2007
Robert M Haralick et al. *Textural Features for Image Classification* IEEE Transactions On Systems Man And Cybernetics, November 1973

Sélection d'un classifieur

Comparaison de classifieurs

Feature \ Algorithm	SVM	Random Forest	Gradient Boosting	KNN	LDA	Régression Logistique
PFTAS	0.821	0.773	0.751	0.763	0.723	0.766
Color statistics	0.823	0.714	0.669	0.762	0.515	0.557
Hu moments	0.276	0.146	0.145	0.262	0.226	0.247
Haralick	0.833	0.715	0.682	0.754	0.612	0.624
Combinaison sans Hu	0.866	0.814	0.772	0.790	0.798	0.815
Combinaison totale	**0.868**	0.819	0.733	0.793	0.785	0.814

Sélection d'un classifieur

Comparaison de kernels SVM

GitHub - gmum/pykernels: Python library for working with kernel methods in machine learning

Feature \ Kernel	Exp	Cauchy	Log	Histogram Intersection	Tanimoto	RBF	Linear
PFTAS	0.820	0.833	0.804	0.754	0.833	0.821	0.772
Color statistics	0.812	0.833	0.772	0.590	0.831	0.823	0.692
Hu moments	0.264	0.259	0.227	0.228	0.281	0.276	0.215
Haralick	0.828	0.850	0.829	0.605	0.833	0.833	0.731
Combinaison totale	0.860	0.855	0.816	0.802	**0.868**	0.868	0.830

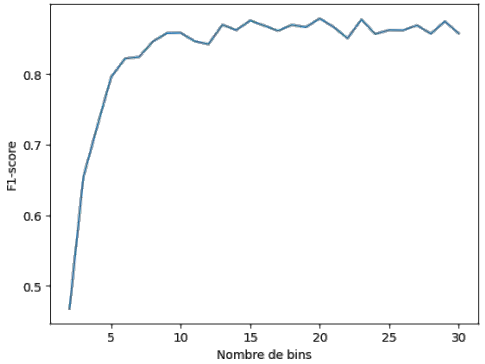
Optimisation de features

Histogramme de couleurs

OpenCV

- Compromis F1-score / nombre de features

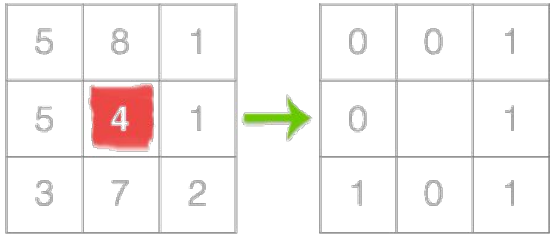
Bins	Features	F1-score
20	8000	0.879
15	3375	0.876
13	2197	0.870



Local Binary Pattern

SKImage

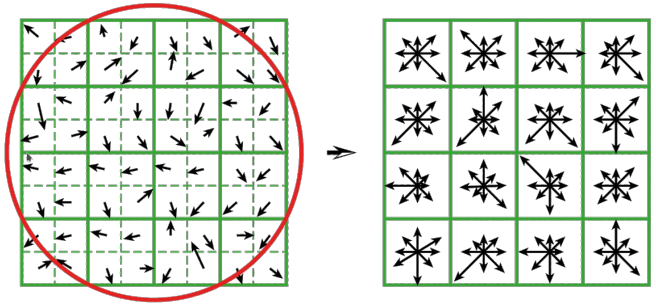
- Seuillage d'un pixel par rapport à *p* voisins dans un rayon *r*
- Histogramme des valeurs
- Meilleur rayon 7 ⇒ **58 features**



Bag of Visual Words / SIFT (Scale Invariant Feature Transform)

OpenCV

- Calcul des descripteurs **SIFT** de chaque image (2 225 354 keypoints trouvés sur train + test)
- K-Means pour trouver les descripteurs les plus représentatifs
- Classification des images en fonction des centroïdes trouvés
- Meilleur score pour k=300 ⇒ **300 features**



T. Ojala et al. *Multiresolution gray-scale and rotation invariant texture classification with local binary patterns*, IEEE Transactions on Pattern Analysis and Machine Intelligence, July 2002

David G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints*, International Journal of Computer Vision, November 2004

Optimisation finale

Combinaison initiale des features

Total : **2764 features** ⇒ F1-Score 0.7947

Réduction du nombre de features globales avec Leave One Out

- Histogramme : 13 bins / 2197 features → 11 bins / 1331 features
- LBP : rayon 7 px / 58 features → 9 px / 74 features

Score final

- F1-Score **0.8026**
- Classifieur SVM avec kernel Tanimoto, C=6, **1914 features** :
  - Parameter-Free Threshold Adjacency Statistics
  - Statistiques des canaux de couleur (moyenne, écart-type, asymétrie, kurtosis)
  - Hu Moments
  - Features de texture Haralick (distance=2, moyenne Point-to-Point et 14 features calculées)
  - Histogramme de couleurs avec 11 bins
  - Local Binary Patterning, avec un rayon de 9 pixels et 72 points
  - Descripteurs SIFT, clusterisés avec 300 centroides