

## GROUPE 14 - Dataset : Good and Bad Eggs Identification Image Dataset - Détection automatique des œufs sains et défectueux

Lien du dataset : <https://data.mendeley.com/datasets/mdty358x8m/1>

### Contexte

La qualité des œufs est un enjeu majeur pour la sécurité alimentaire et pour la réduction des pertes dans l'industrie agroalimentaire. Les défauts invisibles à l'œil nu ou mal détectés entraînent des coûts supplémentaires et augmentent les risques sanitaires. L'objectif du projet est de mettre en place un système d'analyse automatique capable de distinguer les œufs sains des œufs défectueux à partir d'images.

### Partie 1 : Prétraitements

- Normalisation des images et correction d'illumination (éclairage souvent variable).
- Filtre médian pour réduire le bruit sans perdre les détails.
- Égalisation adaptative (CLAHE) pour faire ressortir fissures et tâches.
- Redimensionnement cohérent (ex: 224×224).

### Partie 2 : Extracteurs classiques

L'objectif est de capturer les motifs locaux liés aux défauts:

- **HOG** pour détecter les fissures et variations de contour.
- **LBP multi-échelle** pour analyser la texture de la coquille.
- **GLCM** pour mesurer la rugosité et les patterns des taches.
- **Haralick + moments de Hu** pour décrire les formes irrégulières.
- **Histogrammes HSV** pour repérer les zones de décoloration.

### Partie 3 : Extracteurs profonds

Produire des embeddings à partir de modèles modernes:

- **EfficientNet-B0 / B2** (compact et performant).
- **ConvNeXt-Tiny**.
- **Swin Transformer** (vision globale utile pour repérer les fissures).
- **RegNetY**.

Extraction juste avant la classification finale.

### Partie 4 : Extracteurs hybrides

Créer des représentations enrichies:

- **CNN + GLCM** pour fusionner vision globale et texture.

- **Transformer + HOG** pour combiner attention et détection de bord.
- **Autoencoder convolutionnel** pour apprendre un espace compact des images.
- **VQ-VAE** pour capturer des motifs discrets complexes.
- **Masked Autoencoder (MAE)** pour apprendre des représentations robustes à partir de masques.

## Partie 5 : Réduction de dimension

Comparer deux approches:

### Approches statistiques

- **PCA.**
- **ICA.**
- **TruncatedSVD.**

### Approches deep learning

- **Autoencoder à goulot d'étranglement** (bottleneck).
- **Parametric UMAP.**
- **Deep Embedding Reduction Network.**

## Partie 6 : Classificateurs

Tester plusieurs modèles adaptés aux vecteurs obtenus:

- **SVM RBF.**
- **Random Forest.**
- **XGBoost.**
- **MLP** (pour embeddings profonds).
- **k-NN** (pour descripteurs bas niveau).
- **Logistic Regression** comme baseline.

## Partie 7 : Évaluation

- Accuracy, F1-score, Recall par classe (sain / défectueux).
- Matrice de confusion pour analyser les erreurs.
- Commentaires sur les situations difficiles:  
fissures fines, variations d'éclairage, coquilles atypiques.

## Livrables attendus

- Notebook complet.
- Tableau de comparaison entre:  
extracteurs classiques, profonds, hybrides, avec et sans réduction de dimension.
- Discussion sur les limites (échantillons ambigus, éclairage variable).
- Proposition d'optimisations pour un déploiement en inspection automatique.