Détection des deepfakes vidéos

Paul Tessé [1,2], Christophe Charrier [1], Emmanuel Giguet [1]

[1] Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, Caen, FRANCE
[2] INSA Rouen Normandie

Stage M2 dans l'équipe SAFE

1er mars – 31 juillet 2023

GREYC
Laboratoire de recherche en sciences du numérique

Normandie Université

UNICAEN

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

cnrs

Sommaire

GREYC Laboratoire de recherche en sciences du numérique

Normandie Université

UNICAEN

ENSI CAEN ÉCOLE PUBLIQUE D'INGÉNIEURS CENTRE DE RECHERCHE

cnrs

1 - Contextualisation

Le GREYC

GREYC
Laboratoire de recherche en sciences du numérique

Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen

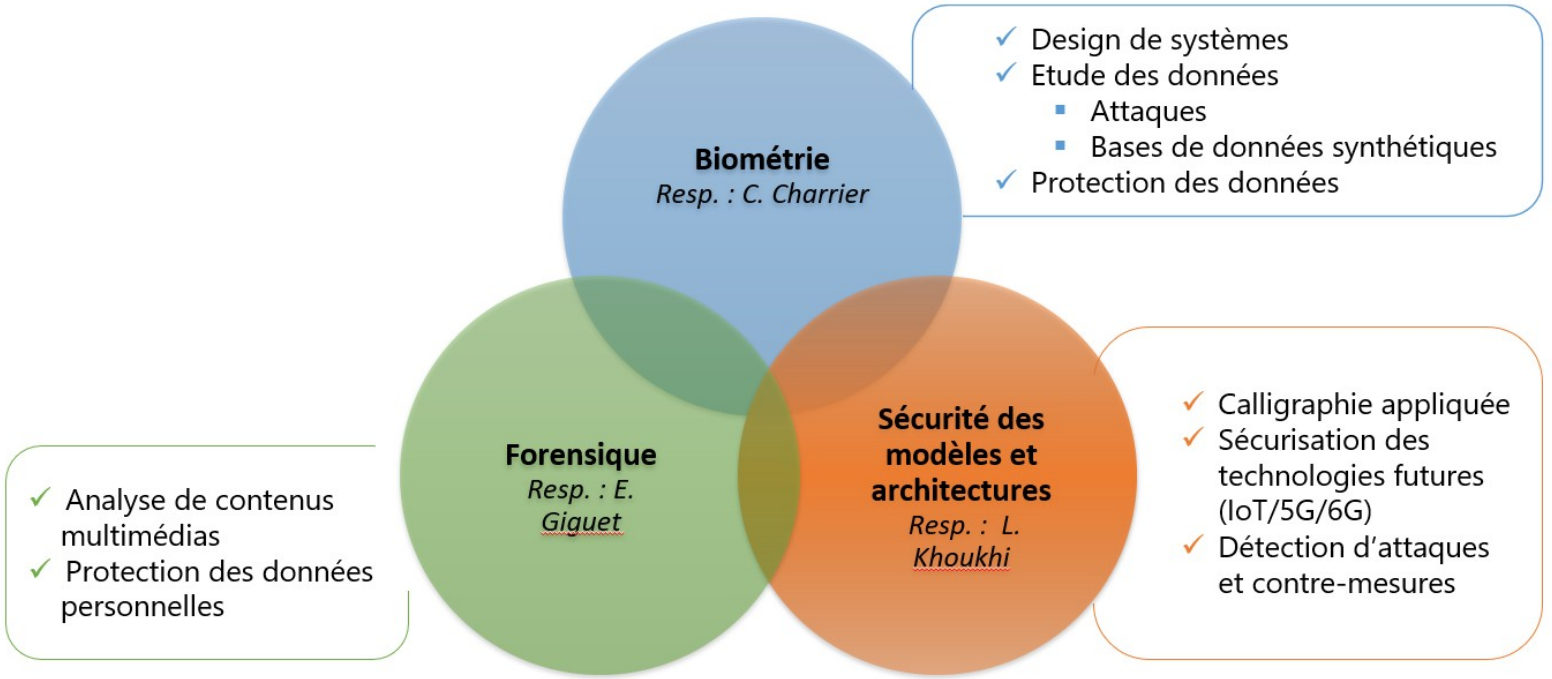
- Unité Mixte de Recherche créée en 1995
- Recherche en sciences du numériques
- Sous tutelle du CNRS, de l'UNICAEN et de l'ENSICAEN
- 6 équipes différentes, soient 180 membres sur 7 sites

cnrs

UNICAEN

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

L’équipe Sécurité, Architecture, Forensique et biomEtrie



- 30 membres dont 12 permanents
- 3 axes de recherche
- Sujet orienté Forensique

Le Deepfake

Réel ou artificiel ?

Le Deepfake

Spécification

Les Enjeux

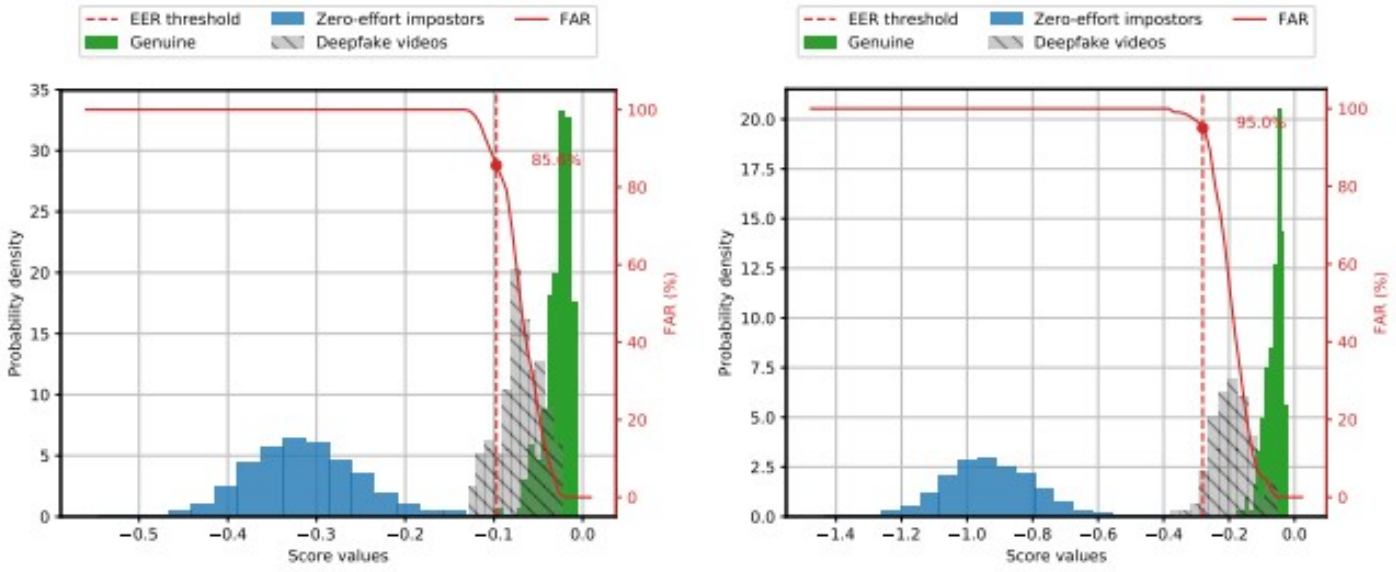
- Usurpation d'identité, fake news, propagande, etc.
- De plus en plus performants et difficiles à détecter
- Démocratisation du deepfake et surtout du face swapping

Les Contraintes

- Face Swapping uniquement
- Pas de référence
- Robuste et généralisable
- **Explicable**

2 – État de l’art

Vulnérabilité des systèmes

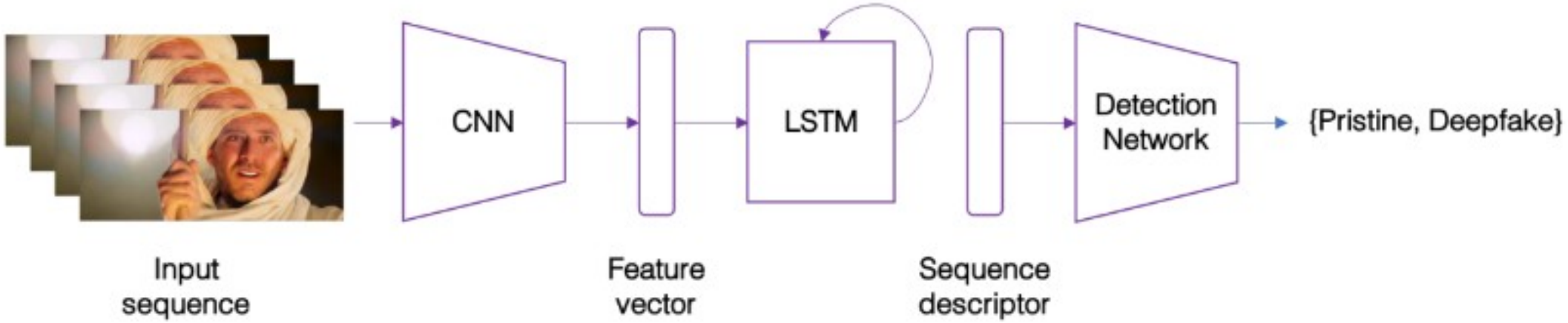


(a) VGG-based face recognition

(b) FaceNet-based face recognition

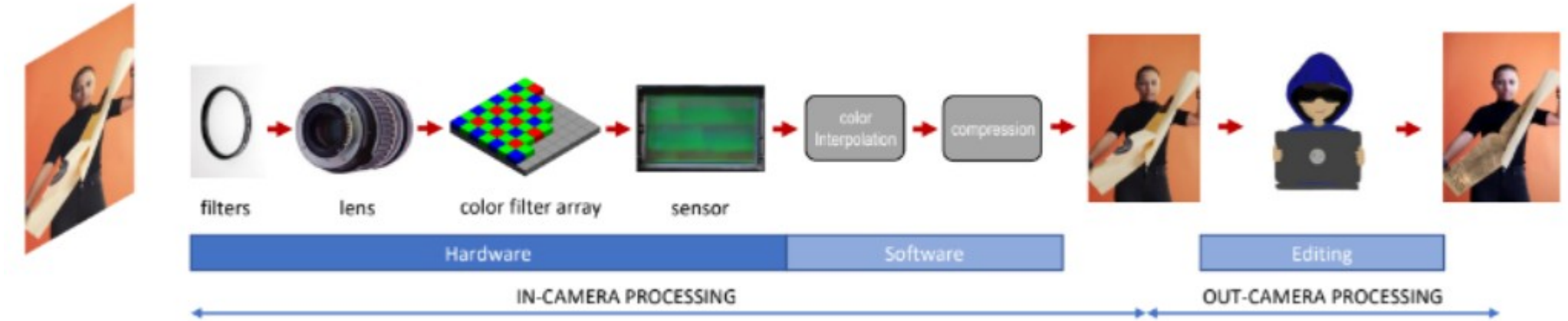
- Modèles à l'état de l'art donc très performants
- Densité de probabilité d'une mesure de similarité
- Equal Error Rate très mauvais et densités très proches
- Deepfakes générés en 2018

Exemple de modèle de détection



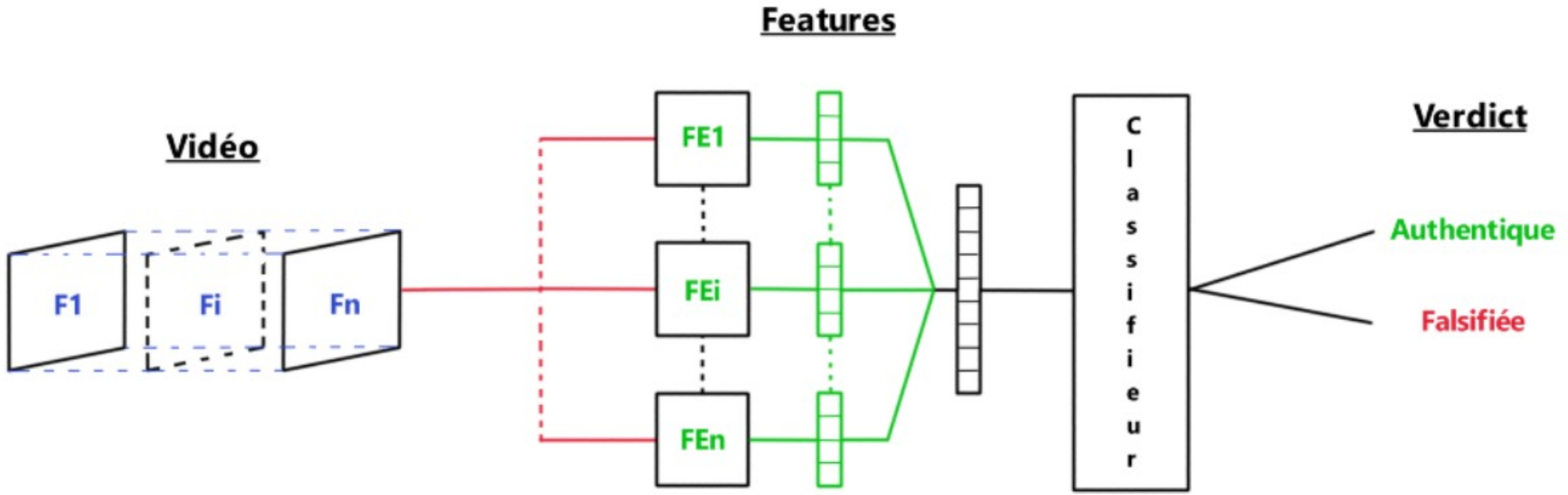
- **Architecture en deux parties**
- **Extraction des caractéristiques visuelles (CNN) & temporelles (LSTM)**
- **Détecteur composé de couches denses**
- **Mauvaise explicabilité des résultats**

Les signaux résiduels



- Caractéristiques intrinsèques des images altérables par les attaques
- Très utilisé en forensique des images/vidéos
- **Explicables et variées combinables à de l'apprentissage profond**

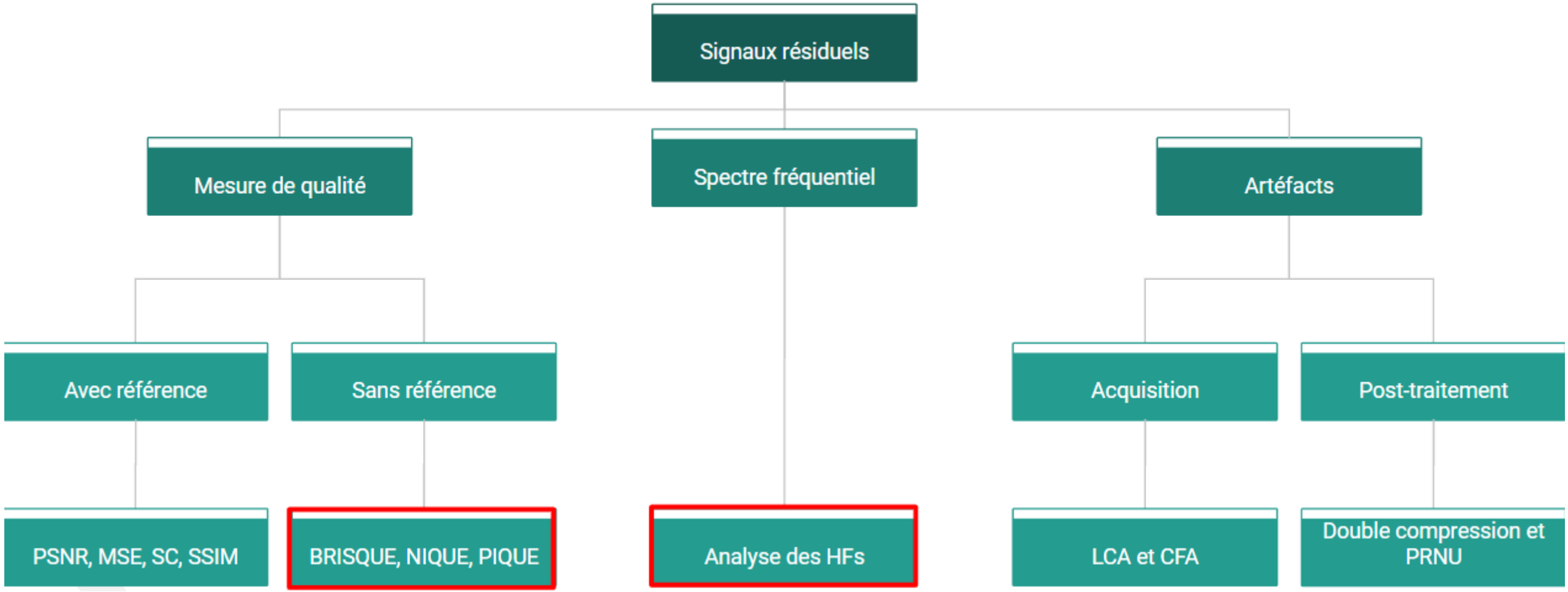
Notre proposition



1. **Extracteurs de caractéristiques basés sur les signaux résiduels**
2. **Concaténation des caractéristiques explicables**
3. **Classifieur profond binaire**

3 – Extracteurs de caractéristiques

Taxonomie

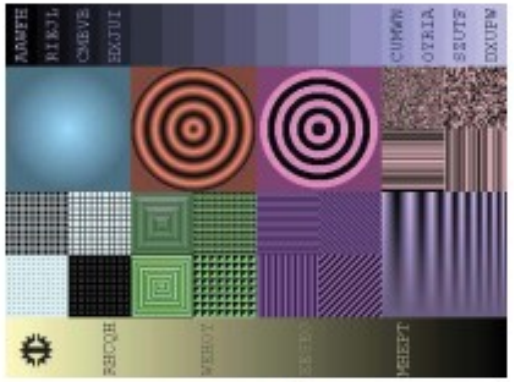

Signaux résiduels
Mesure de qualité
Spectre fréquentiel
Artéfacts
Avec référence
Sans référence
Acquisition
Post-traitement
PSNR, MSE, SC, SSIM
BRISQUE, NIQUE, PIQUE
Analyse des HFs
LCA et CFA
Double compression et PRNU

Démarche ingénieur

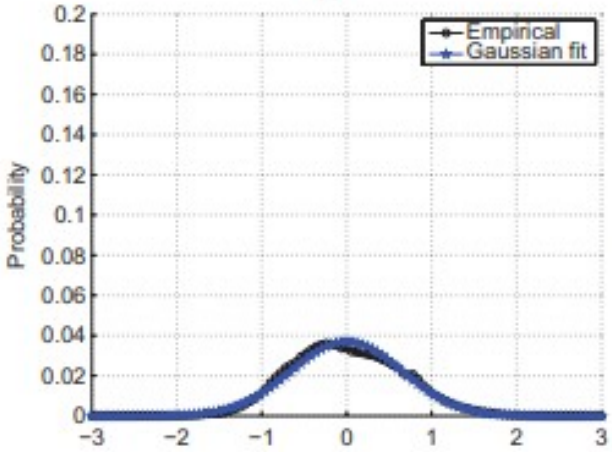
	BRISQUE	Analyse freqs.	LCA	CFA	Blind PRNU
Verdict	Très bon EER avec un SVM donc du potentiel	Très bonne précision avec régression	Précision variable mais bonne	Bonne AUC en blind	Très bonne AUC mais baisse pour compression JPEG
Robustesse/ Généralisation	Baisses de performances récurrentes	Très bons résultats présentés	Sensible aux changements de lentilles et peu de tests réalisés	Résultats variables mais restent bons	Sensible à la compression JPEG
Coût	SVM pré-entraîné et calculs de distributions peu coûteux	Normalisation + DCT donc très peu coûteux	Nombreux calculs de similarité coûteux	Estimations par maximum de vraisemblance donc peu coûteux	Clustering et beaucoup de données pour générer les références
Compatibilité	Score de qualité et paramètres de distributions	Spectre contenant toutes les fréquences	Possibilité de récupérer les valeurs de divergence locales	Nombreuses caractéristiques de distributions	Carte de corrélations

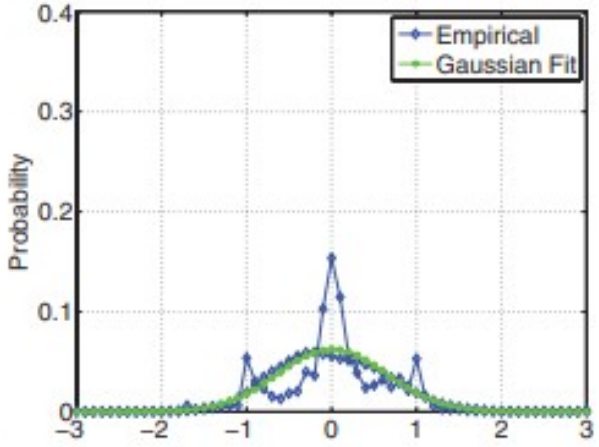
Blind/Referenceless Image Spatial Quality Evaluator

(a)



(b)

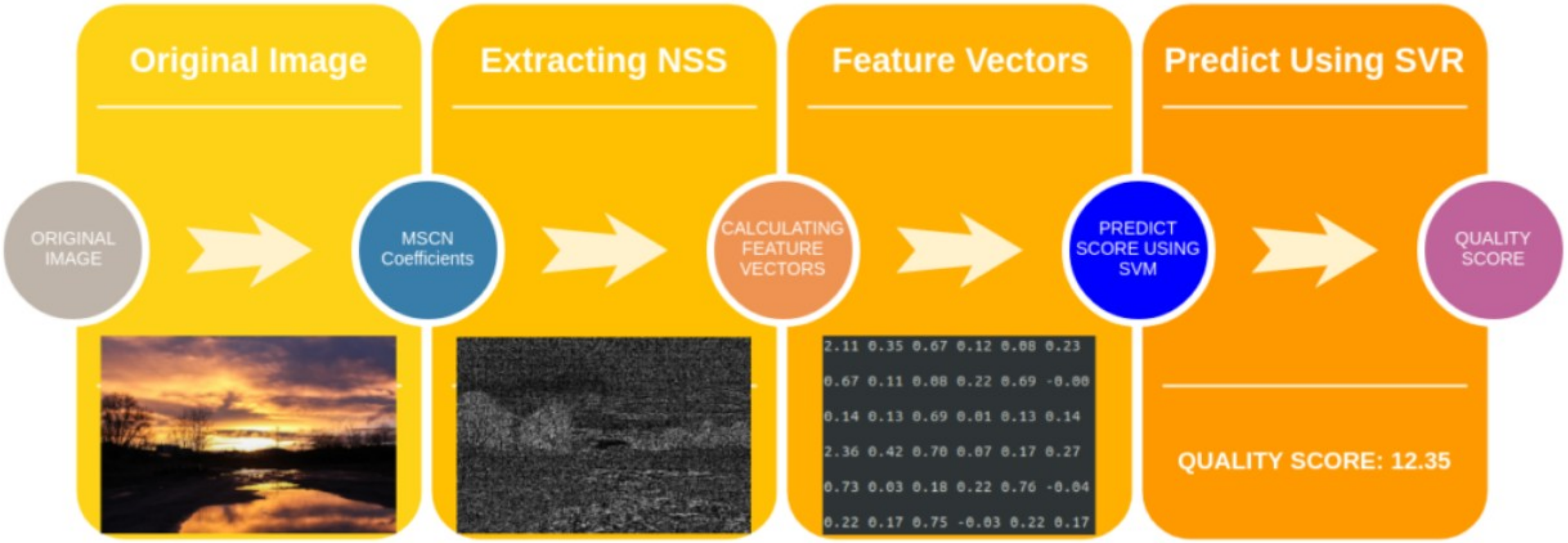




Distribution après MSCN

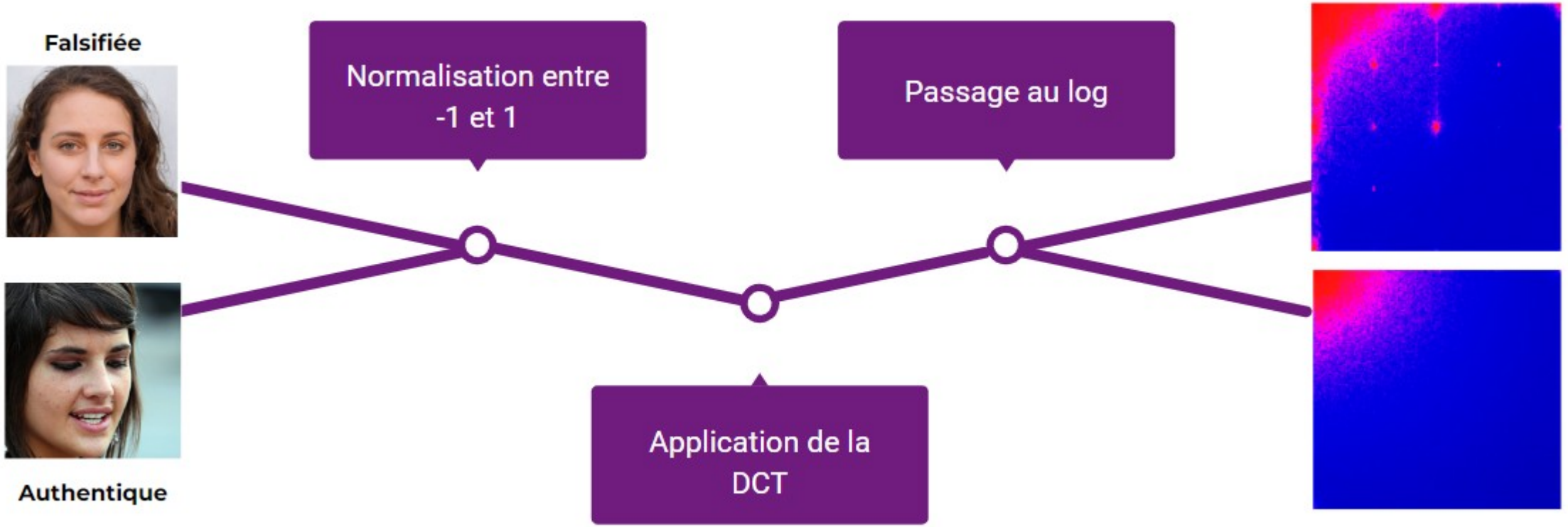
Les coefficients des images **naturelles** après normalisation (Mean Substracted Contrast Normalization) suivent une loi **gaussienne** contrairement aux images artificielles. Ceci peut être dû à du **flou**, de la **compression** ou encore du **bruit** introduit lors du processus de face swapping.

Blind/Referenceless Image Spatial Quality Evaluator

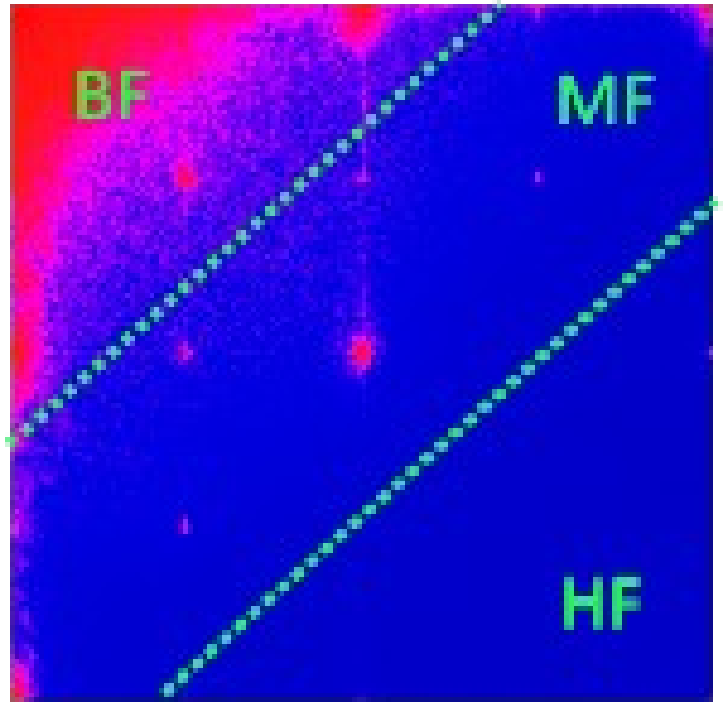


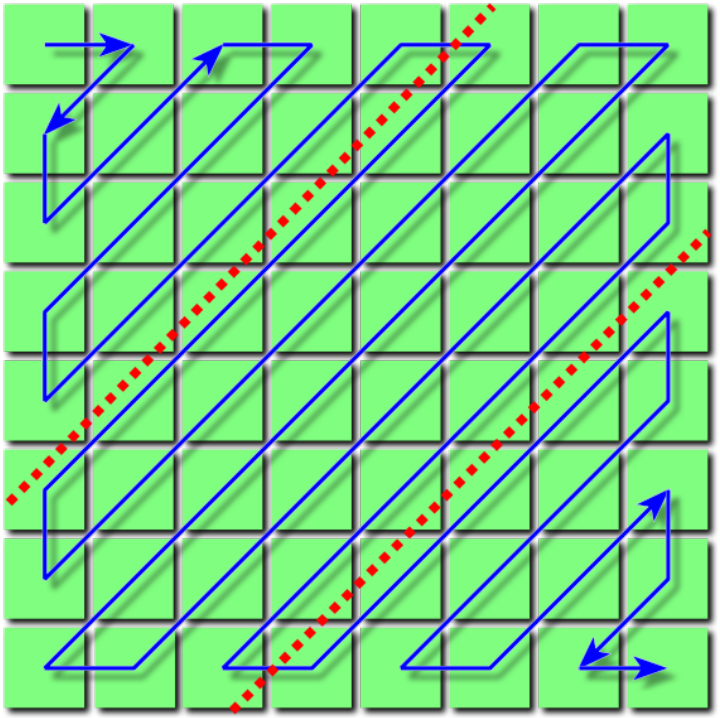
18*2 + 1 caractéristiques décrivant la qualité de l'image

Analyse des fréquences



Analyse des fréquences

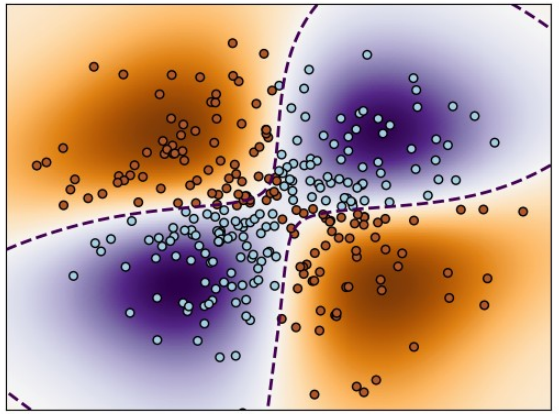




Mean
Std
Q1
Q2
Q3
Skew
Kurt

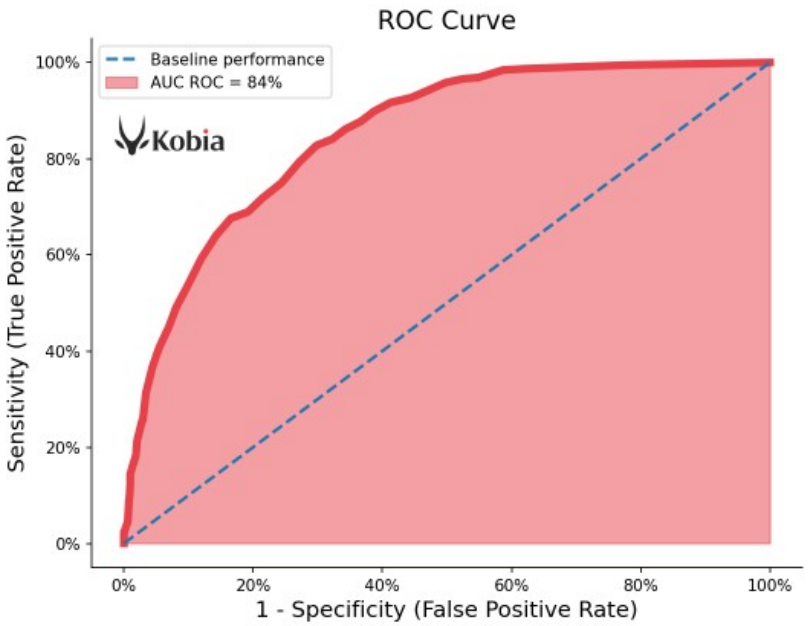
4 – Modèle de détection

Concaténation des caractéristiques



Caractéristiques	Train	Validation	Test	Généralisation
BRISQUE	77,00 %	77,00 %	76,00 %	54,00 %
Ratio fréquentiel	60,00 %	59,00 %	59,00 %	50,00 %
Concaténation	77,00 %	77,00 %	77,00 %	55,00 %

Métriques de référence



$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{all classifications}}$$

$$F1\ Score = 2 \times \frac{recall \times precision}{recall + precision}$$

$$recall = \frac{Vrai\ Positif}{Vrai\ Positif + Faux\ Négatif}$$

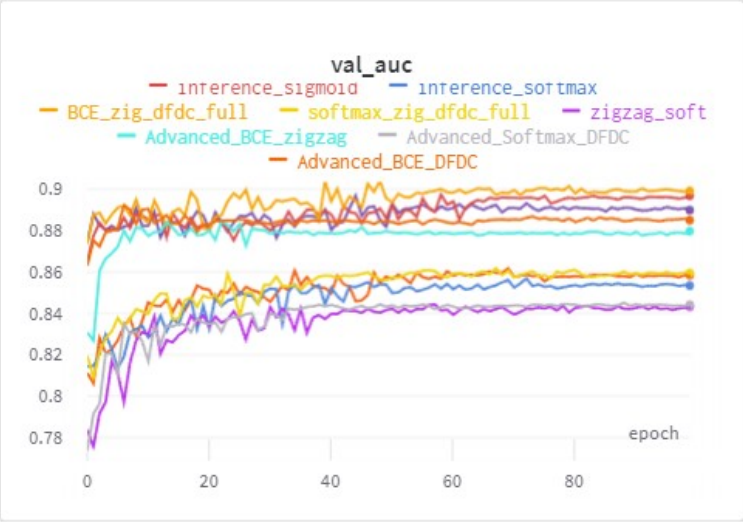
$$precision = \frac{Vrai\ Positif}{Vrai\ Positif + Faux\ Positif}$$

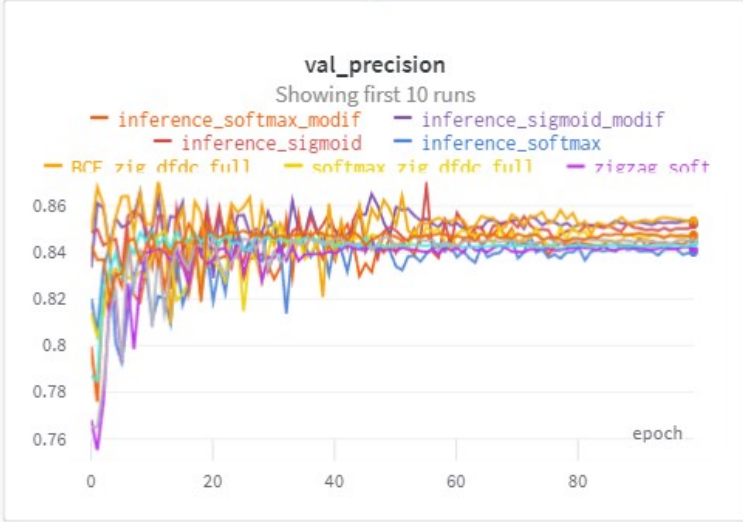
Modèle de référence

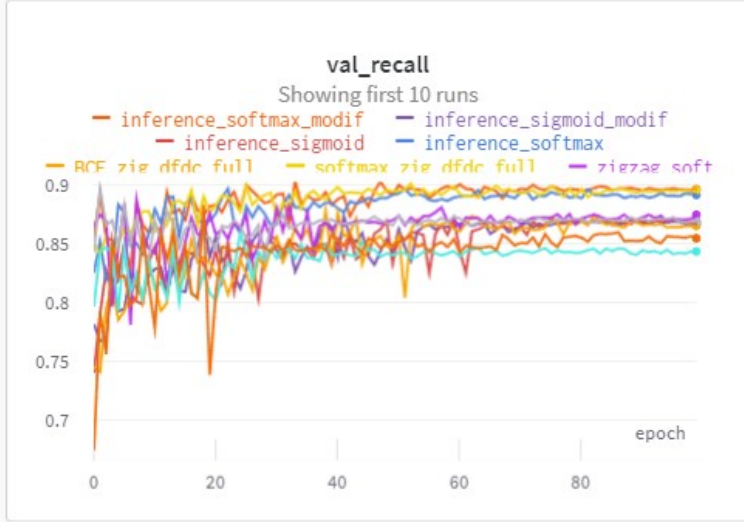
Set/Métriques	F1	Accuracy	AUC	Precision	Recall
Train	87,00 %	88,00 %	89,00 %	86,00 %	95,00 %
Validation	87,00 %	88,00 %	88,00 %	86,00 %	95,00 %
Test	84,00 %	85,00 %	84,00 %	85,00 %	91,00 %
Généralisation	55,00 %	57,00 %	57,00 %	54,00 %	94,00 %

1. **Benchmark Pycaret**
2. **Mise en place LDA**
3. **GridSearch effectué pour hyperparamètres**

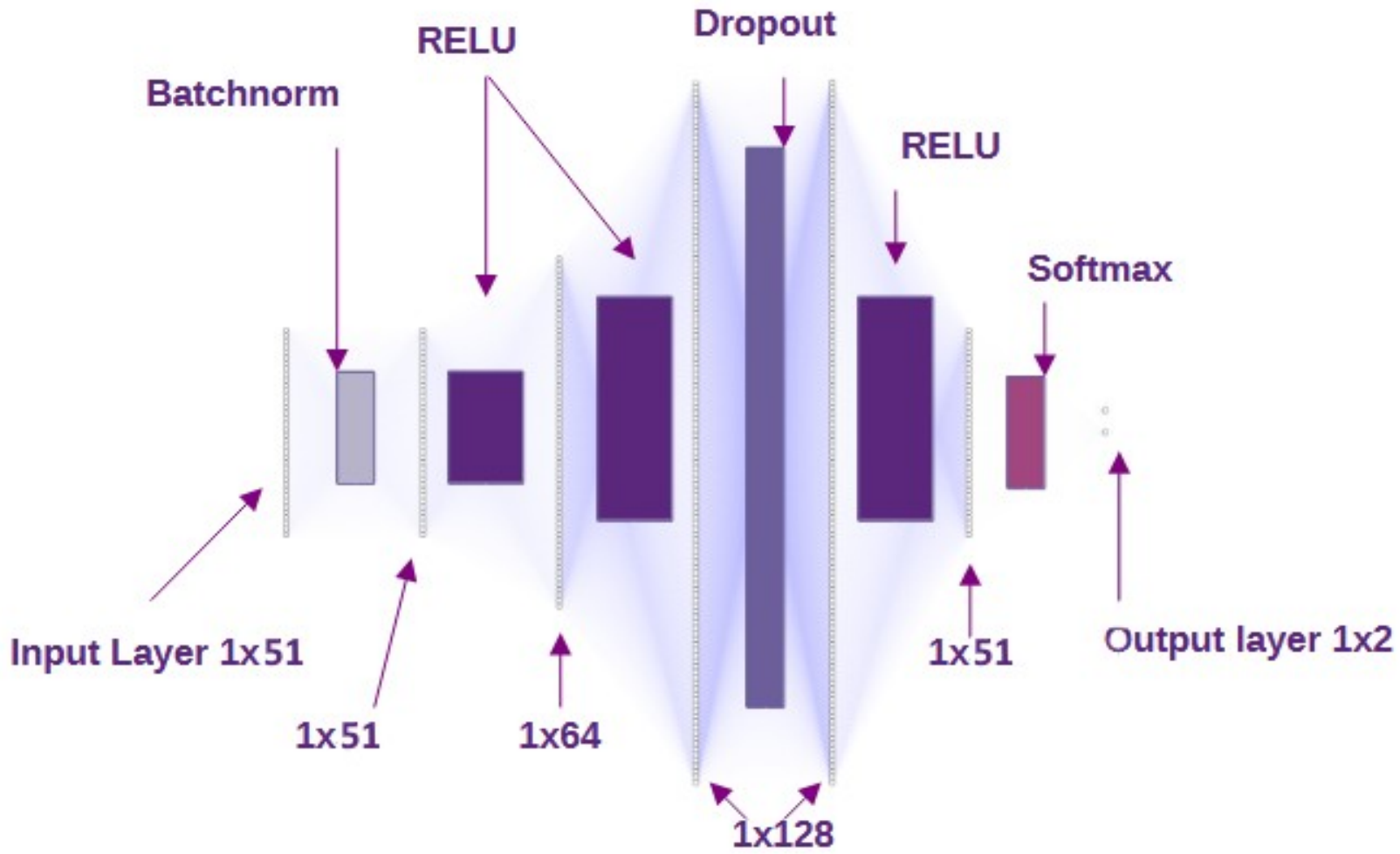
Recherches modèle final







Modèle Deep Learning final



CrossEntropy Loss

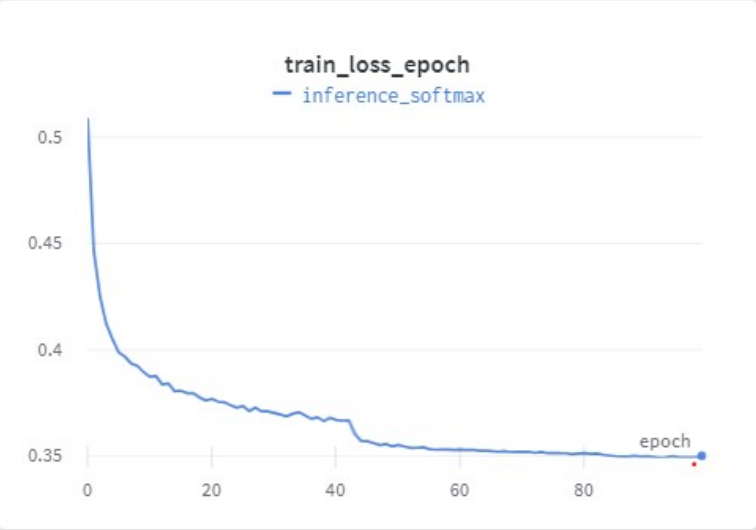
Epochs = 100

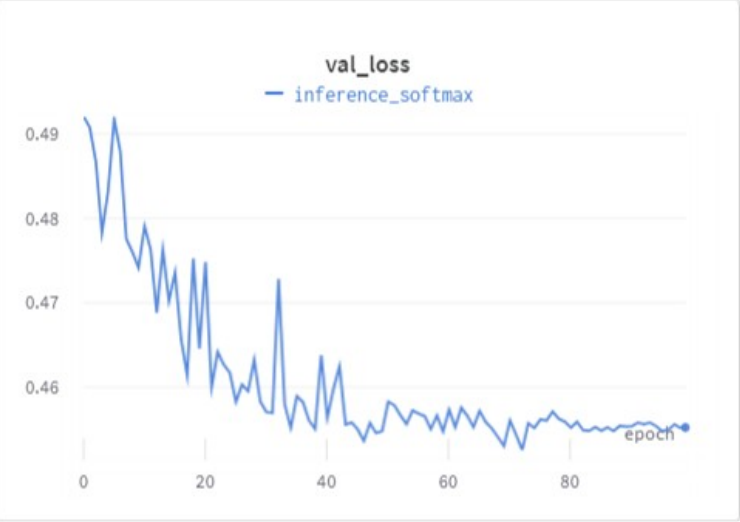
Batchsize = 1024

Lr adaptatif partant de 5e-3

Résultats finaux

Set/Métriques	F1	Accuracy	AUC	Precision	Recall
Train	96,00 %	96,00 %	96,00 %	96,00 %	96,00 %
Validation	87,00 %	86,00 %	86,00 %	85,00 %	90,00 %
Test	87,00 %	86,00 %	86,00 %	82,00 %	93,00 %
Généralisation	61,00 %	53,00 %	53,00 %	53,00 %	74,00 %





5 - Conclusion

Conclusion - Performances

	Mon Modèle	CNN sur le flux optique	CNNs en cascade	CNNs récurrents
Accuracy	96/87/87 %	80 % (test)	X	97/97/97 %
AUC	86 % (test)	X	91 % (test)	X

Conclusion – Cahier des charges

- Résultats encourageants avec un modèle en boîte grise
- Bonne robustesse (cf test) mais généralisation à améliorer
- Modèle léger et rapide donc peu coûteux et plus respectueux du DD

Conclusion – Ce qu'il reste à faire

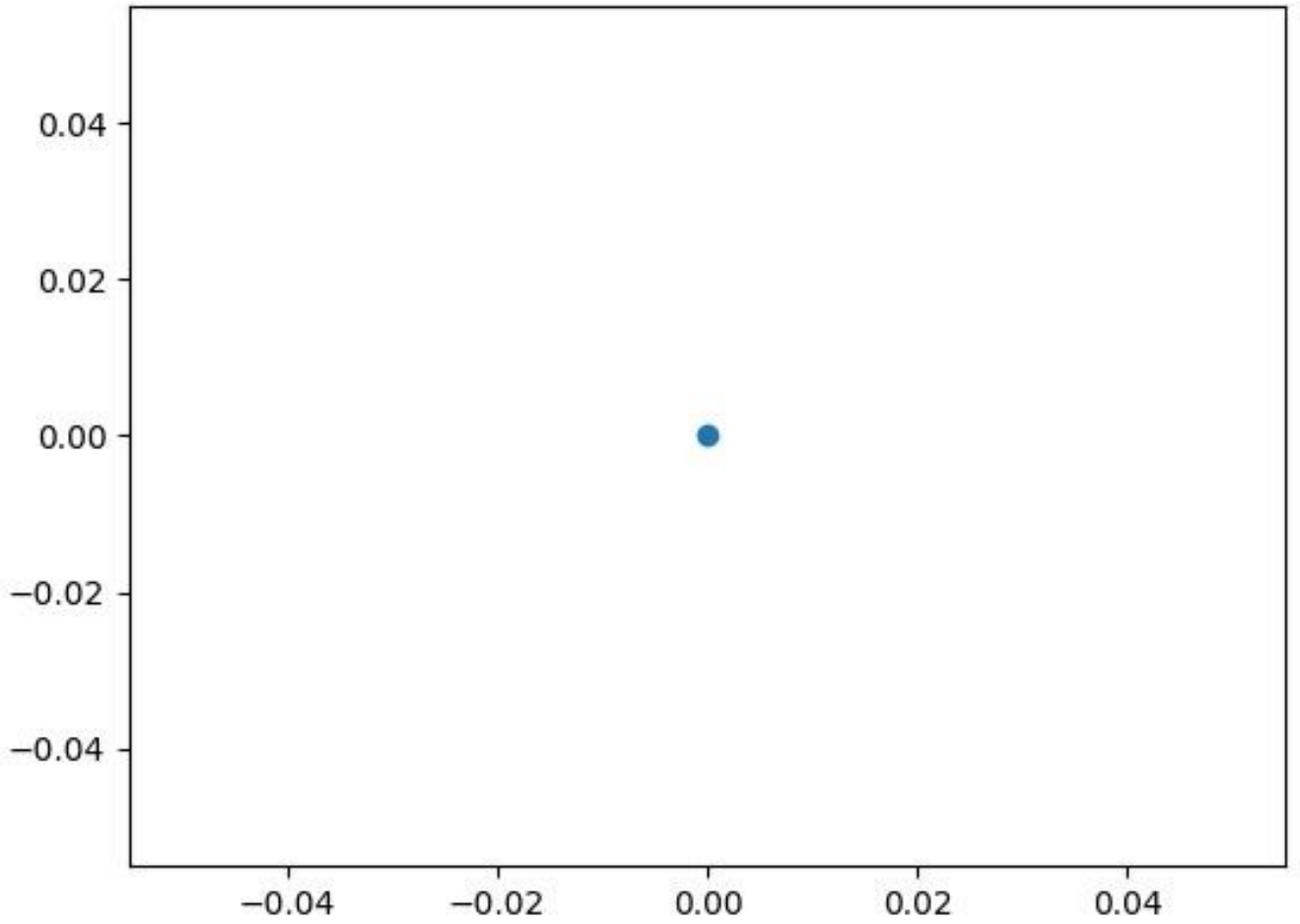
- Intégrer de nouveaux signaux résiduels
- Entraîner un réseau boîte noire pour comparatif
- Attention, LSTM, CNN avec visualisation des filtres, etc.

Bibliographie

- Javier Galbally et Sébastien Marcel. Face anti-spoofing based on general image quality assessment. Proceedings - International Conference on Pattern Recognition, pages 1173-1178, 08 2014.
- Pavel Korshunov, Sébastien Marcel, DeepFakes: a New Threat to Face Recognition? Assessment and Detection., arXiv, 2018
- Anish Mittal, Anush Krishna Moorthy, et Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing, 21(12): 4695-4708, 2012.
- Owen Mayer et Matthew C. Stamm. Accurate and efficient image forgery detection using lateral chromatic aberration. IEEE Transactions on Information Forensics and Security, 13(7) : 1762-1777, 2018.
- Luisa Verdoliva. Media forensics and deepfakes : An overview. IEEE Journal of Selected Topics in Signal Processing, 14 : 910-932, 2020.
- Amerini, Irene & Galteri, Leonardo & Caldelli, Roberto & Bimbo, Alberto, Deepfake Video Detection through Optical Flow Based CNN, 2019
- Joel Frank, Thorsten Eisenhofer, Lea Schonherr, Asja Fischer, Dorothea Kolossa, Thorsten Holz, Leveraging Frequency Analysis for Deep Fake Image Recognition, 2020

Démonstration

Démo animée

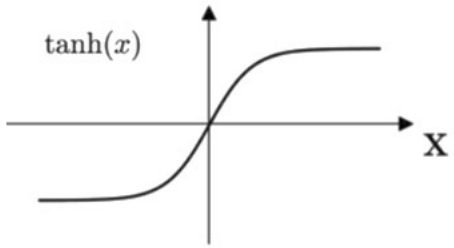


Métriques de référence

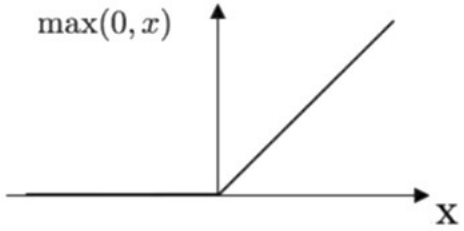
- Recall : à quel point notre modèle prend le risque de rater un deepfake
- Precision : plus elle est élevée et moins le modèle prend des vidéos authentiques pour des deepfakes
- F1-score : comme l'accuracy mais robuste au déséquilibre des classes
- Accuracy : mesure le taux de prédictions correctes en accordant autant d'importance aux éléments négatifs que positifs
- AUC : mesure la capacité à maximiser aussi bien la capacité à correctement reconnaître les éléments de la classe positive que de la classe négative

Fonctions d’activation

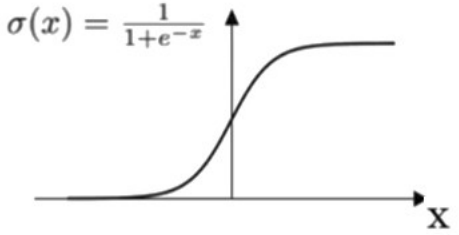
Tanh



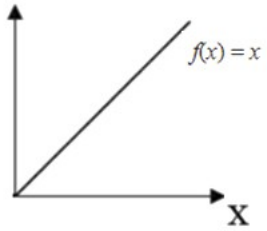
ReLU



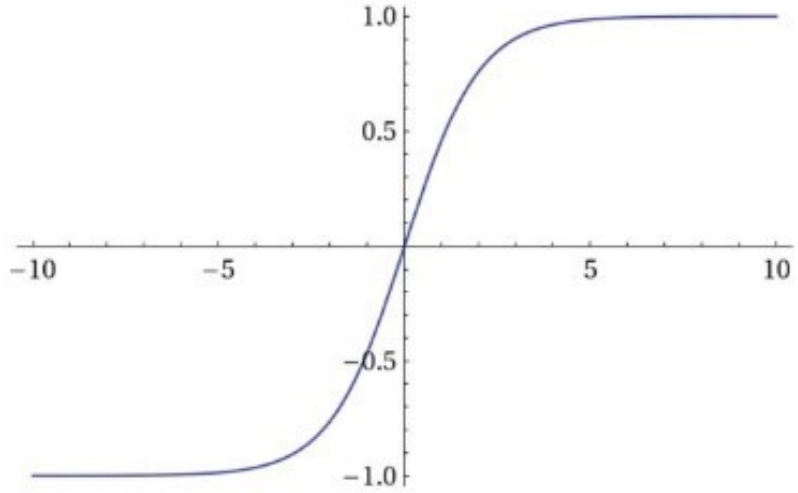
Sigmoid



Linear

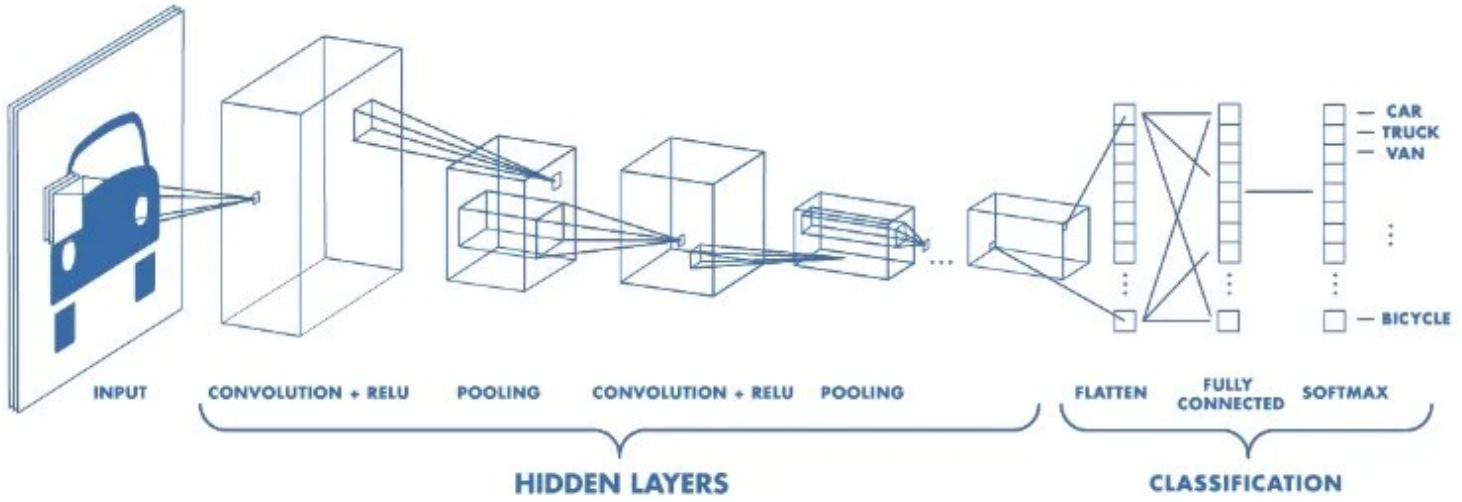


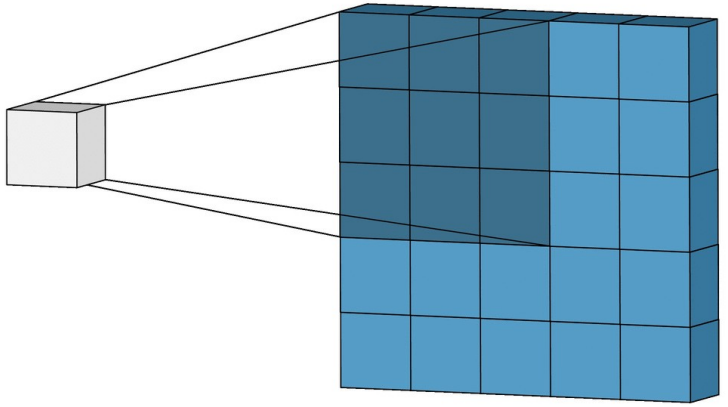
Softmax Activation Function



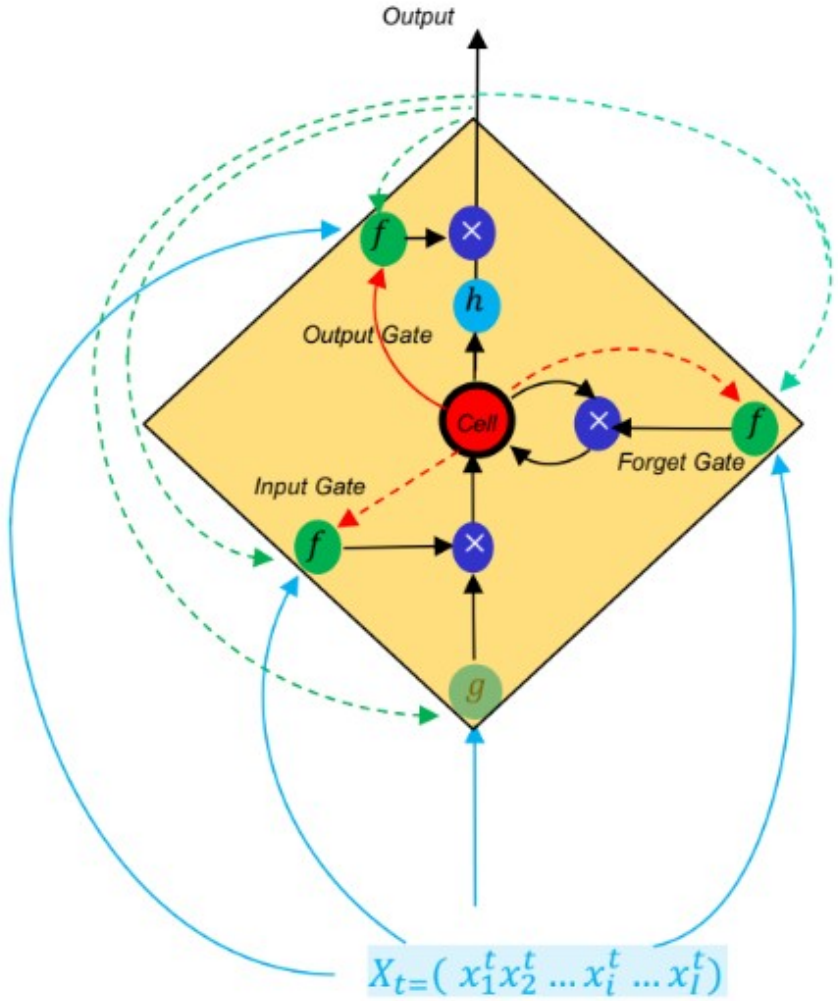
$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Convolutional Neural Network



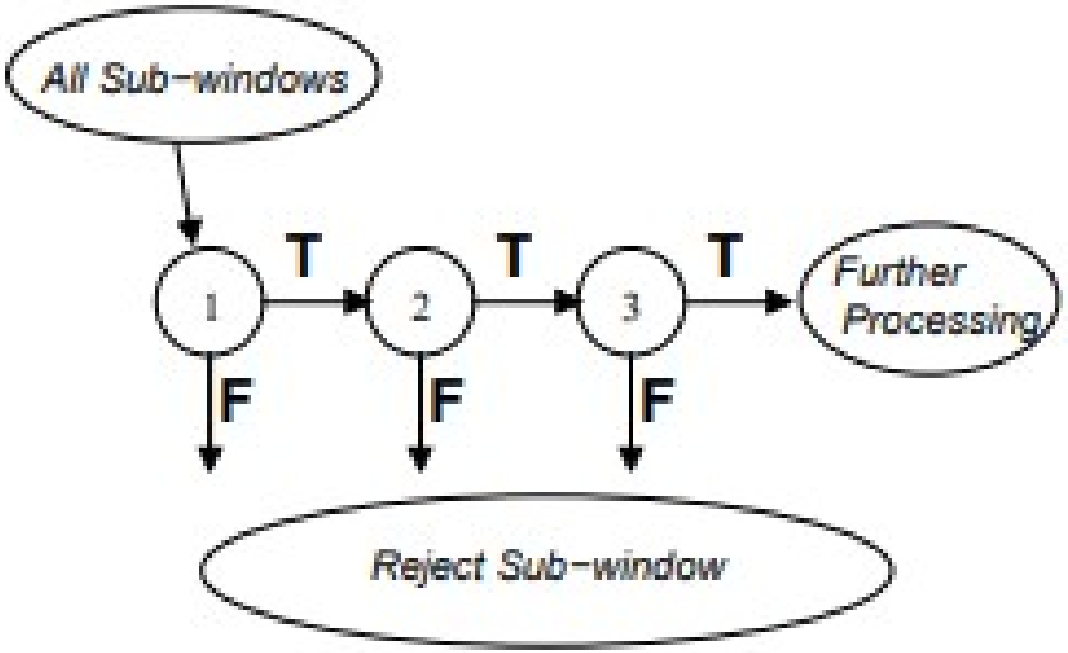


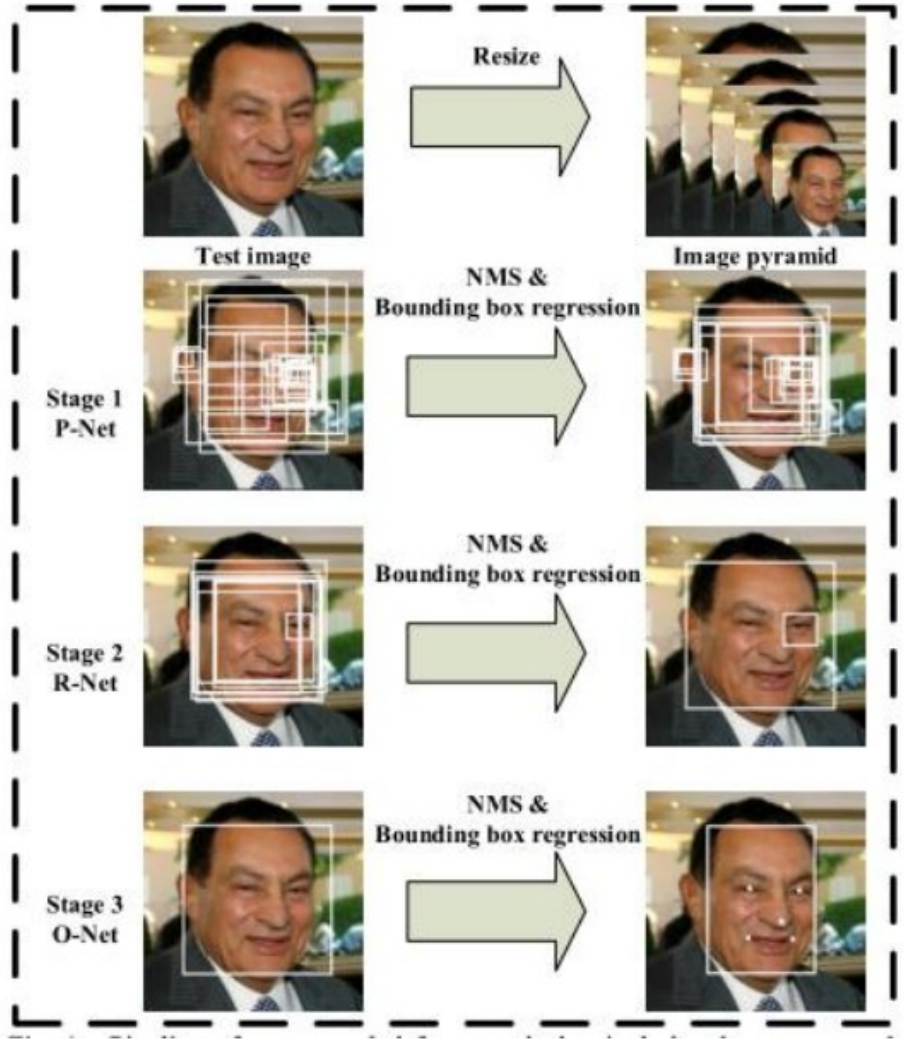
Long Short-Term Memory



- f there are 3 recurrent control gates with sigmoid activation functions
- g 1 standard recurrent unit with sigmoid or *tanh*
- C memory cells (C=1)
- there are M memory blocs
- h h is preferably a *tanh* activation

Extraction des visages

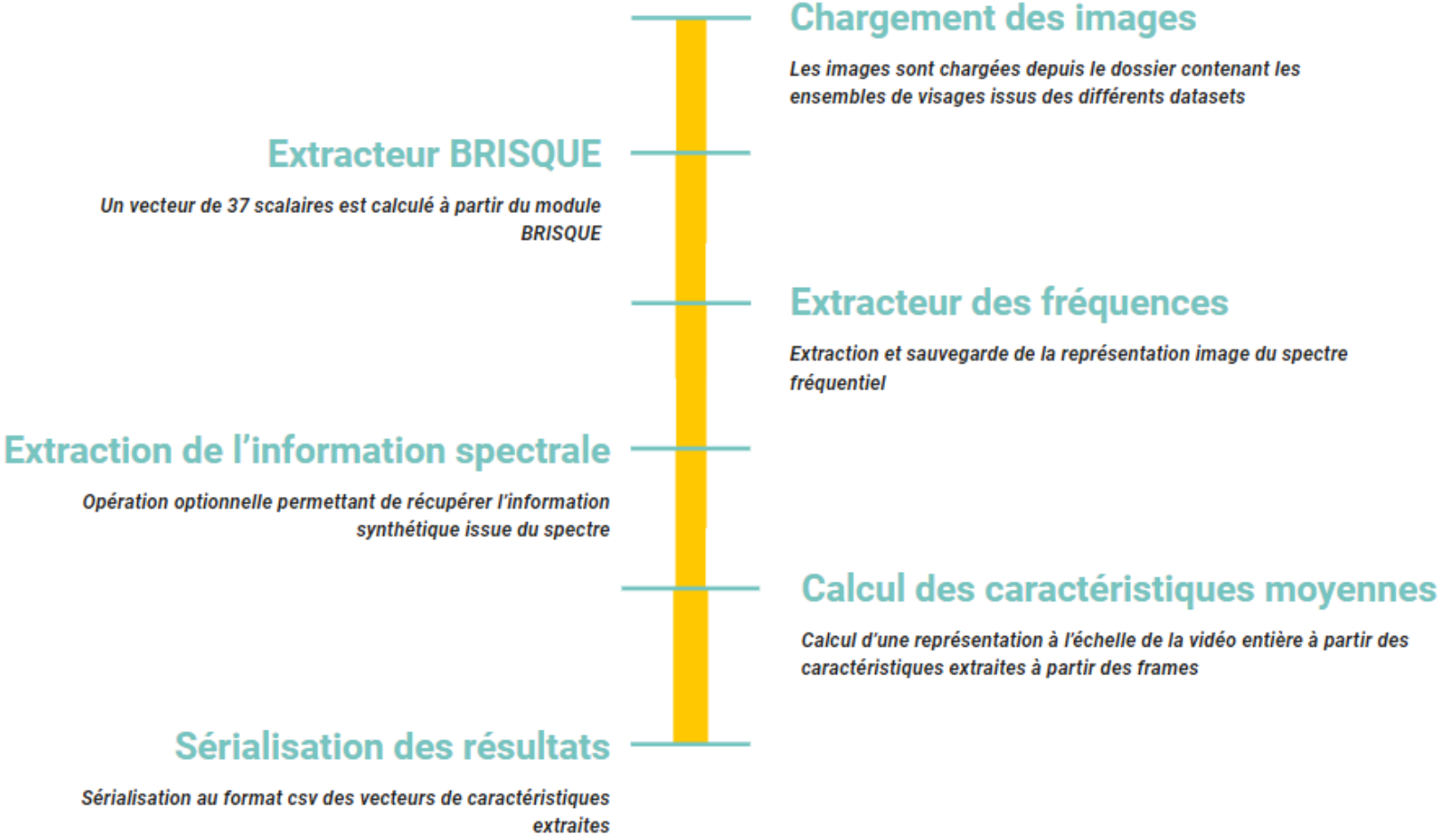




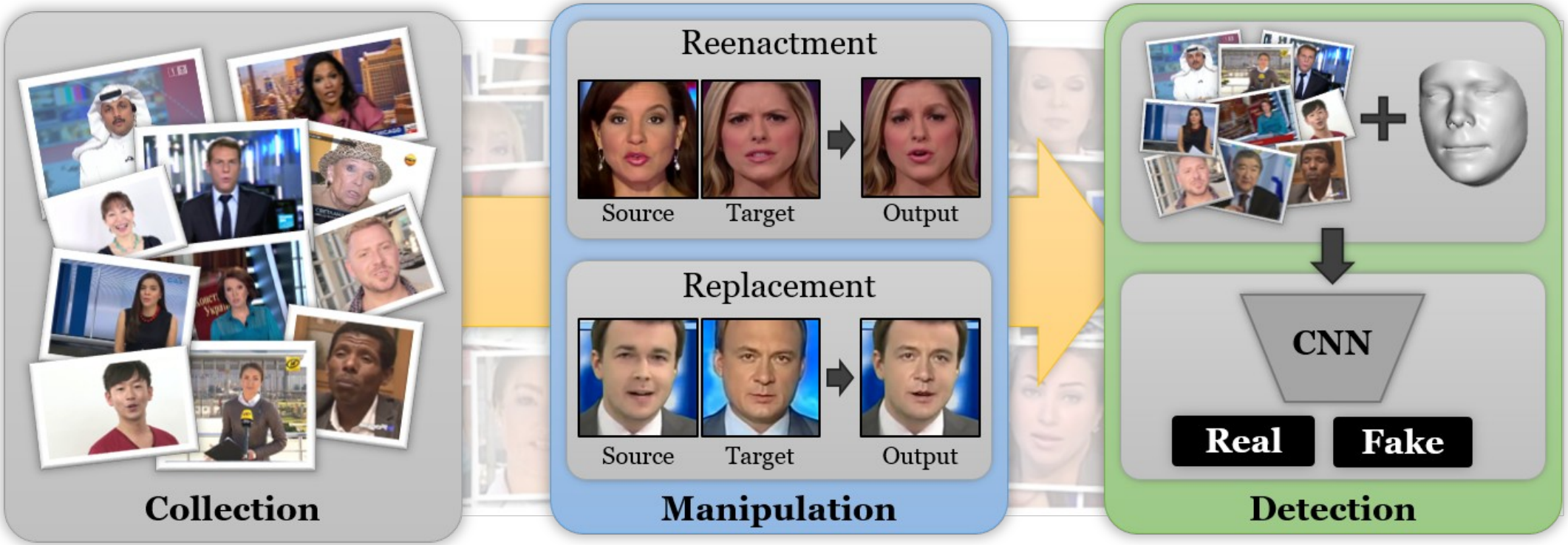
Extraction des visages

	Haar	MTCNN
Précision	Bons résultats	Très bons résultats
Multi-visages	Permet d'extraire deux visages	Permet d'extraire deux visages
Explicabilité	Machine learning basé sur Adaboost donc verdict lisible	Caractéristiques relativement explicables par l'étude des filtres des CNNs donc verdict relativement lisible
Ergonomie	Modèle pré-entraîné mais plus difficile à prendre en main et à ajuster (marges autour du visage)	Modèle clés en mains fourni par FaceNet très simple d'utilisation et adaptable facilement
Coût	Relativement coûteux en temps de calcul mais peu coûteux en poids	Peu coûteux en temps de calcul mais modèle lourd à cause des nombreux poids des CNNs

Pipeline d’extraction



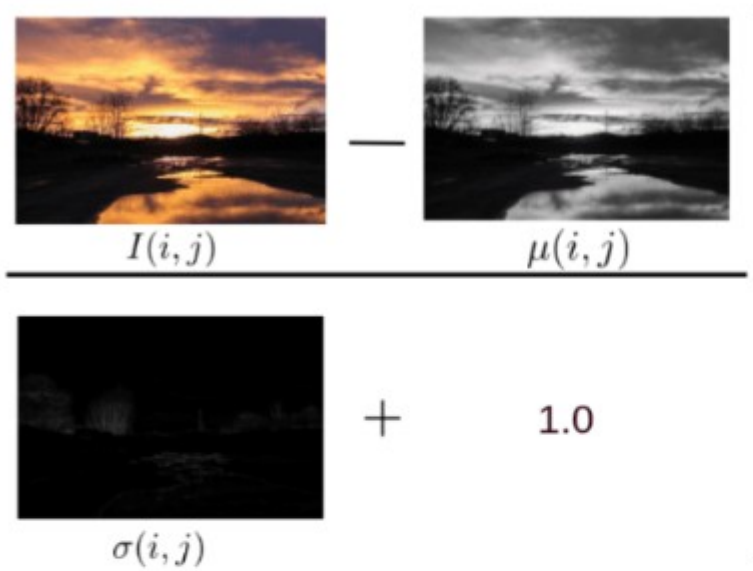
Données


Collection
Reenactment
Source
Target
Output
Replacement
Source
Target
Output
Manipulation
CNN
Real
Fake
Detection

Les données

	VidTIMIT	DeepfakeTIMIT	FF++	Celeb -DF	DFDC	Nb frames
Train	210	160	100/299	X	371/384	53811/56502
Validation	105	80	50/149	X	185/190	24447/26546
Test	105	80	50/149	X	186/192	24765/25422
Généralisation	X	X	X	51/52	X	20694/20905

Blind/Referenceless Image Spatial Quality Evaluator



MSCN Coefficients

$\hat{I}(i,j)$

Î image normalisée

GGD(Î)

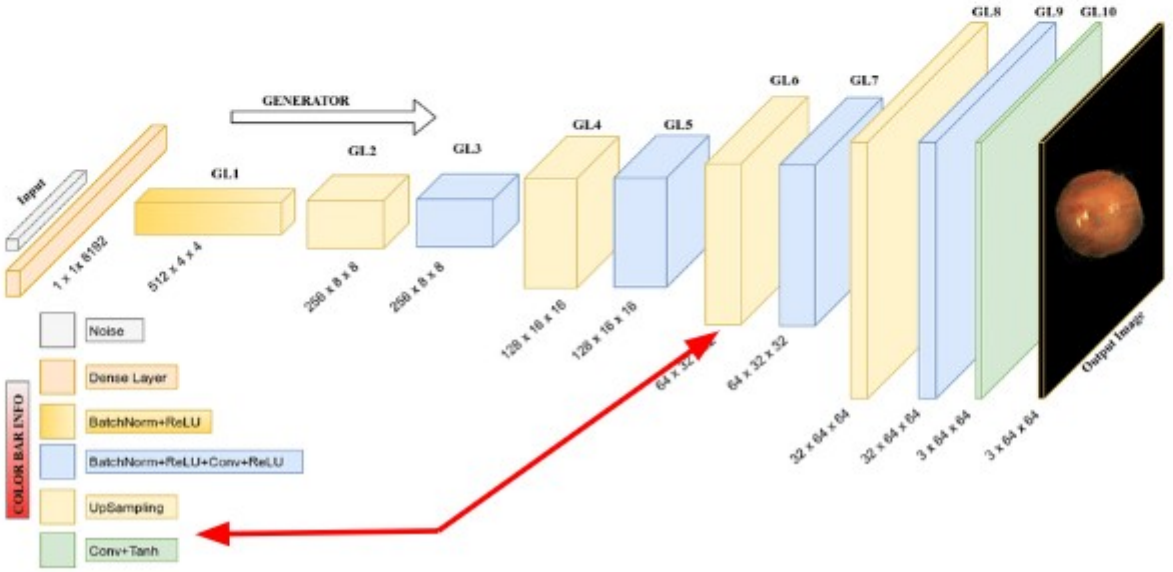
Relations de Voisinage

$$H(i,j) = \hat{I}(i,j)\hat{I}(i,j+1)$$
$$V(i,j) = \hat{I}(i,j)\hat{I}(i+1,j)$$
$$D1(i,j) = \hat{I}(i,j)\hat{I}(i+1,j+1)$$
$$D2(i,j) = \hat{I}(i,j)\hat{I}(i+1,j-1)$$

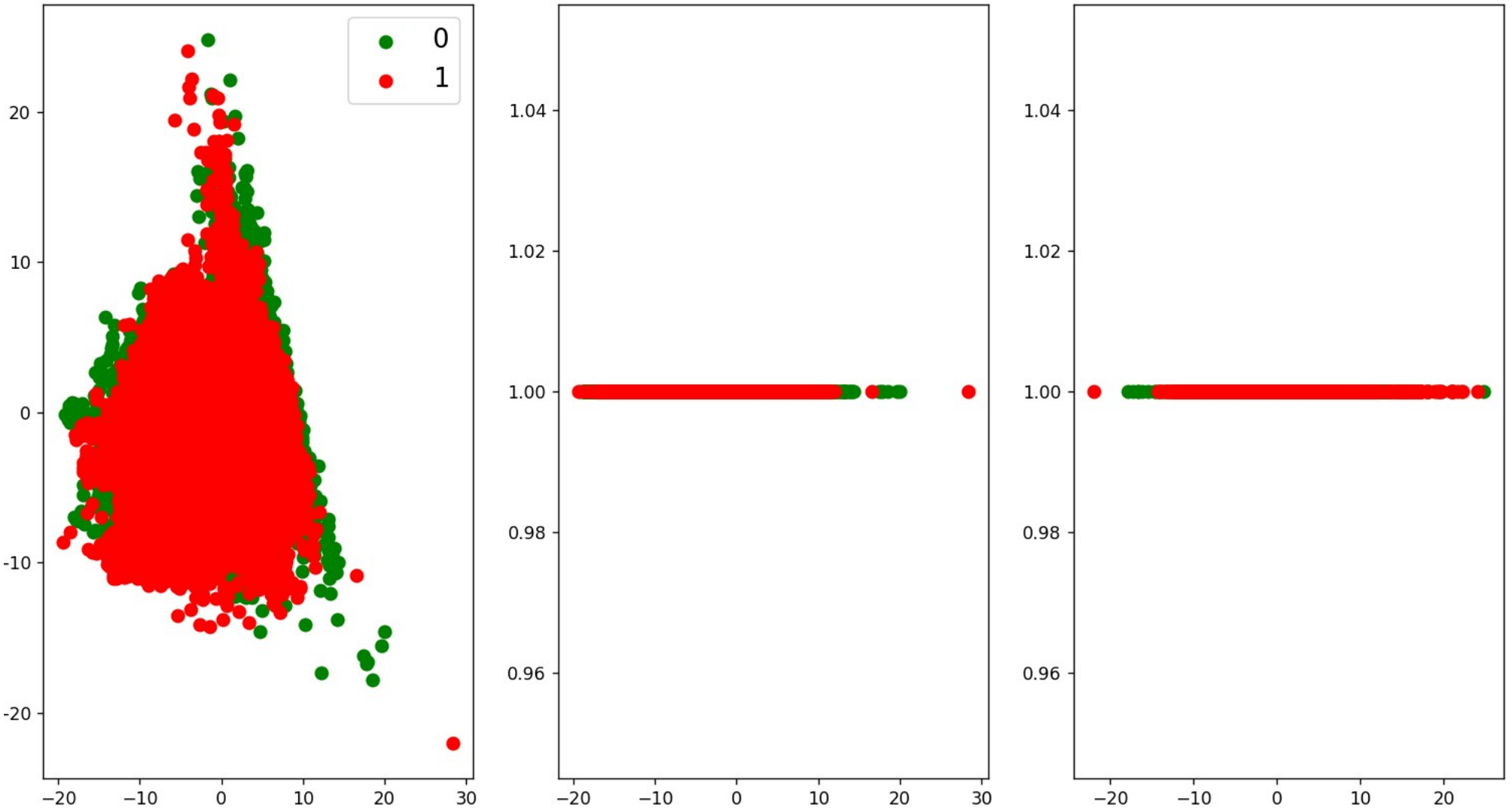
AGGD()

Analyse des hautes fréquences

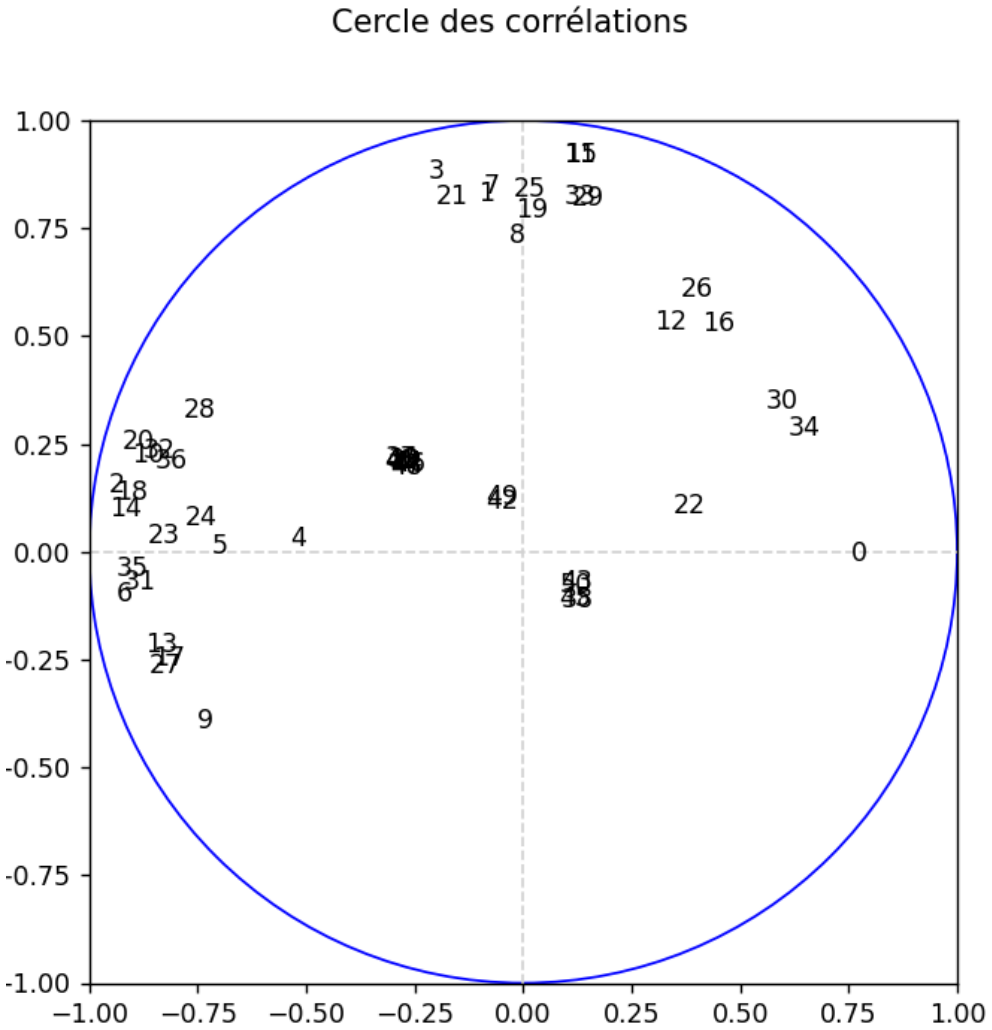
FFHQ Spectrum | FFHQ | StyleGAN | StyleGAN Spectrum



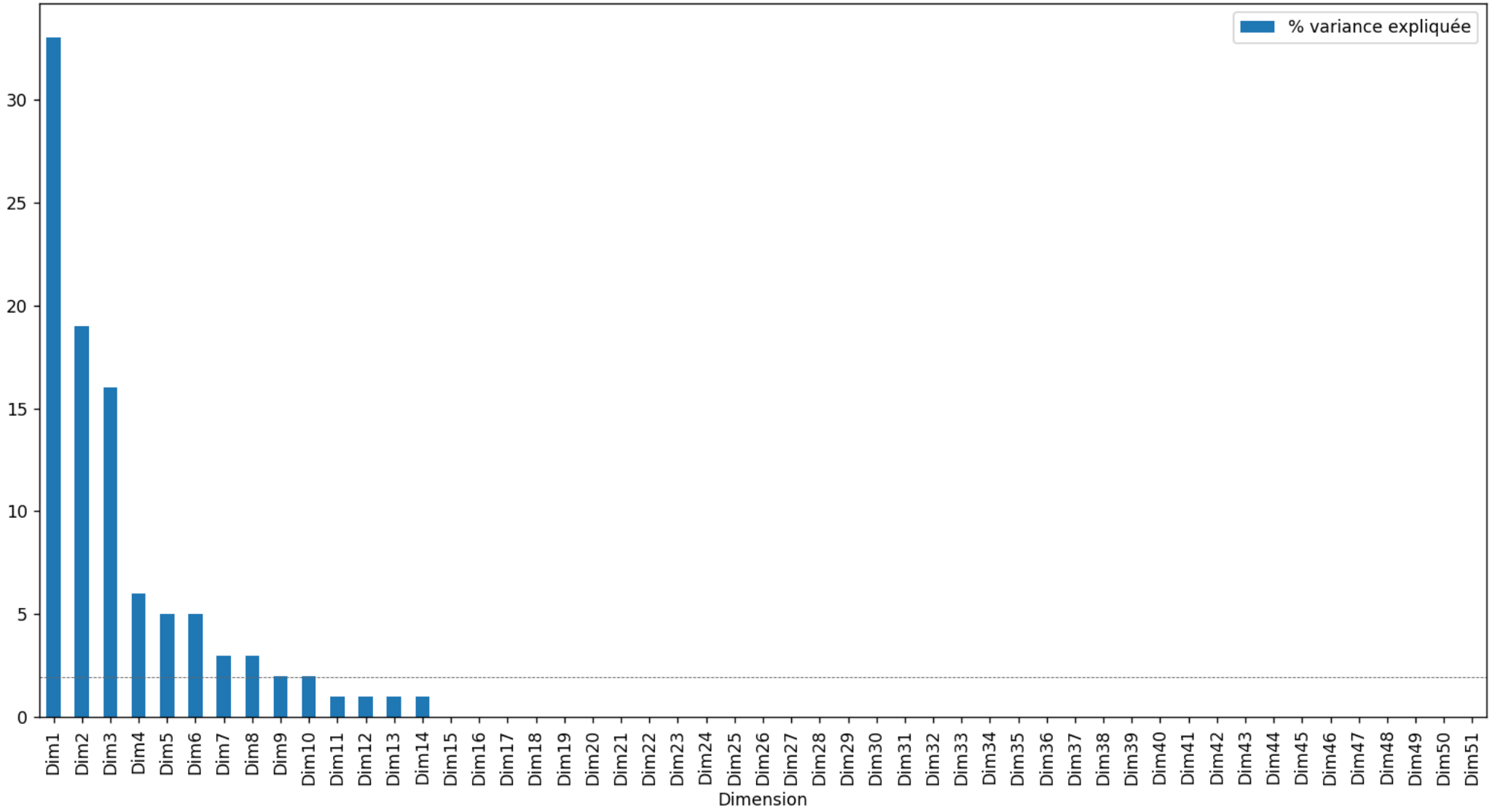
Visualisation



Visualisation



Visualisation



Point de départ deep échelle vidéos (moyenne)

Set/Métriques	F1	Accuracy	AUC	Precision	Recall
Train	81,00 %	78,00 %	78,00 %	81,00 %	84,00 %
Validation	83,00 %	80,00 %	78,00 %	82,00 %	85,00 %
Test	84,00 %	80,00 %	77,00 %	80,00 %	89,00 %
Généralisation	68,00 %	53,00 %	52,00 %	52,00 %	100,00 %

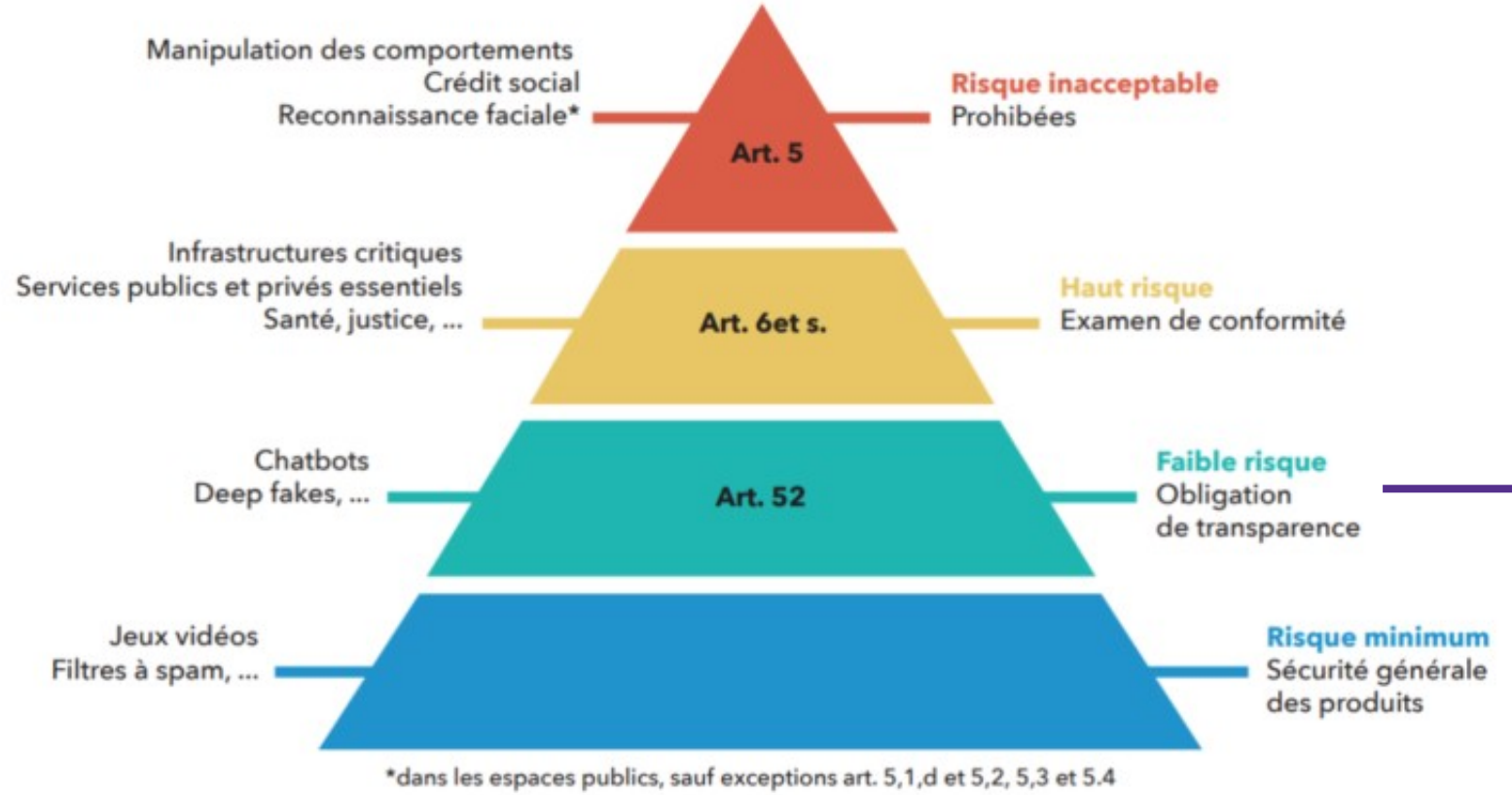
Test final avec quartiles et décimation par 10

Set/Métriques	F1	Accuracy	AUC	Precision	Recall
Train	84,00 %	83,00 %	83,00 %	83,00 %	88,00 %
Validation	86,00 %	84,00 %	83,00 %	83,00 %	89,00 %
Test	87,00 %	86,00 %	85,00 %	86,00 %	89,00 %
Généralisation	68,00 %	53,00 %	52,00 %	52,00 %	98,00 %

Échelles d'analyse

	Échelle vidéo	Échelle frames
Précision	Compression de l'information	Précis car toutes les informations sont analysées sans bruit
Explicabilité	Verdict global	Possibilité de cibler les frames attaquées
Temps d'entraînement	Moins de données (plus de 100 000 à 1500)	100 fois plus de données
Temps d'inférence	Une seule prédiction	Autant de prédictions que de frames

Réglementation de l'IA



→ **Explicabilité**