



Classification et reconstruction de fragments de fichiers binaires

pour l'investigation numérique

Encadrants de stage :

Emmanuel Giguët, Tristan Benoit

Professeur encadrant :

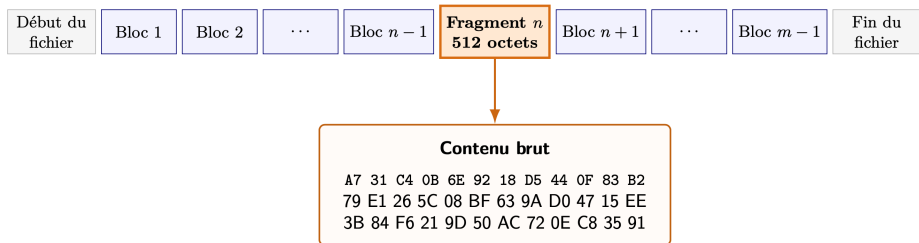
Patrick Lacharme

Nora JAFJAF

Master 2 Cybersécurité

2025–2026

Contexte du stage



- **Fragment** : suite contiguë d'octets extraite d'un fichier.
- **Analyse de fragments** : identifier leur type à partir de leur contenu binaire.
- **Reconstruction** : retrouver les fragments appartenant au même fichier et leur ordre d'origine.
- **Difficultés** : absence d'indices exploitables

Problématique centrale

Comment utiliser des méthodes d'apprentissage automatique pour identifier, puis retrouver l'organisation de fragments binaires courts lorsque leur contexte d'origine n'est plus disponible ?

1

Classifier

Quel est l'effet de la **taille** et du **type de fichier** sur la classification ?

2

Comprendre les erreurs

Les erreurs s'expliquent-elles par la **structure interne** des formats composites, notamment le PDF ?

3

Reconstruire

Peut-on exploiter les **frontières des fragments** pour retrouver leur ordre ?

Modèle	Principe
SVM	Sépare les classes à partir d'une frontière de décision.
MLP	Réseau entièrement connecté
CNN	Apprend des motifs locaux dans une séquence d'octets ou de bits.
TCN	Utilise des convolutions dilatées pour capturer des dépendances à différentes échelles.
Random Forest	Combine plusieurs arbres de décision entraînés indépendamment.
Gradient Boosting	Construit successivement des arbres, chacun cherchant à corriger les erreurs des précédents.

État de l'art : classification de fragments

Méthode	Auteurs	Représentation	Modèle	Jeu / taille	Accuracy
Sceadan	Fitzgerald et al. (2013)	Descripteurs statistiques	SVM	Corpus privé / 512 octets	–
FiFTy	Mittal et al. (2020)	Octets bruts	CNN 1D	FFT-75 / 512 octets	65,6 %
ByteNet	Liu et al. (2024)	Octets 1D + image 2D	Réseau à deux branches	FFT-75 / 512 octets	73,2 %
Hierarchical FFC	Zou et Liu (2026)	Octets bruts	CNN 1D spécialisés	FFT-75 / 512 octets	76,3 %

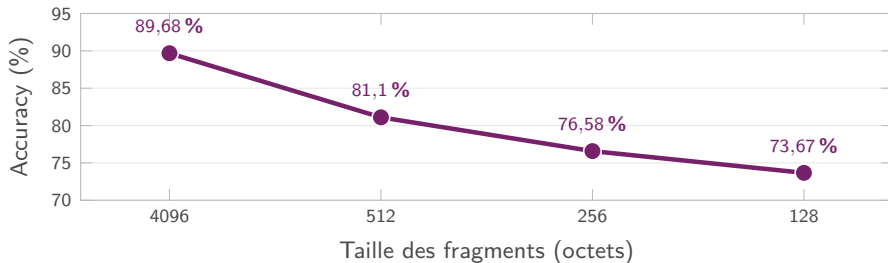
1. Classifier les fragments : corpus étudiés

	Corpus maison	FFT-75
Nombre de classes	7 classes	75 types de fichiers
Formats	TXT, DOC, JPG, PDF, AES-256-CBC, Base64 et ZIP	Textes, images, archives, exécutables, formats composites, etc.
Données sources	3 500 fichiers, soit 500 par classe	Fragments directement fournis dans le jeu de données
Volume	102 600 à 104 820 fragments au total, soit environ 15 000 par classe	102 400 fragments par type, soit 7,68 millions au total
Taille des fragments	4 096, 512, 256 et 128 octets	512 et 4 096 octets disponibles
Sources	GovDocs1 et jeux publics Kaggle pour les images et les PDF	FFT-75, publié avec FiFTy et distribué par IEEE DataPort

1. Classifier : protocole expérimental

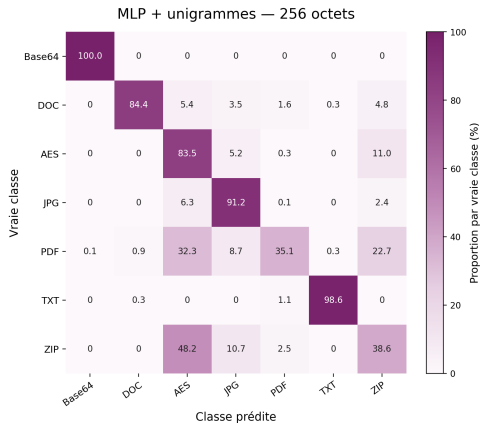
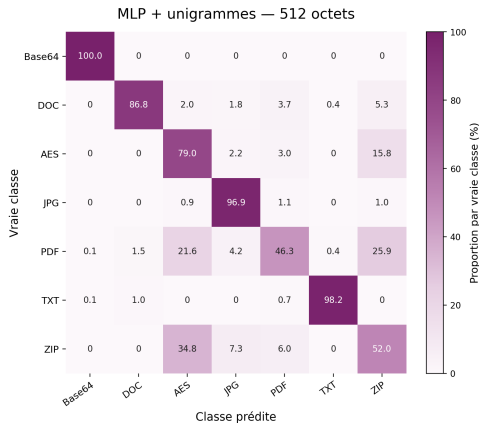
- ▶ **Données** : 3 500 fichiers sources répartis en 7 classes.
- ▶ **Séparation entraînement/test** : 80 % des fichiers pour l'entraînement et 20 % pour le test, avec une graine fixée à 42.
- ▶ **Séparation par fichier source** : les fragments issus d'un même fichier ne se retrouvent jamais à la fois dans l'entraînement et dans le test.
- ▶ **Fragmentation** : découpage des fichiers en fragments.
- ▶ **Tailles étudiées** : 4 096, 512, 256 et 128 octets.
- ▶ **Évaluation** : comparaison de plusieurs représentations et modèles pour chaque taille de fragment.

1. Classifier : effet de la taille des fragments

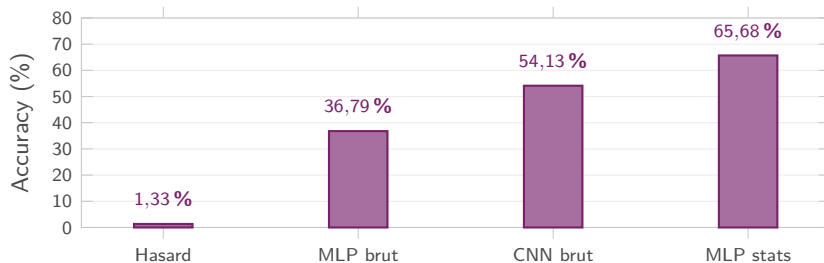


Taille	Meilleur modèle	Représentation
4 096	Gradient Boosting	Unigrammes + entropie
512	MLP	Unigrammes
256	MLP	Unigrammes
128	SVM	Trigrammes hashés + entropie

1. Classifier : difficultés selon le contenu

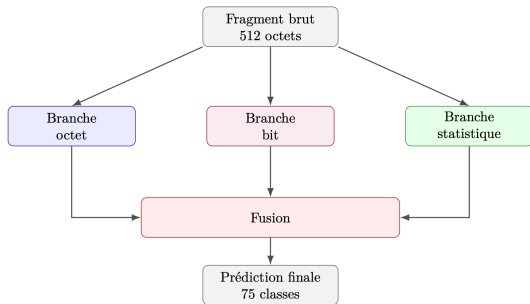


1. Classifier : premières méthodes sur FFT-75



Méthode	Caractéristiques
MLP brut	Valeurs des 512 octets
CNN brut	Motifs locaux dans la séquence d'octets
MLP stats	Unigrammes, entropie et variance

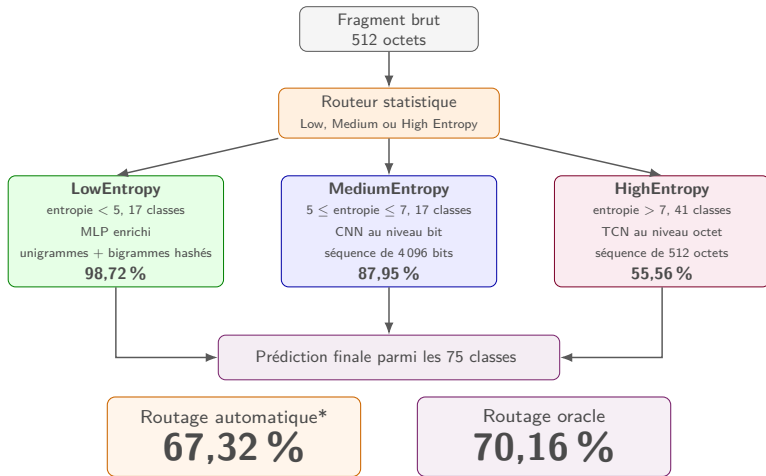
1. Classifier : modèle tri-stream



Le modèle apprend le poids attribué à chacune des trois branches avant la prédiction finale.

Configuration	Accuracy
Branche octet	70,89 %
Branche bit	67,03 %
Branche statistique	58,98 %
Octet + statistique	71,00 %
Octet + bit + statistique	71,12 %

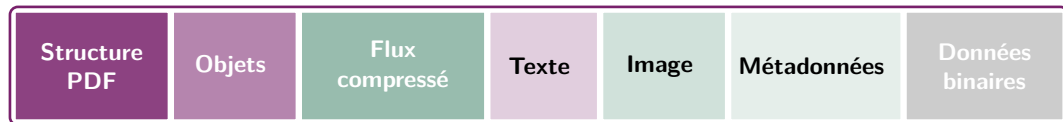
1. Classifier : modèle hiérarchique



*Routage automatique : performance estimée.

2. Comprendre les erreurs : contenu interne d'un fragment PDF

Contenus possibles d'un fragment PDF



Deux fragments PDF peuvent contenir des informations de nature très différente.

2. Comprendre les erreurs : description des fragments PDF

1

Vocabulaire structurel PDF

Extraction des chaînes ASCII visibles dans un corpus externe de fichiers PDF.

- ▶ obj, endobj
- ▶ stream, endstream
- ▶ xref, trailer
- ▶ /Type, /Length, /Filter

2

Descripteurs statistiques

Description des fragments PDF de FFT-75 selon leur contenu interne.

- ▶ Entropie
- ▶ Tokens PDF : marqueurs structurels
- ▶ Entropie par blocs : variations locales

2. Comprendre les erreurs : profils internes obtenus

3

Regroupement par GMM

Quatre profils obtenus automatiquement à partir des descripteurs.

4

Analyse

Comparaison avec les prédictions du tri-stream précédent.

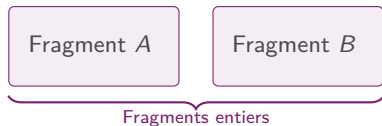
Profil	Entropie	Structure observable	Reconnaissance par le tri-stream
0	Très élevée : 7,546	Aucun token PDF détecté	Faible
1	Très élevée : 7,559	Quelques tokens PDF : 1,767 en moyenne	Faible
2	Plus faible	100 % d'octets imprimables, nombreux tokens	Très élevée
3	Intermédiaire	Structure mixte et fortes variations locales	Élevée

3. Reconstruire : retrouver le fragment suivant

Retrouver le vrai successeur dans le **pool complet du document**.

Un modèle Random Forest est entraîné sur 81 PDF, puis évalué sur des documents distincts.

Caractéristiques globales

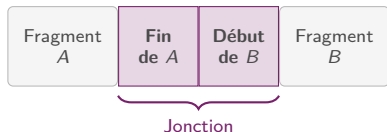


- Entropie et octets imprimables
- Séparateurs et tokens PDF
- Différences entre A et B

Top-1 sur le pool complet

3,0 %

Caractéristiques de frontière



- Fenêtres de 32, 64 et 128 octets
- Fin de A et début de B
- Entropie

Top-1 sur le pool complet

5,9 %

Conclusion

- ▶ Difficulté pour les fragments **courts, compressés ou peu structurés**.
- ▶ Contribution principale : analyse du **contenu interne des fragments PDF** et de son influence sur les erreurs.
- ▶ Exploration de la reconstruction : **signal d'adjacence présent, mais encore insuffisant**.

Merci pour votre attention

